Interpretabilnost modela dubokoga učenja vrjednovanjem mapa relevantnosti

Vukadin, Davor

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Davor Vukadin

INTERPRETABILNOST MODELA DUBOKOGA UČENJA VRJEDNOVANJEM MAPA RELEVANTNOSTI

DOKTORSKI RAD

Zagreb, 2024.

Sveučilište u Zagrebu
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Davor Vukadin

INTERPRETABILNOST MODELA DUBOKOGA UČENJA VRJEDNOVANJEM MAPA RELEVANTNOSTI

DOKTORSKI RAD

Mentor: Izv. prof. dr. sc. Marin Šilić

Zagreb, 2024.

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Davor Vukadin

INTERPRETABILITY OF DEEP LEARNING MODELS BY EVALUATING RELEVANCE MAPS

DOCTORAL THESIS

Supervisor: Associate Professor Marin Šilić, PhD

Zagreb, 2024.

Doktorski rad izrađen je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva, na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave

Mentor: izv. prof. dr. sc. Marin Šilić

Doktorski rad ima: 109 stranica

Doktorski rad br.: _____

O mentoru

Marin Šilić je rođen 1983. godine u Sarajevu u Bosni i Hercegovini. Osnovno obrazovanje pohađao je u Makarskoj u Hrvatskoj. Diplomirao je 2007. godine na Sveučilištu u Zagrebu, Fakultetu elektrotehnike i računarstva. Kao student bio je primatelj stipendije Ministarstva Znanosti Republike Hrvatske i dio naprednog programa završetka diplomskog studija s posebnim fokusom na istraživački rad. Nakon diplome, od strane fakulteta dodijeljena mu je nagrada "Josip Lončar" kao jednom on najboljih studenata generacije u području računarstva. Od 2007. godine zaposlen je kao doktorand na Sveučilištu u Zagrebu, Fakultetu elektrotehnike i računarstva. Tijekom doktorskog studija boravio je 6 mjeseci na znanstvenom usavršavanju u New Yorku u kompaniji Google kao član Google Docs tima radeći na razvoju sustava Google Spreadsheets. Doktorirao je na Sveučilištu u Zagrebu, Fakultetu Elektrotehnike i Računarstva 2013. godine gdje je trenutno zaposlen kao izvanredni profesor. Rezultate svojih istraživanja opisao je u radovima koji su objavljeni u uglednim strukovnim časopisima kao što su IEEE Transactions on Services Computing, IEEE Transactions od Dependable and Secure Computing, Journal of Systems and Software, Knowledge Based Systems, ACM Transactions on Intelligent Systems and Technology i IEEE Access. Nadalje, svoja istraživanja objavio je i na prestižnim skupovima istraživača područja programskog inženjerstva ACM SIGSOFT Symposium on the Foundations of Software Engineering i IEEE International Conference on Software Quality, Reliability and Security. Sudjelovao je kao voditelj i suradnik na više stručnih i znanstvenih projekata na FER-u gdje je do sada uspješno mentorirao tri doktorske disertacije. Fokus njegovog istraživanja trenutno su računarstvo zasnovano na uslugama, sustavi za preporučivanje, analiza velikih skupova podataka, umjetna inteligenciju i strojno učenje te previđanje pogrešaka programskog koda. Član je strukovne udruge IEEE.

About the Supervisor

Marin Šilić was born in 1983. in Sarajevo in Bosnia and Herzegovina. He attended his basic education in Makarska in Croatia. He graduated in 2007. from the Faculty of Electrical Engineering and Computing, University of Zagreb. As a student he received a scholarship from the Croatian Ministry of Science and was part of an advanced program with special emphasis on the research work. Upon graduation he was awarded the "Josip Lončar" as one of the best graduating students in computing of his generation. He has worked at the Faculty of Electrical Engineering and Computing, University of Zagreb since 2007. During his Ph.D. he spent 6 months in New York, US at Google working as part of the Google Docs team focusing on the development of the Google Spreadsheets List View System. He got his Ph.D. from the Faculty of Electrical Engineering and Computing, University of Zagreb in 2013. and is currently em-

ployed there as an Associate Professor. He has published several papers in IEEE Transactions on Services Computing, IEEE Transactions on Dependable and Secure Computing, Journal of Systems and Software, Knowledge-Based Systems, ACM Transactions on Intelligent Systems and Technology and IEEE Access. Also, he has published his research results at the ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering and at the IEEE International Conference on Software Quality, Reliability and Security. He has been a lead and an associate on multiple practical and scientific projects at the FER where he has also successfully mentored three Ph.D. candidates. His research interests include service-oriented computing, recommender systems, big data analysis, artificial intelligence, machine learning, and software defect prediction. He is a member of IEEE.

Zahvale

Zahvaljujem se svojoj djevojci Lauri i našem mačiću Tiboru na pružanju neograničene količine potpore i ljubavi tijekom izrade ove doktorske disertacije. Bez njih ona nikada ne bi bila dovršena.

Također se zahvaljujem i svojoj obitelji: sestri Emi, majci Gordani, ocu Andriji te psu Loli na velikoj podršci tijekom cijelog života.

Zahvaljujem se svojemu mentoru, izv. prof. dr. sc. Marinu Šiliću na iznimnom i stručnom vodstvu kako u pisanju radova i ove disertacije, tako i u profesionalnom svijetu.

Ovaj rad posvećujem svojoj djevojci Lauri.

Sažetak

Recentan napredak radnih značajki dubokih neuronskih mreža doveo je do razvoja brojnih sustava koji nadmašuju čak i ljudske mogućnosti. Međutim, neprozirnosti modela zaslužnih za ovaj napredak često ograničava njihovo korištenje u osjetljivim područjima gdje su objašnjivost i transparentnost modela od iznimne važnosti. Tijekom godina, istraživači su predložili brojne algoritme koji za cilj imaju povećati razumijevanje odluke pojedinog modela dubokog učenja. Jedna od najpopularnijih metoda je i metoda slojevite propagacije relevantnosti. Ona dodjeljuje mapu relevantnosti promatranom primjeru na temelju dekompozicije nelinearnog klasifikatora za svaku ulaznu značajku. S porastom broja predloženih atribucijskih metoda, pojavljuje se potreba za objektivnim načinom vrjednovanja njihove kvalitete. Iz tog razloga, predložene su brojne metrike koje procjenjuju različita svojstva atribucijskih metoda poput vjernosti, robusnosti i lokalizacije. Nažalost, nije postignut konsenzus oko optimalne metrike u svakoj situaciji. Iz tog razloga, istraživači često posežu za korištenjem više metrika odjednom kako bi procijenili kvalitetu atribucijskih mapa. Međutim, ne postoji jasan način kako agregirati rezultate svih korištenih metrika u jednu, sveobuhvatnu ocjenu. Osim objektivne procjene kvalitete same atribucijske metode, važno je analizirati i poravnanje između procjene kvalitete metrike s ljudskom procjenom kvalitete kako bi se osiguralo da će metoda koja ostvaruje bolji rezultat na odabranoj metrici zaista više pridonositi razumijevanju odluke modela krajnjem korisniku. U sklopu ove disertacije predložena je nova metoda za produkciju mapa relevantnosti koja rješava nedostatke u trenutnim metodama slojevite propagacije relevantnosti te se izravno primjenjuje na najnovije duboke arhitekture. Također je predložena i nova metrika za vrjednovanje kvalitete atribucijskih metoda koja pruža jednu, sveobuhvatnu ocjenu analizirajući istovremeno komponente vjernosti, robusnosti i lokalizacije. Naposljetku, predlaže se metoda za vrjednovanje poravnanja između metrika i ljudske percepcije s ciljem isticanja nedostataka u području vrjednovanja mapa relevantnosti.

Ključne riječi: objašnjiva umjetna inteligencija, duboko učenje, transformer, vrjednovanje atribucijskih metoda, ljudska percepcija

Extended Abstract

Interpretability of deep learning models by evaluating relevance maps

Deep neural networks (DNNs) have become essential for processing various types of data, largely due to advancements in deep learning hardware. This progress has enabled training models on an unprecedented scale, both in parameter count and dataset size. Constant scaling of these variables has led to state-of-the-art models that outperform human abilities on numerous tasks, such as image classification, reinforcement learning, and natural language understanding.

The presence of numerous parameters and multiple nonlinear layers contributes to the opaque nature of these models, making it difficult to explain why certain outputs occur. This opacity motivates researchers, especially in fields where model explainability is crucial—such as aerospace, medicine, or finance—to opt for smaller, often linear, models that are transparent in their decision processes.

AI systems rely on models that learn through exposure to large datasets, rather than explicit human programming. This learning process allows models to independently detect patterns and relationships within data.

During training, these models can uncover complex correlations between input features, like clinical symptoms, and can subsequently make predictions or decisions, such as medical diagnoses. Such models are often highly complex, particularly in deep learning, with recent models involving over a trillion interacting parameters. This complexity makes it nearly impossible, even for experts, to identify the exact mechanisms behind model outputs.

In these cases, the reasoning behind AI-driven decisions may remain unclear to both users and those affected by these systems. This opacity is often termed the "black-box" effect, where the model's inner workings are not transparent. This lack of transparency can lead to either misplaced trust from users or excessive reliance on these systems. Both outcomes are problematic: undue trust may result in unconditional belief in a system's accuracy, while over-reliance may lead to delegating critical decisions to models that could make inaccurate predictions. Consequently, these limitations can have significant negative implications, especially in sensitive areas.

Explainable Artificial Intelligence (XAI) refers to the ability of AI systems to provide clear, understandable explanations for their actions and decisions. The main goal of XAI is to clarify the behaviour of these systems by explaining the underlying mechanisms and decision-making processes, helping users understand how and why certain decisions are made.

XAI approaches are broadly categorized into two types: self-interpretable models and post hoc explanations. Self-interpretable models are designed to be inherently understandable in their decision-making processes. In contrast, post hoc explanations offer the possibility of explaining model behaviour without introducing specific architecture modifications.

Self-interpretable models use transparent algorithms or mechanisms that simplify parts of the model, allowing users to see how inputs influence outputs and identify key predictive features.

However, due to the complexity of deep learning models, fully explaining their inner workings in a simplified way is often impractical. For complex systems, post hoc methods are usually more suitable, generating explanations after decisions are made, either as global or local explanations.

Global explanations provide an overall understanding of the AI model's decision-making behaviour, covering general patterns and insights. Local explanations, in contrast, focus on specific outputs, explaining the model's reasoning for individual cases.

This dissertation focuses on post hoc local methods.

With the rise of XAI methods, there is an increasing need for effective metrics to evaluate these methods. This need is recognized not only in AI but also by researchers in human-computer interaction (HCI). In evaluating metrics, the objective is to measure the quality and effectiveness of different explanation methods accurately.

Traditional performance metrics assess prediction accuracy and computational complexity, but evaluating interpretability is challenging. This challenge results in a wide range of explanation methods, as XAI lacks consensus on standardized evaluation metrics, often relying on researchers' subjective assessments based on specific examples. Such subjective evaluation is inadequate for robust quality assessment, which has prompted a focus on objectively validating and comparing explanation methods.

While humans can subjectively assess relevance maps, quantitative measures are essential for objective comparison and ranking. This work emphasizes the quantitative assessment of explanation quality for specific cases and neural networks, a complex task influenced by factors such as network architecture and training data.

Relevance maps should reflect the model's perspective, not strictly adhere to human quality judgments. For instance, an object's segmentation map may miss key evidence outside the object's area. Therefore, various metrics have been developed to assess aspects like fidelity, robustness, and localization. However, there is no established approach for effectively combining these aspects to identify the best attribution method, complicating comprehensive comparisons.

In addition, there is a significant gap in research on evaluating the quality of the evaluation metrics themselves. Since these metrics aim to avoid subjective assessment of relevance maps, there is an increasing need to validate the alignment between these metrics and human perception to ensure that high metric scores correlate with perceived quality from the end-user's perspective.

The content of the dissertation is briefly summarized below.

In Chapter 2, the method of layered relevance propagation, which serves as the foundation

for the proposed method for producing relevance maps, is described in detail. Frequently used rules within this method are thoroughly explained. Additionally, an overview of the application of this method to different architectures, tasks, and fields, such as medicine, is provided. Relevant approaches from the field of layered relevance propagation, as well as other fields related to relevance map production, are presented, and their limitations are listed. The end of the chapter provides an overview of explainability in Transformer architecture, linking existing research to layered relevance propagation and describing their limitations. The main limitation of existing Transformer explainability methods is the reliance on the self-attention values as the primary means for calculating the final attribution maps. This fact results in reduced interpretability, as the resolution of such attribution maps depends on the number of image patches the model uses. For example, a standard Visual Transformer with a patch size of 16×16 pixels processing an image of 224×224 pixels would produce 14×14 patches, which are then scaled to the original image resolution in the output of the attribution method. In the case of a Vision Transformer with a smaller number of larger patches designed for speed and efficiency, attribution methods that operate at the patch level reduce interpretability. The proposed method, on the other hand, is applied to several different architectures, and the relevance values themselves are propagated to the input image, producing a map at the pixel level, a property that no other method in current research offers.

Chapter 3 offers a comprehensive review of the field of evaluating methods for producing relevance maps. This chapter describes approaches to evaluating various components of attribution methods, such as faithfulness, robustness, and localization. Next, it details the most relevant approaches in these categories for the proposed evaluation metric for attribution methods and highlights the shortcomings of each metric. Finally, a global shortcoming in current research is presented, namely the lack of a defined method for combining multiple metrics into a single comprehensive result.

Chapters 4, 5, and 6 present the author's scientific contributions in this dissertation.

In Chapter 4, a proposed method for producing relevance maps based on the layered propagation of relative absolute relevance magnitude is described. The motivation for and solution offered by the proposed method are provided. An efficient way to implement the proposed method within modern deep learning frameworks is explained. Finally, the details of the proposed method's implementation for layers specific to modern architectures are presented.

Chapter 5 includes the motivation and a detailed overview of two crucial components of the proposed metric for evaluating methods for relevance map production, based on global attribution evaluation. The local consistency component includes two subcomponents: faithfulness and robustness, which are calculated innovatively within the proposed method. Unlike previous research that relied either on ranking input features provided by the initial relevance map or on using a random subset of features for perturbations, the proposed approach uses a gradient-

based method at each step to identify regions that will result in the most or least significant change in the model's output. The second component, called contrastiveness, complements the previous component by assessing the attribution method's ability to identify the target class in a mosaic of randomly selected classes. Unlike previous research, the contrastiveness component is designed to allow for soft penalization of attribution methods, depending on the complexity of the observed example.

Furthermore, Chapter 6 provides an overview of the proposed method in the sparsely researched area of alignment methods for evaluating attribution metrics with human quality perception. This chapter begins with a description of previous research and its individual shortcomings. This dissertation proposes a new method for evaluating metrics, where human annotators directly compare the outputs of attribution methods between the original model and the model optimized according to the observed metric. If the annotators consistently prefer the optimized version over the original in a significant number of examples after being presented with both maps, it can be concluded that the observed evaluation metric is well-aligned with human perception. To carry out this process, different proxy metric loss functions for the backward pass and optimization of model parameters were defined to improve the results achieved on a particular metric. To preserve the performance characteristics of the original model, an additional cosine similarity term was added to the total loss, parallel to the loss used in the area of model distillation. After fine-tuning the model using the previously mentioned loss functions, 200 examples were generated for each unique combination of model, attribution method, and metric. Since this research includes two models, two attribution methods, and five metrics, it results in a total of 20 experiments and 4000 examples. The attribution maps of the optimized model, according to the observed metric, should provide a clearer understanding of the model's decision.

Chapters 7 and 8 present the results and discussion of the proposed scientific contributions.

Chapter 7 presents experiments related to the proposed attribution method and evaluation metric. Quantitative experiments include the evaluation of the proposed attribution method and other commonly used methods and those achieving high performance in recent research. These experiments were conducted on two publicly available datasets, and the significance was verified using statistical tests. The proposed method achieves significantly superior results in the experiments using the proposed metric for comprehensive evaluation. Additionally, the advantages and disadvantages of individual methods are highlighted through the employed metric. In addition to the proposed metric, standard metrics from the categories of faithfulness, robustness, and localization described in Chapter 3 were used, on which the proposed method achieves consistently high results. An ablation experiment was conducted to emphasize the benefits the proposed attribution method brings to Transformer explainability. Stopping the attribution propagation before the input pixels and producing relevance maps at the patch level significantly

reduces the attribution method performance. Performance also significantly degrades when only query and key attributions are propagated. Exclusive propagation of relevance values through the self-attention block leads to a moderate decrease in results. This ablation study strongly indicates that relevance propagation through all operations within the Visual Transformer is crucial for achieving high-quality attributions. Qualitative experiments in the chapter visually confirm the advantages of the proposed attribution method, especially in resource-constrained scenarios, highlighting its benefits when using Visual Transformer models with a lower inherent number of image patches. Finally, the robustness of the proposed metric in distinguishing high- and low-quality examples of relevance maps is confirmed. The presented results are published in a scientific paper [1], and the developed code is publicly available [2].

Chapter 8 shows the results of experiments related to the proposed method for aligning evaluation metrics of attribution methods with human quality perception. Experiments were conducted on five commonly used metrics within three categories, a publicly available dataset, and two attribution metrics with two commonly used deep architectures. In the first part of the results, statistically significant improvements in attribution method scores on evaluation metrics after fine-tuning the model on the metrics are highlighted. The values reach nearly ideal results for fidelity metrics, while robustness metrics show an order-of-magnitude improvement. However, from the results of the user study, expert annotators do not observe significant differences between attribution maps of the original and fine-tuned models. In some cases of robustness metrics, the quality of fine-tuned attribution maps is significantly degraded. The localization category offers mixed results, where fine-tuned GradCAM attribution maps improve quality, while GuidedBackprop maps show a decline, influenced by the difference in attribution sparsity between the two methods. This research is published in a scientific paper [3], and the developed code and dataset, including expert labels, are publicly available [4].

Finally, Chapter 9 concludes this doctoral dissertation, listing its scientific contributions and their significance. Future research is recommended to use the proposed method to assess the alignment between metrics and human perception, ensuring effective contributions to human understanding. Additionally, future research could involve evaluating a larger set of existing metrics using an expanded set of attribution methods and datasets. An interesting direction for future studies might include training a reward model on pairs of attribution maps from the conducted user study, replacing human annotators in assessing the alignment between human judgment and evaluation metrics. Furthermore, the reward model could be used instead of traditional metrics to guide fine-tuning in a manner well-aligned with human judgment, similar to the approach supported by reinforcement learning used during the fine-tuning of modern large language models.

Keywords: explainable artificial intelligence, deep learning, transformer, evaluation of attribution methods, human perception

Sadržaj

Poglavlje 1

Uvod

Duboke neuronske mreže (engl. *deep neural networks* (DNN)) postale su ključne za obradu različitih vrsta podataka, što je prije svega omogućeno napretkom sklopovlja za duboko učenje. Ovaj razvoj omogućuje učenje modela na dosad neviđenoj skali u vidu broja parametara i veličine podataka za učenje. Konstantno skaliranje ovih dvaju varijabli dovelo je do najsuvremenijih modela koji nadmašuju i ljudske radne karakteristike na velikom broju zadataka, kao što su klasifikacija slika [5] [6] [7] [8], podržano učenje [9] [10] [11] i razumijevanje prirodnog jezika [12] [13] [14] [15] [16]. Prisutnost velikog broja parametara i višestrukih nelinearnih slojeva doprinosi neprozirnoj prirodi ovih modela, što čini razloge za određeni izlaz izazovnim za objašnjenje. Ova činjenica potiče mnoge istraživače, posebno u područjima gdje je objašnjivost modela ključna, kao što su zrakoplovstvo, medicina ili bankarstvo, da koriste manje, često linearne modele koji su transparentni s obzirom na njihov proces odlučivanja.

Sustavi umjetne inteligencije koriste modele koji se razvijaju kroz proces učenja, a ne putem eksplicitnog ljudskog programiranja. Ovaj proces učenja uključuje izlaganje modela velikim skupovima podataka, omogućujući im samostalno učenje obrazaca i odnosa unutar podataka.

Tijekom učenja, ovi modeli mogu otkriti složene korelacije između ulaznih značajki, poput kliničkih simptoma, i mogu naknadno donositi odluke ili predviđanja, kao što su medicinske dijagnoze. Ovi modeli su često vrlo složeni, posebice u području dubokog učenja, uključujući brojne međudjelujuće parametre kojih je u novije vrijeme nerijetko više od bilijun. Složenost ovih modela dostiže toliku razinu da je čak i stručnjacima u području nemoguće otkriti točne mehanizme putem kojih modeli generiraju svoje izlaze [17].

U takvim scenarijima, obrazloženje iza specifičnih odluka vođenih umjetnom inteligencijom može ostati nejasno i korisnicima i onima na koje ti sustavi utječu. Ova neprozirnost često se naziva efekt "crne kutije" (engl. *black-box*), gdje unutrašnje ponašanje rada modela nije transparentno. Ovaj nedostatak transparentnosti može rezultirati ili pogrešnim povjerenjem od strane korisnika ili pretjeranim oslanjanjem na ove sustave. Oba ishoda su problematična: neopravdano povjerenje može dovesti do bezuvjetnog vjerovanja u točnost sustava bez propit-

kivanja istinitosti izlaznih informacija, dok pretjerano oslanjanje na ovakve sustave može rezultirati da se donošenje ključnih odluka prepusti potencijalno pogrešnim procjenama modela. Posljedično, ovi nedostaci mogu imati značajne negativne implikacije za pojedince, posebno u osjetljivim područjima.

Osim što neprozirnost sustava umjetne inteligencije komplicira razumijevanje njihovih odluka, također posjeduje izravne i značajne utjecaje na pojedince skrivanjem nedostataka modela kao što su pristranost [18] [19] [20], netočnosti ili takozvane "halucinacije" [21] [22] [23] [24] [25] [26] [27] [28]. Loše dizajnirani, razvijeni ili testirani algoritmi mogu proizvesti rezultate koji su potencijalno diskriminatorni ili štetni.

Na primjer, u kontekstu odabira kandidata za posao, modeli umjetne inteligencije mogu nenamjerno favorizirati kandidate iz određenih demografskih skupina zbog pristranih podataka korištenih za učenje. Ako ti sustavi djeluju kao crna kutija, postaje izazovno razlučiti zašto su određeni kandidati odbijeni ili odabrani, čime se kompliciraju napori za prepoznavanje i rješavanje temeljnih pristranosti.

Slično tome, modeli umjetne inteligencije korišteni za medicinsku dijagnostiku mogu pogrešno dijagnosticirati ili zanemariti stanja kod specifičnih demografskih skupina, opet zbog pristranih podataka za učenje [29]. Kada takvi modeli funkcioniraju kao crne kutije, zdravstveni stručnjaci imaju poteškoća razumjeti razloge iza odluka, što otežava ublažavanje potencijalnih pristranosti i osiguranje pravedne zdravstvene skrbi.

Problemi se protežu i izvan diskriminatornih ishoda. Inherentni nedostatak transparentnosti u sustavima umjetne inteligencije sprječava pojedince pogođene automatiziranim odlukama da shvate logiku i utjecaj tih odluka. Ovo je posebno problematično u područjima poput odobravanja kredita, gdje bankovni klijenti često nemaju uvid u automatizirane procese koji utječu na njihove financijske živote.

Štoviše, posljedice neprozirnih sustava umjetne inteligencije posebno su zabrinjavajuće kada se primjenjuju na vladino donošenje odluka. Automatizirani sustavi za donošenje odluka koje koriste vlade mogu značajno utjecati na pojedince, a njihovo djelovanje i sposobnosti možda nisu u potpunosti razumljive ili adekvatno regulirane postojećim zakonodavstvom. Ovaj nedostatak može spriječiti razumijevanje pojedinaca donesene odluke, ograničavajući njihovu sposobnost da osporavaju ili traže pravni lijek protiv potencijalno nepravednih ili pogrešnih ishoda.

Svi navedeni primjeri naglašavaju iznimnu potrebu za većom transparentnošću i odgovornošću u dizajnu, razvoju i primjeni tehnologija koje uključuju modele umjetne inteligencije kako bi se osigurala pravednost i zaštitila prava pojedinaca.

Objašnjiva umjetna inteligencija (engl *Explainable Artificial Intelligence* (XAI)) odnosi se na sposobnost sustava umjetne inteligencije da pruže jasna i razumljiva objašnjenja za svoje radnje i odluke. Primarni cilj XAI-a je razotkriti ponašanje ovih sustava objašnjavajući temeljne

mehanizme i procese koji pokreću njihovo donošenje odluka. Ovaj pristup ima za cilj premostiti jaz između složenih modela i ljudskog razumijevanja, osiguravajući da korisnici mogu shvatiti kako i zašto su određene odluke donesene.

Pristupi objašnjive umjetne inteligencije mogu se široko kategorizirati u dvije vrste: samointerpretirajuće modele i post hoc objašnjenja. Samointerpretirajući modeli dizajnirani su s ugrađenom interpretabilnošću, čineći njihove procese donošenja odluka inherentno razumljivima. Post hoc objašnjenja, s druge strane, nude mogućnost produkcije objašnjenja bez specifičnih modifikacija u arhitekturi modela.

Samointerpretirajući, ili modeli s efektom "bijele kutije" (engl. *white box*), koriste algoritme koji su jednostavni i transparentni, ili uključuju mehanizam za pojednostavljenje određenog dijela modela [30] [31] [32] [33] [34] [35], omogućujući na taj način korisnicima da vide kako ulazni podaci utječu na izlaze ili ciljne varijable te ključne prediktivne značajke mogu biti identificirane i razumljive korisnicima.

Međutim, složena priroda modela dubokog učenja često otežava stvaranje prikaza koji adekvatno objašnjavaju njihov unutarnji rad bez da postanu jednako složeni kao sami modeli [36]. Ova složenost sugerira da je nerealno očekivati da modeli uvijek budu samointerpretirajući.

Za složene sustave, post hoc pristup je često prikladniji. U ovom pristupu, objašnjenja se generiraju nakon što je model donio odluku te ona mogu biti globalna ili lokalna.

Globalna objašnjenja pružaju sveobuhvatno razumijevanje ponašanja AI modela i procesa donošenja odluka, obuhvaćajući opće obrasce, trendove i uvide koji se široko primjenjuju na djelovanje modela, primjerice kako sustav odabire najbolje kandidate za slobodno radno mjesto [37] [38].

Lokalna objašnjenja, nasuprot tome, fokusiraju se na proces donošenja odluka za specifične izlaze, na primjer zašto je određena prijava za posao odbijena. Umjesto da nude širok pregled, lokalna objašnjenja pojašnjavaju ponašanje modela za pojedinačne slučajeve, pomažući razumjeti razloge iza specifičnih predikcija ili odluka.

Područje istraživanja u ovom radu odnosi se na post hoc lokalne metode.

Različiti algoritmi za interpretacije imaju za cilj objasniti donošenje odluka u DNN-u, međutim, niti jedna metoda ne odgovara u potpunosti svakom zadatku i svakom modelu, zbog čega je ovo područje istraživanja vrlo aktivno s ciljem pronalaska algoritma koji bi adekvatno interpretirao čim veći broj kombinacija. Različiti principi primjenjuju se kao metode interpretabilnosti, kao što je isticanje ulaznih značajki na koje se duboki model uglavnom oslanja za određeni zaključak, bilo korištenjem gradijenata [39] [40] [41][42] [43], perturbacija [44] [45] [46] [47] [48] [49] ili posredničkih objašnjivih modela [50] [51] [52]; vizualizacija indirektnih značajki [39] [53] ili vizualizacija protučinjeničnih primjera [54] [55] [56] [57] [58]; analiziranje podataka za učenje kako bi se procijenio doprinos svakog primjera određenom izlazu modela [59] [60]. Fokus ovog rada je na isticanju ulaznih značajki na koje se model najviše oslanjao

tijekom zaključivanja, propagacijom i praćenjem aktivacija slojeva kroz duboku mrežu.

S porastom metoda objašnjive umjetne inteligencije, rasla je i potreba za razvojem učinkovitih metrika za vrjednovanje metoda objašnjive umjetne inteligencije [61] [62] [63] [64] [65]. Ova potreba prepoznata je ne samo unutar područja umjetne inteligencije, već i od strane istraživača unutar područja interakcije čovjek-računalo (engl. *human-computer interaction* (HCI)). Unutar područja istraživanja metrika za vrjednovanje naglašava se važnost mogućnosti objektivnog određivanja kvalitete i učinkovitosti pojedinih metoda objašnjenja.

Dok tradicionalni pokazatelji radnih značajki postoje za procjenu točnosti predikcija i računalne složenosti, vrjednovanje pomoćnih kriterija kao što su razina i kvaliteta interpretabilnosti predstavlja značajne izazove. Ova teškoća doprinosi širokoj varijaciji u metodama objašnjenja, primarno zato što unutar područja objašnjive umjetne inteligencije XAI još uvijek nije postigao konsenzus o standardiziranim metrikama za vrjednovanje, nego se često oslanja na subjektivnu procjenu istraživača temeljenu na pojedinačnim, uvjerljivim primjerima koji prolaze inicijalni test valjanosti. Vrjednovanje koje se temelji samo na intuiciji istraživača nije dovoljno za robusnu procjenu kvalitete [66] [67] [68]. Iz tog razloga, unutar ovog područja trenutno se stavlja značajan naglasak na vrjednovanje pojedinih aspekata objašnjenja kako bi daljnje istraživanje metoda za produkciju objašnjenja bilo moguće validirati, vrjednovati i objektivno usporediti.

Subjektivnu procjenu mapa relevantnosti mogu provesti ljudi, no kvantitativna je mjera presudna za objektivnu usporedbu i rangiranje različitih metoda njihove produkcije. Naglasak u ovom radu je na kvantitativnoj procjeni kvalitete objašnjenja za specifične primjere i neuronske mreže. Ovaj zadatak je iznimno složen, kako kvaliteta objašnjenja ne ovisi samo o metodi za produkciju mapa relevantnosti, već i o izvedbi klasifikatora na koju utječu čimbenici poput arhitekture same mreže te podaci za učenje.

Mape relevantnosti moraju odražavati perspektivu modela, ali uz to se ne nužno striktno pridržavati ljudske procjene kvalitete. Segmentacijska mapa objekta na slici ne može se kvalificirati kao mapa relevantnosti zato što ona potencijalno previđa ključne dokaze određenog objekta koji se mogu naći samo u okolini samog objekta. Iz tog razloga, razvijene su brojne metrike čiji je cilj ocijeniti različite aspekte atribucijskih metoda kao što su vjernost, robusnost i lokalizacija. Međutim, u trenutnoj literaturi ne postoji uspostavljeni pristup za učinkovito kombiniranje rezultata različitih aspekata kako bi se odredila najbolja metode atribucije, otežavajući sveobuhvatnu usporedbu različitih pristupa.

Osim nedostatka istraživanja u području kombinacije različitih aspekata atribucijskih metoda, velika praznina ostaje u području kvantifikacije kvalitete samih metrika za vrjednovanje. Pošto su te metrike uvedene kao način izbjegavanja subjektivne procjene kvalitete atribucijskih mapa istraživača, sve više se javlja potreba za sličnim zahtjevom, ali u sferi metrika za vrjednovanje. Naime, bitno je vrjednovati poravnanje između navedenih metrika i ljudske percepcije kako bi se garantiralo da rezultati koje određene atribucijske metode ostvare na metrikama za

vrjednovanje pozitivno koreliraju s percipiranom kvalitetom ocijenjenom od strane krajnjeg korisnika.

Sadržaj disertacije ukratko je predstavljen u nastavku.

U poglavlju 2 detaljno se opisuje metoda slojevite propagacije relevantnosti koja služi kao temelj za predloženu metodu za produkciju mapa relevantnosti. Detaljno su opisana često korištena pravila unutar ove metode. Također je dan i pregled primjene metode na različite arhitekture, zadatke i područja poput medicine. Nadalje, prikazani su relevantni pristupi iz područja slojevite propagacije relevantnosti, ali i ostalih područja produkcije mapa relevantnosti, te su navedeni njihovi nedostaci. Kraj poglavlja nudi pregled područja objašnjivosti arhitekture Transformer, vezu postojećih istraživanja uz slojevitu propagaciju relevantnosti te opisuje njihove nedostatke.

Poglavlje 3 nudi temeljiti pregled područja vrjednovanja metoda za produkciju mapa relevantnosti. Ovdje se opisuju pristupi vrjednovanju različitih komponenti atribucijskih metoda, poput vjernosti, robusnosti i lokalizacije. Potom se detaljno opisuju pristupi iz navedenih kategorija koji su najrelevantniji za predloženu metriku za vrjednovanje atribucijskih metoda te su istaknuti nedostaci svake pojedine metrike. Naposljetku se predstavlja globalni nedostatak u trenutnim istraživanjima, a to je nedefiniranost načina kombiniranja skupa metrika u jedan sveobuhvatni rezultat.

Poglavlja 4, 5 i 6 predstavljaju znanstvene doprinose autora ove disertacije.

U poglavlju 4 opisuje se predložena metoda za produkciju mapa releventnosti temeljena na slojevitoj propagaciji relativne apsolutne magnitude relevantnosti. Navedena je motivacija te rješenje koje nudi predložena metoda. Opisan je efikasan način implementacije predložene metode unutar modernih razvojnih okvira za duboko učenje. Naposljetku su prikazani detalji vezani uz implementaciju predložene metode za slojeve specifične modernim arhitekturama.

Poglavlje 5 uključuje motivaciju i detaljan pregled dvaju krucijalnih komponenta predložene metrike za vrjednovanje metoda za produkciju mapa relevantnosti temeljene na globalnom vrjednovanju atribucija. Komponenta lokalne konzistentnosti uključuje dvije pod-komponente: vjernost i robusnost, koje se unutar predložene metode računaju na inovativan način. Druga komponenta nazvana kontrastnost upotpunjuje ranije navedenu komponentu ocjenom sposobnosti atribucijske metode da označi ciljnu klasu u mozaiku načinjenom od više slučajno odabranih klasa.

Nadalje, poglavlje 6 daje pregled predložene metode u oskudno istraženom području metoda za vrjednovanje poravnanja metrika za vrjednovanje atribucijskih metoda s ljudskom percepcijom kvalitete. Poglavlje započinje s opisom prethodnih istraživanja te njihovim pojedinačnim nedostacima. Potom se daje pregled predložene metode te njezinih mehanizama: finopodešavanje modela na samim metrikama za vrjednovanje te korisnička studija, u kojoj pet stručnjaka u području dubokog učenja vrjednuje percipiranu kvalitetu izvornih i fino podešenih

atribucijskih mapa.

Poglavlja 7 i 8 prikazuju rezultate i diskusiju predloženih znanstvenih doprinosa.

U poglavlju 7 prezentiraju se eksperimenti vezani uz predloženu atribucijsku metodu i metriku za vrjednovanje. Kvantitativni eksperimenti uključuju vrjednovanje predložene atribucijske metode i ostalih, često korištenih metoda i metoda koje u recentnim istraživanjima ostvaruju najviše radne značajke. Navedeni eksperimenti izvedeni su na dva javno dostupna skupa podataka te su značajnosti provjerene statističkim testom. Predložena metoda ostvaruje značajno superiorne rezultate u izvedenim eksperimentima koristeći predloženu metriku za sveobuhvatno vrjednovanje. Dodatno, s pomoću same korištene metrike moguće istaknute su prednosti i nedostatke pojedinih metoda. Osim predložene metrike, za vrjednovanje su korištene i standardno korištene metrike iz kategorija vjernosti, robusnosti i lokalizacije opisane u ranijem poglavlju 3 na kojima predložena metoda ostvaruje konzistentno visoke rezultate. Izveden je i ablacijski eksperiment kojim se ističu prednosti koje predložena atribucijska metoda uvodi u području objašnjivosti Transformera. Kvalitativni eksperimenti u nastavku poglavlja ilustriraju i vizualno potvrđuju prednosti predložene atribucijske metode, posebice u slučajevima niske količine resursa, gdje su ukazane prednosti predložene metode u slučaju korištenja modela Vizualnog Transformera s manjim inherentnim brojem komadića slike. Naposljetku, potvrđuje se robusnost predložene metrike u determinaciji visokokvalitetnih te niskokvalitetnih primjera mapa relevantnosti. Prikazani rezultati objavljeni su u obliku znanstvenog članka [1], a izrađen programski kod je javno dostupan [2].

Poglavlje 8 prikazuje rezultate eksperimenata vezanih uz predloženu metodu za vrjednovanje poravnanja metrika za vrjednovanje atribucijskih metoda s ljudskom percepcijom kvalitete. Eksperimenti su izvedeni na pet često korištenih metrika unutar tri kategorije, javno dostupnom skupu podataka te dvije atribucijske metrike uz dvije standardno korištene duboke arhitekture. U prvom dijelu prikazanih rezultata ističu se statistički značajna poboljšanja u ocjenama atribucijskih metoda ostvarenih na metrikama za vrjednovanje nakon finog podešavanja modela na samim metrikama. Rezultati iz kategorija vrijednosti dostižu gotovo idealne vrijednosti dok rezultati kategorije robusnosti dostižu promjenu za red veličine. Međutim, uvidom u rezultate korisničke studije, ispostavlja se da stručni anotatori ne primjećuju značajnu razliku između atribucijskih mapa izvornog i fino podešenog modela. Štoviše, u nekim slučajevima metrike robusnosti, kvaliteta fino podešenih atribucijskih mapa je značajno degradirana. Scenariji lokalizacijske kategorije nude mješovit zaključak, gdje fino podešene atribucijske mape GradCAM metode ostvaruju poboljšanje u kvaliteti dok GuidedBackprop mape ostvaruju pogoršanje. Ovi mješoviti rezultati uzrokovani su razlikom u razini prorijeđenosti atribucija dvije metode. Navedeno istraživanje objavljeno je u obliku znanstvenog članka [3], a izrađen programski kod te skup podataka koji uključuje oznake stručnjaka javno su dostupni [4].

Naposljetku, poglavljem 9 zaključuje se ova doktorska disertacija, navode se njeni znans-

tveni doprinosi te njihova važnost. Također se ističu mogućnosti budućeg istraživanja temeljene na prikazanim doprinosima.

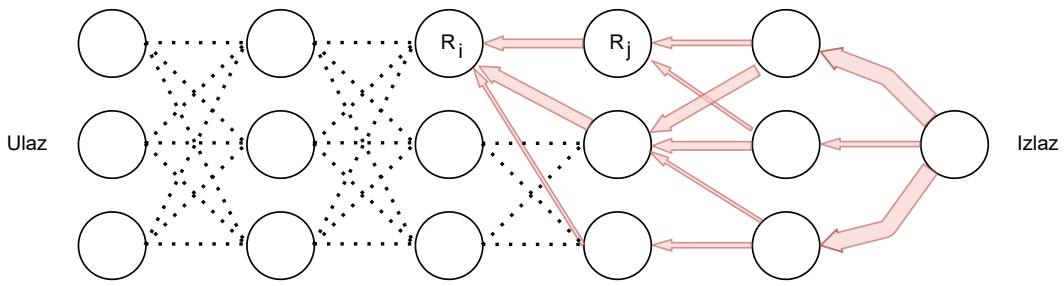
Poglavlje 2

Metode za produkciju mapa relevantnosti

Ovo poglavlje pruža pregled recentnih istraživanja u području post hoc metoda za produkciju mapa relevantnosti. Započinje detaljnim pregledom primarnog područja ovog rada — slojevitoj propagaciji relevantnosti te se ističu nedavni napreci u ovom području. Nakon toga, razmatraju se često korištene i metode s najboljim radnim značajkama temeljene na alternativnim mehanizmima produkcije atribucijskih mapa. Na kraju, detaljno se prikazuju istraživanja specifično vezana uz post hoc metode za produkciju mapa relevantnosti korištenim na sve prisutnijoj Transformer arhitekturi.

