Optimizacija velikih jezičnih modela korištenjem LoRA tehnike za identifikaciju investicijskih prilika

Talan, Fran

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1589

OPTIMIZACIJA VELIKIH JEZIČNIH MODELA KORIŠTENJEM LORA TEHNIKE ZA IDENTIFIKACIJU INVESTICIJSKIH PRILIKA

Fran Talan

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1589

OPTIMIZACIJA VELIKIH JEZIČNIH MODELA KORIŠTENJEM LORA TEHNIKE ZA IDENTIFIKACIJU INVESTICIJSKIH PRILIKA

Fran Talan

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1589

Pristupnik:	**Fran Talan (0036541130)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentor:	doc. dr. sc. Mario Brčić

Zadatak: **Optimizacija velikih jezičnih modela korištenjem LoRA tehnike za identifikaciju investicijskih prilika**

Opis zadatka:

U okviru završnog rada treba istražiti mogućnosti primjene varijanti tehnika adaptacije niskog ranga (engl. Low-Rank Adaptation, LoRA) za fino podešavanje velikih jezičnih modela (engl. Large Language Model, LLM) s ciljem poboljšanja procesa prepoznavanja investicijskih prilika koje proizlaze iz inovativnih događaja interesantnih za tržište dionica. Izvor podataka će biti e-bilteni (engl. newsletter). Odabrati postojeći utrenirani LLM otvorenog koda kao početnu točku razvoja. Čak i uz pretpostavku da model ima dobre performanse u navedenoj domeni problema i namjera, očekuje se da će mu performansa monotono opadati sa vremenom i da ažuriranje informacija u težinama modela finim podešavanjem može kontrirati navedenoj pojavi. Ažuriranje modela treba provesti korištenjem LoRA i LoRA+ tehnika finog podešavanja. Treba provesti eksperiment koji će usporediti neažurirani, LoRA i LoRA+ ažurirane modele kroz njihove performanse u prepoznavanju inovativnih događaja koji donose u prepoznavanju inovativnih događaja kao investicijskih prilika.

Rok za predaju rada: 14. lipnja 2024.

Zahvaljujem mentoru doc. dr. sc. Mariu Brčiću na podršci i stručnom vođenju tijekom izrade ovog rada.

Sadržaj

1 Uvod **7**

1.1 Cilj Rada 8

2 Teorijski Okvir **9**

2.1 Veliki Jezični Modeli (LLM) 9

2.1.1 Definicija i Osnovni Principi 9

2.1.2 Povijest i Razvoj LLM-ova 10

2.1.3 Primjene LLM-ova u Različitim Područjima 12

2.2 Uvod u Predikciju Događaja 13

2.3 Tehnike Adaptacije Niskog Ranga (LoRA) 15

2.3.1 Osnove i Principi Rada LoRA Tehnika 15

2.3.2 Razlike između LoRA i LoRA+ 17

2.3.3 Prednosti Korištenja LoRA za Fino Podešavanje Modela 19

2.3.4 Izazovi i Ograničenja LoRA Tehnika 20

3 Predikcija Događaja Pomoću Jezičnih Modela **21**

3.1 Uvod u Predikciju Događaja 21

3.2 Automatizirana Predikcija Događaja s Jezičnim Modelima 21

3.2.1 Jezični Modeli s Poboljšanim Dohvatom Informacija 21

3.2.2 Eksperimentalni Postav i Rezultati 22

3.2.3 Primjena na Predikciju Investicijskih Signala 22

4 Eksperimentalni Rad **23**

4.1 Prikupljanje i Izrada Dataseta 23

4.2 Trening Modela 24

5 Evaluacija Performansi **28**

5.1 Matrica Konfuzije 28

5.2 Evaluacija Početnog Modela 28

5.3 Evaluacija Istreniranog Modela 29

6 Zaključak **31**

7 Literatura **33**

1 Uvod

U ovom završnom radu istražujemo primjenu tehnika adaptacije niskog ranga (LoRA) za optimizaciju velikih jezičnih modela (LLM) u svrhu identifikacije investicijskih prilika. Veliki jezični modeli, kao što su GPT-2, GPT-3, Mistral, postali su nezaobilazni alati u obradi prirodnog jezika zbog njihove sposobnosti generiranja koherentnih i smislenih tekstova. Međutim, prilagodba ovih modela za specifične zadatke često zahtijeva značajne resurse i vrijeme.