2.1 Slojevita propagacija relevantnosti

Bach i sur. [69] predstavili su algoritam za određivanje relevantnosti određenog ulaznog neurona za izlaz nelinearne neuronske mreže nazvan *slojevita propagacija relevantnosti* (LRP). LRP pretpostavlja da se klasifikator može rastaviti na nekoliko slojeva izračuna, koji mogu biti dijelovi faze ekstrakcije značajki ili dijelovi modela za klasifikaciju koji radi na izračunatim značajkama. Slika 2.1 predstavlja ilustraciju ove metode, gdje je vidljivo kako se relevantnost propagira unatrag od odabranog izlaznog neurona prema ulaznim neuronima.



Slika 2.1: Ilustracija algoritma slojevite propagacije relevantnosti. Svaki neuron određenog sloja redistribuira relevantnost neuronima nižeg sloja proporcionalno procijenjenoj relevantnosti putem LRP pravila.

S obzirom na ulaz x i neuronsku mrežu f, cilj LRP-a je dodijeliti svakoj poziciji ulaza p (na primjer, u slučaju slika, svakom pikselu) ocjenu relevantnosti R_p^0, gdje superskript 0 označava prvi, ulazni sloj. Pretpostavljajući poznavanje mape relevantnosti posljednjeg sloja R_p^l, cilj određene LRP formulacije je opisati kako raspršiti relevantnost izlaznog neurona j na svaki od ulaznih neurona i u sloju prije - $R_{i\leftarrow j}^{(l-1,l)}$, tako da vrijedi sljedeća jednadžba:

$$R_i^{(l-1)} = \sum_{j\in(l)} R_{i\leftarrow j}^{l} \tag{2.1}$$

Pristupajući navedenoj jednadžbi iterativno po slojevima modela, počevši od izlaza prema ulazu, dobivamo relevantnosti svakog ulaznog neurona. Za određeni sloj neuronske mreže s aktivacijskom funkcijom g i ulazom x, aktivacija izlaznog neurona u sljedećem sloju definirana je kao:

$$a_j = g(\sum_i x_i w_{ij}) \tag{2.2}$$

Izraz x_0 postavljen je na 1 tako da w_{0j} predstavlja pristranost izlaznog neurona j. Bach i sur. [69] predlažu nekoliko formula za izračunavanje R^{l-1} iz rezultata relevantnosti sljedećeg sloja R^l.

$$\varepsilon\text{-pravilo (LRP-}\varepsilon\text{):} \qquad R_i^{l-1} = \sum_j \frac{x_i w_{ij}}{\sum_k x_k w_{kj} + \varepsilon} R_j^l \tag{2.3}$$

Ovo pravilo redistribuira relevantnost sljedećeg sloja prema sloju prije na temelju udjela doprinosa svakog ulaznog neurona aktivaciji izlaznog neurona. Uloga parametra ε je apsorbirati dio relevantnosti kada su doprinosi aktivaciji neurona k slabi ili kontradiktorni. Povećanjem vrijednosti parametra ε, samo najznačajniji čimbenici objašnjenja preživljavaju apsorpciju. To obično rezultira atribucijama koje su reducirane u smislu broja istaknutih ulaznih značajki te sadrže manje šuma. Približavanjem vrijednosti 0, uniformnom primjenom ovog pravila kroz mrežu pokazuje se da se produciraju mape relevantnosti ekvivalentne mapama dobivenim Input*Gradient metodom [70].

$$\alpha\beta\text{-pravilo (LRP-}\alpha\beta\text{):} \qquad R_i^{l-1} = \sum_j (\alpha \frac{(x_i w_{ij})^+}{\sum_k (x_k w_{kj})^+} - \beta \frac{(x_i w_{ij})^-}{\sum_k (x_k w_{kj})^-}) R_j^l \tag{2.4}$$

LRP-$\alpha\beta$ odvojeno tretira pozitivne i negativne aktivacije, uvodeći hiperparametre unutar ove metode, α i β koji služe za određivanje relativne važnosti između pozitivne i negativne relevantnosti unutar sloja. Bach i sur. [69] pokazali su da vrijednosti $\alpha = 2$ i $\beta = 1$ proizvode oštre mape relevantnosti bez šuma.

$$\gamma\text{-pravilo (LRP-}\gamma\text{):} \qquad R_i^{l-1} = \sum_j \frac{x_i (w_{ij} + \gamma w_{ij}^+)}{\sum_k x_k (w_{kj} + \gamma w_{kj}^+)} R_j^l \tag{2.5}$$

LRP-γ uvodi još jedno poboljšanje kako bi se dala prednost pozitivnim doprinosima nad negativnim. Parametar γ kontrolira u kojoj mjeri su pozitivni doprinosi favorizirani. Kako se ovaj parametar povećava, tako se negativni doprinosi smanjuju. Dodatno, dominacija pozitivnih doprinosa ograničava veličinu i pozitivne i negativne relevantnosti tijekom faze propagacije, rezultirajući stabilnijim objašnjenjima. Naposljetku, približavanjem parametra γ beskonačnoj vrijednosti ovu metodu čini ekvivalentnom LRP-$\alpha_1\beta_0$ metodi, pravilu z+ [71] i metodi *excitation-backprop* [72].

Kontrastna slojevita propagacija relevantnosti rješava nedostatak primijećen kod većine predloženih LRP pravila, a to je nedostatak konstrastnosti, odnosno sposobnosti metode u isticanju ciljnog objekta ako je model istovremeno izložen nekoliko različitih klasa objekata. Ova metoda mijenja definiciju početne relevantnosti LRP algoritma, to jest relevantnosti izlaznog sloja R_p^l, iz vektora koji sadrži vrijednost izlaza ciljne klase z_t na njezinoj poziciji t i nule za ostale klase u vektor koji je definiran na sljedeći način:

$$R_p^l = \begin{cases} z_t & \text{ako } p = t \\ -\frac{z_t}{N-1} & \text{inače} \end{cases}$$

gdje je N ukupan broj klasa. Na ovaj način, relevantnost ciljne klase odvaja se od ostalih klasa, što dovodi do mogućnosti vizualizacije regija ulazne slike koje odgovaraju odabranim ciljnim klasama.

Predložena LRP pravila široko su prihvaćena u raznim područjima istraživanja strojnog učenja. Uspješno su implementirana u području obrade prirodnog jezika [73], gdje je primjena ovih pravila na mrežu za klasifikaciju teme zadanog teksta dala superiorne rezultate u usporedbi s metodama temeljenima na gradijentima. Ova metoda je naknadno proširena na složenije arhitekture, kao što su modeli *Long Short-Term Memory* (LSTM) [74] i *Gated Recurrent Unit* (GRU) [75]. U tim kontekstima, LRP pravila su značajno nadmašila osnovne pristupe temeljene na gradijentima u zadatku predikcije sentimenta s pet klasa. Autori su riješili izazov dvosmjernih multiplikativnih interakcija tretirajući ulaz za regulaciju kao konstantu tijekom unatražne propagacije relevantnosti [76].

LRP pravila često su primjenjivana u sferi medicine. Yang i sur. [77] primjenjuju algoritam u kontekstu isticanja značajki koje dovode do probabilističke predikcije terapijskih odluka za svakog pojedinog pacijenta. Također vrjednuju objašnjenja usmjerena na značajke generirane od strane LRP-a te ih izlažu kliničkim stručnjacima. Pokazuje se da se značajke koje su identificirane kao relevantne uglavnom slažu s kliničkim znanjem i smjernicama. Böhle i sur. [78] primjenjuju LRP na ulazne slike magnetske rezonance. Pokazali su da metoda uspješno identificira zasićene regije hipokampusa unutar temporalnog režnja slike koje ukazuju na postojanje Alzheimerove bolesti. Dodaju zaključak kako bi LRP metoda mogla pružiti značajnu pomoć prilikom dijagnosticiranja ove bolesti. Eitel i sur. [79] slično koriste 3D konvolucijsku

mrežu kako bi dijagnosticirali multiplu sklerozu koristeći snimke magnetske rezonance. LRP atribucije otkrivaju da se konvolucijska mreža zaista fokusira na pojedinačne lezije, ali također uključuje dodatne informacije poput lokacije lezije, ne-lezionalnih bijelih i sivih područja poput talamusa, koji su uspostavljeni konvencionalni i napredni markeri magnetske rezonance u dijagnostici multiple skleroze. Autori zaključuju kako bi predloženi sustav konvolucijske mreže i LRP objašnjenja mogao poslužiti kao osnova klasifikacijske odluke za kliničku procjenu, provjeru dijagnostički relevantnih značajki i potencijalno prikupljanje novih saznanja o bolesti. Nam i sur. [80] koriste LRP kao metodu odabira relevantnih značajki koje koriste za fino podešavanje konvolucijske mreže i time postižu napredak u radnim značajkama na zadatku klasifikacije signala elektroencefalografije kako bi omogućili korisniku kontrolu određenih uređaja bez stvarnog pomicanja udova.

2.1.1 Relativna Propagacija Atribucije

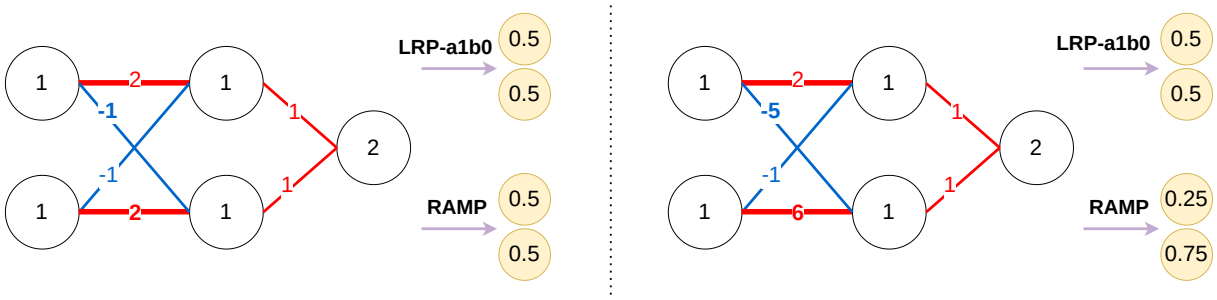
Autori metode Relativna Propagacija Atribucije (engl. *Relative Attributing Propagation* (RAP)) (Nam i sur. [81]) skreću pažnju na ključno ograničenje inherentno uobičajenoj LRP metodologiji prilikom atribuiranja relevancije unatrag na ulazni sloj. Konkretno, identificiraju problem u kojem neuroni mogu primiti kontradiktorne signale relevantnosti, poput jednog značajno pozitivnog neurona i drugog značajno negativnog neurona, iz njihovog sljedećeg sloja. U takvim slučajevima, ti konfliktni signali često se poništavaju, usprkos njihovom značajnom utjecaju na konačni izlaz modela. Iz tog razloga rezultirajuće mape relevantnosti postaju iznimno prorijeđene te su izazovne za interpretaciju, često manifestirajući dominaciju na odabranim lokacijama gdje jedan doprinos, bilo pozitivan ili negativa, pretežno prevladava.

RAP rješava ovaj problem osmišljavanjem metodologije koja usklađuje pozitivne i negativne atribucije iskorištavajući njihove relativne magnitude unutar apsolutne vrijednosti svakog doprinosa. Operativno, RAP provodi dva prolaza LRP-$\alpha_1\beta_0$: jedan za pozitivnu komponentu i drugi za negativnu komponentu sljedećeg sloja. Ti prolazi zatim se ponderiraju prema njihovim relativnim veličinama unutar apsolutne sume i naknadno se agregiraju. Nadalje, kako bi se negirala prekomjerna alokacija relevantnosti prouzrokovana dodatkom izvorno negativne relevantnosti neuronima, RAP metoda dodatno uključuje korektivnu mjeru oduzimanjem srednje razlike između suma relevantnosti prethodnog i trenutnog sloja od svih vrijednosti neurona s aktivacijom različitom od nula, što se izvršava nakon izračuna atribucija svakog sloja.

Formalno, metodologija Relativne Propagacije Atribucije definirana je na sljedeći način:

$$R^{l-1}_{i\in\mathscr{P},\mathscr{N}} = \sum_j (\alpha \frac{(x_i w_{ij})^+}{\sum_k (x_k w_{kj})^+} + \beta \frac{(x_i w_{ij})^-}{\sum_k (x_k w_{kj})^-}) R^l_j \tag{2.6}$$

$$\alpha = \frac{\sum_i (x_i w_{ij})^+}{\sum_i ((x_i w_{ij})^+ + |(x_i w_{ij})^-|)} \qquad \beta = \frac{\sum_i (x_i w_{ij})^-}{\sum_i ((x_i w_{ij})^+ + |(x_i w_{ij})^-|)} \tag{2.7}$$



Slika 2.2: Jednostavan primjer koji ilustrira ključne razlike između LRP-$\alpha_1\beta_0$ i predložene metode. Bijeli krugovi predstavljaju ulaze i aktivacije slojeva, s težinama na povezanim linijama, dok su konačne vrijednosti relevantnosti prikazane u žutim krugovima. U lijevom primjeru, jednake relativne apsolutne magnitude aktivacija rezultiraju sličnim atribucijama između LRP-$\alpha_1\beta_0$ i predložene metode. U desnom primjeru, znatno različite magnitude aktivacija dovode našu metodu integracije informacije o većoj apsolutnoj aktivaciji donjeg skrivenog neurona prilikom produkcije mape relevantnosti.

$$R_i^{l-1} = R_{i\in\mathscr{P},\mathscr{N}}^{l-1} - \Psi_i^{l-1} \tag{2.8}$$

gdje je Ψ_i^{l-1} srednja vrijednost svih neurona s aktivacijom različitom od nula u sloju $l-1$.

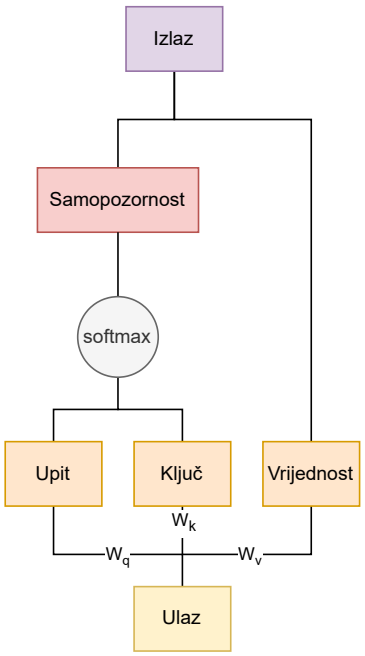
2.2 Objašnjivost Transformera

Transformeri su postali temelj istraživanjima u području strojnog učenja, demonstrirajući vrhunske radne značajke u različitim područjima poput obrade slika [82] [83] [84], razumijevanja prirodnog jezika [85] [15], generiranja slika [86] [87] [88], i podržanog učenja [89]. Njihova sveprisutna primjena ističe nužnost razvoja alata za objašnjivost koji mogu pomoći u ispravljanju modela, osiguravajući pritom pravednost i nepristranost u njihovim primjenama.

Vizualni Transformer [82] je duboki model koji prilagođava izvornu Transformer arhitekturu [90], korištenu u području obrade prirodnog jezika, obradi ulaza u obliku slike. Dok izvorna arhitektura koristi enkoder i dekoder za procesiranje značajki i vršenje predikcije, arhitektura Vizualnog Transformera koristi isključivo enkoder za obradu podataka. U zadatku klasifikacije slika, ulazna slika $x \in \mathbb{R}^{H\times W\times C}$ se inicijalno dijeli na N komadića slika $x_k^i \in \mathbb{R}^{K\times K\times C}$, gdje P odgovara veličini jednog komadića slike u pikselima, a $N = \frac{HW}{P^2}$. Potom se ekstrahirani komadići izravnaju u vektor te se linearno preslikaju u vektor s brojem dimenzija D. Na kraju pretprocesiranja komadićima se dodaju učene pozicijske značajke kako bi se očuvala informacija o njihovoj izvornoj poziciji unutar slike. Ovi komadići se potom se koriste kao ulaz u enkoder na identičan način kao što se tokeni koriste u izvornom Transformeru. Konkatenacija dodatnog, učenog tokena na početak ove sekvence nudi mogućnost klasifikacije nad cijelom sekvencom, slično CLS tokenu unutar BERT arhitekture [85], pošto ovaj token nakon prolaska kroz cijeli model nudi globalnu informaciju i sadržaju slike.

U srcu Transformer arhitekture leže slojevi samopozornosti, koji olakšavaju preslikavanje

upita i parova ključ-vrijednost kako bi se generirao izlaz. Ilustracija ovog mehanizma prikazana je na Slici 2.3. Samopozornost započinje preslikavanjem ulaza sloja u upite, ključeve i vrijednosti $q, k, v \in \mathbb{R}^{N \times D}$ pomoću učenih linearnih transformacija W_q, W_k, W_v. Potom se upiti i ključevi množe te vrijednosti umnožaka prolaze kroz softmax nelinearnost koja efektivno određuje količinu pozornosti koju će svaki komadić slike pridonijeti svakom drugom komadiću slike unutar ulazne sekvence. Tenzor ovih vrijednosti naziva se tenzor samopozornosti. Naposljetku se tenzor samopozornosti množi s vrijednostima iz ranije transformacije kako bi se dobio izlaz sloja samopozornosti. Dodatno, u novijim arhitekturama, uključujući izvorni i Vizualni Transformer, samopozornost se izvodi nekoliko puta paralelno, tvoreći mehanizam višeglave samopozornosti. Nakon sloja samopozornosti, podaci prolaze kroz sloj slojne normalizacije (engl. *layer normalization*) [91], uključuje se rezidualna veza [92] te potom prolaze kroz dva potpuno povezana sloja prije nego ponovno prolaze kroz normalizaciju i rezidual. Ove operacije tvore jedan sloj enkodera transformera. Povezivanje više slojeva zaredom tvori finalnu arhitekturu. Brojna istraživanja iskoristila su mehanizam samopozornosti kako bi izvela relevantne rezultate [90] [93] [94] [95].



Slika 2.3: Ilustracija mehanizma samopozornosti unutar transformer arhitekture. Spajanje linija predstavlja množenje tenzora.

U sferi objašnjive umjetne inteligencije vrijednosti posljednjeg sloja samopozornosti mogu se direktno koristiti kao mapa relevantnosti na razini komadića slike. Primjerice, Ukwuoma i

sur. [96] koriste Vizualni Transformer kao klasifikator na zadacima binarne i višeklasne klasifikacije upale pluća te vizualiziraju vrijednosti samopozornosti u pojedinim glavama modela. Pokazuju da se pozornost fokusira na područja relevantna za klasifikaciju.

Xu i sur. [97] izravno koriste mehanizam pozornosti kako bi prikazali modelu relevantna područja slike prilikom generiranja njenog tekstualnog opisa.

Carion i sur. [98] uče model detekcije objekata pomoću Transformer arhitekture pritom izbjegavajući standardno korištene komponente poput nemaksimalne supresije i temeljnih okvira. Autori koriste samopozornost posljednjeg sloja kako bi kvalitativno pokazali fokus modela na relevantna područja slike.

Caron i sur. [7] predlažu novu metodu za samonadzirano učenje te pokazuju kako modeli producirani ovom metodom učenja nakon finog podešavanja mogu nadmašiti radne značajke potpuno nadziranih modela. Vizualizacijom vrijednost samopozornosti pokazuju da model učen na ovaj način najviše vrijednosti pozornosti drži nad maskom ciljnog objekta.

Vukadin i sur. [93] koriste zadnji sloj pozornosti kako bi kvalitativno vrjednovali sposobnost fino podešenog BERT modela čiji je zadatak klasifikacija semantičkih sekcija životopisa u praćenju tokena povezanih s klasom određene sekcije. Autori potvrđuju da model, primjerice, prilikom klasifikacije sekcije osobnih podataka, najviše pozornosti obraća na tokene klase ime, prezime, adresa, broj mobitela te e-pošte, što odgovara i ljudskom načinu klasifikacije ovog tipa sekcije.

Alternativno, drugi pristupi uključuju kombiniranje izlaza više slojeva pozornosti kako bi se izvelo konačno objašnjenje. Međutim, jednostavno uprosječivanje izlaza sloja samopozornosti po svim slojevima može rezultirati pomućenim signalom relevantnosti jer ova metoda ne uzima u obzir različite uloge svakog sloja samopozornosti [99].

Korisnost izvornog mehanizama pozornosti u pružanju smislenih objašnjenja za odluke modela pomno je proučena od strane Jain i sur. [100]. U svojem istraživanju, autori pokazuju da je korelacija između pozornosti i drugih metoda za produkciju relevantnosti značajki, kao što su metode temeljene na gradijentu i brisanju značajki, slaba kada se koriste povratni enkoderi. Nadalje, autori istražuju suparničke distribucije pozornosti koje, unatoč tome što nude znatno različita objašnjenja za izlaz modela, rezultiraju samo skromno različitim konačnim izlazima. Iako je poznato da mehanizmi pozornosti značajno poboljšavaju radne značajke i skalabilnost modela [101] [102] [103], autori tvrde da njihova učinkovitost u sferi objašnjive umjetne inteligencije ostaje upitna.

Međutim, rad Wiegreffe i sur. [104] osporava prethodno istraživanje u kojem nalaze nekoliko nedostataka kao na primjer činjenicu da distribucija pozornosti nije jednostavno zamjenjiva. Odvajanje ocjena pažnje od modela narušava njegov integritet, budući da su te ocjene generirane od strane ključne komponente optimizirane tijekom inicijalnog učenja modela. Drugo, autori ovog rada napominju da vrijednosti pozornosti ne nude jedino objašnjenje, nego nude jedno

objašnjenje, tako da pronalazak alternativnih distribucija koje proizvode isti izlaz nije neočekivana posljedica te ne isključuje važnost pozornosti kao alata za objašnjenje izlaza mreže.

Osim korištenja zadnjeg sloja samopozornosti kao metode za produkciju mapa relevantnosti, nova istraživanja fokusiraju se na alternativne metode koje nadograđuju ovu osnovnu metodu kroz agregiranje više slojeva samopozornosti, dodatak informacija o relevantnosti iz vrijednosti ili gradijenta slojeva samopozornosti te uključivanjem LRP metode u izračun vrijednosti relevantnosti. Istraživanje u području objašnjivosti Transformera ostaje relativno ograničeno te se u sljedećoj sekciji temeljito opisuju trenutno istražene metode iz ovog područja.

2.2.1 Raspodjela Pozornosti

Raspodjela Pozornosti (engl. *Attention rollout*) (Abnar i sur. [105]) predstavlja intuitivan pristup praćenju protoka informacija koji se širi od ulaznog sloja do vektorskih reprezentacija (engl. *embeddings*) u višim slojevima unutar arhitekture Transformera. U suštini, cilj ove metode je izračunati mehanizam pozornosti od svih položaja u određenom sloju i do svih položaja u sloju j, gdje je $j < i$. Unutar grafa pozornosti, put od čvora v na položaju k u sloju l_i do čvora u na položaju m u sloju l_j obuhvaća niz bridova koji povezuju ta dva čvora. Promatrajući težinu svakog brida kao pokazatelj udjela informacija prenesenih između povezanih čvorova, moguće je procijeniti u kojoj mjeri se informacije s čvora v prenose na čvor u putem određenog puta, množeći težine svih bridova duž tog puta. S obzirom na mogućnost postojanja više različitih puteva između dva čvora unutar grafa pozornosti, izračun ukupne količine prenesenih informacija zahtijeva zbrajanje svih mogućih puteva koji povezuju te čvorove. Kako bi se implementirao opisani mehanizam računanja pozornosti od sloja l_i do sloja l_j, izvodi se rekurzivno množenje matrica pozornosti preko svih slojeva do l_j. Ovaj rekurzivni proces učinkovito obuhvaća kumulativni utjecaj mehanizama pozornosti preko više slojeva, omogućujući time sveobuhvatno razumijevanje protoka informacija unutar arhitekture Transformera. Formalno, Raspodjela Pozornosti je definirana kao:

$$\text{Rollout}(x) = \hat{\mathbf{A}}^{i} \times \hat{\mathbf{A}}^{i-1} \times ... \times \hat{\mathbf{A}}^{j} \tag{2.9}$$

$$\hat{\mathbf{A}}^{b} = \mathbf{I} + \mathbb{E}_h \mathbf{A}^{b} \tag{2.10}$$

Gdje $\mathbf{A}^{b} \in \mathbb{R}^{h \times s \times d_h}$ označava mapu pozornosti u sloju b. $\mathbb{E}_h$ je usrednjena vrijednost mape pozornosti po dimenziji glave operacije pozornosti.

Kvaliteta metode istražena je kroz nekoliko istraživanja. Mondal i sur. [106] uče Vizualni Transformer na zadatku klasifikacije CT slike pluća u tri kategorije: normalna, upala pluća ili COVID-19. Autori koriste metodu raspodjele pozornosti kako bi vizualizirali područja CT snimaka odgovornih za određenu klasifikaciju. Stručno mišljenje radiologa potvrđuje kvalitetu

proizvedenih mapa relevantnosti. Xue i sur. [107] koriste Vizualni Transformer kako bi analizirali fotografije različitih regija usne šupljine i odabrali one koje potencijalno sadrže leziju. Vizualizacije atribucija raspodjele pozornosti potvrđuju kvalitetu fino ugođenog modela. Ha i sur. [108] upotrebljavaju metodu kako bi objasnili klasifikacije slika vlasišta u nekoliko klasa povezanih sa standardnim bolestima i stanjima vezanim uz vlasište. Autori zaključuju da metoda uspješno lokalizira relevantne dijelove slike koji objašnjavaju ciljnu klasifikaciju. Gupta i sur. [109] proširuju originalnu arhitekturu *dropout* slojem baziranim na komadićima slike te koriste raspodjelu pozornosti u zadatku slabo nadzirane lokalizacije objekta.

2.2.2 Interpretabilnost Transformera izvan vizualizacije pozornosti

Chefer i sur. [99] uvode metodu pod nazivom Interpretabilnost Transformera izvan vizualizacije pozornosti (engl. *Transformer Interpretability Beyond Attention Visualization* (TIBAV)) koja proširuje mogućnosti metode Raspodjele Pozornosti uključivanjem LRP-a u izračun raspodjele relevantnosti po sloju samopozornosti.

Nakon što se dobiju ocjene relevantnosti za svaku glavu pozornosti u svakom sloju, one se integriraju kroz graf pozornosti na sličan način kao i u Raspodjeli Pozornosti. Sam proces određivanja atribucija pojedinog sloja uključuje informaciju o gradijentu pozornosti s obzirom na odabrani izlaz te informaciju o relevantnosti koja dolazi iz LRP-$\alpha_1\beta_0$ metode. Kombinacija dvaju pristupa olakšava uklanjanje negativnih doprinosa, nudi mogućnost različite atribucijske mape za različite odabrane klase te poboljšava interpretabilnost i učinkovitost u hvatanju značajnih uzoraka unutar ulaznih podataka.

Uz definiciju $\mathbf{R}^{\mathbf{A}^b}$ kao relevantnost dobivena LRP-$\alpha_1\beta_0$ metodom za svaki sloj samopozornosti b pomnoženu s gradijentom izlazne vrijednosti odabrane klase s obzirom na tenzor samopozornosti svakog sloja $\nabla\mathbf{A}^b$ te činjenicu da se sa svakog sloja dalje propagiraju samo pozitivne vrijednosti relevantnosti, TIBAV metoda je formalno definirana na sljedeći način:

$$\text{TIBAV} = \hat{\mathbf{A}}^1 \times \hat{\mathbf{A}}^2 \times ... \times \hat{\mathbf{A}}^B \tag{2.11}$$

$$\hat{\mathbf{A}}^b = \mathbf{I} + \mathbb{E}_h(\nabla\mathbf{A}^b \odot \mathbf{R}^{\mathbf{A}^b})^+ \tag{2.12}$$

Autori metodu vrjednuju na nizu metrika koristeći dvije verzije Transformer arhitekture, BERT u domeni teksta i Vizualni Transformer u domeni slika. Vrjednovanja pripadaju kategoriji vjernosti, gdje se MoRF ili LeRF načinom maskiraju pojedini tokeni kao ulazi u tekstualni model te se promatra pad, odnosno održavanje točnosti modela; te kategoriji lokalizacije gdje se uspoređuje sklad atribucije ulazne slike ciljnoj segmentacijskoj mapi.

2.2.3 Konzervativna propagacija relevantnosti

Ali i sur. [110] uvode novu metodu za interpretabilnost Transformer modela pod nazivom Konzervativna propagacija relevantnosti (engl. *Conservative propagation*) koristeći Input*Gradient kao osnovnu atribucijsku metodu. Kao što je spomenuto ranije u poglavlju, ova metoda je povezana s LRP-ε metodom za male vrijednosti ovog parametra. Poveznica s LRP-om autorima nudi mogućnost dubljeg pristupa analizi atribucija koje producira Input*Gradient, omogućujući identifikaciju slojeva ili komponenti unutar neuronske mreže gdje se princip očuvanja relevantnosti potencijalno narušava.

Ukoliko se x i y definiraju kao ulaz i izlaz određenog sloja, a o^c izlaz mreže za određenu klasu, Input*Gradient atribucije za ulaz i izlaz sloja se definiraju kao:

$$\mathrm{R}(x) = x\frac{\partial o^c}{\partial x} \tag{2.13}$$

$$\mathrm{R}(y) = y\frac{\partial o^c}{\partial y} \tag{2.14}$$

Uzimajući u obzir pravilo ulančavanja za izračun gradijenata između dva sloja neuronske mreže:

$$\frac{\partial o^c}{\partial x} = \frac{\partial o^c}{\partial y}\frac{\partial y}{\partial x} \tag{2.15}$$

Kombinacijom ovih triju jednadžbi dobiva se sljedeća jednakost:

$$\mathrm{R}(x) = \frac{\partial y}{\partial x}\frac{x}{y}\mathrm{R}(y) \tag{2.16}$$

Autori su koristeći ovu formulaciju otkrili dva sloja unutar Transformer arhitekture koji ne održavaju princip očuvanja relevantnosti $\sum \mathrm{R}(x) = \sum \mathrm{R}(y)$ - propagacija relevantnosti unutar slojeva pozornosti i slojne normalizacije.

Problem očuvanja relevantnosti je za sloj pozornosti riješen tako da su se faktori pozornosti, odnosno izlaz softmax funkcije, unutar sloja promatrali kao konstante te se relevantnost propagira samo kroz vektore vrijednosti unutar sloja samopozornosti. Kod sloja slojne normalizacije standardna devijacija kojom se normaliziraju vrijednosti sloja je promatrana, na sličan način, kao konstanta. Pristup Ali i sur. paralelan je pristupu Arras i sur. [76] koji promatraju propuštajuće vrijednosti (engl. *gating values*) unutar LSTM-a kao konstante.

Slično kao i prethodno istraživanje, autori koriste metrike iz kategorije vjernosti za vrjednovanje predložene metode, gdje MoRF i LeRF načinom maskiraju čvorove unutar grafa u slučaju zadatka obrade grafova te tokene u slučaju obrade teksta.

2.3 Nedostaci trenutnih istraživanja u području slojevite propagacije relevantnosti

Bez obzira na napredak koji ističu prethodno opisane metode, važno je primijetiti i njihove nedostatke.

Metoda Relativne Propagacije Atribucije posjeduje dva krucijalna nedostatka. Prvi se odnosi na samu LRP-$\alpha_1\beta_0$ metodu. Ispitujući jednadžbu propagacije za LRP-$\alpha\beta$ (Jednadžba 2.4), primjećuje se da za pozitivni (α) dio atribucije nazivnik sadrži samo pozitivni dio vrijednosti sljedećeg sloja, a obrnuto, negativni (β) dio sadrži samo zbroj negativnih vrijednosti sloja. Pretpostavlja se da upravo ovo opažanje dovodi do lošeg relativnog dodjeljivanja atribucija kada postoji velika količina i pozitivnih i negativnih aktivacija u sloju. Slika 2.2 sadrži prikaz jednostavnog modela koji ilustrira opisani nedostatak. Oba primjera imaju istu arhitekturu, ulazne, skrivene i izlazne vrijednosti. Međutim, u slučaju mreže na desnoj strani, drugi skriveni neuron ima težine veće magnitude ($[-5,6]$). Koristeći LRP-$\alpha_1\beta_0$ pravilo i propagirajući unazad prema ulazu za obje mreže, dobiva se atribucijska mapa istovjetna lijevom modelu koja iznosi $[0.5,0.5]$, ako se ona normalizira tako da je apsolutni zbroj atribucijskih ocjena jednak 1. Koncept ujednačene atribucije među ulaznim neuronima proturječi stvarnoj dinamici unutar mreže. Ova činjenica postaje očita promatranjem apsolutne magnitude aktivacija u drugom skrivenom neuronu, koje su značajno veće s iznosom 11 u usporedbi s magnitudama prvog neurona od 3. U takvim slučajevima, razumno je očekivati da drugi ulazni neuron nosi više pozitivne važnosti, s njegovom pozitivnom vrijednošću od 6 koja je tri puta veća od pozitivne važnosti prvog neurona od 2. Ova razlika između atribucijske mape i stvarne dinamike sloja povećava se s razlikom između apsolutnih magnituda aktivacija i izlazne vrijednosti neurona. Stoga, učinkovitiji pristup slojevite propagacije trebao bi uzeti u obzir apsolutne magnitude aktivacija unutar neurona.

Drugi nedostatak RAP metode nalazi se u preraspodjeli prekomjerno alocirane relevantnosti. Naime, odluka o ravnomjernom oduzimanju iznosa prekomjerne alokacije je proizvoljna, bez opravdanja od strane autora. Raspršenje viška relevantnosti optimalno bi bilo izvesti tako da raspodijeljene vrijednosti, ponovno, odgovaraju stvarnim mehanizmima zaključivanja unutar modela. Kako bi razriješila ovaj nedostatak, predložena atribucijska metoda u ovoj disertaciji koristi kontrastnu slojevitu propagaciju kako bi se odredila relevantna područja slike.

U sferi objašnjivosti Transformera, sve prethodno spomenute metode upotrebljavaju vrijednosti samopozornosti kao primarno sredstvo za izračunavanje konačnih atribucijskih mapa. Ova činjenica rezultira njihovom smanjenom interpretabilnošću, budući da je razlučivost takvih atribucijskih mapa ovisna o broju komadića slike koje model koristi. Na primjer, standardni Vizualni Transformer s veličinom komadića slike od 16×16 piksela koji obrađuje sliku veličine 224×224 piksela proizveo bi 14×14 komadića, koji se na izlazu atribucijske metode skaliraju na izvornu rezoluciju slike. U slučaju Vision Transformera s manjim brojem većih komadića

dizajniranih za brzinu i učinkovitost, metodama atribucije koje operiraju na razini komadića smanjuje se interpretabilnost.

U trenutnim istraživanjima samo dvije studije izvještavaju o korištenju metoda isključivo temeljenih na LRP-u za objašnjavanje Transformera. Chefer et al. [99] koriste LRP-$\alpha_1\beta_0$ kroz cijeli model, ali prijavljuju loše rezultate i koriste ovu metodu samo kao osnovnu metodu s kojom uspoređuju kvalitetnije pristupe. Ali et al. [110] koriste modificiranu metodu Input*Gradient [70], međutim svoje rezultate ne prijavljuju u domeni računalnog vida, već samo domenama obrade prirodnog jezika i obradi grafova. Također, činjenica da se vrijednosti samopozornosti promatraju kao konstante onemogućuje propagaciju relevantnosti kroz upite i ključeve što odmiče konačne vrijednosti relevantnosti od njihovih stvarnih vrijednosti. Predložena metoda, s druge strane, primjenjuje se na više različitih arhitektura te se same vrijednosti relevantnosti propagiraju do ulazne slike, producirajući mapu na razini piksela, svojstvo koje u trenutnim istraživanjima ne nudi niti jedna druga metoda.

2.4 Ostale post-hoc metode za određivanje mapa relevantnosti

U ovoj sekciji prikazuju se različite metode korištene za izračunavanje mapa relevantnosti. Ove metode služe kao ključni alati u razjašnjavanju istaknutih značajki i značajnih aspekata unutar skupova podataka koji se promatraju. U nastavku ovog rada one će proći rigoroznu evaluaciju unutar eksperimentalnog dijela rada, prikazujući temeljitu komparativnu analizu njihove učinkovitosti i karakteristika izvedbe.

2.4.1 Zasićenje i Vođeni prolaz unatrag

Simonayan i sur. [111] predstavili su metodu Zasićenja (engl. *Saliency*) u svrhu određivanja bitnih područja ulazne slike za klasifikaciju. Metoda je iznimno jednostavna te se bazira na unatražnom prolazu kroz mrežu:

$$\text{Zasićenje}(x) = |\frac{\partial o_i}{\partial x}| \tag{2.17}$$

Gdje je o_i izlazni neuron, odnosno vrijednost klase pod indeksom i za koji je potrebno izračunati objašnjenje. Motivacija računanja zasićenja klase za određenu sliku korištenjem derivacije izlaza za tu klasu jest da veličina derivacije pokazuje koje ulazne neurone treba najmanje promijeniti kako bi se najviše utjecalo na ocjenu klase. Očekuje se da takvi neuroni odgovaraju područjima slike relevantnim za klasifikaciju objekta na slici.

Springenberg i sur. [112] predlažu proširenje ove metode, gdje se prilikom prolaza unatrag pozicije gradijenta koje u tom trenutku ili su u trenutku prolaza unaprijed imale vrijednost manju

od nula postavljaju na nulu kako bi se spriječili negativni utjecaji na samu mapu relevantnosti. Iz ovog razloga navedena metoda producira atribucijske mape bez šuma fokusirane na ciljni objekt.

2.4.2 Dekonvolucija

Zeiler i sur. [39] su u svojem radu predstavili metodu Dekonvolucije (engl. *Deconvolution*) čiji je cilj, koristeći postojeće komponente mreže, invertirati ju te umjesto da se ulazne značajke mapiraju na izlazne, mapirati izlazne značajke na ulazne. Kako bi se izračunala atribucija za određeni izlazni neuron, na početku se sve druge aktivacije u sloju postavljaju na nulu te se one prosljeđuju u prvi sloj mreže, koji je izvornog bio zadnji sloj. Potom se, ovisno o tipu sloja, uzastopno provode operacije proširenja, rektifikacije i transponirane konvolucije kako bi se rekonstruirala aktivacija u sloju ispod onog koja je dovela do odabrane aktivacije. Ovaj proces se ponavlja dok se ne dosegne ulaz u model, odnosno ulazna slika.