Tehnike adaptacije niskog ranga (LoRA) predstavljaju efikasan način za fino podešavanje velikih modela bez potrebe za opsežnim resursima. LoRA tehnike omogućuju smanjenje broja parametara potrebnih za prilagodbu modela, čineći proces bržim i manje zahtjevnim. U ovom radu fokusiramo se na primjenu LoRA i LoRA+ tehnika za identifikaciju investicijskih prilika koje proizlaze iz inovativnih događaja interesantnih za tržište dionica.

Podaci za trening modela prikupljeni su iz e-biltena (engl. newsletter), koji sadrže informacije o inovacijama i događajima relevantnim za tržište dionica. Proces prikupljanja podataka uključivao je automatsko preuzimanje vijesti iz nekoliko pouzdanih izvora, filtriranje prema ključnim riječima vezanim za inovacije i financijska tržišta, te ručno pregledavanje i označavanje relevantnih članaka. Osim e-biltena, korišteni su i podaci s web stranica kao što su BBC, The Guardian, TechCrunch, TechRadar, Wccftech i Reuters kako bi se osigurala raznovrsnost i sveobuhvatnost skupa podataka.

Odabrali smo GPT-2, model otvorenog koda, kao početnu točku za razvoj zbog njegove široke dostupnosti i razumnih zahtjeva za računalne resurse. Eksperimentom ćemo usporediti performanse neažuriranog modela, modela ažuriranog LoRA tehnikama, i modela ažuriranog LoRA+ tehnikama.

1.1 Cilj Rada

Cilj ovog rada je istražiti efikasnost tehnika LoRA i LoRA+ za fino podešavanje velikih jezičnih modela u kontekstu prepoznavanja investicijskih prilika. Specifični ciljevi uključuju:

- **Istražiti trenutne tehnike adaptacije niskog ranga (LoRA) i njihove primjene.** Pregled relevantnih istraživanja i analiza prednosti i nedostataka postojećih metoda.
- **Prikupiti i pripremiti podatke iz e-biltena za trening modela.** Proces prikupljanja podataka, njihovo čišćenje i priprema za korištenje u treningu modela.
- **Implementirati i trenirati GPT-2 model koristeći LoRA i LoRA+ tehnike.** Detaljan opis implementacije, korištenih alata i parametara treninga.
- **Evaluirati performanse neažuriranih i ažuriranih modela.** Usporedba performansi modela kroz mjerne metrike i analiza dobivenih rezultata.

U nastavku rada, nakon uvoda, pružamo teorijski okvir koji uključuje pregled velikih jezičnih modela i tehnika adaptacije niskog ranga. Zatim slijedi poglavlje o predikciji događaja pomoću jezičnih modela, gdje detaljno razmatramo metode i pristupe za automatsku predikciju događaja. Nakon toga prikazujemo eksperimentalni rad koji obuhvaća prikupljanje i izradu dataseta te trening modela. U dijelu evaluacije performansi analiziramo rezultate korištenjem matrice konfuzije. Rad završavamo diskusijom dobivenih rezultata i zaključkom.

2 Teorijski Okvir

U ovom poglavlju obrađujemo ključne koncepte vezane uz velike jezične modele (LLM) i tehnike adaptacije niskog ranga (LoRA).

2.1 Veliki Jezični Modeli (LLM)

Veliki jezični modeli (LLM) su napredni modeli strojnog učenja osposobljeni za obradu i generiranje prirodnog jezika. Primjeri uključuju GPT-2/3.5/4, LLaMA, Mistral i druge.