2.4.3 DeepLIFT

DeepLIFT je predstavljen od strane Shrikumar i sur. [70] te je njihova metoda atribucije bazirana na objašnjavanju razlike između izlaza za dani ulazni primjer i referentnog izlaza u smislu razlike između ulaza i nekog referentnog ulaza. Referentni ulaz opisuje odabrani ili neutralni ulaz koji se odabire na temelju specifičnih potreba problema koji se rješava. Formalno, uzevši određeni izlazni neuron t, uz definiciju $x_1, x_2, ..., x_n$, kao vrijednosti slojeva potrebnih za izračunavanje vrijednosti t, autori DeepLIFT-a tretiraju Δt kao razliku između izlazne vrijednosti i referentne vrijednost t_0. Ova metoda potom prilaže ocjene relevantnosti $C_{\Delta x, \Delta_t}$ svakoj razlici vrijednosti unutar sloja te njegovoj referentnoj vrijednosti Δx_i. Drugim riječima, za svaki pomak od referentne vrijednosti ulaza, ova metoda kvantificira razinu važnosti za odmak od referentne vrijednosti izlaza.

2.4.4 Integrated Gradients

Sundararajan i sur. [42] u svojem radu predstavljaju metodu Integriranih Gradijenata (engl. *Integrated Gradients*), gdje s pomoću dvaju aksioma za koje smatraju da atribucijske metode moraju zadovoljiti, definiraju novu metodu koja bez modifikacije izvorne mreže te nekoliko unatražnih poziva zadovoljavaju ova svojstva. Prvi aksiom je vezan uz osjetljivost metode, svojstvo koje je definirano kao sposobnost metode da za svaki ulaz i referentni ulaz koji se razlikuju u jednoj značajki, a produciraju različite predikcije, ta značajka mora imati relevantnost različitu od nule. Kao primjer metode koja krši ovo pravilo uzimaju gradijent, odnosno metodu zasićenja opisanu ranije u sekciji 2.4.1. Ako se za primjer uzme mreža s jednim ulazom te

ReLU nelinearnošću: $f(x) = 1 - \text{ReLU}(1-x)$, za referentnu vrijednost od $x = 0$, a ulaznom vrijednošću od $x = 1$, izlazne vrijednosti su 0 za referentni i 1 za ulazni primjer. Ako se za ovaj ulaz primijeni metoda zasićenja, relevantnost ulazne značajke će biti 0, pošto je to vrijednost gradijenta za sve vrijednosti jednake ili veće od 1. Nedostatak gradijentne metode primijećen je i u drugim radovima [70]. Kršenje ovog svojstva kod atribucijskih metoda dovodi do fokusa na irelevantne dijelove slike, odnosno unošenje šuma u samu atribucijsku mapu. Drugi aksiom nosi se s invarijantnošću na implementaciju modela, koji tvrdi da ako su izlazi dvaju modela identični za sve ulaze u model, tada i atribucije moraju biti identične između dva modela. Ako metoda atribucije ne zadovoljava invarijantnost implementacije, producirane atribucije će potencijalno biti osjetljive na nevažne aspekte modela. Na primjer, ako promatrani model ima više stupnjeva slobode nego što je potrebno kako bi predstavili željenu funkciju, tada mogu postojati dva skupa vrijednosti parametara modela koji dovode do iste funkcije. Postupak učenja može konvergirati u bilo koji skup vrijednosti, primjerice ovisno o inicijalizaciji. Metode koje ne zadovoljavaju ovo svojstvo bi za dva modela dala različite atribucijske mape, usprkos pokazanoj identičnoj funkcionalnosti modela. Autori rješavaju oba problema tako da definiraju atribucijsku mapu kao sumu gradijenata izlaza modela na putu od referentnog ulaza do trenutnog ulaza:

$$\text{IntegratedGradients}(x) = (x_i - x_i^r) \int_{\alpha=0}^{1} \frac{\partial F(x^r + \alpha(x - x^r))}{\partial x_i} \tag{2.18}$$

Gdje je x ulaz u model, x^r referentni ulaz, a i indeks ulaznog neurona.

2.4.5 SmoothGrad

SmoothGrad je metoda predložena od strane Smilkov i sur. [113] koji ističu mogućnost da se šum u mapama relevantnosti baziranim na gradijentima može objasniti naglim, za atribucije beznačajnim, fluktuacijama u gradijentima na malim skalama unutar slike. Pošto se tijekom učenja modela ni na koji način ne osigurava blagi prijelaz vrijednosti gradijenta između susjednih mjesta ulaza, ovakvo ponašanje je očekivano. Iz tog razloga, autori predlažu da se umjesto korištenja gradijenta koji će biti manje interpretabilan za bilo koji ulaz zbog inherentnih fluktuacija, koristi usrednjena vrijednosti gradijenta s obzirom na ulaz koji je zaglađen dodatkom šuma iz normalne distribucije. Formalno, SmoothGrad je definiran na sljedeći način:

$$\text{SmoothGrad}(x) = \frac{1}{n} \sum_{i}^{n} S(x + \mathcal{N}(0, \sigma^2)) \tag{2.19}$$

2.4.6 GradCAM, GuidedGradCAM i GradCAM++

Selvaraju i sur. [114] u svojem su radu predložili metodu pod nazivom GradCAM. Autori ističu nedostatak metoda poput Dekonvolucije [39] koje ne mogu vizualizirati mape relevantnosti specifične za jednu klasu unutar slike, već neovisno o klasi odabranoj kao traženu, ove metode pro-

duciraju vrlo slična objašnjenja. Kako bi riješili ovaj problem, autori proširuju postojeću Class Activation Mapping (CAM) metodu [115]. CAM metoda mijenja finalne potpuno-povezane slojeve unutar konvolucijske mreže za dodatne konvolucijske slojeve, sloj globalnog sažimanja i jedan linearni sloj. Ova kombinacija slojeva omogućuje izračun CAM-a za klasu c:

$$\text{CAM}_c(x,y) = \sum_k w_k^c f_k(x,y) \tag{2.20}$$

x i y ovdje predstavljaju koordinate izlaza posljednjeg konvolucijskog sloja $f_k(x,y)$, odnosno ulaz u sloj globalnog sažimanja, a w_k^c predstavlja težinu određene sažete mape za klasu c dobiven iz posljednjeg linearnog sloja nad kojim se provodi softmax operacija. Nedostatak CAM-a je u striktno traženoj arhitekturi promatrane mreže. Mape značajki koje se sumiraju moraju izravno prethoditi softmax-u, tako da je ova metoda primjenjiva samo na posebne vrste konvolucijskih arhitektura. Takve arhitekture imaju potencijalno inferiorne radne značajke u odnosu na ostale arhitekture ili jednostavno mogu biti neprimjenjiva druge tipove zadataka, primjerice odgovaranje na pitanja pomoću slika. GradCAM omogućuje kombiniranje mapa značajki pomoću gradijentnog signala bez izmjena u arhitekturi. Za konvolucijske arhitekture potrebne za CAM, on je zapravo poseban slučaj GradCAM-a. Izračun mapa relevantnosti GradCAM metode započinje odabirom klase c te izračunom gradijenta iznosa izlaza za tu klasu o^c s obzirom na aktivacijske mape odabranog konvolucijskog sloja A^k. Kako bi se dobila težina za svaku aktivacijsku mapu α_k^c, ovi gradijenti se globalno usrednjuju operacijom sažimanja:

$$\alpha_k^c = \frac{1}{N}\sum_i\sum_j \frac{\partial o^c}{\partial A_{ij}^k} \tag{2.21}$$

gdje je N umnožak i i j, koordinata koji iteriraju po visini i širini određene mape sloja. Težina α_k^c ovdje predstavlja parcijalnu linearizaciju mreže na slojevima ispod odabranog konvolucijskog sloja. Ova činjenica dovodi do finalne mape relevantnosti za metodu GradCAM:

$$\text{GradCAM}(x) = \text{ReLU}(\sum_k \alpha_k^c A^k) \tag{2.22}$$

Naposljetku, autori uvode dodatno proširenje svoje metode na metodu GuidedGradCAM, koja se jednostavno izvodi umnoškom mapa produciranih GuidedBackpropagation-om i GradCAM-om. GradCAM++ (Chattopadhyay i sur. [116]) uvode proširenje izračuna težina pojedinih filtera A^k na način da u njihov izračun, osim gradijenata kako je to slučaj u GradCAM-u, uključuju i same vrijednosti aktivacija. Težine se definiraju na sljedeći način:

$$\alpha_k^c = \frac{1}{N}\sum_i\sum_j p_{ij}^{kc}\text{ReLU}(\frac{\partial o^c}{\partial A_{ij}^k}) \tag{2.23}$$

gdje je p_{ij}^{kc} definiran kao:

$$p_{ij}^{kc} = \frac{(\frac{\partial o^c}{\partial A_{ij}^k})^2}{2(\frac{\partial o^c}{\partial A_{ij}^k})^2 + \sum_a \sum_b A^{ab} (\frac{\partial o^c}{\partial A_{ij}^k})^3} \tag{2.24}$$

Ova formulacija omogućuje GradCAM++ metodi bolju sposobnost lokalizacije objekta na slikama gdje se pojavljuje više instanci objekata.

2.4.7 HiResCAM i LayerCAM

HiResCAM (Draelos i sur. [117]) nudi rješenje za ograničenje otkriveno unutar GradCAM metode. Izračun težina α_k^c sa sobom nosi usrednjavanje vrijednosti gradijenta aktivacijskih mapa A određenog sloja. Iako je ovo usrednjavanje bilo motivirano globalnim sažimanjem unutar izvorne CAM metode, ono ograničava preciznost GradCAM metode te iz tog razloga atribucijske mape često izgledaju prošireno oko objekta od interesa, a sama relevantna područja imaju veliku površinu i blage prijelaze. Usrednjavanje gradijenta dovodi do gubitka preciznih informacija o samom gradijentu, ako gradijent posjeduje visoko-polarizirane vrijednosti, usrednjenje će producirati malu pozitivnu ili negativnu težinu. Kako bi se povećala oštrina produciranih mapa i unaprijedila lokalizacijska sposobnost metode izgubljena usrednjavanjem, HiResCAM preskače izračun težina α_k^c, već izravno množi gradijente izlaznog neurona s obzirom na izlaze konvolucijskog sloja s aktivacijskim mapama tog konvolucijskog sloja:

$$\text{HiResCAM}(x) = \text{ReLU}(\sum_k \frac{\partial o^c}{\partial A^k} \odot A^k) \tag{2.25}$$

Vrlo sličnu ideju imali su i autori LayerCAM metode (Jiang i sur. [118]), koji uzimaju samo pozitivne vrijednosti gradijenta kao težine za ponderiranje vrijednosti aktivacijskih mapa A^k. Kao razlog se navodi logika slična metodama Dekonvolucije i GuidedBackpropagation-a, gdje pozitivne vrijednosti, u ovom slučaju gradijenata, pozitivno pridonose vrijednosti finalnog izlaznog neurona umjesto da ju snizuju što na posljetku dovodi do vizualizacije pozitivno atributivnih područja na mapi relevantnosti. Formalno je LayerCAM definiran na sljedeći način:

$$\text{LayerCAM}(x) = \text{ReLU}(\sum_k \text{ReLU}(\frac{\partial o^c}{\partial A^k}) \odot A^k) \tag{2.26}$$

Poglavlje 3

Metrike za vrjednovanje metoda za produkciju mapa relevantnosti

Metrike za vrjednovanje atribucijskih metoda mogu se podijeliti u nekoliko kategorija koje ocjenjuju različite osobine tih metoda. Sljedeće kategorije su značajne.

Vjernost - sastoji se od *korelacije važnosti* - veličina atribucijskih težina trebala bi odražavati važnost ulaznih komponenata i *konzistentnosti polariteta* - znak atribucijskih težina trebao bi ispravno ukazivati na polaritet utjecaja ulaza, odnosno efekte doprinosa ili potiskivanja na predikciju modela, kako je definirano od strane Liua i sur. [119]. Metrike u ovoj kategoriji temelje se na promatranju promjene izlaza modela prilikom permutacije ulaza na temelju zadane mape atribucije. Neke metrike ciljaju izračunati korelaciju između izvornog vjerojatnosnog izlaza modela i perturbiranog izlaza [120] [121] [122] [123] [41] [124] [125] ili korelaciju između pada vjerojatnosti i atribucijskih bodova na različitim točkama [121] [120], dok drugi promatraju promjenu u radnim značajkama modela nakon što se perturbira ulaz [126] [127] [128].

Robusnost - ovaj kriterij istražuje stabilnost atribucija kada su izložene manjim promjenama u ulazu, uz pretpostavku da se izlaz modela približno ne mijenja. Jedan aspekt robusnosti uključuje traženje sličnih primjera koji bi trebali proizvesti slična objašnjenja [125], što u osnovi ocjenjuje dosljednost objašnjenja kod sličnih instanci. Drugi aspekt uključuje hvatanje najznačajnijih varijacija u mapi atribucija kada se perturbira ulaz [123], s ciljem identifikacije regija u ulaznom prostoru gdje su predikcije modela najosjetljivije na promjene. Dodatno, postoje metrike koje mjere vjerojatnost da ulazi s identičnim atribucijama daju istu predikciju [129], pružajući uvide u dosljednost predikcija modela preko varijacija ulaza.

Lokalizacija - metrike iz ove kategorije imaju za cilj kvantificirati usklađenost između najviših ocjena u generiranoj mapi relevantnosti i odabranog objekta u usporedbi s označenom segmentacijskom maskom. Ocjenjujući sklad između atribucijske mape i unaprijed određene segmentacijske mape pruža se vrijedan uvid u vjernost prostorne lokalizacije atribucija [72] [130] [131] [132].

Sljedeća sekcija ističe tri najrelevantnija pristupa za vrjednovanje metoda objašnjavanja temeljenih na atribuciji s obzirom na predloženo rješenje u nastavku rada. Kroz detaljno razumijevanje ovih metrika za vrjednovanje i njihovih ograničenja, cilj ovog rada je razviti metriku za sveobuhvatnu procjenu radnih značajki atribucijskih metoda.

3.1 Pristup iz kategorije vjernosti

Samek i sur. [133] predlažu metriku za vrjednovanje metoda objašnjavanja temeljenih na atribuciji, nadograđujući pristup koji su predstavili Bach i sur. [69]. U izvornoj metodi, atribucije generirane različitim tehnikama koriste se za vođenje perturbacija regija unutar ulazne slike. Konkretno, za dani model f, proces uključuje iteriranje kroz L koraka, pri čemu svaki korak označava indeks k. U svakom koraku, prvih k regija dimenzija 9×9 ulazne slike, sortiranih prema njihovim atribucijskim vrijednostima, podvrgnute su perturbaciji zamjenom s nasumično odabranim vrijednostima iz uniformne distribucije.

Tijekom svakog koraka procesa, računa se razlika u vrijednosti izlaznog neurona između izvornog ulaza i perturbiranog ulaza u koraku k, što se označava kao x^k_{MoRF}, gdje "MoRF" označava "najrelevantniji prvo" (engl. *Most Relevant First*). Te razlike se akumuliraju kroz sve korake i uprosječuju preko cijelog skupa podataka. Veća osjetljivost modela na perturbacije u najrelevantnijim regijama rezultirala bi značajnijim smanjenjem vrijednosti izlaznog neurona, što dovodi do veće "površine iznad krivulje perturbacije MoRF" (engl. *area over the MoRF perturbation curve* (AOPC)).

Stoga, metrika AOPC služi kao mjera koliko su perturbacije primijenjene na najrelevantnije regije ulazne slike utjecale na izlazni rezultat modela. Veća vrijednost AOPC-a ukazuje na veću osjetljivost modela na promjene u tim ključnim regijama, a s time i veću točnost atribucijske metode u identifikaciji relevantnih regija. AOPC je definiran kao:

$$\text{AOPC} = \frac{1}{L+1}\frac{1}{N}\sum_{n}^{N}\sum_{k}^{L}(f(x^0_{n_{MoRF}}) - f(x^k_{n_{MoRF}})) \tag{3.1}$$

Osim metrike AOPC, autori uvode još jednu evaluacijsku metriku nazvanu "površina između krivulja perturbacije" (engl. *area between perturbation curves* (ABPC)). Ova metrika ima za cilj pružiti dodatne uvide u radne značajke atribucijskih metoda razmatrajući razliku između vrijednosti izlaznih neurona za slike modificirane korištenjem dvije različite strategije perturbacije: najrelevantnije prvo (MoRF) i najmanje relevantne prvo (LeRF). AOPC je definiran na sljedeći način:

$$\text{ABPC} = \frac{1}{L+1}\frac{1}{N}\sum_{n}^{N}\sum_{k}^{L}(f(x^k_{n_{LeRF}}) - f(x^k_{n_{MoRF}})) \tag{3.2}$$

U slučaju LeRF-a, proces perturbacije započinje s najmanje relevantnim regijama slike, što idealno rezultira minimalnim promjenama u relevantnim informacijama za određenu klasu unutar slike, pri čemu rezultati izlaznih neurona ostaju stabilni i bliski izvornim vrijednostima za male vrijednosti k. Kako k pristupa ukupnom broju koraka L, očekuje se da će vrijednosti izlaznih neurona brže opadati. Ovo ponašanje odražava očekivanje da će perturbiranje najmanje relevantnih regija slike imati minimalan utjecaj na klasifikaciju modela.

Usporedbom rezultata izlaznih neurona dobivenih perturbacijama primijenjenim korištenjem strategija MoRF-a i LeRF-a, metrika ABPC pruža uvide u to koliko učinkovito metoda atribucije odabire i najrelevantnije i najmanje relevantne dijelove slike. Autori provode ove perturbacije na prvih 100 regija slike, rezultirajući zamjenom približno 15.7% slike.

U recentnom istraživanju, Rong i sur. [127] razvijaju novi pristup pod nazivom Brisanje i rješavanje pristranosti (engl. *Remove and Debias* (ROAD)) koji se nadovezuje na rad Samek i sur. Njihova metrika usvaja temeljni princip maskiranja ulaznih značajki kategoriziranih kao visoko relevantne (MoRF) ili nisko relevantne (LeRF), što je određeno inicijalnom atribucijskom mapom. Međutim, njihov doprinos leži u uvođenju nove tehnike perturbacije, koja uključuje uporabu linearnih kombinacija susjednih piksela.

Korištenjem linearnih kombinacija susjednih piksela, učinkovito poboljšavaju konzistentnost rangiranja različitih atribucijskih metoda između dva načina maskiranja - MoRF i LeRF.

3.1.1 Nedostaci metrika iz kategorije vjernosti

Među metrikama iz kategorije vjernosti, međutim, postoji nedostatak istaknut od strane Ju i sur. [134]. Metrike vrjednovanja temeljene na perturbacijama, poput AOPC(k), mogu se smatrati oblikom atribucijske metode. Primjerice, AOPC(k) koji ocjenjuje važnost $k\%$ značajki može se promatrati ako atribucijska metoda koja računa vrijednosti relevantnosti za $k\%$ značajki. Ako se radi o perturbacijama koji uključuju postavljanje ulazne vrijednosti na nulu te se k suzi na jednu značajku, AOPC postaje ekvivalentan osnovnoj atribucijskoj metodi poznatoj kao Izostavi Jedan (engl. *leave-one-out*) [135]. Stoga, ovaj način vrjednovanja atribucijskih metoda se zapravo svodi na vrjednovanje sličnosti dvaju atribucijskih metoda.

Nadalje, važno je primijetiti mogućnost velike varijabilnosti u magnitudama atribucijske mape koje i dalje produciraju isti redoslijed sortiranih ulaznih značajki. Ta varijabilnost može značajno promijeniti ljudski percipiranu relativnu važnost ulaznih značajki, unatoč održavanju identičnih ocjena prilikom korištenja metrika za vrjednovanje ove kategorije.

Osim toga, ključno je prepoznati činjenicu da atribucijske metode nisu nužno zadužene za razjašnjavanje načina na koji bi ulaz trebao biti modificiran kako bi se postigao određeni učinak na izlaz neuronske mreže. Umjesto toga, njihov je primarni cilj točno pripisati važnost ulaznih značajki za **trenutačni** primjer ulaza. Ovaj temeljni aspekt čini se zanemaren u trenutnim istraživanjima.

3.2 Pristup iz kategorije robusnosti

Alvarez i sur. [136] istražuju aspekt robusnosti metoda atribucije, ističući važnost zadovoljenja kriterija da slični ulazi trebaju rezultirati sličnim atribucijskim mapama. U svom istraživanju provode detaljnu analizu ovog koncepta i predlažu novu metriku usmjerenu na kvantifikaciju ovog svojstva tako da ispituju kako varijacije u ulaznim podacima, koje dovode do određene predikcije, utječu na odgovarajuće atribucije koje pruža određena metoda za produkciju mapa relevantnosti.

U njihovoj studiji, robusnost je definirana u smislu dosljednosti atribucija kada ulazni podaci podliježu blagim promjenama, odnosno perturbacijama, dok se izlaz modela ili ne mijenja uopće ili mijenja samo marginalno. Ova ideja robusnosti ključna je za osiguravanje pouzdanosti i stabilnosti metoda atribucije u različitim primjenama, gdje bi male promjene u ulaznim podacima idealno rezultirale marginalnim promjenama u promatranom izlazu modela, ili se promjene ne bi niti dogodile.

Za kvantifikaciju robusnosti, autori uvode koncept Lipschitz kontinuiteta, koji se procjenjuje na lokalnoj razini kroz rješavanje određenog optimizacijskog problema za svaki primjer od interesa x_{orig}. Ovaj pristup omogućuje im procjenu osjetljivosti metode atribucije na varijacije u ulaznim podacima, pružajući vrijedne uvide u svojstva njezine robusnosti.

$$\hat{L}(x_{orig}) = \underset{x_{pert} \in \aleph_\varepsilon(x_{orig})}{argmax} \frac{||f(x_{orig}) - f(x_{pert})||_2}{||x_{orig} - x_{pert}||_2} \tag{3.3}$$

$$\aleph_\varepsilon(x_{orig}) = \{||x_{orig} - x_{pert}||_2 \leq \varepsilon; x_{pert} \in \mathbb{R}^n\} \tag{3.4}$$

Optimizirana vrijednost $\hat{L}(x_i)$ je u ovom radu gledana kao kvantifikacija mjere robusnosti.

3.2.1 Nedostatak metrika iz kategorije robusnosti

Duboki modeli često posjeduju nedostatke u robusnosti na male perturbacije u ulazu, svojstvo koje vrlo lako pokazuju ljudi te se često oslanjaju na korelacije koje mogu naizgled biti neosnovane. Primjerice, čak i tekstovi sličnog značenja mogu potaknuti potpuno različite procese zaključivanja unutar dubokih modela. Dobro poznata ranjivost dubokih modela je njihova osjetljivost na suparničke primjere [137] [138], gdje suptilne promjene u ulaznim podacima, neprimjetne ljudima, mogu dovesti do toga da model donese pogrešnu predikciju. Uspjeh suparničkih napada na duboke modele ističe kako slični ulazi mogu rezultirati iznimno različitim procesima zaključivanja unutar tih modela. Ovu činjenicu prepoznaju i Ju i sur. u svojem istraživanju [134].

Glavna razlika između metoda napada na atribucijsku metodu i napada na model leži u njihovim zahtjevima unutar ovog konteksta. Metode napada na atribuciju zahtijevaju od mo-

dela da proizvede identične predikcije za suparničke primjere. Međutim, ovaj ograničavajući faktor često nije dovoljan da bi se model prisilio da ispravno zaključuje, uzimajući u obzir da duboki modeli inherentno komprimiraju kompleksne izračune u ograničene klase predikcija. Primjerice, prilikom binarne klasifikacije, model s potpuno slučajno odabranom metodom zaključivanja ima 50% vjerojatnosti da će odabrati određenu klasu, a uz zbog slučajnog odabira metode zaključivanja imati značajno različitu atribucijsku mapu.

Razlika u vrijednostima relevantnosti ne ukazuje nužno na nepouzdanost atribucijske metode, već može prikazivati stvarne promjene u procesu zaključivanja modela. Pouzdana atribucijska metoda trebala bi proizvesti različite relevantnosti za izvorne i suparničke primjere, odražavajući stvarni proces zaključivanja modela. Nažalost, ova ključna razlika često se zanemaruje u trenutnim istraživanjima koja iz tog razloga dolaze do zaključka da su trenutne atribucijske metode krhke i nepouzdane.

Dodatno, bitno je primijetiti kako je "idealna" atribucijska metoda za metrike iz ove kategorije jednostavna konstanta, koja se ne mijenja ovisno o ulaznim perturbacijama. To dodatno narušava pouzdanost i valjanost takvih metrika.

3.3 Pristup iz kategorije lokalizacije

Arias-Duart i sur. [139] predstavljaju metriku nazvanu Focus koja u navedenoj kategorizaciji pripada metrikama lokalizacije. Ova metrika koristi kompozicije slika iz skupa podataka, nazvane mozaici, kako bi istražila sposobnost atribucijske metode da alocira relevantnost na odabrane regije mozaika.

Konstrukcija svakog mozaika m uključuje odabir klase koja služi kao fokalna točka za vrjednovanje atribucije. Pritom se odabiru dvije slike koje pripadaju toj odabranoj klasi te se one označavaju kao pozitivi (p_1, p_2), predstavljajući instance koje bi metoda atribucije trebala precizno istaknuti. Dodatno, slučajnim odabirom se odabiru i dvije slike koje ne pripadaju početnoj klasi nasumično te se one označavaju kao negativi (n_1, n_2), pružajući kontrast pozitivnim instancama.

Nakon odabira ove četiri slike, one se spajaju u jednu kompozitnu sliku, formirajući matricu slika s dvije slike u svakom retku i stupcu. Raspored ovih slika unutar mreže odabire se slučajno za svaki primjer, osiguravajući varijabilnost i eliminirajući pristranosti u procesu kompozicije. Ovaj proces amalgamacije ima za cilj simulirati scenarije u stvarnom svijetu gdje multiple instance različitih klasa koegzistiraju unutar jednog konteksta.

Srž Focus metrike leži u njezinoj sposobnosti kvantificiranja točnosti i preciznosti metoda atribucije u razgraničavanju ciljnih regija unutar mozaika. Analizom distribucije i intenziteta relevantnosti dodijeljene pozitivnim instancama u odnosu na negativne instance, Focus metrika pruža uvide u sposobnost metode atribucije u isticanju, odnosno lokalizaciji, bitnih značajki povezanih s odabranom klasom. Zaključno, Focus metrika definirana je na sljedeći način:

$$F(m) = \frac{R(p_1) + R(p_2)}{R(m)} \tag{3.5}$$

R je suma relevantnosti na pozicijama odabrane, odnosno pozitivne klase, unutar određenog kvadranta ili cijelog mozaika.

3.3.1 Nedostatak metrika iz kategorije lokalizacije

Srž pouzdane atribucijske metode prema metrici Focus leži u njezinoj sposobnosti da pretežito istakne značajnost dvije pozitivne slike povezane s ciljnom klasom. Međutim, unutar ove metrike pojavljuje se problem sličan problemu opisanom u metrikama iz kategorije vjernosti. Naime, postoje višestruke atribucija koje, unatoč njihovoj značajnoj raznolikosti, rezultiraju identičnim Focus rezultatom. Primjerice, ako razmotrimo atribucijsku mapu koja precizno razgraničava dva kvadranta koja opisuju pozitive unutar kreiranog mozaika te dodjeljuje relevantnost jedan središnjem pikselu svake od njih poprima maksimalan rezultat od strane Focus metrike. Paradoksalno, alternativna atribucijska strategija koja dodjeljuje ocjenu 1 cijelim kvadrantima obuhvaćajući pozitivne segmente unutar mozaika također bi osigurala isti maksimalni Focus rezultat. Ovaj paradoks ističe potrebu za inicijalnom procjenom kvalitete atribucije vezane uz određeni kvadrant te potom kombinaciju te procjene s metrikom lokalizacije.

3.4 Zaključak o nedostacima metrika za vrjednovanje

Prethodno opisani izazovi u evaluaciji metoda atribucije doveli su do nedostatka suglasnosti među istraživačima o optimalnoj metrici za vrjednovanje atribucijskih metoda. Stoga istraživači često koriste više metrika kako bi ocijenili različite aspekte atribucijskih metoda. Učinkovito kombiniranje ovih različitih rezultata u jedan, sveobuhvatan rezultat ostaje neistraženo područje, te ova činjenica ograničava široku prihvaćenost pojedinih metrika i zahtijeva da njihovi korisnici eksperimentiraju i subjektivno odaberu najprikladniju metriku za svoj model ili specifičan problem.

Poglavlje 4

Predložena metoda za produkciju mapa relevantnosti temeljena na slojevitoj propagaciji relativne apsolutne magnitude relevantnosti

Potaknuti spomenutim nedostacima trenutnih pravila slojevite propagacije relevantnosti, u ovom poglavlju se predstavlja novo pravilo nazvano Slojevita propagacija relativne apsolutne magnitude relevantnosti (engl. *Relative Absolute Magnitude Layer-Wise Relevance Propagation* (RAMP)). Ono rješava problem sukobljenih relevantnosti koje se nalaze u većini LRP pravila promatrajući samo pozitivne dijelove aktivacije neurona, slično kao LRP-$\alpha_1\beta_0$ pravilo. Međutim, kako bi se riješila netočna relativna atribucija zbog varijabilnih magnituda aktivacija unutar neurona istog sloja, kao normalizacijski faktor se koristi apsolutni **završni** izlaz svakog neurona. Kako bi se producirale kontrastne atribucijske mape, koristi se ideja iz rada Gu i sur. [140] te se kao početna atribucija zadnjeg sloja postavlja 1 za ciljnu klasu i $-\frac{1}{N-1}$ za ostale klase, gdje je N ukupan broj klasa. Primjena ovog novog pravila rezultira prorijeđenim i kontrastnim atribucijskim mapama bez šuma. Novo pravilo LRP-a definirano je sljedećom formulom:

$$R_i^{l-1} = \sum_j \frac{(x_i w_{ij})^+}{|\sum_k x_k w_{kj}| + \varepsilon} R_j^l \tag{4.1}$$

Vraćajući se na primjer prikazan na Slici 2.2, u prethodnom poglavlju je ilustrirano kako LRP-$\alpha_1\beta_0$ dodjeljuje istu atribucijsku mapu svakom od ulaznih neurona, neovisno o relativnim apsolutnim magnitudama aktivacija u prvom i drugom skrivenom neuronu, koseći se sa stvarnom slikom zaključivanja mreže. Primjenom predloženog pravila na lijevu mrežu, dobivaju

se isti rezultati atribucije kao i metodom LRP-$\alpha_1\beta_0$, budući da obje grane imaju jednake apsolutne magnitude od 3. Međutim, prilikom primijene metode na desnu mrežu, rezultirajuće atribucije postaju $[0.25, 0.75]$, što predstavlja razliku od 50% u atribucijskoj razlici za svaki od ulaznih neurona u usporedbi s atribucijom LRP-$\alpha_1\beta_0$, budući da druga grana sada ima znatno veću apsolutnu magnitudu od 11 što dovodi do njezinog većeg utjecaja na konačnu atribucijsku mapu. Razlika u magnitudi između dviju grana pojačava ovaj učinak. Ako se ta razlika učini još većom, na primjer, postavljanjem težina drugog skrivenog neurona na $[-17, 18]$, LRP-$\alpha_1\beta_0$ zadržava iste atribute, dok predložena metoda sada dodjeljuje 9 puta više relevantnosti drugom ulaznom neuronu, budući da je njegov apsolutni pozitivni utjecaj u mreži 18, u usporedbi s 2 prvog neurona, što dovodi do vjernijeg prikaza unutarnjih mehanizama mreže.

Ovo novo pravilo se može efikasno implementirati putem automatske diferencijacije, funkcionalnost koja je dostupna u većini razvojnih okvira za duboko učenje. Primjer implementacije predložene metode u pseudokodu nalik PyTorch implementaciji prikazan je u Algoritmu 4.1.

```
def backprop_relevance(prev_rel, input, module):
    """
    prev_rel: tensor
        Tenzor vrijednosti relevantnosti sljedećeg sloja
    input: tensor
        Tenzor ulaza u trenutni sloj
    module: nn.Module
        Implementacija trenutnog sloja, zajedno
        korištenim parametrima sloja (ukoliko postoje)
    """

    # Izracun izlaza
    h = module(input)

    # Izracun apsolutne vrijednosti i izlaza
    input_abs = input.abs()
    module_abs = abs_module_params(module)
    ha = module_abs(input_abs)

    # Unatrazni prolazkako bise izracunala
    # atribucija trenutnog sloja
    rel_scaling = prev_rel / (h.abs() + 1e-9)
    rel = autograd.grad(ha + h, x, rel_scaling)

    return rel
```

Algoritam 4.1: Pseudokod predložene metode slojevite propagacija relativne apsolutne magnitude relevantnosti

S obzirom na to da se predložena metoda temelji na LRP-u, primjena unutar složenijih

modela zahtjeva implementaciju ovog pravila u određene slojeve specifične promatranoj arhitekturi. Naime, integracija metode se postiže implementacijom pravila za **rezidualne veze** i **normalizaciju nad grupom** unutar ResNet arhitekture te implementacijom pravila za **blokove samopozornosti** i **slojnu normalizaciju** unutar Transformer arhitekture.

U slučaju **rezidualnih veza** unutar ResNet-a, pravilo LRP-a mora biti definirano za dva ključna stupnja u komputacijskom grafu. Prvo, ključno je uspostaviti pravilo za početno razdvajanje rezidualnih i standardnih komponenti rezidualne veze. Ovo početno razdvajanje postavlja temelje za naknadno širenje relevantnosti unutar rezidualnog bloka. Drugo, pravilo treba specificirati za fazu rekombinacije na kraju rezidualnog bloka. Ova faza rekombinacije kombinira izlaze razdvojenih komponenti kako bi se proizveo konačni izlaz. Za postizanje toga, pravilo kroz ovaj tip operacije je jednostavno definirano kao dva prolaska kroz predloženo pravilo. Prvi prolaz odnosi se na standardni dio veze, dok se drugi prolaz odnosi na rezidualni dio veze. Nakon toga, standardni dio prolazi kroz mrežu na uobičajen način sve dok ne dosegne početnu točku razdvajanja reziduala. U ovoj fazi, preostali dio pravila reziduala primjenjuje se, što uključuje sumiranje dviju relevantnosti koje pružaju odgovarajuće komponente rezidualne veze. Ovim pristupom osigurava se da se relevantnost širi kroz cijeli rezidualni blok, hvatajući doprinose kako standardne, tako i rezidualne komponente unutar samog bloka.

Prilikom implementacije pravila za **normalizaciju nad grupom** naučenu srednju vrijednost i skaliranje standardne devijacije te množenje težinom i dodatak pristranosti, promatraju se kao dva odvojena sloja operacija. Ovo rezultira time da se predloženo pravilo primjenjuje na slojeviti način kako bi se producirala relevantnost ulaza u ovaj sloj mreže.

Slojna normalizacija uključuje tretiranje srednje vrijednosti i varijance ulaza korišteni za samu normalizaciju kao konstante te se njih izdvaja iz komputacijskog grafa. Nakon toga, predloženo pravilo se primjenjuje kako bi se izračunala relevantnost ulaza, slično kao i kod normalizacije nad grupom.

Blok samopozornosti sadrži više podoperacija, pri čemu svaka zahtijeva zaseban unatražni prolaz pomoću LRP-a: množenje između upita i ključeva unutar operacije samopozornosti, softmax operacija koja računa vrijednosti samopozornosti te množenje tih vrijednosti pozornosti s vrijednostima dobivenim transformacijom ulaza kako bi se dobio konačni izlaz bloka. Slično pristupu kod implementacije za rezidualnu vezu, provode se zasebni unatražni prolazi za svaki ulaz operacije. Konkretno, jedan prolaz za upite i drugi za ključeve u množenju samopozornosti te odvojene prolaze za matricu vrijednosti pozornosti i vrijednosti transformiranog ulaza u operaciji množenja koja proizvodi konačni izlaz. Za operaciju softmax-a, ponovno se koristi predloženo pravilo. Međutim, u ovom slučaju, nazivnik softmaxa koji sadrži zbroj eksponenciranih ulaza odvaja se od računalnog grafa, slično kao što su srednje vrijednosti i varijance bile odvojene u sloju slojne normalizacije. Kako bi se naposljetku kombinirale relevantnosti upita, ključeva i vrijednosti, kao što je bio slučaj kod rezidualne veze, njihovi utjecaju se zbrajaju.

Poglavlje 5

Predložena metrika za vrjednovanje metoda za produkciju mapa relevantnosti temeljena na globalnom vrjednovanju atribucija

Kako bi se riješili postojeći nedostaci u vrjednovanju mapa relevantnosti, u ovom radu se predlaže sveobuhvatna metrika za procjenu kvalitete određene atribucijske metode m_A promatranjem nekoliko čimbenika istovremeno i njihovo kombiniranje u jedan rezultat. Ti čimbenici su **Lokalna konzistentnost** i **Kontrastnost**.

Proces računanja metrike započinje nasumičnim odabirom četiri slike iz skupa podataka. Od navedene četiri slike, ponovno nasumičnim odabirom se određuje jedna slika kao pozitivna slika. Nakon toga, proces se nastavlja izračunom lokalne konzistentnosti za pozitivnu sliku.

5.1 Lokalna konzistentnost

Ovo je čimbenik koji nije istražen u postojećoj literaturi, budući da se većina istraživanja uglavnom usredotočila na dva poželjna aspekta metoda atribucije: vjernost i robusnost. Međutim, predložena metrika mjeri oba faktora na nov način.

Za razliku od prethodnih istraživanja koja su se oslanjala ili na rangiranje ulaznih značajki koje pruža početna mapa relevantnosti ili su koristile nasumični podskup značajki za perturbacije, predloženi pristup koristi metodu temeljenu na gradijentu u svakom koraku kako bi identificirao područja koja će rezultirati najznačajnijom ili najmanje značajnom promjenom u izlazu modela. Konkretno, u svakom koraku (označenom kao t), odabire se k posto po vrijednosti najvećih ulaznih značajki na temelju gradijentnog pristupa.

U sklopu eksperimenata unutar ovog istraživanja koristi se maskiranje vrijednošću nula kao

odabranom perturbacijskom metodom, služeći se apsolutnom vrijednosti produkta između vrijednosti ulaza i gradijenta funkcije gubitka maskiranja s obzirom na ulaz. Funkcija gubitka maskiranja jednostavno je definirana kao apsolutna vrijednost ciljnog izlaza. Ova strategija temelji se na opservaciji da veće vrijednosti gradijenta u odnosu na ulaz odgovaraju područjima slike koja najviše utječu na izlaz modela. Množenje tih vrijednosti s ulazom pruža procjenu utjecaja postavljanja tih pozicija na nulu na izlaz modela. Ovaj alternativni pristup rješava ograničenje u postojećim istraživanjima o metrikama za vrjednovanje, gdje se za atribucijsku mapu, za koju se očekuje da objasni relevantne značajke za određeni ulazni primjer, također očekuje da bude iskoristiva za potpuno nepovezani cilj - identificiranje ulaznih značajki koje uzrokuju najveću promjenu u izlazu modela. Iako se ove ideje mogu djelomično preklapati za neke ulazne značajke, one su različite i trebale bi se tretirati kao takve. Korištenjem predložene metrike ovaj problem se može ublažiti i osigurati jasna razlika između ova dva cilja.

Predložena metrika za vrjednovanje uključuje provođenje niza perturbacija na ulazni model tijekom T koraka. Za vrijeme tih perturbacija koriste se dvije različite strategije za odabir značajki: u prvom prolazu pristup najrelevantnije prvo (MoRF), a u drugom pristup najmanje relevantno prvo (LeRF).

Izvođenje ovih dviju strategija rezultira generiranjem dva različita niza vrijednosti izlaza modela. Prvi niz odgovara koračnoj perturbaciji najutjecajnijih značajki, prateći strategiju MoRF. U ovom nizu, očekuje se da će izlazna vrijednost o_t pri svakom koraku pokazati brzi pad u usporedbi s početnom izlaznom vrijednosti o_{init}, odražavajući značajan utjecaj perturbiranih značajki na predikciju modela. Nasuprot tome, drugi niz odgovara perturbaciji najmanje utjecajnih značajki, prema strategiji LeRF. Ovdje se očekuje da će vrijednost izlaza opadati sporijim tempom, odražavajući manji utjecaj ovih značajki na predikcije modela.

U sljedećem dijelu analize, generiraju se atribucijske mape za svaki perturbirani ulazni podatak pri svakom koraku, razmatrajući oba načina maskiranja - MoRF i LeRF. Potom se kvantificira stupanj sličnosti između inicijalne atribucijske mape i atribucija generiranih kroz strategije perturbacije MoRF i LeRF s pomoću sljedeće definicije sličnosti:

$$\text{sim}_t^{Mo/Le} = 1 - \frac{||A_{init} - A_t^{Mo/Le}||_1}{||A_{init}||_1 + ||A_t^{Mo/Le}||_1} \tag{5.1}$$

gdje $||A||_1$ dezignira L^1-norm izravnatog tenzora.

Ova operacija rezultira s dvije sekvence koje opisuju koliko su slične atribucijske mape u svakom koraku perturbacije, kada se perturbiraju najvažnije značajke i kada se perturbiraju najmanje važne značajke. Slično kao i izlaz, očekuje se da će sličnost između početne atribucije i atribucije trenutnog koraka brzo pasti kada se perturbiraju najutjecajnije značajke, i obrnuto, mijenjati se sporije kada se radi suprotno. Nadalje, razlika između ove dvije sekvence - d^A trebala bi korelirati s razlikom između dvije sekvence vrijednosti izlaza - d^o. S obzirom na to da

se maskiraju ključni dokazi povezane s određenom klasom, očekuje se da će atribucijska mapa postupno odstupati od svojeg početnog stanja kako se ti dokazi uklanjaju, rezultirajući smanjenjem vrijednosti izlaza. Obrnuto, kada se perturbiraju najmanje važne značajke i zadržava samo najvažniji dokaz u ulazu, karta atribucije trebala bi ostati relativno stabilna, dok će vrijednost izlaza ostati bliska svojoj početnoj vrijednosti.

Kako bi se izračunala korelacija između ove dvije krivulje, započinje se normalizacijom svih vrijednosti izlaza s pomoću početne vrijednosti izlaza. Ovaj korak normalizacije ključan je kako bi se osigurala normaliziranost metrike između različitih primjera, jer različiti ulazi mogu imati različite početne izlazne vrijednosti. Normalizacija vrijednosti pomaže u sprječavanju pristranosti prema klasama s prosječno višim ili nižim konačnim izlazima. Nakon provedbe normalizacije, nastavlja se s izračunom komponente robusnosti kao dijela mjere lokalne konzistentnosti predložene metrike:

$$\mathbf{LC}_R(m_A) = 1 - \frac{2||d^o - d^A||_1}{||d^o||_1 + ||d^A||_1} \tag{5.2}$$

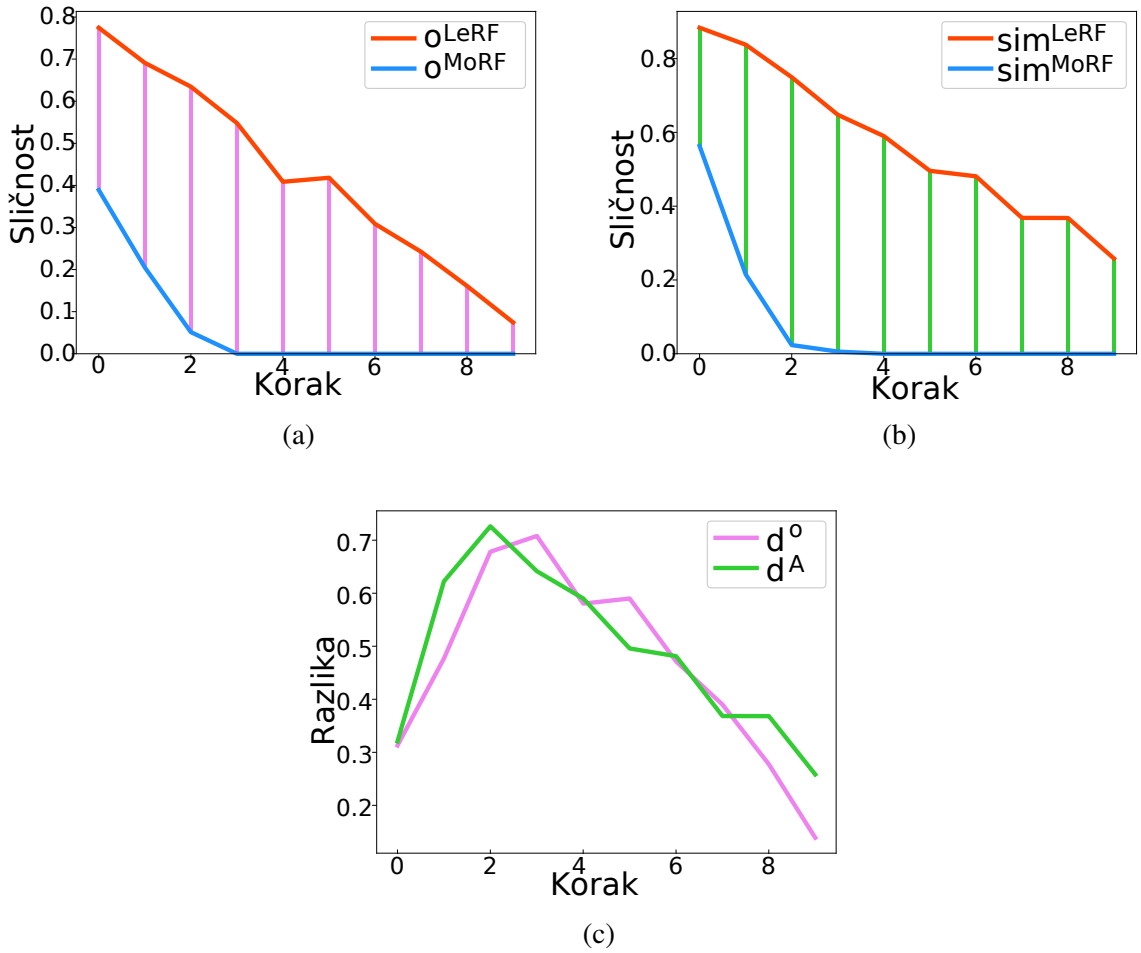
$$d_t^o = o_t^{Le} - o_t^{Mo} \tag{5.3}$$

$$d_t^A = sim_t^{Le} - sim_t^{Mo} \tag{5.4}$$

Ova mjera producira vrijednosti između -1 i 1, što ukazuje na kvalitetu lokalne robusnosti atribucijske metode. Primjer spomenutih krivulja dobivenih opisanim dijelom predložene metrike na nasumičnom primjeru iz ImageNet skupa podataka prikazan je na Slici 5.1.

Za procjenu aspekta vjernosti mjere lokalne konzistentnosti, koriste se zabilježene vrijednosti apsolutnih umnožaka gradijenta vrijednosti izlaza s obzirom na ulaz te samog ulaza za svaki korak. Intuitivno, prilikom maskiranja najmanje važnih ulaznih značajki, najviše vrijednosti u tim mapama ostaju stabilne i usredotočene oko klasificiranog objekta tijekom postupka maskiranja. Nasuprot tome, prilikom maskiranja najvažnijih značajki, visoke vrijednosti unutar zabilježenih mapa kreću se po slici kako se postupno maskiraju, rezultirajući time da model dobiva manje i manje dokaza za ciljnu klasu.

Kao rezultat toga, kumulativna suma ovih koračnih mapa, dobivenih slijednim maskiranjem najmanje važnih područja, koncentrirana je na ciljni objekt (Slika 5.2b). Naspram tome, kumulativne sume mapa dobivenih sekvencijalnim maskiranjem najvažnijih područja pokazuju manje fokusa na ciljno područje te posjeduju više uniformnu distribuciju rasprostranjenu na pozadinu slike, odnosno dijelu slike koji proizvodi manje aktivacije unutar mreže s obzirom na ciljnu klasu (Slika 5.2c). Razlika između ovih dviju mapa rezultira konačnom mapom, nazvanom "kombinirana mapa utjecaja" (I_c) (Slika 5.2d), gdje pozitivne vrijednosti ukazuju na područja visokog utjecaja, a negativne vrijednosti na područja niskog utjecaja. Ova mapa se



Slika 5.1: Krivulje promjene za slučajno odabrani primjer iz ImageNet skupa podataka, VGG model, RAMP atribucijska metoda: **(a)** MoRF i LeRF krivulje promjene za izlaz modela **(b)** MoRF i LeRF krivulje promjene za atribucijske mape **(c)** Krivulje razlike za izlaz modela i atribucijske mape. Manja razlika u ovim krivuljama indikacija je atribucijske metode koja je lokalno robusnija.

koristi za procjenu vjernosti promatrane atribucijske metode. Množenjem predznaka kombinirane mape utjecaja s izvornom atribucijom dobivaju se razložena točno procijenjena područja visokog utjecaja (pozitivni rezultat atribucije i pozitivan predznak utjecaja) od nepravilno procijenjenih područja, gdje je atribucija dala visok rezultat relevantnosti, ali je predznak utjecaja bio negativan.

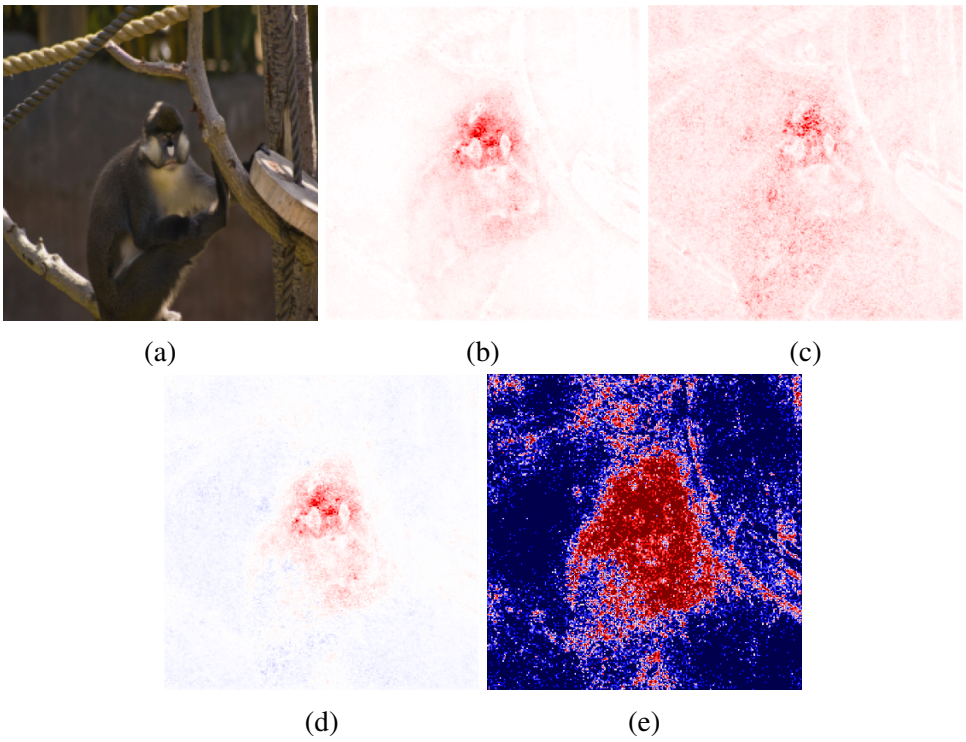
Komponenta vjernosti predložene mjere lokalne konzistentnosti definirana je kao:

$$\mathbf{LC}_F(m_A) = \text{sum}\{\frac{A_{init} \cdot \text{sign}(I_c)}{||A_{init}||_1}\} \tag{5.5}$$

gdje sum označava sumu po svim dimenzijama tenzora.

Predznak kombinirane mape utjecaja (Slika 5.2e) množi se s inicijalnom mapom relevantnosti te se ovaj umnožak globalno sumira te normalizira zbrojem inicijalne mape relevantnosti. To rezultira ocjenom koja se nalazi između -1 i 1, ovisno o kvaliteti lokalne vjernosti atribucij-

ske metode.



(a) (b) (c)

(d) (e)

Slika 5.2: Kombinirana mapa utjecaja promjene za slučajno odabrani primjer iz ImageNet skupa podataka: **(a)** Izvorna slika **(b)** Zbroj apsolutnih produkata ulaza i gradijenta po koracima prilikom sekvencijalnog maskiranja nisko utjecajnih značajki **(c)** Zbroj apsolutnih produkata ulaza i gradijenta po koracima prilikom sekvencijalnog maskiranja visoko utjecajnih značajki **(d)** Razlika između (a) i (b) **(e)** Predznak (d)

Na posljetku, **Mjera lokalne konzistencije** definirana je kao pozitivno ograničena srednja vrijednost komponenti robusnosti i vjernosti:

$$\mathbf{LC}(m_A) = (\frac{\mathrm{LC}_R(m_A) + \mathrm{LC}_F(m_A)}{2})^+ \tag{5.6}$$

Faktor lokalne konzistentnosti u predloženoj metrici učinkovito procjenjuje sposobnost atribucijske metode da identificira visoko utjecajna područja unutar slike. Međutim, važno je napomenuti da ova utjecajna područja nisu nužno specifična za određenu klasu, budući da više ciljnih klasa može dijeliti ista utjecajna područja. Kako bi se riješilo ovo ograničenje i pružilo mjeru kvalitete utjecajnih područja specifičnih za određenu klasu, predloženoj metrici vrjednuje se i dodatni faktor nazvan **Kontrastnost**.

5.2 Kontrastnost

Inspirirana metrikom Focus [139], predložena komponenta kontrastnosti se razlikuje u tome što koristi četiri slike - jednu pozitivnu i tri negativne, uz njihove predviđene odgovarajuće klase

(c_p za pozitivnu i c_{n_i} za negativne). Budući da postoji samo jedan pozitivan i tri negativna kvadranta, oznake ne trebaju biti poznate unaprijed kako bi se uzorkovala dva pozitiva, kao što je bio slučaj kod metrike Focus. Nakon što se dobiju softmax izlazi modela s^p za pozitivnu sliku, sve četiri slike se kombiniraju u mozaik. Položaji pozitivnih i negativnih slika u mozaiku nasumično se biraju kako bi se spriječila pristranost određenoj poziciji unutar mozaika. Potom se definira *mapa ocjene* S_{mozaik} istog oblika kao i ulazni mozaik, pri čemu su vrijednosti unutar pozitivnog kvadranta postavljene na 1, a unutar drugih kvadranta, indeksiranih s i kao $\frac{2s^p[c_{n_i}]}{s^p[c_p]} - 1$. Ova definicija postavlja vrijednosti u negativnim kvadrantima na -1 ako se njihova odgovarajuća negativna klasa ne nalazi u pozitivnoj slici, i obrnuto, za slične ili identične klase vrijednosti se približavaju 1. Ekstreman slučaj je ako se uzorkuje više slika s istom predviđenom klasom. U tom slučaju, svi kvadranti mape ocjene koji sadrže tu klasu postavili bi se na 1 ako je njihova klasa odabrana kao pozitiv. Na ovaj način, atribucijska metoda se ne kažnjava ako se dio njezine atribucije nalazi u kvadrantu koji sadrži slične ili identične klase. Konačno, producira se atribucijska mapa mozaika A_{mozaik} određenom atribucijskom metodom gdje je zatražena atribucija za pozitivnu klasu te se nakon toga računa **ocjena kontrastnosti**:

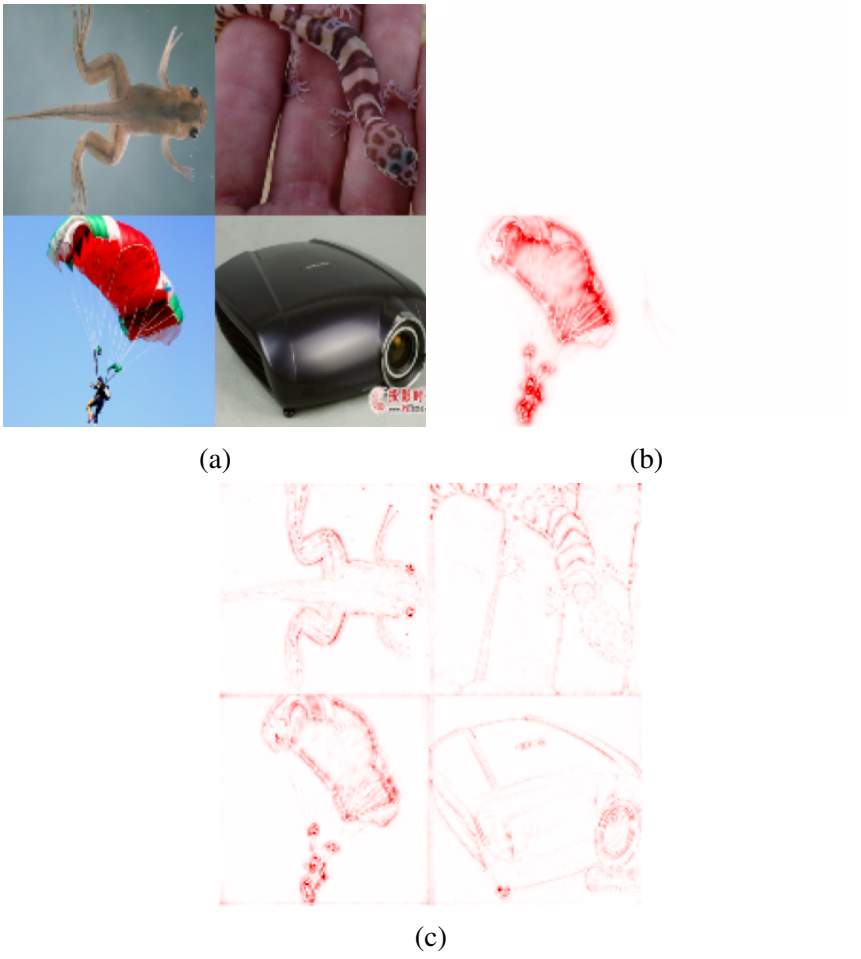
$$\mathbf{C}(m_A) = \text{sum}\{\frac{A_{mosaic} \cdot S_m}{||A_{mosaic}||_1}\}^+ \tag{5.7}$$

Ocjene kontrastnosti blizu 1 ukazuju na bolju sposobnost stvaranja kontrastnih mapa relevantnosti. Primjeri visokih i niskih ocjena kontrastnosti za mozaik sliku na Slici 5.3a prikazani su na Slici 5.3b - predložena RAMP metoda i 5.3c - LRP-$\alpha_2\beta_1$ metoda.

Konačno, kako bi ispunila zahtjeve lokalne konzistentnosti i kontrastnosti, pouzdana atribucijska metoda mora pokazati visoke rezultate za oba faktora. Stoga, konačna metrika GAE-a za atribucijsku metodu atribucije m_A određuje se kao produkt ova dva faktora:

$$\mathbf{GAE}(m_A) = \text{LC}(m_A) \cdot \text{C}(m_A) \tag{5.8}$$

U svim prethodno navedenim izračunima koriste se pozitivni dijelovi mapa atribucije, koje su normalizirane tako da najveća vrijednost odgovara 1. Fokusiranjem na pozitivne dijelove i normalizacijom atribucija, osigurava se dosljednost i usporedivost rezultata za različite ulazne primjere, omogućavajući točnu procjenu relevantnosti i utjecaja različitih značajki u atribucijskim mapama.



(a) (b)

(c)

Slika 5.3: **(a)** Primjer mozaik slike, dobivene uzorkovanjem četiri slike iz skupa podataka ImageNet **(b)** Atribucijska mapa RAMP metode (samo pozitivne vrijednosti) uz odabir donje lijeve slike kao ciljne klase **(c)** Atribucijska mapa LRP-$\alpha_2\beta_1$ (samo pozitivne vrijednosti) - primjer atribucijske metode s lošim kontrastnim svojstvima

Poglavlje 6

Predložena metoda za vrjednovanje poravnanja metrika za vrjednovanje atribucijskih metoda s ljudskom percepcijom kvalitete

Moderni sustavi umjetne inteligencije često koriste modele koji su potpuno neprozirni s obzirom na proces donošenja odluka. Ova neprozirnost je posebno problematična u sigurnosno kritičnim domenama gdje posljedice odluka autonomnih sustava mogu biti značajne, čime se naglašava potreba za transparentnim i pouzdanim sustavima umjetne inteligencije.

Objašnjiva umjetna inteligencija (engl. *Explainable Artificial Intelligence* (XAI)) nastoji riješiti izazove interpretabilnosti razvijanjem metoda koje čine procese donošenja odluka interpretabilnima. Kako je područje XAI-a raslo, predložene su razne metode interpretacije odluka modela, od kojih svaka ima svoj pristup pojašnjavanju načina na koji modeli donose odluke. U skladu s tim, predložene su i brojne metrike za procjenu učinkovitosti metoda u pojašnjavanju predviđanja modela. Pretpostavka je da viši rezultati na ovim metrikama za vrjednovanje koreliraju s većom procijenjenom kvalitetom od strane korisnika. Međutim, područje istraživanja koje se bavi usklađenošću ovih metrika za vrjednovanje s ljudskom percepcijom je ograničeno. Unatoč njihovoj širokoj upotrebi u suvremenim istraživanjima atribucije, odnos između metrika i stvarne ljudske interpretabilnosti ostaje nedovoljno istražen [141] [99].

U ovom kontekstu, ovaj rad predstavlja pionirsku korisničku studiju koja uspoređuje atribucijske mape producirane od strane prednaučenih modela s onima proizvedenim od strane modela fino podešenih na temelju ispitanih metrika za vrjednovanje. Predloženo rješenje nudi nepristranu metodologiju za procjenu učinkovitosti metrike u vođenju modela prema proizvodnji atribucijskih mapa usklađenih s ljudskom percepcijom.

6.1 Povezana istraživanja

Hedström i sur. definiraju dva kritična kriterija neuspjeha koje metrike za vrjednovanje trebaju izbjegavati. Prvi kriterij je da promatrana metrika mora biti otporna na manje perturbacije u svojim ulaznim parametrima. Ovo svojstvo je slično svojstvu stabilnosti, odnosno robusnosti atribucijskih metoda, promatrano u istraživanjima Agarwal i sur. [142] te Montavon i sur. [123], te se usklađuje s konceptom Lipschitz kontinuiteta kojeg su istaknuli Alvarez-Melis i sur. [121] te Yeh i sur. [125]. Slično istraživanjima koje su proveli Brunke i sur. [143], Brocki i sur. [144] te Rong i sur. [127], autori ovog rada mjere ranjivost metrike na varijacije, odnosno "male smetnje".

Nadalje, autori vrjednuju otpornost metrike s obzirom na drugi kriterij neuspjeha: metrika za vrjednovanje bi trebala proizvesti značajnu promjenu u procjeni kvalitete atribucije prilikom izlaganja disruptivnim perturbacijama svojih ulaznih parametara. Ovaj drugi kriterij neuspjeha uvodi provjeru reaktivnosti na značajne promjene. On se proteže istraživanjima poput Rieger i sur. [124] te Arias-Duart i sur. [139] koja ukazuju na to da bi se rezultati metrike za vrjednovanje trebali značajno razlikovati kada su generirani iz nasumične atribucijske metode ili nasumično inicijaliziranog modela.

Međutim, automatsko vrjednovanje metrika za vrjednovanje atribucijskih metoda, umjesto ponude rješenja na pitanje: "Koja metrika dobro odgovara ljudskoj percepciji?", samo prosljeđuje upit na višu razinu. Ovaj proces ostavlja nesigurnost u pogledu toga odgovara li vrjednovanje metrika doista ljudskoj percepciji.

Druga istraživanja su pokušala povezati vrjednovanje metrika i ljudskog označavanja. Nguyen i sur. [145] provode korisničku studiju usmjerenu na zadatak unaprijednog predviđanja. U ovom istraživanju, označivačima se daje uputa predviđanja izlaza modela na temelju vizualizacije izlaza atribucijske metode. Nakon provedbe korisničke studije, autori istraživanja su računali korelacije između točnosti korisnika u navedenom zadatku i mjera dobivenih od strane tri metrike za vrjednovanje: IoU [146], WSL [146], i Pointing Game [72]. Njihovi zaključci otkrivaju svega marginalnu pozitivnu korelaciju s ljudskom točnošću.

Međutim, koreliranje ljudske izvedbe na zasebnom zadatku s kvalitetom atribucijskih metoda procjenjuje vezu između ljudske percepcije i atribucija samo neizravno. U ovom pristupu, anotatori ne uspoređuju izravno dvije atribucijske mape; umjesto toga, vrjednovanje se fokusira na to kako atribucijske mape utječu na ljudsku izvedbu u spomenutom zadatku.

U ovom radu se predlaže nova metoda vrjednovanja metrika, gdje ljudski označivači izravno uspoređuju izlaze atribucijskih metoda između izvornog modela i modela optimiziranog prema promatranoj metrici. Ako označivači, nakon što im se prezentiraju obje mape, dosljedno preferiraju optimiziranu verziju u značajnom broju primjera, moguće je zaključiti da je promatrana metrika za vrjednovanje dobro usklađena s ljudskom percepcijom.

Kako bi se ovaj proces izveo, potrebno je definirati funkcije gubitka za unatražni prolaz i optimizaciju parametara modela za poboljšanje rezultata ostvarenih na određenoj metrici uz očuvanje radnih značajki izvornog modela. Atribucijske mape optimiziranog modela bi, prema promatranoj metrici, trebale pružati jasnije razumijevanje odluke modela.

6.2 Predložena metoda

U ovom poglavlju najprije se predstavlja predložena metoda za optimizaciju modela, a zatim slijedi detaljan opis korisničke studije provedene za vrjednovanje kvalitete atribucijskih mapa.

6.2.1 Optimizacija modela korištenjem metrika za vrjednovanje

Budući da većina implementacija spomenutih metrika za vrjednovanje ne podržava prolaz unatrag radi optimizacije parametara modela zbog složenosti formulacije ili računalne nepraktičnosti same metrike, unutar ovog istraživanja uvode se surogatne funkcije gubitka za svaku pojedinu metriku. Minimizacijom ovih funkcija gubitka izravno se poboljšavaju rezultati modela na izvornoj metrici za vrjednovanje.

6.2.2 Brisanje i rješavanje pristranosti te Iterativno brisanje značajki

Način rada metrike Brisanja i rješavanja pristranosti (engl. *Remove and Debias* (ROAD)) opisan je u ranijem Poglavlju 3.1. Tijekom eksperimenata vezanih uz vrjednovanje poravnanja između metrika i ljudske percepcije koristi se MoRF verzija ove metrike čiji je finalni rezultat dodatno invertiran oduzimajući vrijednost metrike od broja jedan kako bi smjer poboljšanja kvalitete atribucija odgovarao smjeru pozitivnog povećanja same metrike te isto tako odgovarao smjeru metrike navedene u nastavku. Ova inverzija ni na koji način ne utječe na samu metriku, već je ova izmjena načinjena radi mogućnosti definiranje jedne surogatne funkcije gubitka za metrike iz kategorije vjernosti.

U sklopu ovog istraživanja, uključuje se i još jedna metoda iz područja vjernosti pod nazivom Iterativno brisanje značajki (engl. *Iterative Removal Of Features* (IROF)) [124]. Autori ovog rada su motivirani činjenicom da visokodimenzionalni ulazi u model, poput slika, imaju visoko korelirane značajke. Prema tome, maskiranje jednog piksela neće napraviti značajnu razliku u izlazu modela te je efikasnije maskirati veću površinu slike odjednom i promatrati utjecaj na odluku modela. Stoga koriste superpiksele dobivene segmentacijskom metodom Jednostavnog linearnog iterativnog grupiranja (engl. *Simple Linear Iterative Clustering (SLIC)*) [147] koji se iterativno maskiraju prema srednjoj vrijednosti atribucije svakog segmenta te se bilježi promjena izlaza modela. Finalna ocjena definirana je kao površina iznad niza omjera izlaznih vrijednosti izvornog izlaza i izlaza dobivenih za maskirane ulaze. Ovim mehanizmom

maskiranja postiže se brza metoda koja zahtijeva manje računalnih resursa u usporedbi s klasično korištenim metrikama vjernosti.

Kako bi se povećao rezultat na metrikama ROAD i IROF, bitno je povećati razliku između izlaza izvornog modela i izlaza fino podešenog modela $f_{student}$ za određenu ciljnu klasu y. Umjesto istovremenog maksimiziranja ove razlike kroz sve korake, račun se pojednostavljuje tako što se inicijalno uzorkuje indeks koraka s. Potom se producira maskirana ulazna slika koristeći funkciju perturbacije $p^s_{ROAD/IROF}$ za taj određeni korak. Nakon toga, dobivaju se izlazi za perturbirani ulaz i neperturbirani ulaz x jednostavnim prolazom unaprijed kroz model. Konačno, minimizira se omjer između ova dva izlaza:

$$\mathscr{L}^{ROAD/IROF}_{metric} = \frac{f_{student}(p^s_{ROAD/IROF}(x))_y}{f_{student}(x)_y} \tag{6.1}$$

6.2.3 Relativna stabilnost ulaza

Metrika Relativne stabilnosti ulaza (engl. *Relative Input Stability* (RIS)) predložena od strane Agarwal i sur. [142] unaprjeđuje ideju Alvarez-Melis i sur. [136]. Autori ovog rada predlažu korištenje maksimalnog omjera normaliziranih udaljenosti između izvorne atribucijske mape i atribucije slučajno odabranih susjeda te normalizirane udaljenosti između ulaznih značajki i ulaznih značajki susjeda.

Optimizacija RIS metrike uključuje minimizaciju maksimalne udaljenosti između izvorne atribucijske mape a^o i skupa atribucijskih mapa proizvedenim uvođenjem malih perturbacija u izvorni ulaz x. U ovom slučaju koristi se strategiju uzorkovanja jednog perturbiranog primjera x^p te se potom smanjuje udaljenost između odgovarajuće perturbirane atribucijske karte a^p i izvorne atribucijske karte a^o kroz unatražni prolaz primijenjen izravno na samu atribucijsku metodu. Ova udaljenost je normalizirana udaljenošću između izvornog i perturbiranog ulaza. Koristi se L1 norma kao mjera udaljenosti:

$$\mathscr{L}^{RIS}_{metric} = \frac{||\frac{a^p - a^o}{a^o}||_1}{||\frac{x^p - x^o}{x^o}||_1} \tag{6.2}$$

6.2.4 Focus i Točnost mase relevantnosti

Način rada metrike Focus opisan je u ranijem Poglavlju 3.3. U ovom istraživanju uvodi se modifikacija na opisani proces konstrukcije mozaika. Naime, izmijenjena verzija najprije uzorkuje četiri slike iz skupa podataka te nad njima računa softmax izlaz modela. Jedna klasa se odabire kao pozitivna (c_p), a ostale kao negativne (c_{n_i}). Stvara se mozaik te se tvori i ciljna mapa, gdje se vrijednost svakog kvadranta postavlja na $\frac{s^p[c_{n_i}]}{s^p[c_p]}$, gdje je s^p softmax distribucija pozitivnog kvadranta. Dobivena normalizirana ciljna mapa m^{gt} se zatim množi s atribucijom mozaika a te se sumira kako bi se dobio konačni rezultat. Ova modificirana metoda eliminira potrebu za pret-

hodnim označavanjem i uzima u obzir meke predikcije modela, omogućujući blago kažnjavanje ako su pozitivne i negativne klase slične.

S druge strane, metrika Točnost mase relevantnosti (engl. *Relevance Mass Accuracy* (RMA)) definirana je jednostavno kao omjer količine atribucije određene slike m^{gt} koja prekriva ciljnu mapu a. Kao ciljne mape koriste se gotove segmentacijske mape dobivene iz ImageNet-S skupa podataka [148].

$$\mathscr{L}_{metric}^{Focus/RMA} = 1 - \frac{||m^{gt} \cdot a||_1}{||a||_1} \tag{6.3}$$

6.2.5 Održavanje blizine izvornom modelu

U finalnom koraku predložene metode, koristi se gubitak kosinusne sličnosti $\mathscr{L}_{cosine}$ kako bi se održala usklađenost između radnih značajki fino podešenog modela i izvornog modela. Ova funkcija gubitka, nadahnuta destilacijom znanja [149], održava izlaze fino podešenog modela blizu izvornog. Optimizacijom ponderirane sume gubitka kosinusne sličnosti i gubitka metrike, koristeći hiperparametar α kako bi se prilagodila snaga signala učenja metrike, osigurava se minimalno odstupanje fino podešenog modela od izvornog, uz adekvatnu optimizaciju rezultata modela postignutom na promatranoj metrici:

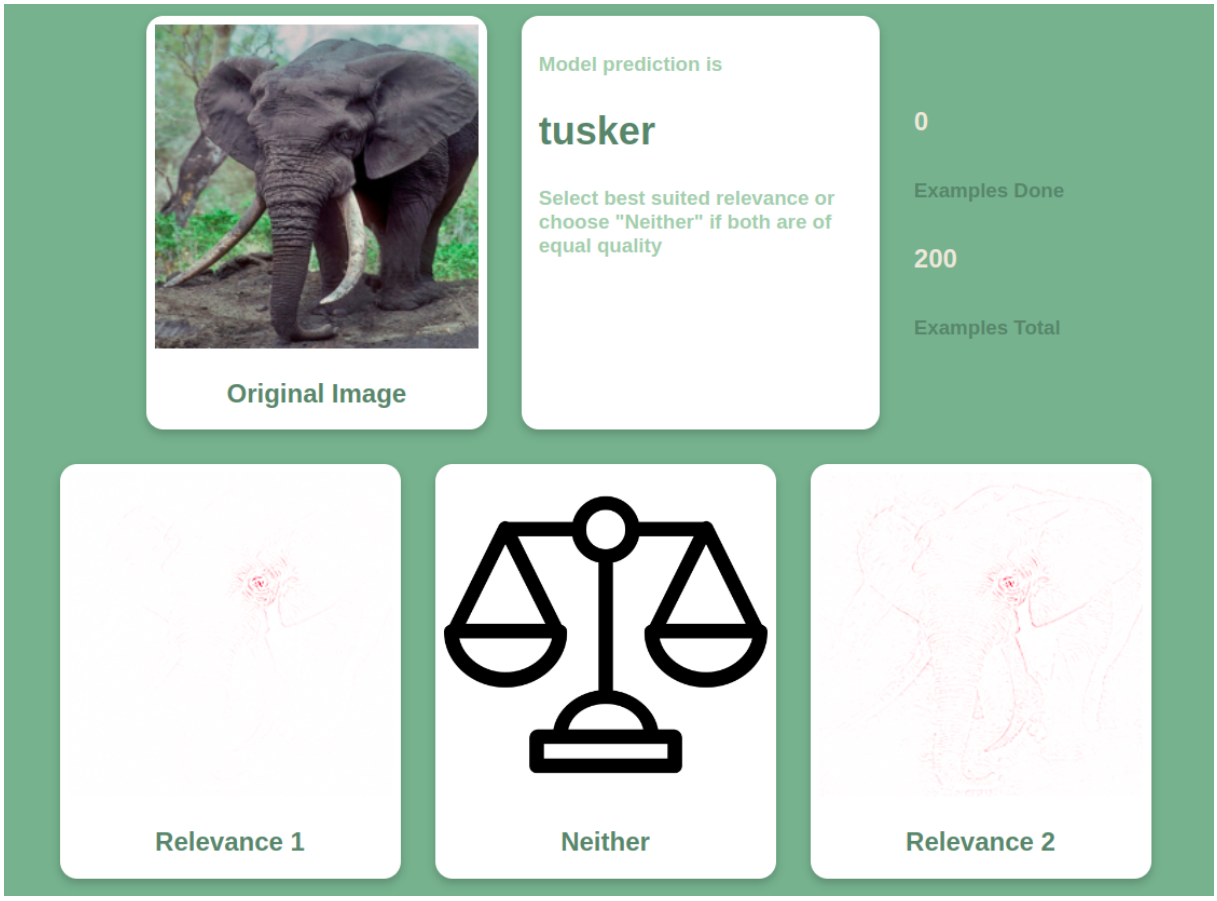
$$\mathscr{L}_{combined} = \mathscr{L}_{cosine} + \alpha \mathscr{L}_{metric} \tag{6.4}$$

Nakon fino podešavanja modela pomoću prethodno navedenih funkcija gubitka, generirano je 200 primjera za svaku jedinstvenu kombinaciju modela, atribucijske metode i metrike. Kako ovo istraživanje obuhvaća dva modela, dvije atribucijske metode i pet metrika, to rezultira s ukupno 20 eksperimenata i 4000 primjera. Za svaki primjer zabilježena je izvorna slika, predikcija modela, izvorna atribucijska mapa i atribucijska mapa fino podešenog modela. Važno je napomenuti da su u eksperimentu razmatrane samo slike s točnim predikcijama oba modela s obzirom na klasu definiranu unutar ImageNet skupa podataka.

6.2.6 Korisnička studija

Pet stručnjaka za duboko učenje sudjelovalo je u korisničkoj studiji. U sklopu studije, razvijena je web aplikacija za korisničko označavanje, prikazana na Slici 6.1. Za svaki primjer, korisnici su na raspolaganju imali izvornu sliku i predikciju modela, te su potom bili upućeni u odabir između *izvorne* (engl. *original*) atribucijske mape, *fino podešene* (engl. *fine-tuned*) atribucijske mape, ili opcije *niti jedna* (engl. *neither*). Korisnici su odabirali atribuciju za koju smatraju da bolje predstavlja predikciju modela. Ako su obje mape bile jednake kvalitete ili ih nije bilo moguće razlikovati, korisnicima je rečeno da odaberu opciju *niti jedna*.