2.1.1 Definicija i Osnovni Principi

Veliki jezični modeli (LLM) predstavljaju jednu od najnaprednijih tehnologija u području obrade prirodnog jezika (NLP). Ovi modeli koriste duboke neuronske mreže za učenje složenih uzoraka i struktura unutar jezičnih podataka, omogućujući im generiranje koherentnih i smislenih tekstova te razumijevanje konteksta u kojem se koriste riječi.

LLM-ovi se oslanjaju na veliku količinu podataka za treniranje, često uključujući milijarde riječi iz različitih izvora kao što su knjige, članci, web stranice i druge tekstualne baze podataka. Kroz proces treniranja, modeli uče statističke veze između riječi, fraza i rečenica, što im omogućuje da predviđaju sljedeću riječ u nizu ili generiraju odgovore na postavljena pitanja.[5]

Osnovni principi rada LLM-ova uključuju:

- **Učenje iz velikih količina podataka:** LLM-ovi su trenirani na ogromnim korpusima tekstualnih podataka, što im omogućuje učenje širokog spektra jezičnih struktura i konteksta.
- **Korištenje dubokih neuronskih mreža:** Modeli koriste slojevite neuronske mreže koje omogućuju obradu složenih jezičnih uzoraka. Svaki sloj u mreži uči različite aspekte jezika, od jednostavnih morfoloških obrazaca do složenih semantičkih veza.
- **Transformerska arhitektura:** Većina modernih LLM-ova koristi transformersku arhitekturu, koja omogućuje paralelno procesiranje podataka i efikasno učenje dugoročnih ovisnosti u jeziku.
- **Samopouzdanje u generiranju:** LLM-ovi koriste različite tehnike za procjenu vlastite sigurnosti u generiranju odgovora ili predviđanju riječi, čime se postiže veća preciznost i koherentnost u izlazima.

Transformerska arhitektura, koja je prvi put predstavljena u radu ”Attention is All You Need” od strane Vaswanija i suradnika (2017)[9], postala je temelj za većinu modernih LLM-ova. Ova arhitektura koristi mehanizam pozornosti (attention mechanism) koji omogućuje modelu da se fokusira na relevantne dijelove ulaznog teksta dok generira izlaz.

Jedan od ključnih aspekata LLM-ova je njihova sposobnost generiranja tekstova koji su često teško razlikovati od onih koje su napisali ljudi. Ovo postignuće je rezultat sofisticiranog treniranja i fine-tuninga na specifičnim zadacima, što omogućuje modelima da razumiju i repliciraju složene jezične obrasce.

Uz generiranje teksta, LLM-ovi su također vrlo efikasni u raznim drugim zadacima NLP-a, kao što su prevođenje jezika, sažimanje dokumenata, odgovaranje na pitanja, analiza sentimenta i prepoznavanje imenovanih entiteta. Ovi zadaci zahtijevaju duboko razumijevanje jezika i konteksta, što je upravo ono što LLM-ovi pružaju.

U zaključku, veliki jezični modeli predstavljaju značajan napredak u obradi prirodnog jezika, omogućujući raznovrsne primjene koje su prije bile nezamislive. Njihova sposobnost učenja iz ogromnih količina podataka i generiranja koherentnog teksta čini ih nezamjenjivim alatima u modernom NLP-u.

2.1.2 Povijest i Razvoj LLM-ova

Razvoj velikih jezičnih modela (LLM) prošao je kroz nekoliko ključnih faza, počevši od jednostavnijih modela koji su se fokusirali na reprezentaciju riječi, do složenih modela sposobnih za generiranje koherentnih i smislenih tekstova.