Osim toga, kako bi se ublažila pristranost, pozicije izvornih i fino podešenih atribucijskih mapa bile su nasumično dodijeljene za svaki primjer prije nego što su prikazane korisniku. Ovaj pristup imao je za cilj osigurati nepristranu procjenu kvalitete svakog primjera tako što sprječava korisnike da preferiraju određenu poziciju.



Slika 6.1: Primjer iz korisničke studije. Izvorna slika i predikcija modela prikazani su u prvom redu. U sljedećem redu vizualizirane su atribucijske mape izvornog modela za ciljnu klasu i mapa fino podešenog modela te opcija "niti jedna".

Poglavlje 7

Rezultati predložene atribucijske metode i metrike za vrjednovanje

U ovom poglavlju provodi se vrjednovanje predložene atribucijske metode s pomoću predložene metrike za vrjednovanje metoda za produkciju mapa relevantnosti. Cilj eksperimenata je sveobuhvatno procijeniti radne značajke predložene atribucijske metode te trenutno najboljih (engl. *state-of-the-art*) i najčešće korištenih metoda u ovom području istraživanja. Rezultati su prikazani na raznim modelima različitih arhitektura te skupovima podataka. Radi pružanja kompletne analize, u rezultate su uključena i vrjednovanja dobivena postojećim, često korištenim metrikama za vrjednovanje iz kategorija vjernosti, robusnosti i lokalizacije. Dodatno, provodi se i ablacijska studija na modelu Vizualnog Transformera kako bi se pokazale ključne prednosti predložene atribucijske metode u usporedbi s postojećim metodama. Nadalje, prikazuju se kvalitativni eksperimenti kako bi se i vizualno potvrdila prednost RAMP metode atribucije u raznim scenarijima. Na kraju, kvalitativno se prikazuje kako predložena GAE metrika učinkovito razlikuje između visokokvalitetnih i niskokvalitetnih atribucijskih mapa.

7.1 Kvantitativni eksperimenti

Predložena RAMP metoda vrjednuje se na dva javno dostupna skupa podataka za klasifikaciju slika - ImageNet (Deng i sur. [150]) i PascalVOC2012 (Everingham i sur. [151]), koristeći GAE metriku za vrjednovanje. Broj koraka za postupak maskiranja prilikom izračuna lokalne konzistentnosti postavljen je na $T = 10$ u svim eksperimentima. Za oba skupa podataka vrjednuju se razne atribucijske metode na tri prednaučena modela za klasifikaciju slika: VGG16 [152], ResNet50 [92] i ViT-Base [82]. Značajnost razlike u srednjim vrijednostima ukupnih rezultata provjerena je Wilcoxonovim rang testom s razinom značajnosti postavljenom na $p = 0.05$.

U prvom kvantitativnom eksperimentu, gdje se kao odabrani model promatra VGG, predložena metoda se uspoređuje s trenutno najboljim i najčešće korištenim atribucijskim metodama

za konvolucijske mreže: Saliency [111], Input*Gradient [70], Deconvolution [39], LRP-ε, DeepLIFT [70], LRP-$\alpha_1\beta_0$, LRP-$\alpha_2\beta_1$, Integrated Gradients [42], SmoothGrad [113], GradCAM [43], HiResCAM [117], LayerCAM [118], GradCAM++ [116], contrastive-LRP (cLRP) [140], GuidedGradCAM [43], Relative Attributing Propagation (RAP) [81] i contrastive-RAP (cRAP), što je proširena verzija RAP-a koja koristi istu metodu dobivanja kontrastivnih mapa opisanu u radu o kontrastnom LRP-u. U eksperiment su također uključene i dvije bazne atribucijske metode (engl. *baseline*): konstantna atribucija koja uvijek daje vrijednost jedan kao mapu atribucije i nasumična atribucija uzorkovana iz normalne distribucije, kako bi se dodatno prikazali nedostaci određenih metrika za vrjednovanje te snage predložene metrike.

Rezultati prvog eksperimenta mogu se pronaći u Tablicama 7.1 i 7.2. Predložena metoda nadmašuje sve navedene metode sa značajnom razlikom na oba skupa podataka, postižući ukupni rezultat od 0.272 na ImageNetu i 0.238 na PascalVOC-u. Slijedi ju GuidedGradCAM s rezultatima od 0.207 i 0.167 na odgovarajućim skupovima podataka te cLRP s rezultatima 0.12 i 0.112. Većina ostalih metoda postiže znatno niže rezultate od ovih metoda, prvenstveno zbog nedostatka kontrasnosti, jer njihove mape atribucije ne variraju ovisno o odabranoj ciljnoj klasi. Ovo zapažanje nedostatka kontrastnosti opisano je i od strane Gu i sur. [140] i Chefer i sur. [99]. Kako bi se prikazala kompletna slika, u rezultate su uključene i prosječne vrijednosti ocjena lokalne konzistentnosti i kontrastnosti. Iako promatranje ovih ocjena zasebno može ponuditi uvid u probleme ili prednosti pojedine atribucijske metode, oslanjanje na samo jedan faktor dovodi do sličnih nedostataka kao prisutnih kod ostalih metrika za vrjednovanje.

U sljedećem eksperimentu procjenjuju se rezultati nekoliko metoda atribucije, uključujući Integrated Gradients, SmoothGrad, GradCAM, HiResCAM, LayerCAM, GradCAM++, cLRP, GuidedGradCAM, RAP, cRAP i predložene nove metode. Ove metode su vrjednovane na ResNet50 modelu na oba skupa podataka, ImageNet i PascalVOC.

Kao što je vidljivo iz Tablica 7.3 i 7.4, predložena metoda pokazuje superiorne performanse u usporedbi s većinom drugih vrjednovanih metoda. Međutim, vrijedi napomenuti da cLRP metoda pokazuje istu razinu rezultata kao i predložena metoda za ovaj specifični model. Pretpostavka je da se bliske radne značajke ova dva pristupa manifestiraju zbog smanjene osjetljivosti modela na problem relativne magnitude atribucije, kao što je prikazano na Slici 2.2. Međutim, precizan razlog ove neosjetljivosti ostaje nepoznat jer je moguće da je to rezultat utjecaja različitih faktora poput rezidualne arhitekture, normalizacije nad grupom ili jednostavno razlikom u hiperparametrima prilikom učenja modela.

Predložena metoda i cLRP metoda nadmašuju sve ostale metode sa statistički značajnom razlikom, pri čemu su njihovi ukupni GAE rezultati više nego dvostruko veći od sljedeće najbolje metode atribucije, koja je u ovom slučaju GuidedGradCAM.

Naposljetku, procjenjuju se radne značajke različitih metoda atribucije na modelu Vizualnog Transformera, uključujući GradCAM, GradCAM++, HiResCAM, LayerCAM, *Last Layer*

Tablica 7.1: GAE mjera ostvarena od strane različitih atribucijskih metoda na ImageNet skupu podataka i VGG arhitekturi

Method name	$\overline{LC}$	$\overline{C}$	$\overline{GAE}$
Constant	0.0	0.0	0.0
Random	0.0	0.0	0.0
Saliency	0.0	0.011	0.0
Input*Gradient	0.076	0.016	0.0
Deconvolution	0.0	0.0	0.0
LRP-ε	0.117	0.022	0.001
DeepLIFT	0.159	0.08	0.011
LRP-$\alpha_1\beta_0$	0.318	0.003	0.001
LRP-$\alpha_2\beta_1$	0.307	0.016	0.006
Integrated Gradients	0.139	0.022	0.002
SmoothGrad	0.004	0.016	0.0
GradCAM	0.132	0.424	0.06
HiResCAM	0.107	0.342	0.038
LayerCAM	0.079	0.089	0.008
GradCAM++	0.093	0.056	0.006
GuidedGradCAM	0.281	0.665	0.207
cLRP	0.3	0.407	0.12
RAP	0.383	0.202	0.097
cRAP	0.106	0.542	0.053
RAMP	0.396	0.676	**0.272**

Tablica 7.2: GAE mjera ostvarena od strane različitih atribucijskih metoda na PascalVOC skupu podataka i VGG arhitekturi

Method name	$\overline{LC}$	$\overline{C}$	$\overline{GAE}$
Constant	0.0	0.001	0.0
Random	0.0	0.001	0.0
Saliency	0.0	0.013	0.0
Input*Gradient	0.085	0.023	0.002
Deconvolution	0.0	0.002	0.0
LRP-ε	0.134	0.027	0.003
DeepLIFT	0.161	0.071	0.01
LRP-$\alpha_1\beta_0$	0.312	0.009	0.002
LRP-$\alpha_2\beta_1$	0.3	0.022	0.006
Integrated Gradients	0.157	0.029	0.004
SmoothGrad	0.002	0.018	0.0
GradCAM	0.112	0.372	0.042
HiResCAM	0.096	0.287	0.025
LayerCAM	0.069	0.089	0.007
GradCAM++	0.093	0.063	0.006
GuidedGradCAM	0.254	0.606	0.167
cLRP	0.267	0.398	0.112
RAP	0.36	0.242	0.094
cRAP	0.11	0.497	0.054
RAMP	0.399	0.594	**0.238**

Tablica 7.3: GAE mjera ostvarena od strane različitih atribucijskih metoda na ImageNet skupu podataka i ResNet50 arhitekturi

Method name	$\overline{LC}$	$\overline{C}$	$\overline{GAE}$
Integrated Gradients	0.127	0.119	0.015
SmoothGrad	0.003	0.141	0.001
GradCAM	0.181	0.385	0.07
HiResCAM	0.186	0.307	0.056
LayerCAM	0.187	0.299	0.055
GradCAM++	0.24	0.391	0.093
GuidedGradCAM	0.243	0.523	0.127
cLRP	0.358	0.792	**0.283***
RAP	0.004	0.109	0.001
cRAP	0.073	0.599	0.042
RAMP	0.361	0.75	**0.272***

*Razlika u rezultatima nije statistički značajna prema Wilcoxon testu ranga na razini značajnosti p=0.05

Tablica 7.4: GAE mjera ostvarena od strane različitih atribucijskih metoda na PascalVOC skupu podataka i ResNet50 arhitekturi

Method name	$\overline{LC}$	$\overline{C}$	$\overline{GAE}$
Integrated Gradients	0.088	0.101	0.009
SmoothGrad	0.0	0.128	0.0
GradCAM	0.165	0.391	0.07
HiResCAM	0.165	0.299	0.054
LayerCAM	0.157	0.261	0.047
GradCAM++	0.21	0.371	0.082
GuidedGradCAM	0.189	0.576	0.11
cLRP	0.312	0.863	**0.266***
RAP	0.005	0.099	0.002
cRAP	0.058	0.597	0.04
RAMP	0.299	0.828	**0.251***

*Razlika u rezultatima nije statistički značajna prema Wilcoxon testu ranga na razini značajnosti p=0.05

Attention, *Rollout*, *Transformer Interpretability Beyond Attention Visualization* (TIBAV) i predloženu metodu.

U skladu s prethodnim eksperimentima, predložena metoda nadmašuje sve ostale atribucijske metode sa statistički značajnom razlikom te postiže najviši rezultat na novoj metrici za vrjednovanje. Rezultati su prikazani u Tablicama 7.5 i 7.6. Promatrajući prosječne rezultate lokalne konzistentnosti, predložena metoda postiže rezultat koji je više nego dvostruko veći od rezultata sljedeće metode, HiResCAM. Međutim, metode temeljene na gradijentu pokazuju prednost u rezultatu kontrastnosti, nadmašujući predloženu metodu u ovom aspektu. Važno je napomenuti da *Rollout* i *Last Layer Attention*, kako su primijetili Chefer i sur. [99], ne generiraju kontrastne mape, što rezultira dosljedno nižim rezultatima na komponenti kontrastnosti. TIBAV, kao kombinacija metoda temeljenih na gradijentu i *Rollout* metode, po rezultirajućoj mjeri nalazi se između njihovih rezultata za svaku od komponenti metrike.

Važno je istaknuti relativno niži rezultat predložene metode u eksperimentu s Vizualnim Transformerom u usporedbi s prethodna dva eksperimenta. Ova činjenica ukazuje na prostor za poboljšanje predložene metode modifikacijom pravila za slojeve specifične Transformerima ili definiranjem različitih pravila za kontrastnu propagaciju relevantnosti, kao što je primjerice metoda softmax-gradijent slojevita propagacija relevantnosti (engl. *Softmax-Gradient Layer-wise Relevance Propagation* (SGLRP)) [153].

Tablica 7.5: GAE mjera ostvarena od strane različitih atribucijskih metoda na ImageNet skupu podataka i ViT-Base arhitekturi uz veličinu komadića slike od 16 piksela

Method name	$\overline{LC}$	$\overline{C}$	$\overline{GAE}$
GradCAM	0.042	0.651	0.024
GradCAM++	0.003	0.164	0.0
HiResCAM	0.06	0.821	0.047
LayerCAM	0.032	0.745	0.022
LLAttention	0.002	0.0	0.0
Rollout	0.005	0.022	0.0
TIBAV	0.018	0.475	0.006
RAMP	0.162	0.564	**0.075**

Tablica 7.6: GAE mjera ostvarena od strane različitih atribucijskih metoda na PascalVOC skupu podataka i ViT-Base arhitekturi uz veličinu komadića slike od 16 piksela

Method name	$\overline{LC}$	$\overline{C}$	$\overline{GAE}$
GradCAM	0.034	0.53	0.017
GradCAM++	0.002	0.169	0.0
HiResCAM	0.072	0.679	0.045
LayerCAM	0.042	0.557	0.02
LLAttention	0.001	0.008	0.0
Rollout	0.005	0.027	0.0
TIBAV	0.02	0.394	0.008
RAMP	0.164	0.476	**0.072**

Radi cjelovitosti kvantitativnih eksperimenata, također se prikazuju rezultati svih atribucijskih metoda na oba skupa podataka koristeći standardne metrike za vrjednovanje metoda temeljenih na atribucijskim mapama opisanih u Poglavlju 3.1: ROAD (Rong i sur. [127]), Lokalna Lipschitz procjena (Alvarez i sur. [136]) i Focus metrika (Arias-Duart i sur. [139]).

Rezultati dobiveni na skupu podataka ImageNet prikazani su u Tablici 7.7 za VGG, Tablici 7.9 za ResNet50 i naposljetku u Tablici 7.11 za ViT-Base. Rezultati eksperimenata provedenih na skupu podataka PascalVOC mogu se pronaći u Tablici 7.8 za VGG arhitekturu, Tablici 7.10 za ResNet50 te 7.12 za ViT-Base. Rezultati ovog eksperimenta pokazuju da RAMP dosljedno postiže najviše ili je u samom vrhu promatrajući rezultate kroz različite postavke modela i skupova podataka. Međutim, bitno je istaknuti komparabilno niži Focus rezultat u eksperimentima s ViT-Base modelom, što se podudara s nalazima komponente kontrastnosti GAE metrike, dodatno sugerirajući prostor za istraživanje alternativnih metoda izračuna kontrastnih atribucijskih mapa u ovom specifičnom scenariju.

Nadalje, važno je napomenuti kako, iako RAMP postiže izvrsne rezultate na konvencionalnim metrikama, značajne razlike u rezultatima opažene na GAE metrici nisu tako izražene kao u standardno korištenim vrjednovanjima. Ova razlika proizlazi iz inherentnih ograničenja standardno korištenih metrika, kako je opisano u Poglavlju 3. Procjena svake karakteristike (vjernost, robusnost i lokalizacija) zasebno zanemaruje sinergijsku vrijednost koja proizlazi iz njihove kombinirane procjene. Atribucijska metoda može postići visoke rezultate u vjernosti i robusnosti za određenu sliku, ali pokazati lošu komponentu lokalizacije. S druge strane, sljedeća slika može povećati vrijednosti prema lokalizaciji, ali ostvariti značajne nedostatke u komponentama vjernosti ili robusnosti. Kvaliteta atribucijske metode pronalazi se isključivo u

Tablica 7.7: Rezultati raznih atribucijskih metoda ostvarenih na standardnim metrikama vrjednovanim na ImageNet skupu podataka koristeći VGG arhitekturu.

Method name	$\overline{ROAD^{MoRF}(\downarrow)}$	$\overline{ROAD^{LeRF}(\uparrow)}$	$\overline{Lipschitz(\downarrow)}$	$\overline{Focus(\uparrow)}$
Constant	0.461	0.459	**0.0**	0.5
Random	0.585	0.587	1.123	0.5
Saliency	0.398	0.65	0.459	0.549
Input*Gradient	0.352	0.619	0.371	0.547
Deconvolution	0.483	0.639	0.445	0.501
LRP-ε	0.3	0.678	0.341	0.566
DeepLIFT	0.244	0.731	0.291	0.621
LRP-$\alpha_1\beta_0$	**0.199***	0.762	0.126	0.506
LRP-$\alpha_2\beta_1$	0.226	0.776	0.087	0.535
Integrated Gradients	0.302	0.669	0.331	0.559
SmoothGrad	0.28	0.746	0.357	0.561
GradCAM	0.207	**0.798***	0.687	0.844
HiResCAM	0.204	0.777	0.634	0.807
LayerCAM	**0.199***	0.779	0.499	0.642
GradCAM++	0.213	0.758	0.496	0.628
GuidedGradCAM	**0.177***	**0.818***	0.13	**0.886***
cLRP	0.229	0.79	0.167	0.833
RAP	0.216	0.77	0.176	0.482
cRAP	0.302	0.702	0.401	0.833
RAMP	**0.184***	**0.792***	0.23	**0.885***

$\downarrow$ - niži rezultat je bolji; $\uparrow$ - viši rezultat je bolji
*Razlika u rezultatima nije statistički značajna prema Wilcoxon testu ranga na razini značajnosti p=0.05

Tablica 7.8: Rezultati raznih atribucijskih metoda ostvarenih na standardnim metrikama vrjednovanim na PascalVOC skupu podataka koristeći VGG arhitekturu.

Method name	$\overline{ROAD^{MoRF}(\downarrow)}$	$\overline{ROAD^{LeRF}(\uparrow)}$	$\overline{Lipschitz(\downarrow)}$	$\overline{Focus(\uparrow)}$
Constant	0.301	0.302	**0.0**	0.5
Random	0.343	0.338	1.123	0.5
Saliency	0.222	0.401	0.477	0.526
Input*Gradient	0.22	0.361	0.394	0.537
Deconvolution	0.252	0.416	0.47	0.5
LRP-ε	0.166	0.428	0.361	0.537
DeepLIFT	0.108	0.478	0.316	0.594
LRP-$\alpha_1\beta_0$	0.105	0.562	0.124	0.507
LRP-$\alpha_2\beta_1$	0.105	**0.58***	0.089	0.526
Integrated Gradients	0.182	0.419	0.361	0.549
SmoothGrad	0.133	0.533	0.375	0.528
GradCAM	**0.074***	0.511	0.775	0.766
HiResCAM	0.102	0.52	0.73	0.731
LayerCAM	**0.087***	0.53	0.613	0.581
GradCAM++	**0.084***	0.51	0.6	0.572
GuidedGradCAM	**0.071***	**0.591***	0.135	**0.779***
cLRP	**0.062***	**0.559***	0.17	0.755
RAP	0.103	0.481	0.195	0.441
cRAP	0.134	0.468	0.346	0.761
ROAD	**0.071***	**0.563***	0.237	**0.787***

$\downarrow$ - niži rezultat je bolji; $\uparrow$ - viši rezultat je bolji
*Razlika u rezultatima nije statistički značajna prema Wilcoxon testu ranga na razini značajnosti p=0.05

Tablica 7.9: Rezultati raznih atribucijskih metoda ostvarenih na standardnim metrikama vrjednovanim na ImageNet skupu podataka koristeći ResNet50 arhitekturu.

Method name	$\overline{ROAD^{MoRF}(\downarrow)}$	$\overline{ROAD^{LeRF}(\uparrow)}$	$\overline{Lipschitz(\downarrow)}$	$\overline{Focus(\uparrow)}$
Integrated Gradients	0.452	0.773	0.338	0.65
SmoothGrad	0.416	**0.856***	0.336	0.665
GradCAM	0.298	0.792	0.417	0.769
HiResCAM	0.297	0.787	0.42	0.714
LayerCAM	0.305	0.777	0.457	0.685
GradCAM++	0.311	0.777	0.479	747
GuidedGradCAM	0.305	**0.857***	**0.126***	0.803
cLRP	**0.25***	**0.866***	**0.13***	**0.96***
RAP	0.42	0.689	0.814	0.544
cRAP	0.413	0.759	0.744	0.878
RAMP	**0.266***	0.836	0.243	**0.966***

↓ - niži rezultat je bolji; ↑ - viši rezultat je bolji
*Razlika u rezultatima nije statistički značajna prema Wilcoxon testu ranga na razini značajnosti p=0.05

Tablica 7.10: Rezultati raznih atribucijskih metoda ostvarenih na standardnim metrikama vrjednovanim na PascalVOC skupu podataka koristeći ResNet50 arhitekturu.

Method name	$\overline{ROAD^{MoRF}(\downarrow)}$	$\overline{ROAD^{LeRF}(\uparrow)}$	$\overline{Lipschitz(\downarrow)}$	$\overline{Focus(\uparrow)}$
Integrated Gradients	0.248	0.548	0.339	0.639
SmoothGrad	0.187	**0.655***	0.335	0.633
GradCAM	0.152	0.585	0.472	0.735
HiResCAM	0.153	0.567	0.471	0.67
LayerCAM	0.154	0.584	0.495	0.66
GradCAM++	0.146	0.584	0.505	0.733
GuidedGradCAM	0.141	0.63	**0.131***	0.743
cLRP	**0.109***	**0.683***	**0.142***	**0.906***
RAP	0.218	0.413	0.796	0.524
cRAP	0.173	0.566	0.693	0.784
ROAD	**0.113***	**0.659***	0.232	**0.913***

↓ - niži rezultat je bolji; ↑ - viši rezultat je bolji
*Razlika u rezultatima nije statistički značajna prema Wilcoxon testu ranga na razini značajnosti p=0.05

Tablica 7.11: Rezultati raznih atribucijskih metoda ostvarenih na standardnim metrikama vrjednovanim na ImageNet skupu podataka koristeći ViT-Base arhitekturu uz veličinu komadića slike od 16 piksela.

Method name	$\overline{ROAD^{MoRF}(\downarrow)}$	$\overline{ROAD^{LeRF}(\uparrow)}$	$\overline{Lipschitz(\downarrow)}$	$\overline{Focus(\uparrow)}$
GradCAM	**0.317**	**0.82***	0.867	**0.913***
GradCAM++	0.56	0.678	2.556	0.707
HiResCAM	0.383	**0.833***	0.61	**0.913***
LayerCAM	0.399	**0.819***	0.665	0.883
LLAttention	0.406	0.774	0.54	0.472
Rollout	0.414	0.789	1.064	0.496
TIBAV	0.384	**0.813***	0.876	0.798
RAMP	0.457	**0.803***	**0.477**	0.826

↓ - niži rezultat je bolji; ↑ - viši rezultat je bolji
*Razlika u rezultatima nije statistički značajna prema Wilcoxon testu ranga na razini značajnosti p=0.05

Tablica 7.12: Rezultati raznih atribucijskih metoda ostvarenih na standardnim metrikama vrjednovanim na PascalVOC skupu podataka koristeći ViT-Base arhitekturu uz veličinu komadića slike od 16 piksela.

Method name	$\overline{ROAD^{MoRF}(\downarrow)}$	$\overline{ROAD^{LeRF}(\uparrow)}$	$\overline{Lipschitz(\downarrow)}$	$\overline{Focus(\uparrow)}$
GradCAM	**0.131**	**0.642***	0.853	**0.8***
GradCAM++	0.341	0.451	2.434	0.594
HiResCAM	0.162	**0.634***	0.524	**0.785***
LayerCAM	0.182	**0.62***	0.586	0.756
LLAttention	0.208	0.566	0.574	0.485
Rollout	0.232	0.572	1.059	0.487
TIBAV	0.168	**0.609***	0.756	0.677
ROAD	0.195	0.573	**0.46**	0.689

↓ - niži rezultat je bolji; ↑ - viši rezultat je bolji
*Razlika u rezultatima nije statistički značajna prema Wilcoxon testu ranga na razini značajnosti p=0.05

Tablica 7.13: Ablacijski eksperiment predložene metode za ViT-Base arhitekturu uz veličinu komadića slike od 16 piksela.

	ImageNet	PascalVOC
RAMP	**0.075**	**0.072**
Propagacija na razini komadića slike	0.041 (-45.3%)	0.035 (-51.4%)
Relevantnost vrijednosti	0.064 (-14.7%)	0.047 (-34.7%)
Relevantnost upita i ključeva	0.037 (-50.7%)	0.04 (-44.4%)

slučaju kada su sve tri karakteristike istovremeno zadovoljene za svaku sliku.

7.1.1 Ablacijski eksperiment

U sljedećem eksperimentu se ispituju tri modificirane varijante predložene atribucijske metode primijenjene na arhitekturu Vizualnog Transformera kako bi se pokazali nedostaci postojećih metoda te kvantificirao napredak postignut implementacijom predloženih pravila za pojedine slojeve ove mreže. Tri ispitane varijante uključuju:

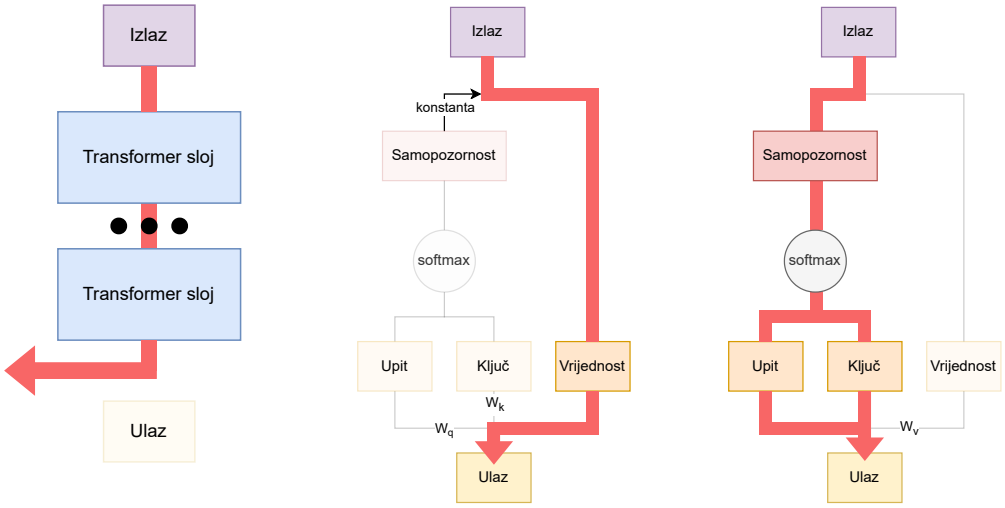
1. Propagacija na razini komadića slike: U ovoj varijanti, propagiraju se atribucije isključivo na razini komadića ulazne slike, zanemarujući relevantnost pojedinačnih ulaznih piksela. Ilustracija propagacije relevantnosti unutar ove varijante prikazana je na Slici 7.1a.

2. Relevantnost vrijednosti: Ova varijanta uključuje propagaciju samo relevantnosti vrijednosti kroz sloj samopozornosti, isključujući relevantnosti upita i ključeva unutar istog sloja. Ovaj pristup odgovara implementaciji propagacije relevantnosti unutar samopozornosti koju su predložili Ali i sur. [110]. Ilustracija propagacije relevantnosti ove varijante unutar sloja samopozornosti prikazana je na Slici 7.1b.

3. Relevantnost upita i ključeva: U ovoj varijanti, fokus se prenosi na propagaciju isključivo relevantnosti upita i ključeva, zanemarujući relevantnost vrijednosti. Ovo oponaša metodologiju TIBAV, koristeći samo vrijednosti samopozornosti dobivene iz množenja upita i ključeva. Ilustracija propagacije relevantnosti ove varijante unutar sloja samopozornosti prikazana je na Slici 7.1c.

Istraživanjem ovih modificiranih varijanti, namjera je istražiti utjecaj različitih razina propagacije atribucija na rezultate predložene metode za arhitekturu Vizualnog Transformera.

Tablica 7.13 prikazuje rezultate ablacijskog eksperimenta na oba skupa podataka. Zaustavljanje propagacije atribucija prije samih ulaznih piksela i produkcija mapa relevantnosti na razini komadića slike značajno smanjuje rezultat atribucijske metode ostvaren na GAE metrici. Značajan dio rezultata gubi se kada se kroz mrežu propagiraju samo atribucije upita i ključeva, što djelomično objašnjava relativno slabiju kvalitetu TIBAV metode u ranijim eksperimentima,



(a) (b) (c)

Slika 7.1: Ilustracija različitih postavki predložene metode unutar ablacijskog eksperimenta: **(a)** Propagacija na razini komadića slike - relevantnost ne dostiže značajke ulazne slike **(b)** Relevantnosti vrijednosti - relevantnost se propagira unatrag samo kroz vrijednosti dok se samopozornost koristi samo kao konstanta, pristup je analogan pristupu Ali i sur. [110] **(c)** Relevantnosti se propagiraju samo kroz samopozornost, odnosno ključeve i upite unutar sloja. Ova pristup je sličan pristupu TIBAV.

što je metoda koja uključuje oba principa, i propagaciju isključivo komadića slike te propagaciju samo upita i ključeva. Isključiva propagacija relevantnosti vrijednosti kroz blok samopozornosti dovodi do umjerenog smanjenja rezultata. Ova ablacijska studija snažno ukazuje na to da je propagacija relevantnosti kroz sve operacije unutar Vizualnog Transformera ključna za postizanje visokokvalitetnih atribucija.

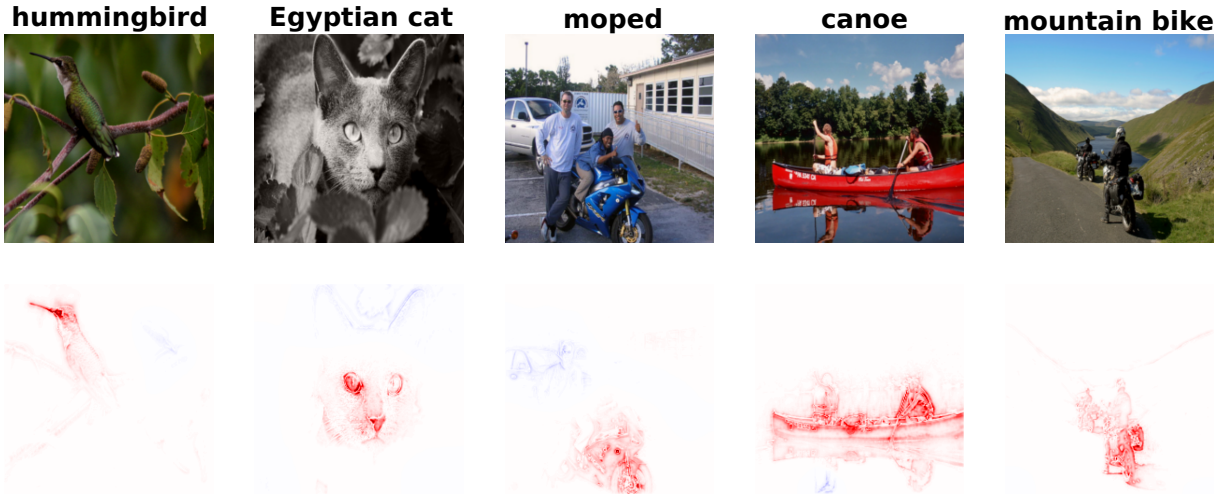
7.2 Kvalitativni eksperimenti

U ovom poglavlju predstavljaju se kvalitativni rezultati predložene metode u različitim scenarijima. Prikazani rezultati predstavljaju standardnu izvedbu naše metode te su svi prikazani primjeri nasumično odabrani.

U prvom kvalitativnom eksperimentu, vizualiziraju se atribucijske mape nekoliko nasumično odabranih slika iz skupova podataka ImageNet i PascalVOC za svaki od vrjednovanih modela. Rezultati za VGG, ResNet50 i ViT-Base prikazani su redom na Slikama 7.2, i 7.5 na ImageNet skupu podataka te Slikama 7.3, 7.4 i 7.6 na PascalVOC skupu podataka. Pored toga, vizualizirane su i atribucijske mape nekoliko primjera koji sadrže više klasa kako bi se adekvatno pokazalo svojstvo kontrastnosti predložene RAMP metode, u usporedbi s nekoliko

drugih metoda iz kvantitativnih testova. Ove vizualizacije prikazane su na Slici 7.7 za VGG, 7.8 za ResNet50 i 7.9 za ViT-Base na skupu podataka ImageNet.

Slika 7.2: RAMP atribucijske mape za VGG arhitekturu na ImageNet skupu podataka.

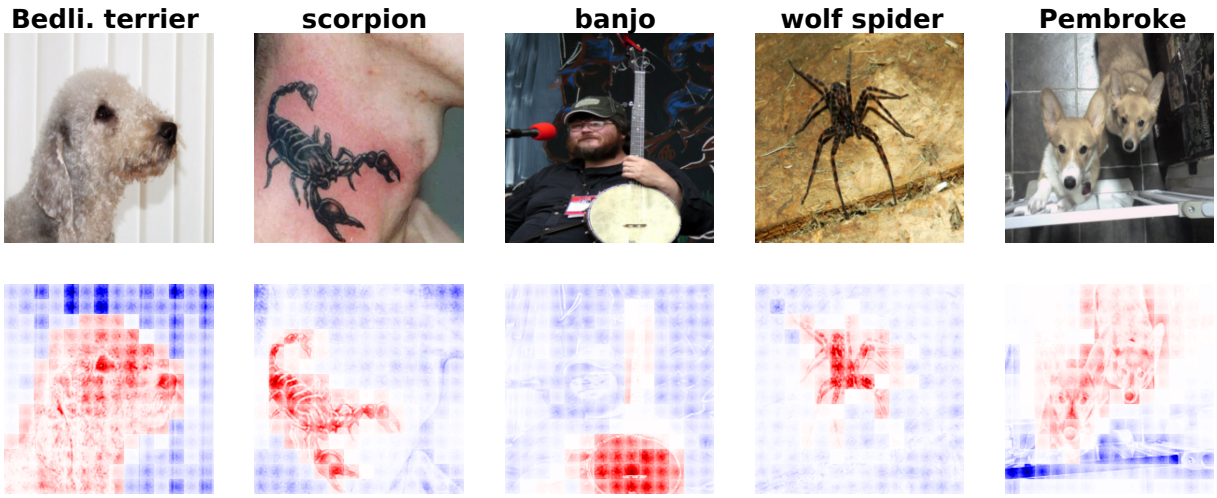


Slika 7.3: RAMP atribucijske mape za VGG arhitekturu na PascalVOC skupu podataka.

U eksperimentu koji uključuje ViT-Base arhitekturu, atribucijske mape predložene metode zadržavaju strukturu komadića slike uzrokovanu arhitekturom modela. Međutim, ističe se kao jedina metoda koja unutar tih komadića čuva prepoznatljive detalje, sve do pojedinačnih ulaznih piksela.

Ova činjenica dovodi do ključne prednosti RAMP metode u usporedbi s drugim atribucijskim metodama u primjeni na arhitekturu Vizualnog Transformera - atribucije na razini piksela. Na slici 7.11 prikazane su mape relevantnosti za dva različita modela, ViT-Base s veličinama komadića slike od 16x16 piksela i veličinom ulazne slike od 384x384 piksela, što čini ukupno 576 korištenih komadića; te ViT-Base s veličinama komadića od 32x32 piksela i veličinom ulazne slike od 224x224 piksela, što je ukupno 49 korištenih komadića. Vizualno se uspoređuju rezul-

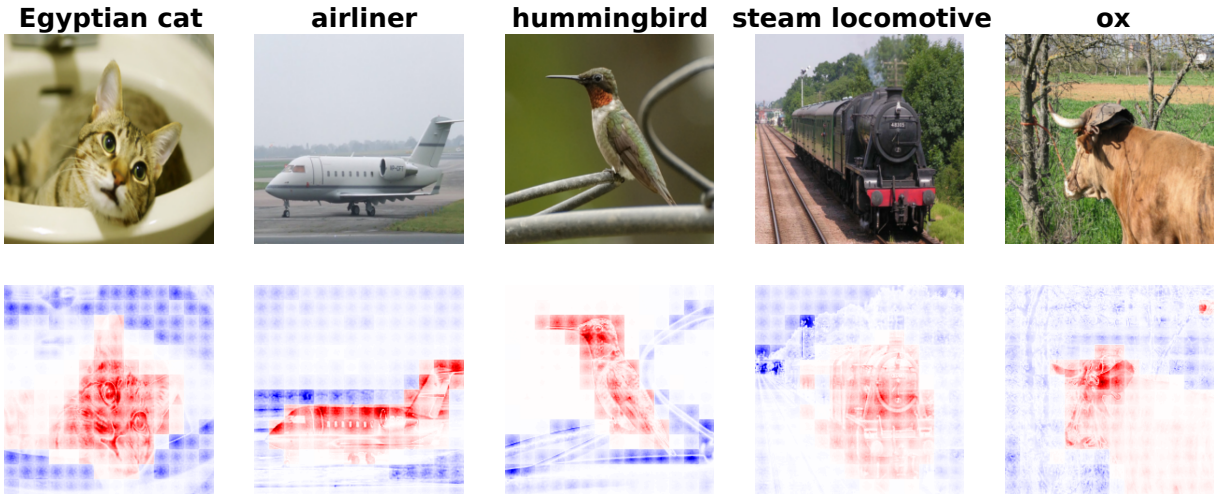
Slika 7.4: RAMP atribucijske mape za ResNet50 arhitekturu na ImageNet (gore) i PascalVOC (dolje) skupu podataka.



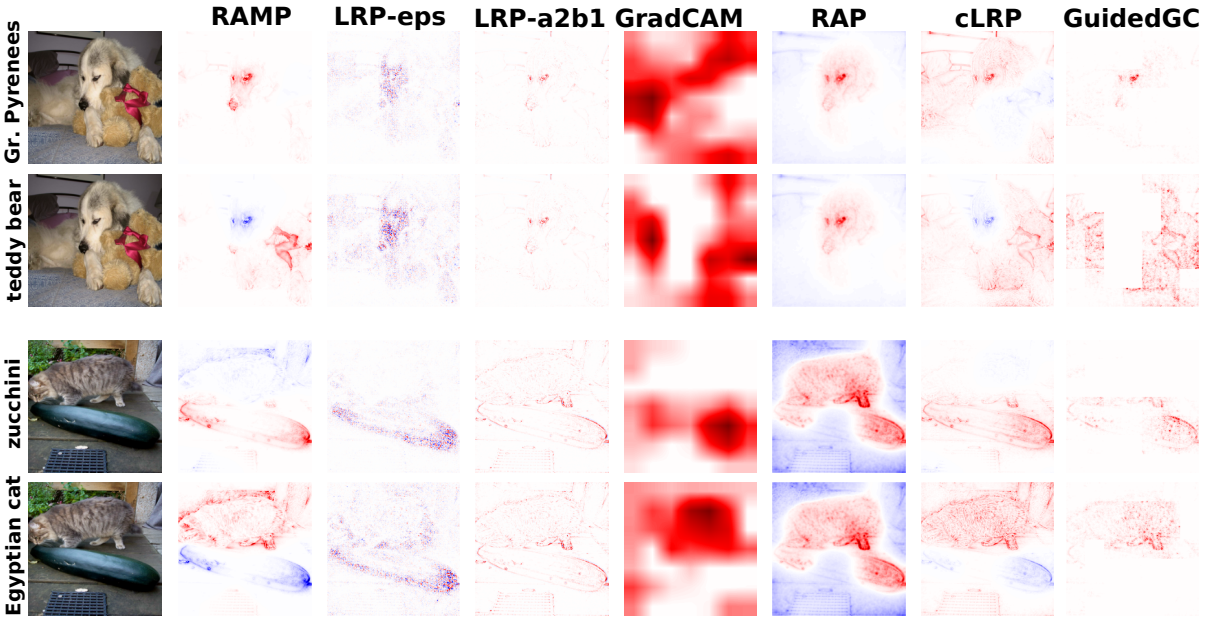
Slika 7.5: RAMP atribucijske mape za ViT-Base arhitekturu uz veličinu komadića slike od 16 piksela na ImageNet skupu podataka.

tati predložene metode s drugom najboljom metodom atribucije iz kvantitativnih eksperimenata - HiResCAM-om.

Očekivano, razlučivost u atribucijskim mapama povećava se s većim ukupnim brojem ko-



Slika 7.6: RAMP atribucijske mape za ViT-Base arhitekturu uz veličinu komadića slike od 16 piksela na PascalVOC skupu podataka.

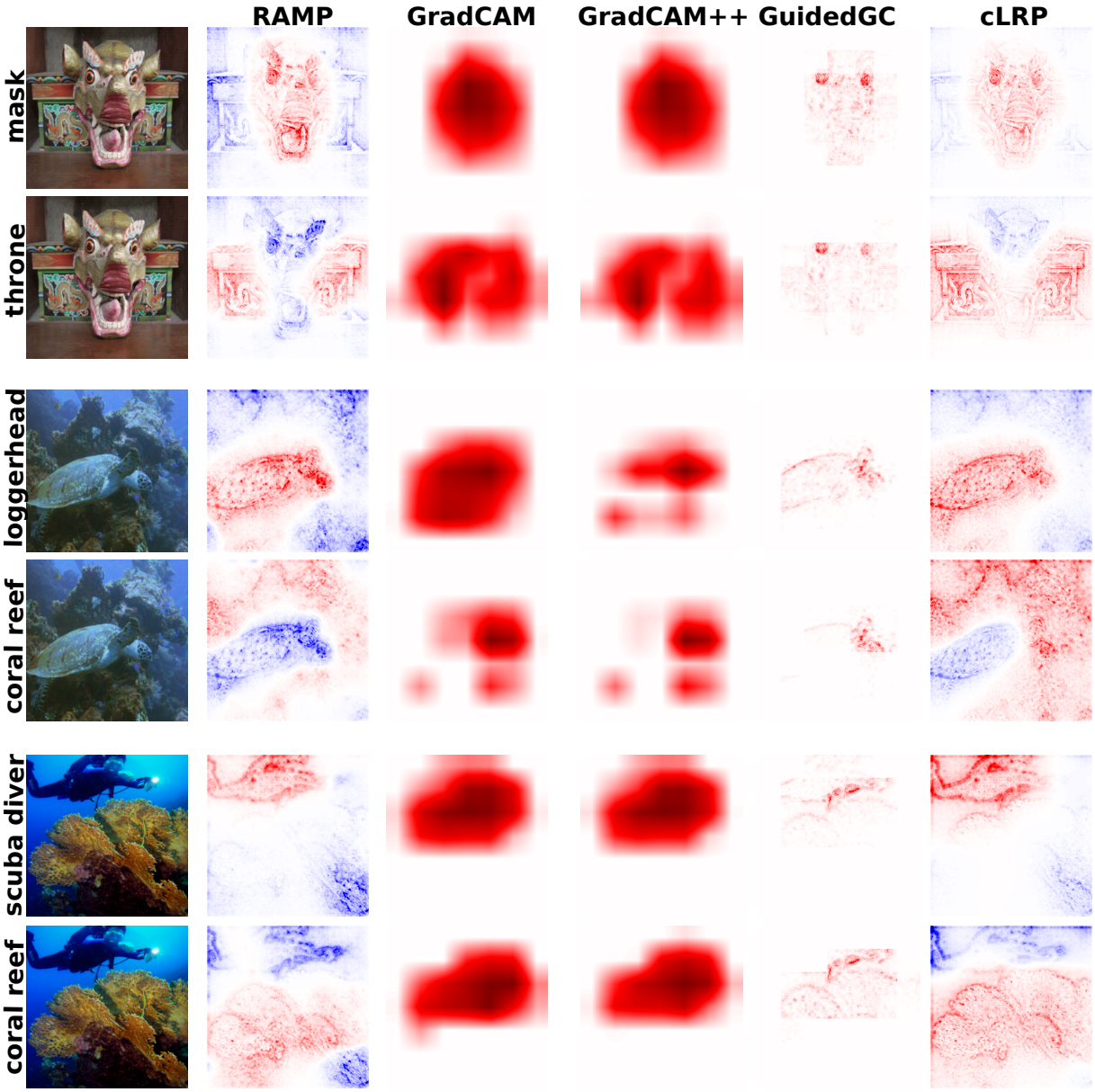


Slika 7.7: Klasno-specifične vizualizacije za različite atribucijske metode - VGG arhitektura i ImageNet skup podataka.

madića slike i smanjuje se s manjim brojem komadića za obje metode. Međutim, predložena metoda pokazuje znatno manju razliku u razlučivosti u usporedbi s HiResCAM-om. Usprkos broju komadića slike koji iznosi manje od jedne desetine, RAMP metoda proizvodi prorijeđene i jasne mape relevantnosti na razini piksela. Ovaj eksperiment ističe prednost korištenja predložene metode u scenarijima s ograničenim komputacijskim resursima.

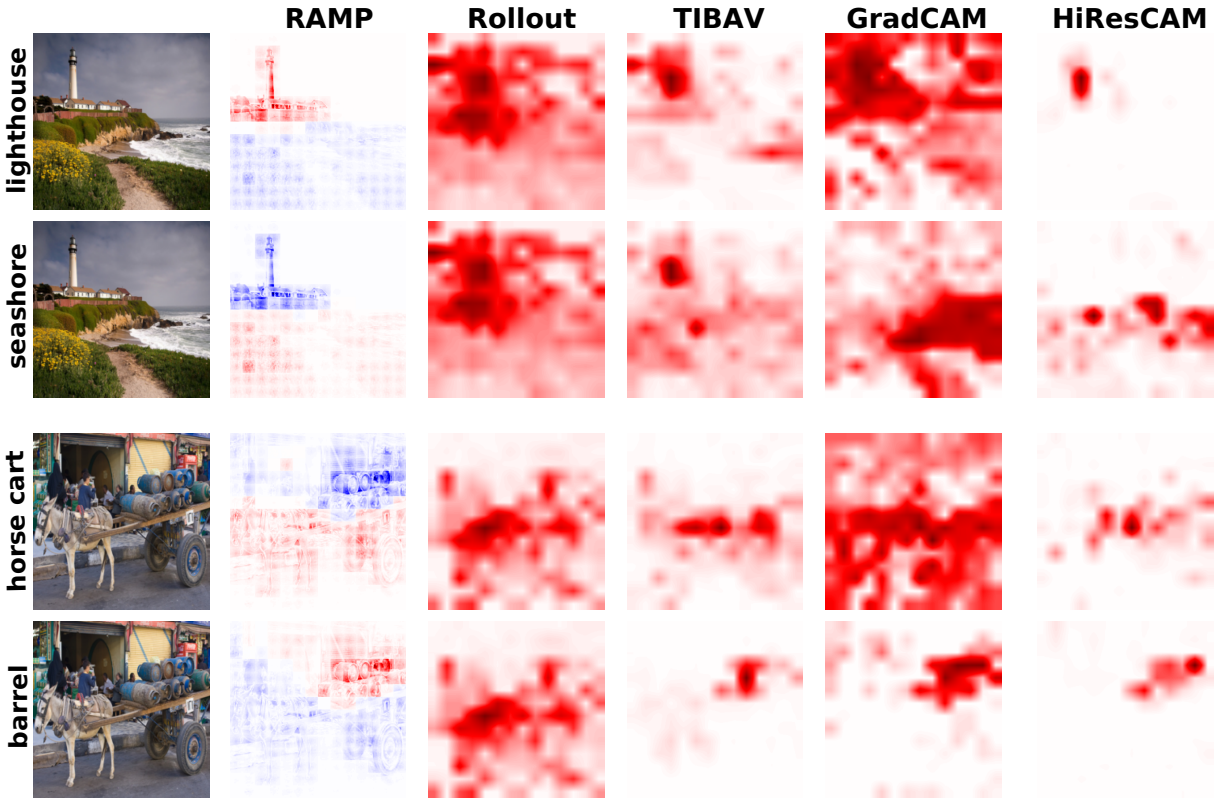
Na kraju, demonstrirana je svestranost predložene metode primjenjujući je, bez ikakvih modifikacija, na BERT [85] model učen na zadatku klasifikacije sentimenta teksta. Slika 7.10 prikazuje vizualizacije atribucija za različite nasumično odabrane recenzije filmova i emisija.

Prezentacijom kvalitativnih eksperimenata uspostavlja se paralela s kvantitativnim eksperimentima predstavljenim u ranijem dijelu ovog poglavlja te se potvrđuju zaključci. Aspekt



Slika 7.8: Klasno-specifične vizualizacije za različite atribucijske metode - ResNet50 arhitektura i ImageNet skup podataka

lokalne konzistentnosti GAE metrike procjenjuje sposobnost metode da točno identificira područja na slici koja pridonose općenitoj aktivaciji modela, očekujući da će superiorne metode prvenstveno biti fokusirane na relevantne regije koje odgovaraju klasama viđenim unutar skupa podataka. Posljedično, prilikom pregleda kontrastnih vizualizacija na Slici 7.7, očekuje se da će se najbolje metode koncentrirati na relevantne objekte, minimizirajući atribuciju na elementima pozadine. Primjetno je da LRP-ε i GradCAM pokazuju tendenciju dodjele značajnog dijela svojih atribucija izvan ključnih objekata interesa (npr. pas i plišani medvjedić u prvom primjeru, te mačka, tikvica i rešetka u drugom primjeru). Nasuprot tome, metode poput RAMP, LRP-$\alpha_2\beta_1$ i GuidedGradCAM dosljedno zadržavaju većinu svojih atribucija na relevantnim objektima. Ova razlika ogleda se u njihovim prosječnim ocjenama na komponenti lokalne konzistentnosti pred-



Slika 7.9: Klasno-specifične vizualizacije za različite atribucijske metode - ViT-Base arhitektura uz veličinu komadića slike od 16 piksela i ImageNet skup podataka

I hate this movie and i never want to see it again ! However , I loved the leading actor so much ! 30 % good , 10 % too predictable , 60 % terrible . I guess davies won ' t be saving dr . feh after all . Pros : The actor performance was admiring , Cons : The script was embarrassing

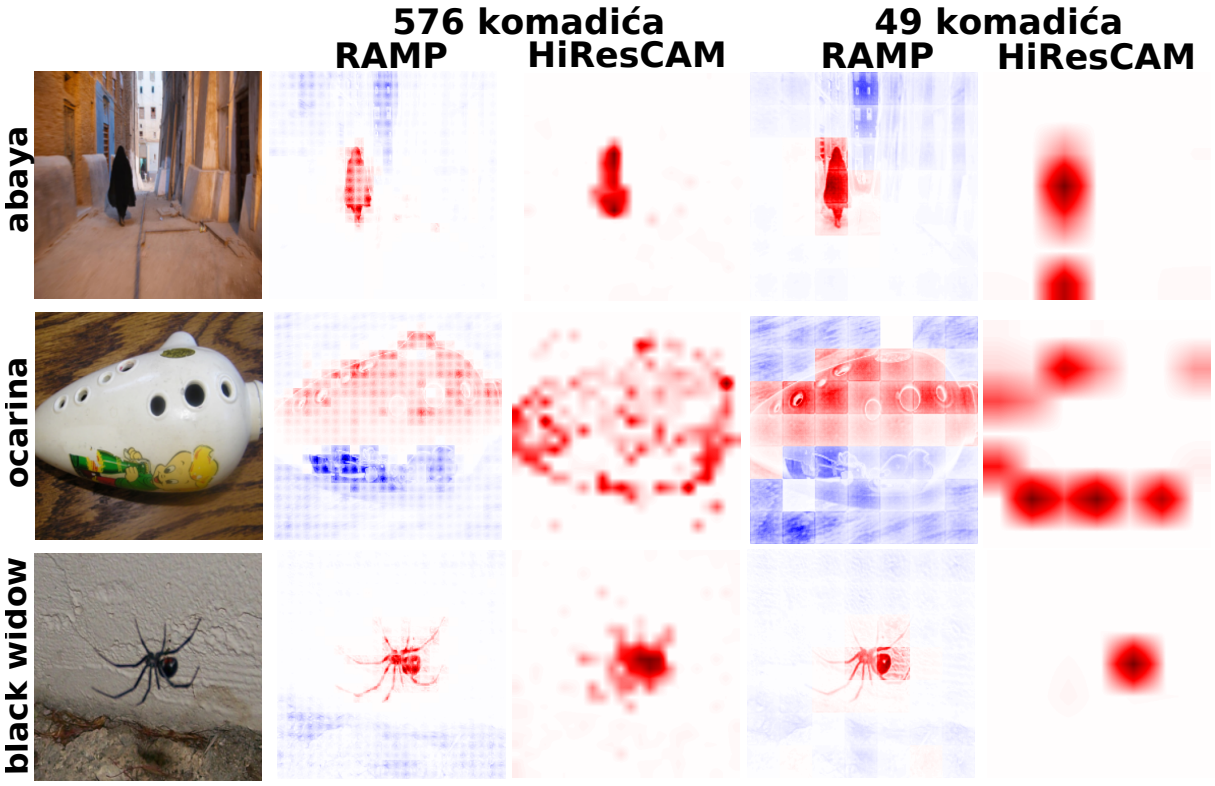
Slika 7.10: RAMP atribucijske mape za BERT model učen na zadatku klasifikacije sentimenta. Tokeni koji pridonose pozitivnom sentimentu označeni su crvenom bojom, tokeni koji pridonose negativnom sentimentu su prikazani plavom bojom.

ložene metrike, pri čemu LRP-ε i GradCAM postižu relativno niže ocjene (0.117 i 0.132) u usporedbi s RAMP, LRP-$\alpha_2\beta_1$ i GuidedGradCAM (0.396, 0.307 i 0.281).

Kako bi se potvrdila kvaliteta komponente kontrastnosti predložene metrike, promatra se razlika u atribucijskim mapama pri odabiru različitih klasa kao ciljne klase prilikom vizualizacije. Ova komponenta se na mapu relevantnosti odražava tako da su atribucije pretežno fokusirane na odabranu ciljnu klasu, umjesto raspršivanja atribucija na sve prisutne klase u slici. Metode koje pokazuju minimalne promjene pri promjeni ciljne klase, kao što su LRP-ε i LRP-$\alpha_2\beta_1$, postižu niže ocjene kontrastnosti na GAE metrici (0.022 i 0.016) u usporedbi s vizualno superiornim metodama, kao što su RAMP i cLRP (0.676 i 0.407).

Slična opažanja vrijede za metode primijenjene na druge evaluirane modele, što dodatno potvrđuje robusnost i primjenjivost predložene metrike na različite arhitekture dubokih modela.

Predložena metoda se ističe kao jedini pristup koji dosljedno producira atribucijske mape bez šuma i visokom komponentom kontrastnosti za svaki vrjednovani model, čime pruža inter-



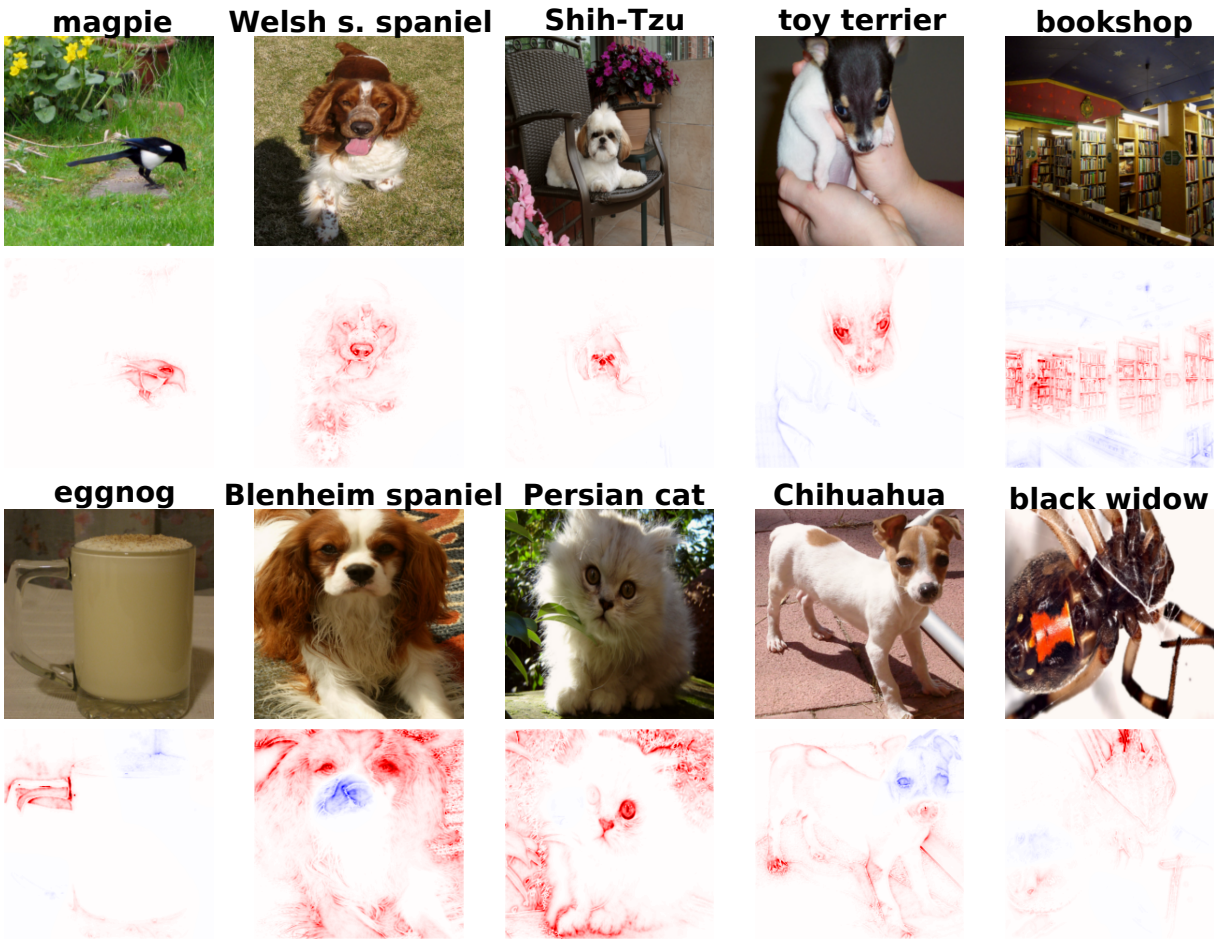
Slika 7.11: Atribucijske mape proizvedene pomoću RAMP i HiResCAM metode za ViT-Base model s veličinom komadića slike 16 i veličinom slike 384 (576 komadića) - prva dva stupca; te ViT-Base s veličinom komadića slike 32 i veličinom slike 224 (49 komadića) - zadnja dva stupca. Predložena metoda pruža atribucije na razini piksela čak i za modele s malim brojem komadića.

pretabilne vizualizacije. Nasuprot tome, metodama LRP-$\alpha_2\beta_1$ i LRP-ε nedostaje svojstvo kontrastnosti, jer pridaju relevantnost neselektivno. Metode tipa CAM, osim GuidedGradCAM-a, imaju problema s izoliranjem ključnih segmenata slike, pridajući značaj objektima u pozadini. Među najboljim često korištenim metodama u eksperimentima s VGG i ResNet50 arhitekturama, GuidedGradCAM i cLRP generiraju prorijeđene mape fokusirane na ciljani objekt. Međutim, GuidedGradCAM uvodi šum u obliku komadića slika, dok cLRP pridaje relevantnost objektima u pozadini, što je moguća posljedica osjetljivosti modela na probleme relativne magnitude atribucije, kao što je prikazano na Slici 2.2.

Mnoštvo nedostataka uočenih kod postojećih metoda, zajedno s njihovom znatno smanjenom prominentnošću, ili potpunim izostankom kod predložene metode, naglašava njezinu superiornost. Nadalje, neusporediva svestranost RAMP metode, koja se može bez problema prilagoditi raznim arhitekturama bez potrebe za modifikacijama, dodatno ističe njezinu nadmoć nad drugim trenutno korištenim metodama za produkciju mapa relevantnosti.

U drugom kvalitativnom eksperimentu, vizualno se analizira podskup primjera slika iz nasumično odabranog skupa od 1024 slike u ImageNet datasetu. Konkretno, kao promatrani model se postavlja VGG model, te se u ovom eksperimentu procjenjuje sposobnost predložene metrike za vrjednovanje da razlikuje između visokokvalitetnih i niskokvalitetnih atribucijskih mapa. U

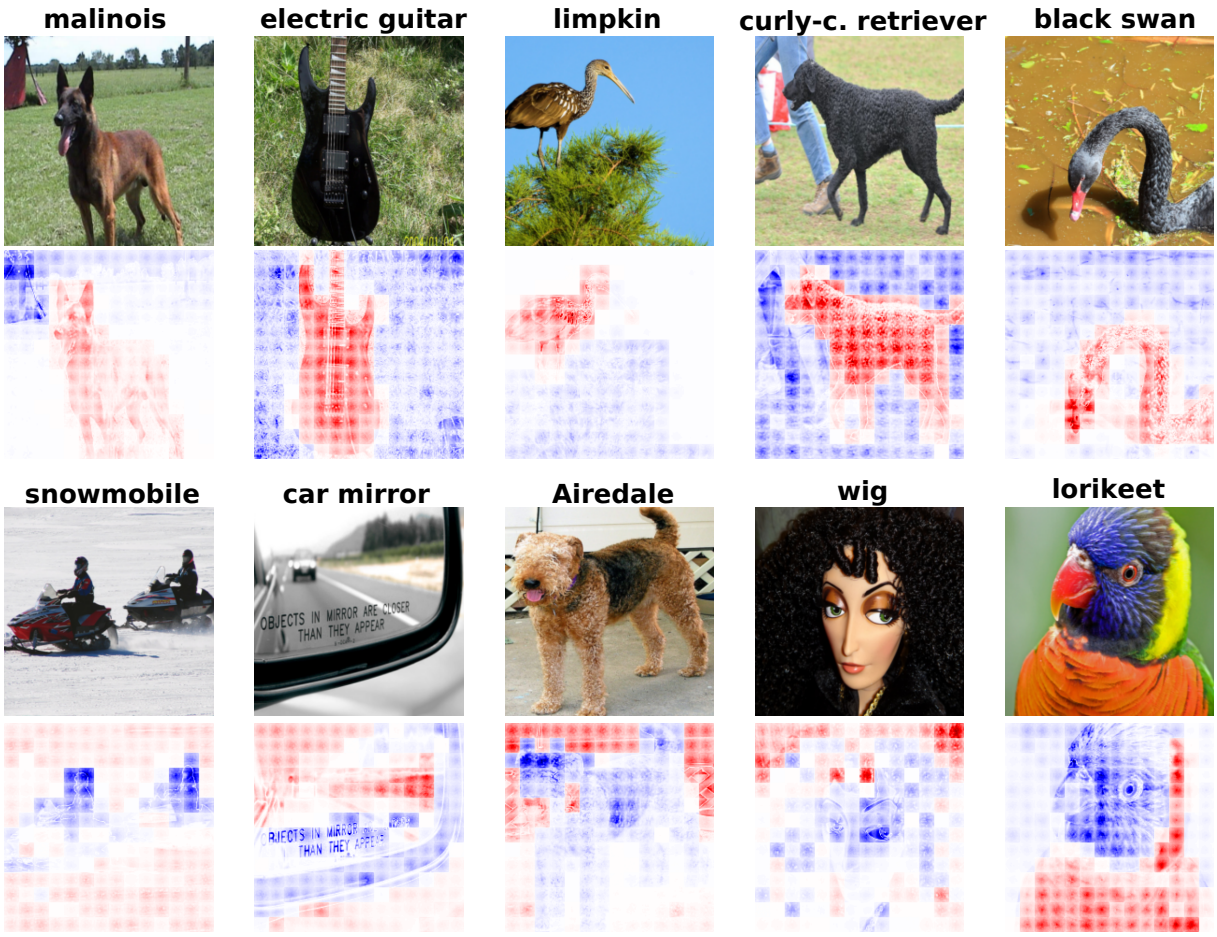
nastavku se prikazuju zaključci nastali vizualizacijom primjera s najvišim i najnižim postignutim ocjenama na GAE metrici.



Slika 7.12: Pet primjera s najvišim ocjenama (gornji red) i najnižim ocjenama (donji red) postignutim na predloženoj GAE metrici od 1024 nasumično odabranih slika unutar ImageNet skupa podataka uz VGG model.

Rezultati prikazani na Slikama 7.12 i 7.13 demonstriraju učinkovitost GAE evaluacijske metrike u razlikovanju između visokokvalitetnih i niskokvalitetnih atribucijskih mapa. Najbolje ocijenjene atribucije pokazuju prorijeđenost, odsutnost šuma i jasan fokus na ciljni objekt, točno obuhvaćajući relevantne značajke koje doprinose odluci modela. Nasuprot tome, najlošije ocijenjene atribucijske mape pokazuju manje poželjne karakteristike. Njih karakterizira raspršena distribucija vrijednosti relevantnosti oko ciljnih objekata te su prošarane šumom. Ove karakteristike ukazuju na nedostatak preciznosti i vjernosti u hvatanju krucijalnih značajki relevantnih za proces donošenja odluka modela.

Vizualnom procjenom primjera s najvišim i najnižim ocjenama, ovaj eksperiment potvrđuje učinkovitost predložene metrike za vrjednovanje u razlikovanju između visokokvalitetnih i niskokvalitetnih mapa atribucija. Ovi rezultati naglašavaju važnost primjene robusne i pouzdane metrike za temeljitu procjenu pojedinačnih snaga i slabosti atribucijskih metoda.



Slika 7.13: Pet primjera s najvišim ocjenama (gornji red) i najnižim ocjenama (donji red) postignutim na predloženoj GAE metrici od 1024 nasumično odabranih slika unutar ImageNet skupa podataka uz ViT-Base model.

Poglavlje 8

Rezultati predložene metode za vrjednovanje poravnanja metrika za vrjednovanje atribucijskih metoda s ljudskom percepcijom kvalitete

U ovom poglavlju prikazuju se rezultati finog podešavanja modela na prethodno opisanim metrikama za vrjednovanje. Nakon toga, detaljno se promatraju rezultati korisničke studije.

8.1 Rezultati finog podešavanja modela na metrikama za vrjednovanje

U eksperimentalnom dijelu fino se podešavaju često korišteni duboki modeli VGG16 [154] i ResNet18 [92] koristeći opisane funkcije gubitka iz poglavlja 6.2.1 na ImageNet skupu podataka [150]. Kao atribucijske metode koriste se standardno korištene metode GradCAM [114] i GuidedBackprop [112]. Korišteni režim učenja modela sastoji se od 10 000 koraka, uz vrjednovanje modela svakih 100 koraka koristeći validacijski skup pomoću izvornih metrika. Tijekom učenja identificira se i bilježi model s najboljim rezultatima na validacijskom skupu, uz to osiguravajući činjenicu da model postiže minimalno 95% identičnih predikcija na ovom skupu u usporedbi s izvornim modelom. Finalni rezultati metrika za vrjednovanje prikazani su na izdvojenom testnom skupu. Parametar α postavljen je na 0.1 u eksperimentima vjernosti, $5e{-}6$ za eksperimente robusnosti i 0.05 za eksperimente lokalizacije. Optimizacija α uključivala je pretraživanje mogućih vrijednosti unutar specifičnih raspona: $1e{-}3$ do 1 za vjernost i lokalizaciju, te $1e{-}8$ do $1e{-}5$ za eksperimente robusnosti, zbog njihovih većih relativnih vrijednosti gubitka. Kako bi se potvrdila značajnost razlike između rezultata svih izvornih i fino podešenih

modela, koristi se Wilcoxonov rang test [155] sa značajnošću postavljenom na $p < 0.01$.

Rezultati su prikazani u Tablici 8.1 za metodu GradCAM i Tablici 8.2 za metodu GuidedBackprop. Prikazani rezultati ističu značajna poboljšanja radnih značajki fino podešenih modela u usporedbi s njihovim izvornim verzijama. U ROAD i IROF eksperimentima, optimizirane varijante dosljedno postižu gotovo savršene rezultate koristeći GradCAM metodu. GradCAM rezultati dosežu 0.998 za VGG i 0.989 za ResNet u ROAD eksperimentima, čineći poboljšanje od 22% i 20%. U IROF eksperimentu fino-podešeni modeli u prosjeku dostižu vrijednosti od 0.985 i 0.948, ostvarujući napredak od čak 59% te 55% nad izvornim modelom. U RIS eksperimentu poboljšanja se očituju smanjenim rezultatom metrike, koji je u slučaju fino podešenog modela sveden na cijelu magnitudu manje u slučaju VGG mreže, smanjujući se s 9.1e5 na 0.5e5, što je smanjenje od 95%, a u slučaju ResNet18 mreže pomak je s izvornih 2.8e5 na 1.5e5, što iznosi 46%. U eksperimentu koji uključuje Focus metriku, nastavlja se superiornost fino podešenih modela. GradCAM postiže rezultate od 0.715, naspram izvornih 0.562 te 0.786 uz početnih 0.552 za VGG i ResNet18 model redom. Ove vrijednosti opisuju povećanje radnih značajki od 27% te 42%, ostvarujući značajan napredak. Naposljetku, u RMA eksperimentu fino podešeni modeli ostvaruju superiorne rezultate od 0.757, uz izvorne rezultate od 0.450 te 0.787 s početnim rezultatima od 0.502. Ovi rezultati signaliziraju značajan napredak od 68% i 57% za obje arhitekture.

U slučaju GuidedBackprop atribucijske metode, fino podešeni modeli također pokazuju značajna povećanja rezultata. U ROAD eksperimentu ova metoda s dostiže gotovo optimalne vrijednosti od 0.909 i 0.965 naspram izvornih 0.574 i 0.555, što je poboljšanje koje iznosi 58% za VGG mrežu te 74% za ResNet18 mrežu. Nadalje, u IROF eksperimentu, fino podešeni modeli nastavljaju s poboljšanjima, gdje prosječne vrijednosti iznose 0.979 za VGG te 0.953 za ResNet18 arhitekture. Ovi iznosi su povećanje od 26% i 33% u odnosu na početne vrijednosti od 0.775 i 0.715. U nastavku, RIS eksperiment nosi poboljšanja od 82% te 91% za fino podešene iznose ostvarene na ovoj metrici od 0.5e6 te 0.2e6 naspram izvornih 2.8e6 te 2.3e6. U eksperimentu koji uključuje Focus metriku, nastavlja se superiornost fino podešenih modela. GuidedBackprop postiže rezultate od 0.547, naspram izvornih 0.397 te 0.592 uz početnih 0.432 za VGG i ResNet18 model redom. Ove vrijednosti opisuju povećanje radnih značajki od 38% te 37%, ostvarujući značajan napredak. Naposljetku, u RMA eksperimentu fino podešeni modeli ostvaruju superiorne rezultate od 0.738, uz izvorne rezultate od 0.547 te 0.827 s početnim rezultatima od 0.540. Ovi rezultati signaliziraju značajan napredak od 35% za obje arhitekture.

8.2 Rezultati korisničke studije

U nastavku eksperimentalnog dijela ovog rada, testira se tvrdnja da atribucijske metode s višim rezultatom ostvarenim na metrici za vrjednovanje daju superiornu mapu atribucije i prema ljud-

Tablica 8.1: Prosječni rezultati GradCAM metode ostvareni na metrikama za vrjednovanje (↑) - viši rezultat je bolji; (↓) - niži rezultat je bolji

	VGG		ResNet	
Metrika	Izvoran	Fino Podešen	Izvoran	Fino Podešen
$ROAD^{inv}_{MoRF}$ (↑)	0.816	**0.998**	0.813	**0.989**
IROF (↑)	0.619	**0.985**	0.613	**0.948**
RIS (↓)	9.1e5	**0.5e5**	2.8e5	**1.5e5**
Focus (↑)	0.562	**0.715**	0.552	**0.786**
RMA (↑)	0.450	**0.757**	0.502	**0.787**

Tablica 8.2: Prosječni rezultati GuidedBackprop metode ostvareni na metrikama za vrjednovanje (↑) - viši rezultat je bolji; (↓) - niži rezultat je bolji

	VGG		ResNet	
Metrika	Izvoran	Fino Podešen	Izvoran	Fino Podešen
$ROAD^{inv}_{MoRF}$ (↑)	0.574	**0.909**	0.555	**0.965**
IROF (↑)	0.775	**0.979**	0.715	**0.953**
RIS (↓)	2.8e6	**0.5e6**	2.3e6	**0.2e6**
Focus (↑)	0.397	**0.547**	0.432	**0.592**
RMA (↑)	0.547	**0.738**	0.540	**0.827**

skoj procjeni kvalitete. Ovo je ispitano u korisničkoj studiji detaljno opisanoj u poglavlju 6.2.6. Ako je ova tvrdnja točna, značajna razlika između izvornih rezultata i fino podešenih rezultata ostvarenih na metrikama trebala bi olakšati anotatorima prepoznavanje mape više kvalitete.

Kako bi se ispitala tvrdnja, u eksperimentu se izlaže 4000 primjera označivačima koji se sastoje od pet stručnjaka iz područja dubokog učenja, pritom osiguravajući da svaki anotator procjenjuje isti skup od 4000 primjera. Ova činjenica omogućuje provođenje analize međuanotatorske suglasnosti (engl. *Inter-Annotator Agreement* (IAA)), što će potvrditi da odluke pojedinih označivača nisu samo odraz njihovih osobnih prosudbi, već proizlaze iz šireg i sveobuhvatnijeg pojma ljudske percepcije.

Anotacije za svaku metriku se agregiraju te se računa Bennett i sur. S ocjena [156] kako bi se kvantificirala suglasnost označivača. Ova ocjena izbjegava nagli pad suglasnosti ako sama distribucija oznaka nije uniformna, što je nedostatak metoda za kvantifikaciju međuanotatorske suglasnosti poput Cohenove kappa ocjene [157]. Račun ove vrijednosti potvrđuje jesu li identificirane razlike u skladu s većinskim konsenzusom. U nastavku se zbrajaju kumulativne anotacije za *izvorna*, *niti jedna* i *fino podešena* opcije u korisničkoj studiji za svaku metriku. Ta-

blica 8.3 prikazuje vrijednosti međuanotatorske suglasnosti i proporcije za svaku opciju. Osim toga, u Tablicama 8.4, 8.5, 8.6 i 8.7 su prikazani rezultati razloženi po specifičnim arhitekturama i atribucijskim metodama.

Tablica 8.3: Agregirani rezultati korisničke studije.

Metrika	IAA	Izvorna	Niti jedan	Fino podešen
ROAD^{inv}_{MoRF}	0.817	0.029	0.927	0.044
IROF	0.901	0.016	0.965	0.019
RIS	0.804	0.224	0.758	0.018
Focus	0.343	0.185	0.606	0.210
RMA	0.322	0.382	0.307	0.311

Tablica 8.4: Rezultati korisničke studije za GradCAM atribucijsku metodu i VGG model.

Metrika	IAA	Izvorna	Niti jedan	Fino podešen
ROAD^{inv}_{MoRF}	0.545	0.078	0.801	0.121
IROF	0.728	0.049	0.901	0.050
RIS	0.848	0.034	0.947	0.019
Focus	0.203	0.082	0.409	0.511
RMA	0.269	0.039	0.338	0.623

Tablica 8.5: Rezultati korisničke studije za GradCAM atribucijsku metodu i ResNet18 model.

Metrika	IAA	Izvorna	Niti jedan	Fino podešen
ROAD^{inv}_{MoRF}	0.852	0.014	0.949	0.037
IROF	0.940	0.008	0.980	0.012
RIS	1.000	0.000	1.000	0.000
Focus	0.575	0.030	0.838	0.132
RMA	0.172	0.155	0.365	0.480

Rezultati za ROAD, IROF i RIS donose jasne zaključke. Unatoč tome što su fino podešeni modeli i atribucijske metode postigli gotovo savršene ili za red veličine bolje rezultate, anotatori nisu percipirali jednako značajna poboljšanja u kvaliteti atribucije. Značajno je da su, uz visoke do gotovo savršene međuanotatorske suglasnosti od 0.817 ostvarenih na ROAD metrici, 0.901 na IROF metrici te 0.804 RIS metrici za vrjednovanje, anotatori odabrali fino podešene mape atribucije samo 4%, 2% i 2% vremena, redom.

Tablica 8.6: Rezultati korisničke studije za GuidedBackprop atribucijsku metodu i VGG model.

Metrika	IAA	Izvorna	Niti jedan	Fino podešen
ROAD^{inv}_{MoRF}	0.961	0.011	0.987	0.002
IROF	0.950	0.006	0.983	0.011
RIS	0.737	0.069	0.899	0.032
Focus	0.376	0.240	0.659	0.102
RMA	0.218	0.482	0.386	0.132

Tablica 8.7: Rezultati korisničke studije za GuidedBackprop atribucijsku metodu i ResNet18 model.

Metrika	IAA	Izvorna	Niti jedan	Fino podešen
ROAD^{inv}_{MoRF}	0.910	0.014	0.970	0.016
IROF	0.985	0.002	0.996	0.003
RIS	0.636	0.792	0.187	0.022
Focus	0.217	0.388	0.517	0.095
RMA	0.630	0.851	0.140	0.009

U slučajevima ROAD i IROF, anotatori su percipirali neznatne promjene u kvaliteti mape atribucije 93% i 97% vremena. Nadalje, RIS metrika pokazala je pad u rezultatima atribucija, s preko 22% prikazanih mapa atribucije izvornog modela koje su odabrane kao kvalitetnije. Promatrajući rezultate unutar Tablice 8.7 vidljivo je kako ova degradacija u kvalitete atribucijskih mapa dolazi primarno iz eksperimenta vezanog uz ResNet18 model te GuidedBackprop atribucijsku metodu, gdje su označivači odabrali opciju "izvorna" u čak 79% primjera.

Focus i RMA metrike pokazuju značajno različite rezultate u usporedbi s prethodnim metodama, s nižim međuanotatorskim suglasnostima u rasponu od 0.22 do 0.42. To ukazuje na svega srednju visinu suglasnosti među anotatorima, što implicira raznolike anotacije različitih korisnika za značajan dio ne-*niti jedna* anotacija u eksperimentima s GradCAM i GuidedBackprop metodama. Ovu činjenicu potvrđuju i Slike 8.2 te 8.3 koje prikazuju distribuciju konfliktnih oznaka u svakom od eksperimenata. Konfliktni primjeri su oni gdje oznaka na određenom primjeru uključuje obje opcije, i "izvornu" i "fino podešenu", sugerirajući postojanje subjektivnog elementa u procjeni kvalitete mape, gdje je jedna skupina označivača preferirala mape koje pokrivaju veći dio objekta te važan kontekst u blizini objekta, a druga skupina preferira mape koje su koncentrirane na temeljni dio ciljnog objekta. Rezultati ukazuju da se ovi konflikti pojavljuju prilikom finog podešavanja modela na lokalizacijskim metrikama, kako je čak 87% ukupnog broja konfliktnih primjera vezano uz lokalizacijske eksperimente.

Izbori *niti jedna* smanjuju se, dosežući 61% za Focus i 31% za RMA. *Fino podešena* ano-

tacije su značajne s 21% za Focus i 31% za RMA, dok su *izvorna* anotacije značajne s 19% u Focus eksperimentima i 38% u RMA eksperimentima. Ovo ukazuje na koegzistenciju kako više, tako i niže kvalitetnih mapa koje proizlaze iz procesa finog podešavanja. Promatranjem razloženih rezultata za GradCAM i GuidedBackprop atribucijske metode vidljivo je kako su anotatori u slučaju GradCAM metode češće odabirali atribucije fino podešenog modela, 51% u slučaju Focus metrike te 62% u slučaju RMA metrike uz korištenje VGG modela i 13% u slučaju Focus metrike te 48% uz korištenje ResNet18 modela, uz umjerene do niže vrijednosti suglasnosti od 0.58 do 0.17 među anotatorima što, ponovno, signalizira relativno nisku, ili nepostojeću razinu poboljšanja kvalitete atribucija u nekim primjerima. S druge strane, promatrajući metodu GuidedBackprop moguće je primijetiti značajan pad kvalitete atribucija zbog čega anotatori odabiru izvorne atribucije čak 24% puta za Focus metriku, 48% za RMA metriku tijekom korištenja VGG modela i 39% uz Focus metriku i 85% puta uz RMA metriku koristeći ResNet18 model. Suglasnost anotatora nalazi se u rasponu od 0.63 do 0.22 u ovim eksperimentima.

Na kraju se vizualiziraju atribucijske mape u svim navedenim eksperimentima kako bi se i na kvalitativan način prikazale razlike u atribucijama proizvedenim od strane modela fino podešenih na metrikama za vrjednovanje te se potvrdili rezultati iz navedenih kvantitativnih eksperimenata unutar korisničke studije.

Na Slici 8.4 prikazane su izvorne i fino podešene atribucijske mape iz GradCAM eksperimenta koristeći VGG model fino podešen na metrikama iz kategorije vjernosti, ROAD i IROF. Uspoređujući dva stupca atribucija teško je razaznati značajnu, te u nekim slučajevima i bilo kakvu, razliku u kvaliteti objašnjenja koju nude dva modela. Ovo opažanje slaže se sa zaključcima iz kvantitativnog eksperimenta.

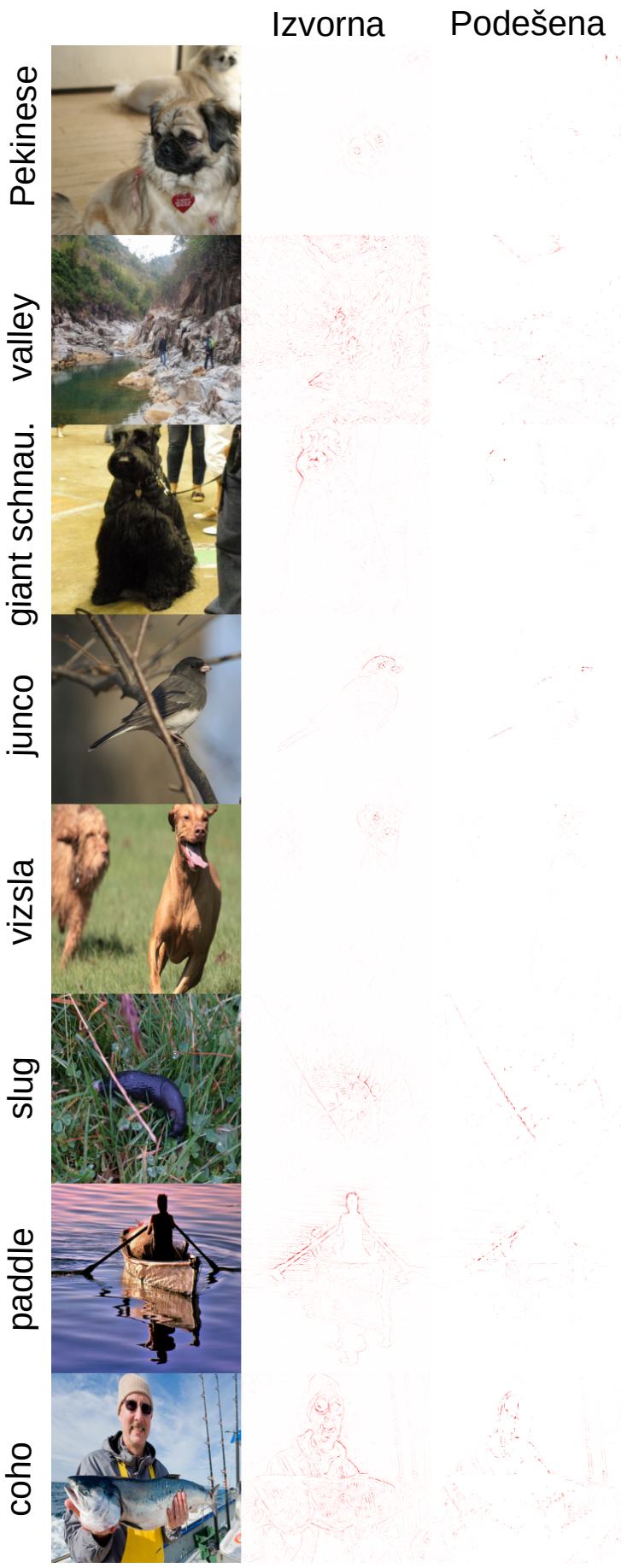
U sljedećoj vizualizaciji prikazanoj na Slici 8.5 na isti način prikazane su atribucije GuidedBackprop metode dobivene izvornim VGG modelom te njegovom fino podešenom varijantom na metrikama vjernosti. Ponovno, teško je izdvojiti značajne promjene unutar samih mapa što podupire zaključke iz kvantitativnih eksperimenata.

Za eksperimente iz kategorije robusnosti, gdje su se modeli fino podešavali na RIS metrici za vrjednovanje, atribucije GradCAM i GuidedBackprop metode koristeći VGG model ne mijenjaju se značajno, što je vidljivo iz Slika 8.6 i 8.7. Međutim, promatrajući Sliku 8.1 koja je vezana uz fino podešavanje GuidedBackprop atribucija za ResNet18 model, može se primjetiti značajna degradacija vizualiziranih atribucija, gdje one povećavaju svoju prorijeđenost i time smanjuju interpretabilnost. Pretpostavlja se da je degradacija rezultat prethodno spomenutog nedostatka ove metrike robusnosti, gdje je pokazano kako je idealna atribucijska metoda za ovu metriku konstanta. Fino podešavanje modela na RIS metrici ostavlja samo najznačajnije značajke koje je teško promijeniti malim perturbacijama, međutim, ovo dovodi do znatnog povećanja stupnja prorijeđenosti same mape, približavajući ponašanje metode "idealnoj" konstantnoj

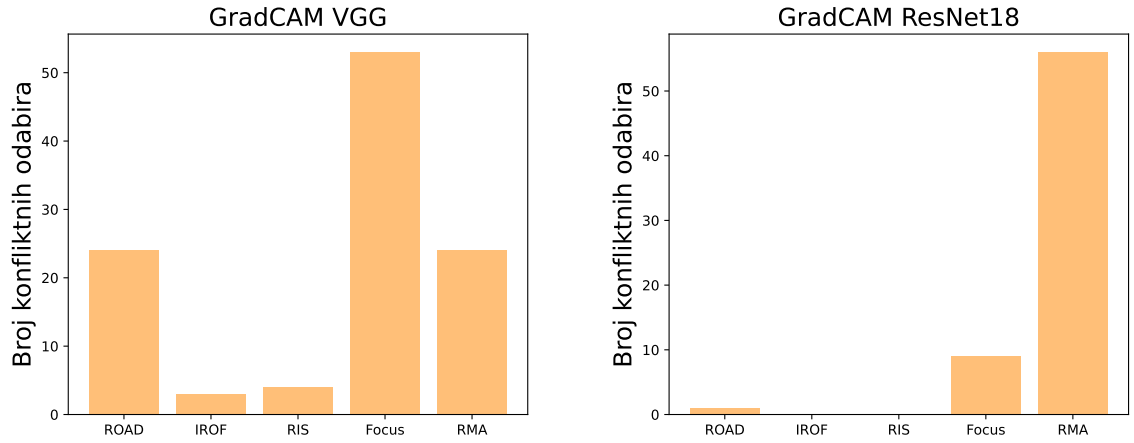
atribuciji, nauštrb procijenjene kvalitete od strane anotatora.

Naposljetku, predstavlja se pregled metoda fino podešenih na metrikama iz kategorije lokalizacija na Slikama 8.8 za GradCAM metodu i VGG model te 8.9 za GuidedBackprop metodu i ResNet18 model. Pregled vizualizacija pokazuje da obje atribucijske metode prikazuju poboljšanu sposobnost lokalizacije ciljnog objekta prilikom primjene na fino podešene modele. Za razliku od širenja atribucije preko cijele slike, GradCAM, na fino podešenom modelu, koncentrira najrelevantniji dio na poziciju ciljnog objekta. Slično tome, GuidedBackprop smanjuje atribuciju na područjima izvan granica ciljnog objekta. Ovo je prednost za metode koje generiraju atribucijske mape koje nisu prorijeđene, simulirajući mapu segmentacije, kao što je GradCAM metoda.

Međutim, daljnja optimizacija metrika lokalizacije za metode koje već u početku generiraju prorijeđene atribucije, poput GuidedBackprop-a, dodatno povećava njihovu razinu prorijeđenosti. To čini rezultirajuću mapu manje razumljivom anotatorima, što je u skladu s ranijim kvantitativnim eksperimentima.

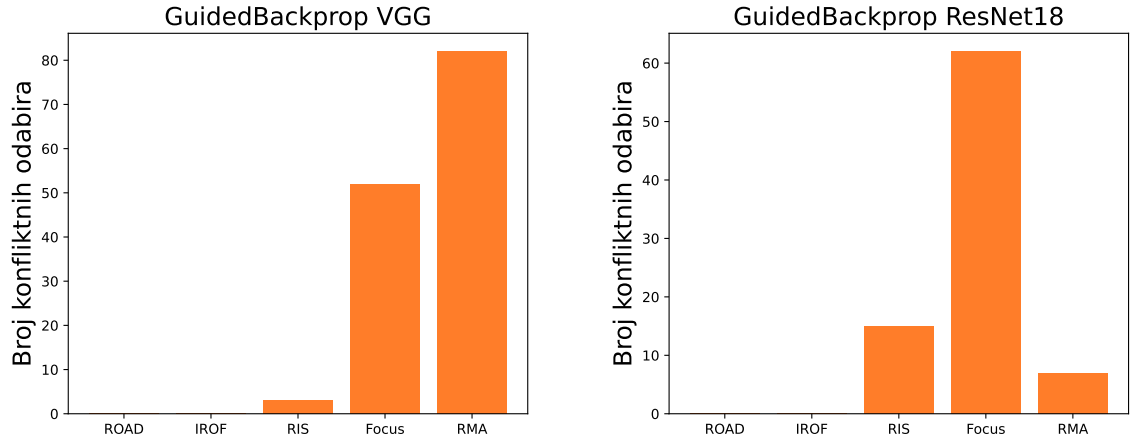


Slika 8.1: GuidedBackprop atribucijske mape iz eksperimenata robusnosti uz ResNet18 model gdje su svi anotatori označili opciju "izvorna". Optimizacija metrika robusnosti u ovom scenariju značajno degradira kvalitetu atribucijskih mapa.



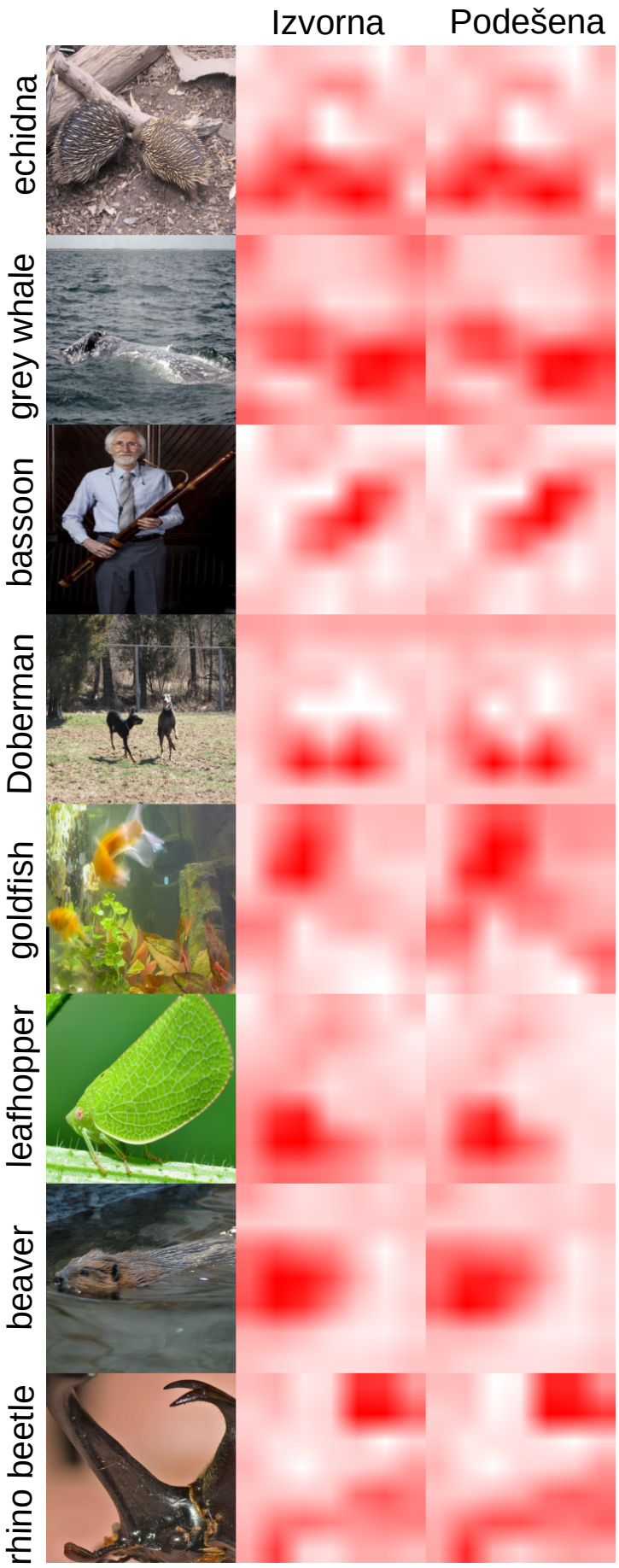
(a) (b)

Slika 8.2: **(a)** Frekvencija pojavljivanja konfliktnih oznaka u eksperimentu koji uključuje GradCAM atribucijsku metodu i VGG mrežu **(b)** Frekvencija pojavljivanja konfliktnih oznaka u eksperimentu koji uključuje GradCAM atribucijsku metodu i ResNet18 mrežu

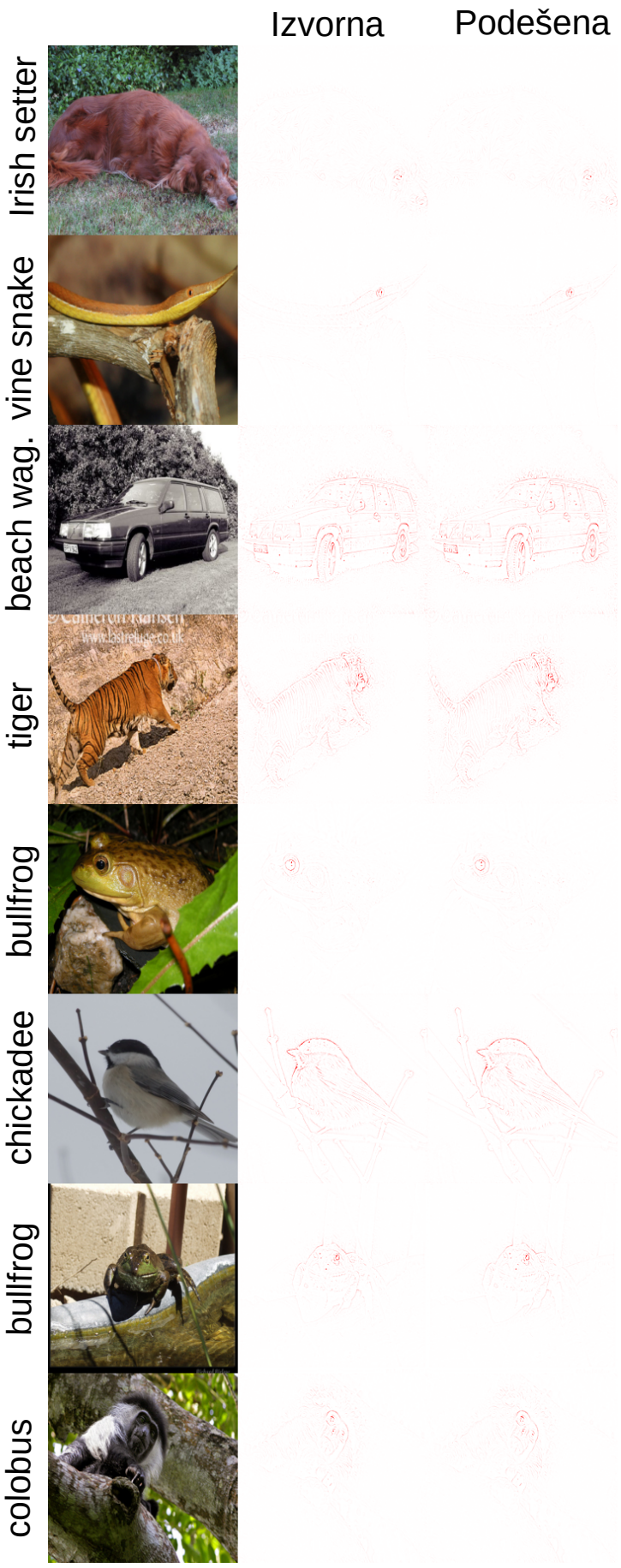


(a) (b)

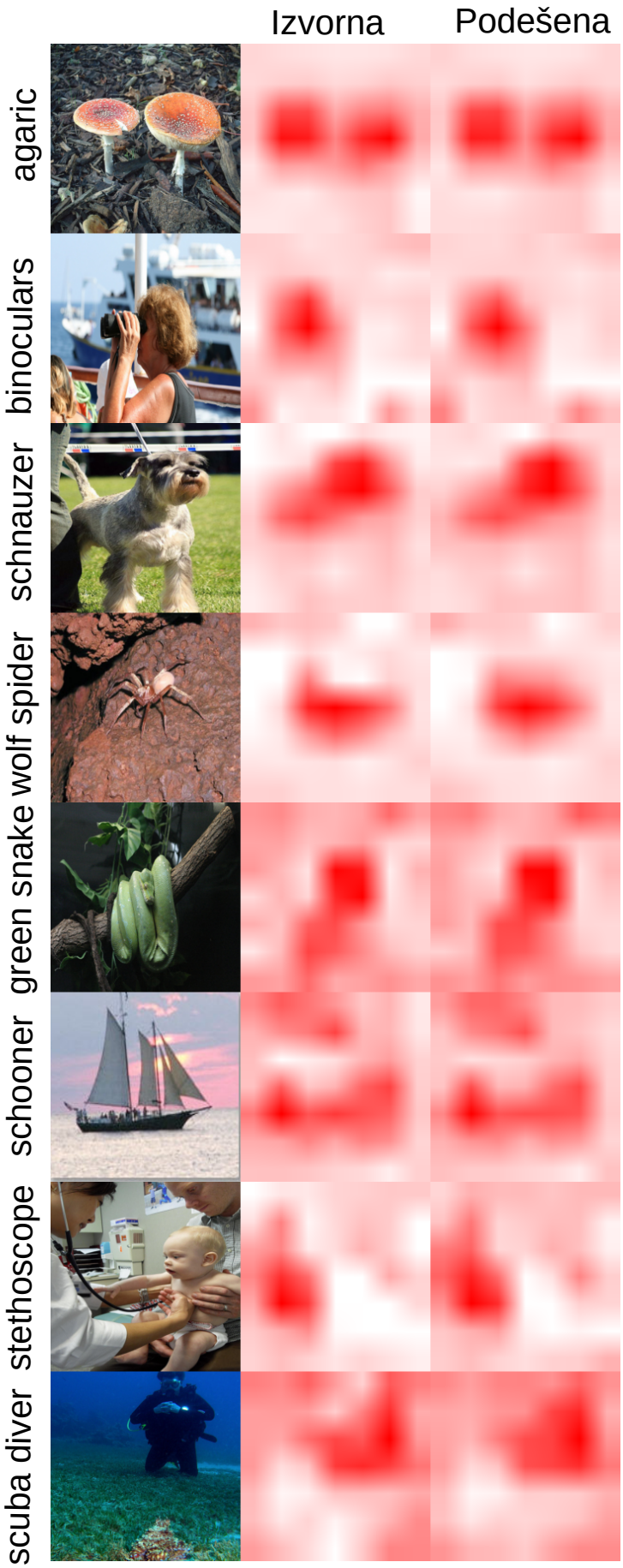
Slika 8.3: **(a)** Frekvencija pojavljivanja konfliktnih oznaka u eksperimentu koji uključuje GuidedBackprop atribucijsku metodu i VGG mrežu **(b)** Frekvencija pojavljivanja konfliktnih oznaka u eksperimentu koji uključuje GuidedBackprop atribucijsku metodu i ResNet18 mrežu



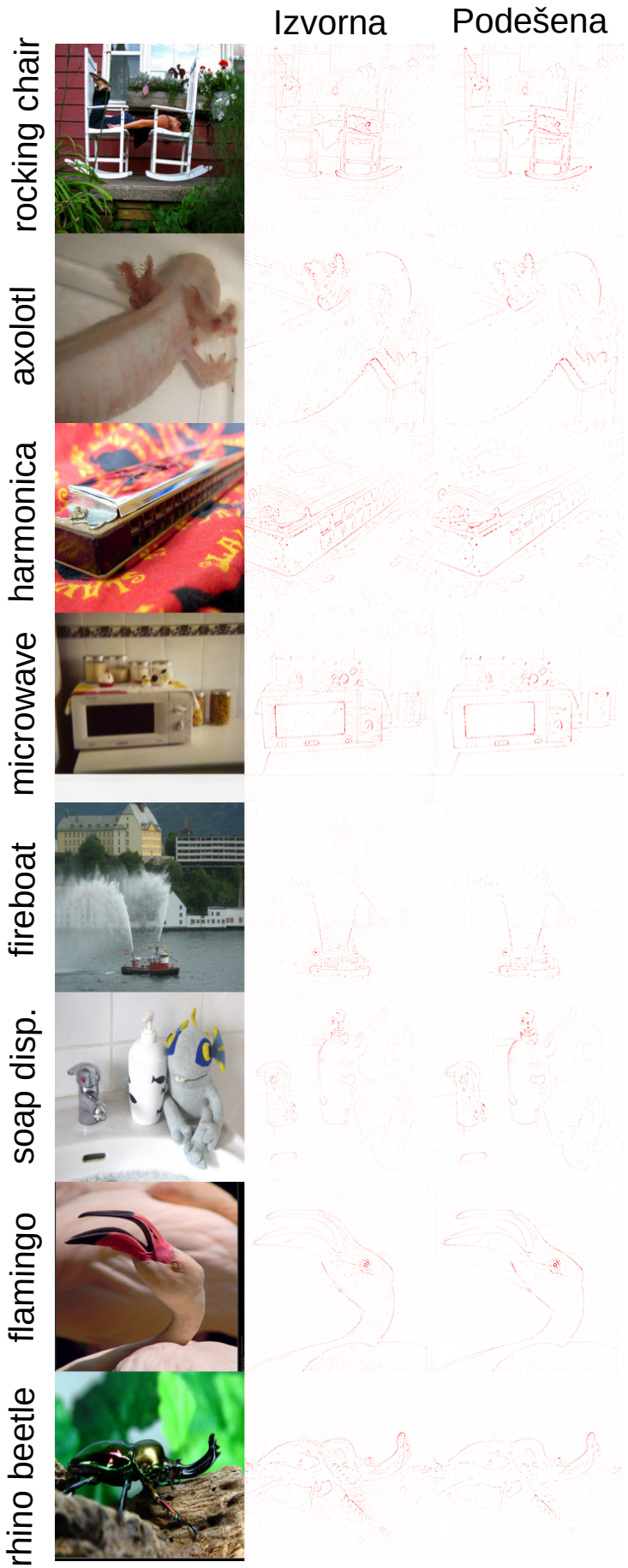
Slika 8.4: GradCAM atribucijske mape iz eksperimenata vjernosti uz VGG model gdje su svi anotatori označili opciju "niti jedna". Optimizacija metrika vjernosti u ovom scenariju ne mijenja kvalitetu atribucijskih mapa.



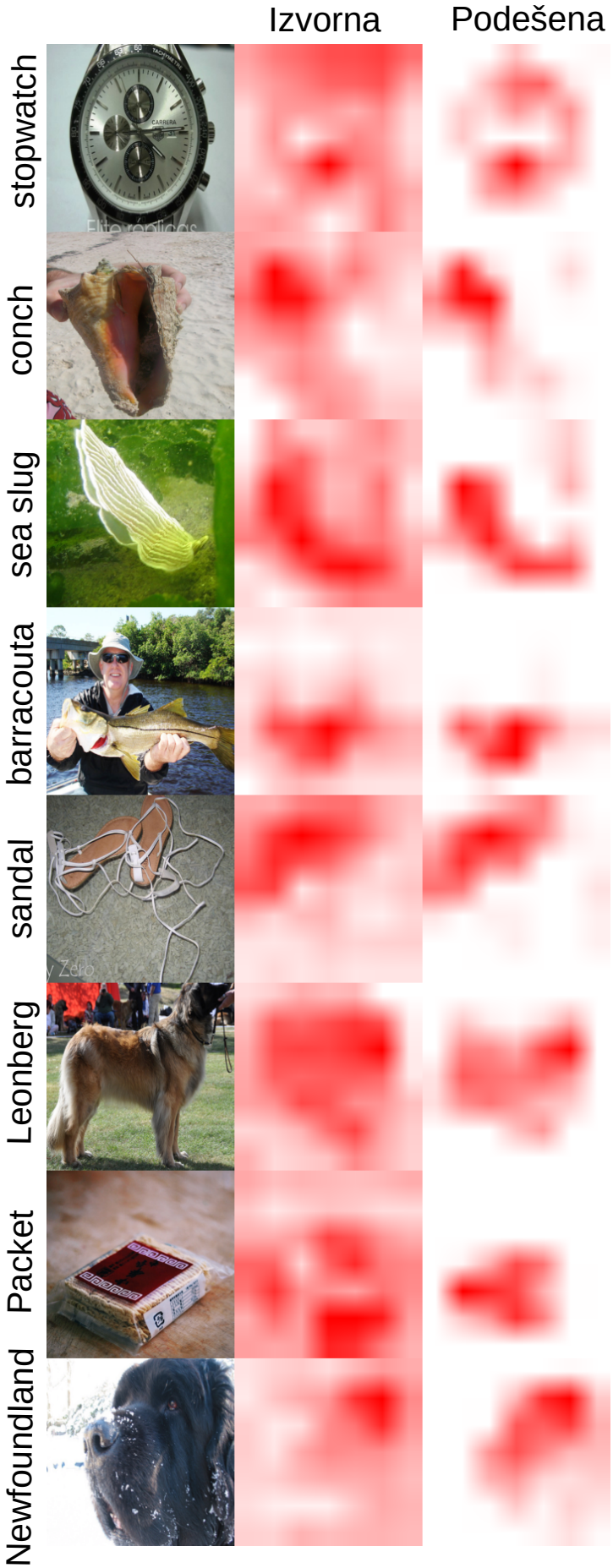
Slika 8.5: GuidedBackprop atribucijske mape iz eksperimenata vjernosti uz VGG model gdje su svi anotatori označili opciju "niti jedna". Optimizacija metrika vjernosti u ovom scenariju ne mijenja kvalitetu atribucijskih mapa.



Slika 8.6: GradCAM atribucijske mape iz eksperimenata robusnosti uz VGG model gdje su svi anotatori označili opciju "niti jedna". Optimizacija metrika robusnosti u ovom scenariju ne mijenja kvalitetu atribucijskih mapa.



Slika 8.7: GuidedBackprop atribucijske mape iz eksperimenata robusnosti uz VGG model gdje su svi anotatori označili opciju "niti jedna". Optimizacija metrika robusnosti u ovom scenariju ne mijenja kvalitetu atribucijskih mapa.



Slika 8.8: GradCAM atribucijske mape iz eksperimenata lokalizacije gdje su svi anotatori označili fino podešen primjer kao kvalitetniji. Optimizacija metrika lokalizacije poboljšava kvalitetu atribucijskih mapa.

Slika 8.9: GuidedBackprop atribucijske mape iz eksperimenata lokalizacije gdje su svi anotatori označili izvorni primjer kao kvalitetniji. Optimizacija metrika lokalizacije dodatno prorjeđuje rezultirajuće atribucije, čineći ih manje interpretabilnima.

Poglavlje 9

Zaključak

Zaključno, ovaj rad donosi nekoliko značajnih doprinosa. Prvo, predlaže se novo pravilo slojevite propagacije relevantnosti, nazvano Slojevita propagacija relativne apsolutne magnitude relevantnosti (engl. ***R**elative **A**bsolute **M**agnitude Layer-Wise **R**elevance **P**ropagation* (RAMP)), koje učinkovito rješava problem netočne relativne atribucije između neurona unutar istog sloja s različitim apsolutnim vrijednostima aktivacija. Ovo pravilo se u eksperimentalnom dijelu rada bez većih modifikacija primjenjuje na tri različite arhitekture, uključujući i recentnu arhitekturu Vizualnog Transformera. Štoviše, predlaže se i proširenje ovog pristupa na različite tipove podataka i zadataka, kao što je klasifikacija teksta, otvarajući brojne mogućnosti za buduća istraživanja.

Drugo, predlaže se nova metrika za vrjednovanje atribucijskih metoda pod nazivom Globalno vrjednovanje atribucije (engl. ***G**lobal **A**ttribution **E**valuation* (GAE)), koja pruža novu perspektivu na procjenu komponenata vjernosti, robusnosti i lokalizacije metoda za produkciju mapa relevantnosti. Za razliku od prethodnih istraživanja koji koriste više metrika bez jasne metodologije za kombiniranje rezultata, predložena metrika kombinira maskiranje temeljeno na gradijentu izlaza te aspekt lokalizacije, pružajući sveobuhvatno vrjednovanje s pomoću jedne metrike. Kroz opsežne eksperimente, procjenjuju se različite atribucijske metode, otkrivajući njihove pojedinačne prednosti i slabosti.

Kvantitativnom evaluacijom više atribucijskih metoda na različitim arhitekturama i skupovima podataka, ustanovljuje se nadmoć predloženog pristupa u odnosu na najnaprednije i često korištene metode u ovom području. Nadalje, kvalitativni eksperimenti provedeni na predloženoj atribucijskoj metodi i metrici za vrjednovanje dodatno naglašavaju prednosti doprinosa.

Naposljetku, predložen je i novi pristup koji uključuje fino podešavanje modela s pomoću samih metrika za vrjednovanje atribucijskih metoda, što rezultira značajno poboljšanim mapama relevantnosti prema tim metrikama. Međutim, prilikom usporedbe produciranih mapa tijekom korisničke studije, stručni anotatori teško su uočavali bilo kakve promjene u kategorijama metrika vjernosti i robusnosti. U nekim slučajevima, optimizacija za robusnost smanjila

je kvalitetu atribucija. Nasuprot tome, optimizacija modela na metrikama lokalizacije izazvala je znatne promjene, s fino podešenim GradCAM mapama koje su dosljedno ocijenjene kao kvalitetnije, dok su GuidedBackprop mape doživjele značajno pogoršanje.

Ovi nalazi otkrivaju izazove s trenutno korištenim automatskim metrikama za vrjednovanje mapa relevantnosti čime se naglašava upitna pouzdanost prijavljenih radnih značajki atribucijskih metoda u trenutnim istraživanjima. U budućim istraživanjima preporučuje se korištenje predložene metode za procjenu usklađenosti između metrika i ljudske percepcije, osiguravajući učinkovite doprinose ljudskom razumijevanju. Osim toga, budućem istraživanju ostavlja se i vrjednovanje većeg skupa postojećih metrika uz korištenje proširenog skupa atribucijskih metoda i skupova podataka. Zanimljiv smjer za buduća istraživanja mogao bi uključivati učenje modela nagrade na parovima atribucijskih mapa iz izvedene korisničke studije, zamjenjujući ljudske anotatore u procjeni usklađenosti između ljudske prosudbe i metrika za vrjednovanje. Nadalje, model nagrade mogao bi se koristiti umjesto tradicionalnih metrika za vođenje finog podešavanja na način koji je dobro usklađen s ljudskom prosudbom, slično pristupu podržanog učenja koji se koristi tijekom finog podešavanja modernih velikih jezičnih modela [158].

Literatura

[1]Vukadin, D., Afri ć, P., Šilić, M., Delač, G., "Advancing attribution-based neural network explainability through relative absolute magnitude layer-wise relevance propagation and multi-component evaluation", ACM Trans. Intell. Syst. Technol., Vol. 15, No. 3, apr 2024, dostupno na: https://doi.org/10.1145/3649458

[2]Vukadin, D., "Advancing Attribution-Based Explainability through Multi-Component Evaluation and Relative Absolute Magnitude Propagation", dostupno na: https://github.com/davor10105/relative-absolute-magnitude-propagation Apr. 2024.

[3]Vukadin, D., Silic, M., Delac, G., Vladimir, K., "Evaluating harmony: Neural network explanation metrics and human perception", in 47th MIPRO ICT and Electronics Convention, MIPRO 2024, Opatija, Croatia, May 20-24, 2024. IEEE, 2024, str. 13–18.

[4]Vukadin, D., "Evaluating Harmony: Neural Network Explanation Metrics and Human Perception", dostupno na: https://github.com/davor10105/harmony_app May 2024.

[5]Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., "ImageNet large scale visual recognition challenge", International Journal of Computer Vision, Vol. 115, No. 3, Apr. 2015, str. 211–252.

[6]He, K., Zhang, X., Ren, S., Sun, J., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. IEEE Computer Society, 2015, str. 1026–1034, dostupno na: https://doi.org/10.1109/ICCV.2015.123

[7]Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., "Emerging properties in self-supervised vision transformers", in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 2021, str. 9630–9640, dostupno na: https://doi.org/10.1109/ICCV48922.2021.00951

[8]Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., "Learning transferable visual models from natural language supervision", in Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, ser. Proceedings of Machine Learning Research, Meila, M., Zhang, T., (ur.), Vol. 139. PMLR, 2021, str. 8748–8763, dostupno na: http://proceedings.mlr.press/v139/radford21a.html

[9]Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M. A., "Playing atari with deep reinforcement learning", CoRR, Vol. abs/1312.5602, 2013, dostupno na: http://arxiv.org/abs/1312.5602

[10]Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., "Mastering the game of go with deep neural networks and tree search", Nature, Vol. 529, No. 7587, Jan. 2016, str. 484–489, dostupno na: https://doi.org/10.1038/nature16961

[11]Li, S., Puig, X., Paxton, C., Du, Y., Wang, C., Fan, L., Chen, T., Huang, D., Akyürek, E., Anandkumar, A., Andreas, J., Mordatch, I., Torralba, A., Zhu, Y., "Pre-trained language models for interactive decision-making", in Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., (ur.), 2022, dostupno na: http://papers.nips.cc/paper_files/paper/2022/hash/ca3b1f24fc0238edf5ed1ad226b9d655-Abstract-Conference.html

[12]He, P., Liu, X., Gao, J., Chen, W., "Deberta: decoding-enhanced bert with disentangled attention", in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. Virtual Event, Austria: OpenReview.net, 2021, dostupno na: https://openreview.net/forum?id=XPZIaotutsD

[13]Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., "Language models are few-shot learners", in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Larochelle, H., Ranzato, M., Hadsell, R., Balcan,

M., Lin, H., (ur.), 2020, dostupno na: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[14]Gu, A., Dao, T., "Mamba: Linear-time sequence modeling with selective state spaces", CoRR, Vol. abs/2312.00752, 2023, dostupno na: https://doi.org/10.48550/arXiv.2312.00752

[15]OpenAI, "GPT-4 technical report", CoRR, Vol. abs/2303.08774, 2023, dostupno na: https://doi.org/10.48550/arXiv.2303.08774

[16]Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de Las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W. E., "Mixtral of experts", CoRR, Vol. abs/2401.04088, 2024, dostupno na: https://doi.org/10.48550/arXiv.2401.04088

[17]Peters, U., "Explainable ai lacks regulative reasons: why ai and human decision-making are not equally opaque", AI and Ethics, Vol. 3, No. 3, Sep. 2022, str. 963–974, dostupno na: http://dx.doi.org/10.1007/s43681-022-00217-w

[18]Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., "A survey on bias and fairness in machine learning", ACM Comput. Surv., Vol. 54, No. 6, 2022, str. 115:1–115:35, dostupno na: https://doi.org/10.1145/3457607

[19]Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M., Chadha, A. S., Mavridis, N., "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare", npj Digit. Medicine, Vol. 3, 2020, dostupno na: https://doi.org/10.1038/s41746-020-0288-5

[20]Mikolajczyk-Barela, A., Grochowski, M., "A survey on bias in machine learning research", CoRR, Vol. abs/2308.11254, 2023, dostupno na: https://doi.org/10.48550/arXiv.2308.11254

[21]Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., Fung, P., "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity", in Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023, Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., Krisnadhi,

A. A., (ur.). Association for Computational Linguistics, 2023, str. 675–718, dostupno na: https://doi.org/10.18653/v1/2023.ijcnlp-main.45

[22]Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., Martins, A. F. T., "Hallucinations in large multilingual translation models", CoRR, Vol. abs/2303.16104, 2023, dostupno na: https://doi.org/10.48550/arXiv.2303.16104

[23]Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P., "Survey of hallucination in natural language generation", ACM Comput. Surv., Vol. 55, No. 12, 2023, str. 248:1–248:38, dostupno na: https://doi.org/10.1145/3571730

[24]Rawte, V., Sheth, A. P., Das, A., "A survey of hallucination in large foundation models", CoRR, Vol. abs/2309.05922, 2023, dostupno na: https://doi.org/10.48550/arXiv.2309.05922

[25]Liu, Y., Yao, Y., Ton, J., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., Li, H., "Trustworthy llms: a survey and guideline for evaluating large language models' alignment", CoRR, Vol. abs/2308.05374, 2023, dostupno na: https://doi.org/10.48550/arXiv.2308.05374

[26]Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., Shi, S., "Siren's song in the AI ocean: A survey on hallucination in large language models", CoRR, Vol. abs/2309.01219, 2023, dostupno na: https://doi.org/10.48550/arXiv.2309.01219

[27]Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Gao, W., Hu, X., Qi, Z., Wang, Y., Yang, L., Wang, J., Xie, X., Zhang, Z., Zhang, Y., "Survey on factuality in large language models: Knowledge, retrieval and domain-specificity", CoRR, Vol. abs/2310.07521, 2023, dostupno na: https://doi.org/10.48550/arXiv.2310.07521

[28]Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions", CoRR, Vol. abs/2311.05232, 2023, dostupno na: https://doi.org/10.48550/arXiv.2311.05232

[29]Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N. K., "Bias and fairness in large language models: A survey", CoRR, Vol. abs/2309.00770, 2023, dostupno na: https://doi.org/10.48550/arXiv.2309.00770

[30]Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J., "This looks like that: Deep learning for interpretable image recognition", in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems

2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., Garnett, R., (ur.), 2019, str. 8928–8939, dostupno na: https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html

[31]Nauta, M., van Bree, R., Seifert, C., "Neural prototype trees for interpretable fine-grained image recognition", in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021, str. 14 933–14 943, dostupno na: https://openaccess.thecvf.com/content/CVPR2021/html/Nauta_Neural_Prototype_Trees_for_Interpretable_Fine-Grained_Image_Recognition_CVPR_2021_paper.html

[32]Brendel, W., Bethge, M., "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet", in 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019, dostupno na: https://openreview.net/forum?id=SkfMWhAqYQ

[33]Wang, W., Han, C., Zhou, T., Liu, D., "Visual recognition with deep nearest centroids", in The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023, dostupno na: https://openreview.net/pdf?id=CsKwavjr7A

[34]Donnelly, J., Barnett, A. J., Chen, C., "Deformable protopnet: An interpretable image classifier using deformable prototypes", in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 2022, str. 10 255–10 265, dostupno na: https://doi.org/10.1109/CVPR52688.2022.01002

[35]Böhle, M., Fritz, M., Schiele, B., "B-cos networks: Alignment is all we need for interpretability", in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 2022, str. 10 319–10 328, dostupno na: https://doi.org/10.1109/CVPR52688.2022.01008

[36]Lipton, Z. C., "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.", Queue, Vol. 16, No. 3, Jun. 2018, str. 31–57, dostupno na: http://dx.doi.org/10.1145/3236386.3241340

[37]Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A. J., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S., "From local explanations to global understanding with explainable AI for trees", Nat. Mach. Intell., Vol. 2, No. 1, 2020, str. 56–67, dostupno na: https://doi.org/10.1038/s42256-019-0138-9

[38]Azzolin, S., Longa, A., Barbiero, P., Liò, P., Passerini, A., “Global explainability of gnns via logic combination of learned concepts”, in The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023, dostupno na: https://openreview.net/pdf?id=OTbRTIY4YS

[39]Zeiler, M. D., Fergus, R., “Visualizing and understanding convolutional networks”, in Computer Vision - ECCV 2014 - 13th European Conference, ser. Lecture Notes in Computer Science, Vol. 8689. Zurich, Switzerland: Springer, 2014, str. 818–833, dostupno na: https://doi.org/10.1007/978-3-319-10590-1_53

[40]Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., Wattenberg, M., “Smoothgrad: removing noise by adding noise”, CoRR, Vol. abs/1706.03825, 2017, dostupno na: http://arxiv.org/abs/1706.03825

[41]Ancona, M., Ceolini, E., Öztireli, C., Gross, M. (2018) Towards better understanding of gradient-based attribution methods for deep neural networks. ICLR. Vancouver, BC, Canada, dostupno na: https://openreview.net/forum?id=Sy21R9JAW

[42]Sundararajan, M., Taly, A., Yan, Q., “Axiomatic attribution for deep networks”, in Proceedings of the 34th International Conference on Machine Learning, ICML 2017, ser. Proceedings of Machine Learning Research, Vol. 70. Sydney, NSW, Australia: PMLR, 2017, str. 3319–3328, dostupno na: http://proceedings.mlr.press/v70/sundararajan17a.html

[43]Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., “Grad-CAM: Visual explanations from deep networks via gradient-based localization”, International Journal of Computer Vision, Vol. 128, No. 2, oct 2019, str. 336–359, dostupno na: https://doi.org/10.1007%2Fs11263-019-01228-7

[44]Petsiuk, V., Das, A., Saenko, K., “RISE: randomized input sampling for explanation of black-box models”, in British Machine Vision Conference 2018, BMVC 2018. Newcastle, UK: BMVA Press, 2018, str. 151, dostupno na: http://bmvc2018.org/contents/papers/1064.pdf

[45]Fong, R., Patrick, M., Vedaldi, A., “Understanding deep networks via extremal perturbations and smooth masks”, in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. IEEE, 2019, str. 2950–2958, dostupno na: https://doi.org/10.1109/ICCV.2019.00304

[46]Fong, R. C., Vedaldi, A., “Interpretable explanations of black boxes by meaningful perturbation”, in IEEE International Conference on Computer Vision, ICCV 2017,

Venice, Italy, October 22-29, 2017. IEEE Computer Society, 2017, str. 3449–3457, dostupno na: https://doi.org/10.1109/ICCV.2017.371

[47]Petsiuk, V., Das, A., Saenko, K., “RISE: randomized input sampling for explanation of black-box models”, in British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. BMVA Press, 2018, str. 151, dostupno na: http://bmvc2018.org/contents/papers/1064.pdf

[48]Yang, Q., Zhu, X., Fwu, J., Ye, Y., You, G., Zhu, Y., “MFPP: morphological fragmental perturbation pyramid for black-box model explanations”, in 25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021. IEEE, 2020, str. 1376–1383, dostupno na: https://doi.org/10.1109/ICPR48806.2021.9413046

[49]Dabkowski, P., Gal, Y., “Real time image saliency for black box classifiers”, in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., Garnett, R., (ur.), 2017, str. 6967–6976, dostupno na: https://proceedings.neurips.cc/paper/2017/hash/0060ef47b12160b9198302ebdb144dcf-Abstract.html

[50]Ribeiro, M. T., Singh, S., Guestrin, C., “"why should I trust you?": Explaining the predictions of any classifier”, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM, 2016, str. 1135–1144, dostupno na: https://doi.org/10.1145/2939672.2939778

[51]Wu, Z., D’Oosterlinck, K., Geiger, A., Zur, A., Potts, C., “Causal proxy models for concept-based model explanations”, in International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, ser. Proceedings of Machine Learning Research, Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., (ur.), Vol. 202. PMLR, 2023, str. 37 313–37 334, dostupno na: https://proceedings.mlr.press/v202/wu23b.html

[52]Wood-Doughty, Z., Cachola, I., Dredze, M., “Proxy model explanations for time series rnns”, in 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021, Pasadena, CA, USA, December 13-16, 2021, Wani, M. A., Sethi, I. K., Shi, W., Qu, G., Raicu, D. S., Jin, R., (ur.). IEEE, 2021, str. 698–703, dostupno na: https://doi.org/10.1109/ICMLA52953.2021.00117

[53]Patil, M. S., Främling, K., “Do intermediate feature coalitions aid explainability of black-box models?”, in Explainable Artificial Intelligence - First World Conference,

xAI 2023, Lisbon, Portugal, July 26-28, 2023, Proceedings, Part I, ser. Communications in Computer and Information Science, Longo, L., (ur.), Vol. 1901. Springer, 2023, str. 115–130, dostupno na: https://doi.org/10.1007/978-3-031-44064-9_7

[54]Wachter, S., Mittelstadt, B., Russell, C., "Counterfactual explanations without opening the black box: Automated decisions and the gdpr", Harv. JL & Tech., Vol. 31, 2017, str. 841.

[55]Asher, N., de Lara, L., Paul, S., Russell, C., "Counterfactual models for fair and adequate explanations", Mach. Learn. Knowl. Extr., Vol. 4, No. 2, 2022, str. 316–349, dostupno na: https://doi.org/10.3390/make4020014

[56]Byrne, R. M. J., "Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning", in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, Kraus, S., (ur.). ijcai.org, 2019, str. 6276–6282, dostupno na: https://doi.org/10.24963/ijcai.2019/876

[57]Chou, Y., Moreira, C., Bruza, P., Ouyang, C., Jorge, J. A., "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications", Inf. Fusion, Vol. 81, 2022, str. 59–83, dostupno na: https://doi.org/10.1016/j.inffus.2021.11.003

[58]Dandl, S., Molnar, C., Binder, M., Bischl, B., "Multi-objective counterfactual explanations", in Parallel Problem Solving from Nature - PPSN XVI - 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I, ser. Lecture Notes in Computer Science, Bäck, T., Preuss, M., Deutz, A. H., Wang, H., Doerr, C., Emmerich, M. T. M., Trautmann, H., (ur.), Vol. 12269. Springer, 2020, str. 448–469, dostupno na: https://doi.org/10.1007/978-3-030-58112-1_31

[59]Koh, P. W., Liang, P., "Understanding black-box predictions via influence functions", in Proceedings of the 34th International Conference on Machine Learning, ICML 2017, ser. Proceedings of Machine Learning Research, Vol. 70. Sydney, NSW, Australia: PMLR, 2017, str. 1885–1894, dostupno na: http://proceedings.mlr.press/v70/koh17a.html

[60]Koh, P. W., Ang, K., Teo, H. H. K., Liang, P., "On the accuracy of influence functions for measuring group effects", in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., Garnett, R., (ur.),

2019, str. 5255–5265, dostupno na: https://proceedings.neurips.cc/paper/2019/hash/a78482ce76496fcf49085f2190e675b4-Abstract.html

[61]Adadi, A., Berrada, M., "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)", IEEE Access, Vol. 6, 2018, str. 52 138–52 160, dostupno na: https://doi.org/10.1109/ACCESS.2018.2870052

[62]Leavitt, M. L., Morcos, A. S., "Towards falsifiable interpretability research", CoRR, Vol. abs/2010.12016, 2020, dostupno na: https://arxiv.org/abs/2010.12016

[63]Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., Doshi-Velez, F., "An evaluation of the human-interpretability of explanation", CoRR, Vol. abs/1902.00006, 2019, dostupno na: http://arxiv.org/abs/1902.00006

[64]Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI", Inf. Fusion, Vol. 58, 2020, str. 82–115, dostupno na: https://doi.org/10.1016/j.inffus.2019.12.012

[65]Burkart, N., Huber, M. F., "A survey on the explainability of supervised machine learning", J. Artif. Intell. Res., Vol. 70, 2021, str. 245–317, dostupno na: https://doi.org/10.1613/jair.1.12228

[66]Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C., "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI", ACM Comput. Surv., Vol. 55, No. 13s, 2023, str. 295:1–295:42, dostupno na: https://doi.org/10.1145/3583558

[67]Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., Kim, B., "Sanity checks for saliency maps", in Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., (ur.), 2018, str. 9525–9536, dostupno na: https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[68]Jacovi, A., Goldberg, Y., "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?", in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. R., (ur.).

Association for Computational Linguistics, 2020, str. 4198–4205, dostupno na: https://doi.org/10.18653/v1/2020.acl-main.386

[69]Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation", PLOS ONE, Vol. 10, No. 7, jul 2015, str. e0130140, dostupno na: https://doi.org/10.1371%2Fjournal.pone.0130140

[70]Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A., "Not just a black box: Learning important features through propagating activation differences", CoRR, Vol. abs/1605.01713, 2016, dostupno na: http://arxiv.org/abs/1605.01713

[71]Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.-R., "Explaining nonlinear classification decisions with deep taylor decomposition", Pattern Recognition, Vol. 65, May 2017, str. 211–222, dostupno na: https://doi.org/10.1016/j.patcog.2016.11.008

[72]Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S., "Top-down neural attention by excitation backprop", International Journal of Computer Vision, Vol. 126, No. 10, 2018, str. 1084–1102.

[73]Arras, L., Horn, F., Montavon, G., Müller, K.-R., Samek, W., "Explaining predictions of non-linear classifiers in NLP", in Proceedings of the 1st Workshop on Representation Learning for NLP, Blunsom, P., Cho, K., Cohen, S., Grefenstette, E., Hermann, K. M., Rimell, L., Weston, J., Yih, S. W.-t., (ur.). Berlin, Germany: Association for Computational Linguistics, Aug. 2016, str. 1–7, dostupno na: https://aclanthology.org/W16-1601

[74]Gers, F. A., Schmidhuber, J., Cummins, F. A., "Learning to forget: Continual prediction with LSTM", Neural Comput., Vol. 12, No. 10, 2000, str. 2451–2471, dostupno na: https://doi.org/10.1162/089976600300015015

[75]Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., "Learning phrase representations using RNN encoder-decoder for statistical machine translation", in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, Moschitti, A., Pang, B., Daelemans, W., (ur.). ACL, 2014, str. 1724–1734, dostupno na: https://doi.org/10.3115/v1/d14-1179

[76]Arras, L., Montavon, G., Müller, K.-R., Samek, W., "Explaining recurrent neural network predictions in sentiment analysis", in Proceedings of the 8th Workshop on

Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, str. 159–168, dostupno na: https://aclanthology.org/W17-5221

[77]Yang, Y., Tresp, V., Wunderle, M., Fasching, P. A., "Explaining therapy predictions with layer-wise relevance propagation in neural networks", in 2018 IEEE International Conference on Healthcare Informatics (ICHI), 2018, str. 152-162.

[78]Böhle, M., Eitel, F., Weygandt, M., Ritter, K., "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification", Frontiers in Aging Neuroscience, Vol. 11, 2019, dostupno na: https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194

[79]Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A. U., Ruprecht, K., Giess, R. M., Kuchling, J., Asseyer, S., Weygandt, M., Haynes, J.-D., Scheel, M., Paul, F., Ritter, K., "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation", NeuroImage: Clinical, Vol. 24, 2019, str. 102003, dostupno na: https://www.sciencedirect.com/science/article/pii/S2213158219303535

[80]Nam, H., Kim, J.-M., Choi, W., Bak, S., Kam, T.-E., "The effects of layer-wise relevance propagation-based feature selection for eeg classification: a comparative study on multiple datasets", Frontiers in Human Neuroscience, Vol. 17, 2023, dostupno na: https://www.frontiersin.org/articles/10.3389/fnhum.2023.1205881

[81]Nam, W., Gur, S., Choi, J., Wolf, L., Lee, S., "Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks", in The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. New York, NY, USA: AAAI Press, 2020, str. 2501–2508, dostupno na: http://arxiv.org/abs/1904.00605

[82]Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale", in 9th International Conference on Learning Representations, ICLR 2021. Virtual Event, Austria: OpenReview.net, 2021, dostupno na: https://openreview.net/forum?id=YicbFdNTTy

[83]Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows", in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 2021, str. 9992–10 002, dostupno na: https://doi.org/10.1109/ICCV48922.2021.00986

[84]Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W. T., "Unsupervised semantic segmentation by distilling feature correspondences", in The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022, dostupno na: https://openreview.net/forum?id=SaKO6z6Hl0c

[85]Devlin, J., Chang, M., Lee, K., Toutanova, K., "BERT: pre-training of deep bidirectional transformers for language understanding", in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1. Minneapolis, MN, USA: Association for Computational Linguistics, 2019, str. 4171–4186, dostupno na: https://doi.org/10.18653/v1/n19-1423

[86]Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., "Zero-shot text-to-image generation", in Proceedings of the 38th International Conference on Machine Learning, ICML 2021, ser. Proceedings of Machine Learning Research, Meila, M., Zhang, T., (ur.), Vol. 139. Virtual Event: PMLR, 2021, str. 8821–8831, dostupno na: http://proceedings.mlr.press/v139/ramesh21a.html

[87]Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., "High-resolution image synthesis with latent diffusion models", in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 2022, str. 10 674–10 685, dostupno na: https://doi.org/10.1109/CVPR52688.2022.01042

[88]Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R., "Scaling rectified flow transformers for high-resolution image synthesis", CoRR, Vol. abs/2403.03206, 2024, dostupno na: https://doi.org/10.48550/arXiv.2403.03206

[89]Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., Mordatch, I., "Decision transformer: Reinforcement learning via sequence modeling", in Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021. Virtual Event: NeurIPS, 2021, str. 15 084–15 097, dostupno na: https://openreview.net/pdf?id=a7APmM4B9d

[90]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., "Attention is all you need", in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, CA, USA: NeurIPS, 2017, str. 5998–6008, dostupno na: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[91]Ba, L. J., Kiros, J. R., Hinton, G. E., "Layer normalization", CoRR, Vol. abs/1607.06450, 2016, dostupno na: http://arxiv.org/abs/1607.06450

[92]He, K., Zhang, X., Ren, S., Sun, J., "Deep residual learning for image recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, Nevada: CVPR, 2016, str. 770–778.

[93]Vukadin, D., Kurdija, A. S., Dela č, G., Šilić, M., "Information extraction from free-form cv documents in multiple languages", IEEE Access, Vol. 9, 2021, str. 84 559-84 575.

[94]Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., "End-to-end object detection with transformers", in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, ser. Lecture Notes in Computer Science, Vedaldi, A., Bischof, H., Brox, T., Frahm, J., (ur.), Vol. 12346. Glasgow, UK: Springer, 2020, str. 213–229, dostupno na: https://doi.org/10.1007/978-3-030-58452-8_13

[95]Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., Bengio, Y., "Show, attend and tell: Neural image caption generation with visual attention", in Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, ser. JMLR Workshop and Conference Proceedings, Vol. 37. Lille, France: JMLR.org, 2015, str. 2048–2057, dostupno na: http://proceedings.mlr.press/v37/xuc15.html

[96]Ukwuoma, C. C., Qin, Z., Belal Bin Heyat, M., Akhtar, F., Bamisile, O., Muaad, A. Y., Addo, D., Al-antari, M. A., "A hybrid explainable ensemble transformer encoder for pneumonia identification from chest x-ray images", Journal of Advanced Research, Vol. 48, Jun. 2023, str. 191–211, dostupno na: http://dx.doi.org/10.1016/j.jare.2022.08.021

[97]Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., Bengio, Y., "Show, attend and tell: Neural image caption generation with visual attention", in Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, ser. JMLR Workshop and Conference Proceedings, Bach, F. R., Blei, D. M., (ur.), Vol. 37. JMLR.org, 2015, str. 2048–2057, dostupno na: http://proceedings.mlr.press/v37/xuc15.html

[98]Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., "End-to-end object detection with transformers", in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I, ser. Lecture Notes in Computer Science, Vedaldi, A., Bischof, H., Brox, T., Frahm, J., (ur.), Vol. 12346. Springer, 2020, str. 213–229, dostupno na: https://doi.org/10.1007/978-3-030-58452-8_13

[99]Chefer, H., Gur, S., Wolf, L., "Transformer interpretability beyond attention visualization", in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. Virtual Event: Computer Vision Foundation / IEEE, 2021, str. 782–791, dostupno na: https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html

[100]Jain, S., Wallace, B. C., "Attention is not explanation", in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Burstein, J., Doran, C., Solorio, T., (ur.). Association for Computational Linguistics, 2019, str. 3543–3556, dostupno na: https://doi.org/10.18653/v1/n19-1357

[101]Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L., "Scaling vision transformers", in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 2022, str. 1204–1213, dostupno na: https://doi.org/10.1109/CVPR52688.2022.01179

[102]Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Ruiz, C. R., Minderer, M., Puigcerver, J., Evci, U., Kumar, M., van Steenkiste, S., Elsayed, G. F., Mahendran, A., Yu, F., Oliver, A., Huot, F., Bastings, J., Collier, M., Gritsenko, A. A., Birodkar, V., Vasconcelos, C. N., Tay, Y., Mensink, T., Kolesnikov, A., Pavetic, F., Tran, D., Kipf, T., Lucic, M., Zhai, X., Keysers, D., Harmsen, J. J., Houlsby, N., "Scaling vision transformers to 22 billion parameters", in International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, ser. Proceedings of Machine Learning Research, Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., (ur.), Vol. 202. PMLR, 2023, str. 7480–7512, dostupno na: https://proceedings.mlr.press/v202/dehghani23a.html

[103]Huang, G., Fu, H., Bors, A. G., "Masked image residual learning for scaling deeper vision transformers", in Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023,

New Orleans, LA, USA, December 10 - 16, 2023, Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., (ur.), 2023, dostupno na: http://papers.nips.cc/paper_files/paper/2023/hash/b3bac97f3227c52c0179a6d967480867-Abstract-Conference.html

[104]Wiegreffe, S., Pinter, Y., "Attention is not not explanation", in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Inui, K., Jiang, J., Ng, V., Wan, X., (ur.). Association for Computational Linguistics, 2019, str. 11–20, dostupno na: https://doi.org/10.18653/v1/D19-1002

[105]Abnar, S., Zuidema, W., "Quantifying attention flow in transformers", in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, Jul. 2020, str. 4190–4197, dostupno na: https://aclanthology.org/2020.acl-main.385

[106]Mondal, A. K., Bhattacharjee, A., Singla, P., Prathosh, A. P., "xvitcos: Explainable vision transformer based covid-19 screening using radiography", IEEE Journal of Translational Engineering in Health and Medicine, Vol. 10, 2022, str. 1-10.

[107]Xue, Z., Yu, K., Pearlman, P. C., Pal, A., Chen, T.-C., Hua, C.-H., Kang, C. J., Chien, C.-Y., Tsai, M.-H., Wang, C.-P., Chaturvedi, A. K., Antani, S., "Automatic detection of oral lesion measurement ruler toward computer-aided image-based oral cancer screening", in 2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), 2022, str. 3218-3221.

[108]Ha, C., Go, T., Choi, W., "Intelligent healthcare platform for diagnosis of scalp and hair disorders", Applied Sciences, Vol. 14, No. 5, 2024, dostupno na: https://www.mdpi.com/2076-3417/14/5/1734

[109]Gupta, S., Lakhotia, S., Rawat, A., Tallamraju, R., "Vitol: Vision transformer for weakly supervised object localization", in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022. IEEE, 2022, str. 4100–4109, dostupno na: https://doi.org/10.1109/CVPRW56347.2022.00455

[110]Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K., Wolf, L., "XAI for transformers: Better explanations through conservative propagation", in International Conference on Machine Learning, ICML 2022, ser. Proceedings of Machine Learning Research, Vol. 162. Baltimore, Maryland, USA: PMLR, 2022, str. 435–451, dostupno na: https://proceedings.mlr.press/v162/ali22a.html

[111]Simonyan, K., Vedaldi, A., Zisserman, A., "Deep inside convolutional networks: Visualising image classification models and saliency maps", in 2nd International Conference on Learning Representations, ICLR 2014, Bengio, Y., LeCun, Y., (ur.). Banff, AB, Canada: ICLR, 2014, dostupno na: http://arxiv.org/abs/1312.6034

[112]Springenberg, J. T., Dosovitskiy, A., Brox, T., Riedmiller, M. A., "Striving for simplicity: The all convolutional net", in 3rd International Conference on Learning Representations, ICLR 2015, Bengio, Y., LeCun, Y., (ur.). San Diego, CA, USA: ICLR, 2015, dostupno na: http://arxiv.org/abs/1412.6806

[113]Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., Wattenberg, M., "Smoothgrad: removing noise by adding noise", CoRR, Vol. abs/1706.03825, 2017, dostupno na: http://arxiv.org/abs/1706.03825

[114]Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., "Grad-CAM: Visual explanations from deep networks via gradient-based localization", International Journal of Computer Vision, Vol. 128, No. 2, oct 2019, str. 336–359, dostupno na: https://doi.org/10.1007%2Fs11263-019-01228-7

[115]Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A., "Learning deep features for discriminative localization", CoRR, Vol. abs/1512.04150, 2015, dostupno na: http://arxiv.org/abs/1512.04150

[116]Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V. N., "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks", in 2018 IEEE winter conference on applications of computer vision (WACV), IEEE. Lake Tahoe, NV, USA: IEEE Workshop on Applications of Computer Vision, 2018, str. 839–847.

[117]Draelos, R. L., Carin, L. (2021) Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. CoRR.

[118]Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., Wei, Y., "Layercam: Exploring hierarchical class activation maps for localization", IEEE Transactions on Image Processing, Vol. 30, 2021, str. 5875–5888.

[119]Liu, Y., Li, H., Guo, Y., Kong, C., Li, J., Wang, S., "Rethinking attention-model explainability through faithfulness violation test", in International Conference on Machine Learning, PMLR. 1 W Pratt St, Baltimore, MD 21201, USA: ICML, 2022, str. 13 807–13 824.

[120]Bhatt, U., Weller, A., Moura, J. M. F., "Evaluating and aggregating feature-based model explanations", in Proceedings of the Twenty-Ninth International Joint Conference on

Artificial Intelligence. Yokohama, Japan: IJCAI, 2020, str. 3016—3022, dostupno na: https://arxiv.org/abs/2005.00631

[121]Alvarez Melis, D., Jaakkola, T., "Towards robust interpretability with self-explaining neural networks", in Advances in Neural Information Processing Systems, Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., (ur.), Vol. 31. Montréal, Canada: Curran Associates, Inc., 2018, dostupno na: https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf

[122]Nguyen, A., Martínez, M. R., "On quantitative aspects of model interpretability", CoRR, Vol. abs/2007.07584, 2020, dostupno na: https://arxiv.org/abs/2007.07584

[123]Montavon, G., Samek, W., Müller, K.-R., "Methods for interpreting and understanding deep neural networks", Digital Signal Processing, Vol. 73, feb 2018, str. 1–15, dostupno na: https://doi.org/10.1016%2Fj.dsp.2017.10.011

[124]Rieger, L., Hansen, L. K., "IROF: a low resource evaluation metric for explanation methods", CoRR, Vol. abs/2003.08747, 2020, dostupno na: https://arxiv.org/abs/2003.08747

[125]Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., Ravikumar, P. K., "On the (in) fidelity and sensitivity of explanations", Advances in Neural Information Processing Systems, Vol. 32, 2019, str. 10 967–10 978.

[126]Arya, V., Bellamy, R. K. E., Chen, P., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilovic, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J. T., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., Zhang, Y. (2019) One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. CoRR, dostupno na: http://arxiv.org/abs/1909.03012

[127]Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E., "A consistent and efficient evaluation strategy for attribution methods", in Proceedings of Machine Learning Research. Baltimore, Maryland, USA: PMLR, 2022, str. 18 770–18 795.

[128]Lee, J., Cho, H., Pyun, Y., Kang, S., Nam, H., "Heatmap assisted accuracy score evaluation method for machine-centric explainable deep neural networks", IEEE Access, Vol. 10, 2022, str. 64 832–64 849.

[129]Dasgupta, S., Frost, N., Moshkovitz, M., "Framework for evaluating faithfulness of local explanations", in Proceedings of Machine Learning Research. Baltimore, Maryland, USA: PMLR, 2022, str. 4794–4815, dostupno na: https://arxiv.org/abs/2202.00734

[130]Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., Lapuschkin, S., "Towards best practice in explaining neural network decisions with lrp", in 2020 International Joint Conference on Neural Networks (IJCNN), IEEE. Glasgow, United Kingdom: IJCNN, 2020, str. 1–7.

[131]Arras, L., Osman, A., Samek, W., "Ground truth evaluation of neural network explanations with clevr-xai", Information Fusion, Vol. 81, No. 3, 2020, str. 14–40, dostupno na: https://arxiv.org/abs/2003.07258

[132]Theiner, J., Müller-Budack, E., Ewerth, R., "Interpretable semantic photo geolocation", in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, IEEE. Waikoloa, HI: WACV, 2022, str. 750–760.

[133]Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., "Evaluating the visualization of what a deep neural network has learned", IEEE transactions on neural networks and learning systems, Vol. 28, No. 11, 2016, str. 2660–2673.

[134]Ju, Y., Zhang, Y., Yang, Z., Jiang, Z., Liu, K., Zhao, J., "Logic traps in evaluating attribution scores", in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022, str. 5911–5922.

[135]Li, J., Monroe, W., Jurafsky, D., "Understanding neural networks through representation erasure", CoRR, Vol. abs/1612.08220, 2016, dostupno na: http://arxiv.org/abs/1612.08220

[136]Alvarez-Melis, D., Jaakkola, T. S., "On the robustness of interpretability methods", CoRR, Vol. abs/1806.08049, 2018, dostupno na: http://arxiv.org/abs/1806.08049

[137]Papernot, N., McDaniel, P. D., Jha, S., Fredrikson, M., Celik, Z. B., Swami, A., "The limitations of deep learning in adversarial settings", in IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016. IEEE, 2016, str. 372–387, dostupno na: https://doi.org/10.1109/EuroSP.2016.36

[138]Goodfellow, I. J., Shlens, J., Szegedy, C., "Explaining and harnessing adversarial examples", in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Bengio, Y., LeCun, Y., (ur.), 2015, dostupno na: http://arxiv.org/abs/1412.6572

[139]Arias-Duart, A., Parés, F., Garcia-Gasulla, D., Giménez-Ábalos, V., "Focus! rating xai methods and finding biases", in 2022 IEEE International Conference on Fuzzy Systems, IEEE. Padua, Italy: FUZZ-IEEE, 2022, str. 1–8.

[140]Gu, J., Yang, Y., Tresp, V., "Understanding individual decisions of cnns via contrastive backpropagation", in Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Springer. Perth, Australia: ACCV, 2019, str. 119–134.

[141]Zhang, H., Torres, F., Sicre, R., Avrithis, Y., Ayache, S., "Opti-CAM: Optimizing saliency maps for interpretability", dostupno na: http://arxiv.org/abs/2301.07002 ArXiv:2301.07002 [cs]. Feb. 2024.

[142]Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., Lakkaraju, H., "Rethinking stability for attribution-based explanations", in ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data, 2022, dostupno na: https://openreview.net/forum?id=BfxZAuWOg9

[143]Brunke, L., Agrawal, P., George, N., "Evaluating input perturbation methods for interpreting cnns and saliency map comparison", in Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part I, ser. Lecture Notes in Computer Science, Bartoli, A., Fusiello, A., (ur.), Vol. 12535. Springer, 2020, str. 120–134, dostupno na: https://doi.org/10.1007/978-3-030-66415-2_8

[144]Brocki, L., Chung, N. C., "Evaluation of interpretability methods and perturbation artifacts in deep neural networks", CoRR, Vol. abs/2203.02928, 2022, dostupno na: https://doi.org/10.48550/arXiv.2203.02928

[145]Nguyen, G., Kim, D., Nguyen, A., "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores", in Advances in Neural Information Processing Systems, Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J. W., (ur.), Vol. 34. Curran Associates, Inc., 2021, str. 26 422–26 436, dostupno na: https://proceedings.neurips.cc/paper_files/paper/2021/file/de043a5e421240eb846da8effe472ff1-Paper.pdf

[146]Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., "Learning deep features for discriminative localization", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, str. 2921–2929.

[147]Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., "Slic superpixels compared to state-of-the-art superpixel methods", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 11, Nov. 2012, str. 2274–2282, dostupno na: http://dx.doi.org/10.1109/TPAMI.2012.120

[148]Gao, S., Li, Z.-Y., Yang, M.-H., Cheng, M.-M., Han, J., Torr, P., "Large-scale unsupervised semantic segmentation", IEEE Transactions on Pattern Analysis and

Machine Intelligence, Vol. 45, No. 6, Jun. 2023, str. 7457–7476, dostupno na: http://dx.doi.org/10.1109/TPAMI.2022.3218275

[149]Li, W.-H., Liu, X., Bilen, H., "Universal representation learning from multiple domains for few-shot classification", in IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, str. 9526-9535.

[150]Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., "Imagenet: A large-scale hierarchical image database", in 2009 IEEE conference on computer vision and pattern recognition, Ieee. Miami, Florida, USA: Conference on Computer Vision and Pattern Recognition, 2009, str. 248–255.

[151]Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A., "The pascal visual object classes (voc) challenge", International Journal of Computer Vision, Vol. 88, No. 2, Jun. 2010, str. 303–338.

[152]Krizhevsky, A., Sutskever, I., Hinton, G. E., "Imagenet classification with deep convolutional neural networks", in Advances in Neural Information Processing Systems, Pereira, F., Burges, C., Bottou, L., Weinberger, K., (ur.), Vol. 25. Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2012.

[153]Iwana, B. K., Kuroki, R., Uchida, S., "Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation", in 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019. IEEE, 2019, str. 4176–4185, dostupno na: https://doi.org/10.1109/ICCVW.2019.00513

[154]Simonyan, K., Zisserman, A., "Very deep convolutional networks for large-scale image recognition", in International Conference on Learning Representations. San Diego, CA, USA: International Conference on Learning Representations, 2015.

[155]Wilcoxon, F., "Individual comparisons by ranking methods", Biometrics Bulletin, Vol. 1, No. 6, Dec. 1945, str. 80, dostupno na: http://dx.doi.org/10.2307/3001968

[156]Bennett, E. M., Alpert, R., Goldstein, A. C., "Communications through limited response questioning", Public Opinion Quarterly, Vol. 18, No. 3, 1954, str. 303, dostupno na: http://dx.doi.org/10.1086/266520

[157]Cohen, J., "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.", Psychological Bulletin, Vol. 70, No. 4, 1968, str. 213–220, dostupno na: http://dx.doi.org/10.1037/h0026256

[158]Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., Lowe, R., "Training language models to follow instructions with human feedback", in Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., (ur.), 2022, dostupno na: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

Popis slika

Popis tablica

Životopis

Davor Vukadin rođen je 1996. godine u Zagrebu u Hrvatskoj gdje je završio Gimnaziju Lucijana Vranjanina. 2015. upisuje preddiplomski studij računarske znanosti u Zagrebu na Fakultetu Elektrotehnike i Računarstva, Sveučilišta u Zagrebu, kojeg završava 2018. godine. Diplomski studij završava 2020. godine uz najviše pohvale *Summa Cum Laude* kako je po uspjehu bio u najboljih 1% generacije. Tijekom studija radio je kao mladi istraživač u *Ericsson Nikola Tesla d.d.* i *TakeLab*-u te kao podatkovni znanstvenik u *DataBlast d.o.o.* 2020. godine zaposlen je na poziciji istraživača na Fakultetu Elektrotehnike i Računarstva, Sveučilišta u Zagrebu gdje je radio na projektu *VODIME - vode Imotske krajine*. Također radi u vlastitom obrtu *KVIKI MIC*. Za vrijeme rada na fakultetu sudjelovao je u izvođenju nekoliko kolegija. 2022. zapošljava se kao podatkovni znanstvenik u *ScanDoc d.o.o.* U siječnju 2024. se zapošljava kao senior inženjer umjetne inteligencije u *iOLAP d.o.o.* Konačno, u rujnu 2024. se zapošljava kao istraživač na Fakultetu Elektrotehnike i Računarstva, Sveučilišta u Zagrebu. Objavio je radove u časopisima *ACM Transactions on Intelligent Systems and Technology* i *IEEE Access* te na konferenciji *MIPRO*. Njegovi znanstveni interesi fokusirani su na područje objašnjive umjetne inteligencije, vrjednovanju metoda unutar ovog područja te primjeni istih na najnovije arhitekture unutar područja dubokog učenja.

Popis objavljenih djela

Radovi u časopisima

1. Vukadin, D., Afri ć, P., Šilić, M. i Delač, G., "Advancing Attribution-Based Neural Network Explainability through Relative Absolute Magnitude Layer-Wise Relevance Propagation and Multi-Component Evaluation", ACM Transactions on Intelligent Systems and Technology, Vol. 15, Issue 3, travanj 2024.
2. Vukadin, D., Satja Kurdija, A., Dela č, G., i Šilić, M., "Information Extraction From Free-Form CV Documents in Multiple Languages", IEEE Access, Vol. 9, Lipanj 2021, str. 84559-84575
3. Afri ć, P., Vukadin, D., Šilić, M. i Delač, G., "Empirical Study: How Issue Classification Influences Software Defect Prediction", IEEE Access, Vol. 11, Veljača 2023, str. 11732-

11748

Radovi na konferencijama

1.Vukadin, D., Šili ć, M., Delač, G. i Vladimir, K., "Evaluating Harmony: Neural Network Explanation Metrics and Human Perception", Proceedings of the International Conference on Computers in Technical Systems MIPRO 2024 Opatija, Opatija, Hrvatska, svibanj 2024., str. 13-18

2.Džida, M., Vukadin, D., Šili ć, M., Delač, G. i Vladimir, K., "An Overview of State-of-the-art Solutions for Scene Text Detection", Proceedings of the International Conference on Computers in Technical Systems MIPRO 2023 Opatija, Opatija, Hrvatska, svibanj 2023., str. 947-952

Biography

Davor Vukadin was born in 1996 in Zagreb, Croatia, where he completed Lucijan Vranjanin High School. In 2015, he enrolled in the undergraduate program in computer science at the Faculty of Electrical Engineering and Computing, University of Zagreb, which he completed in 2018. He graduated with a master's degree in 2020 with the highest honors, *Summa Cum Laude*, as he was among the top 1% of his class. During his studies, he worked as a junior researcher at *Ericsson Nikola Tesla d.d.* and *TakeLab*, as well as a data scientist at *DataBlast d.o.o.* In 2020, he was employed as a researcher at the Faculty of Electrical Engineering and Computing, University of Zagreb, where he worked on the project *VODIME - waters of the Imotski region*. He also runs his own business, *KVIKI MIC*. While working at the faculty, he participated in teaching several courses. In 2022, he was hired as a data scientist at *ScanDoc d.o.o.* In January of 2024, he was employed as a senior artificial intelligence engineer at *iOLAP d.o.o.* Finally, in September of 2024, he was employed as a researcher at the Faculty of Electrical Engineering and Computing, University of Zagreb. He has published papers in the journals *ACM Transactions on Intelligent Systems and Technology* and *IEEE Access*, as well as at the *MIPRO* conference. His scientific interests are focused on the field of explainable artificial intelligence, the evaluation of methods within this field, and their application to the latest architectures in the field of deep learning.