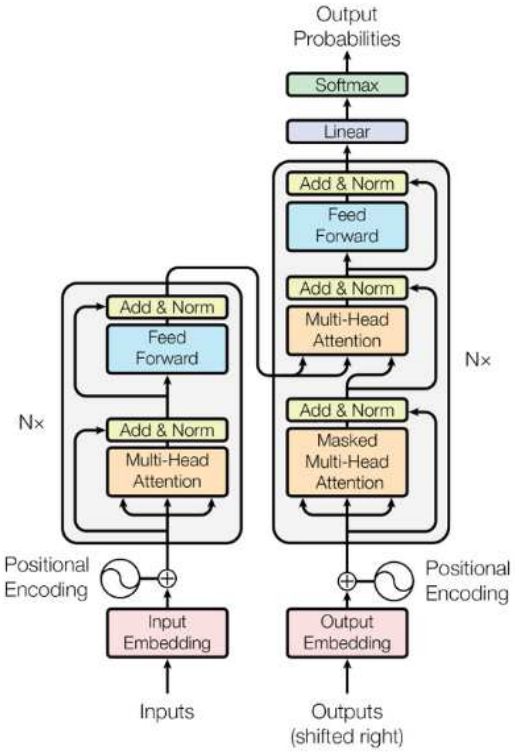
Početni razvoj: Word2Vec i GloVe

Jedan od prvih značajnih koraka u području obrade prirodnog jezika dogodio se s razvojem Word2Vec modela. Word2Vec koristi dvostruki pristup, model kontinualne vreće riječi (CBOW) i model preskakanja grama (Skip-gram), za učenje vektorskih reprezentacija riječi. Ove reprezentacije omogućuju modelima da uhvate semantičke sličnosti između riječi.

GloVe (Global Vectors for Word Representation) je još jedan značajan model za učenje vektorskih reprezentacija riječi, razvijen 2014. godine. GloVe koristi globalnu matricu supojavljivanja riječi za generiranje ugrađivanja riječi, što omogućuje bolju reprezentaciju riječi u različitim kontekstima.

Revolucija transformera

Transformerska arhitektura bila je revolucionarni korak u razvoju LLM-ova. Transformeri su zamijenili tradicionalne rekurentne neuronske mreže (RNN) i LSTM u mnogim NLP zadacima zbog svoje sposobnosti paralelnog procesiranja podataka i efikasnog učenja dugoročnih ovisnosti u jeziku. Ključna komponenta transformera je mehanizam pozornosti (attention mechanism).



Slika 1: Prikaz arhitekture transformerskog modela[9]

Napredni LLM-ovi: Od BERT-a do GPT-4

BERT (Bidirectional Encoder Representations from Transformers), razvijen od strane Googlea 2018. godine, koristi dvosmjerni pristup kodiranju, što mu omogućuje razumijevanje konteksta riječi iz oba smjera tijekom treniranja. Ovo svojstvo omogućuje BERT-u postizanje vrhunskih rezultata u raznim NLP zadacima.

GPT-2 i GPT-3, razvijeni od strane OpenAI, predstavili su nove standarde u generiranju koherentnih tekstova. GPT-3, sa svojih 175 milijardi parametara, pokazao je izuzetne sposobnosti u razumijevanju i generiranju prirodnog jezika.

Najnoviji napredak u LLM-ovima uključuje modele poput GPT-4 i Mistral, koji dodatno povećavaju broj parametara i poboljšavaju sposobnosti modela. GPT-4 koristi napredne tehnike optimizacije za postizanje još boljih performansi, dok Mistral predstavlja specijalizirani model dizajniran za visoku učinkovitost u specifičnim zadacima.[8]

Napredak u prilagodbi modela: LoRA tehnike

Razvoj tehnika za adaptaciju modela, poput LoRA (Low-Rank Adaptation), omogućio je efikasno fino podešavanje velikih modela uz smanjenje broja potrebnih parametara. LoRA tehnike koriste prilagodbu samo određenih slojeva modela, čime se značajno smanjuje vrijeme treninga i potrebni resursi.

Budući razvoj LLM-ova usmjeren je na povećanje efikasnosti i smanjenje računalnih zahtjeva, uz zadržavanje visoke razine performansi. Tehnike poput LoRA igraju ključnu ulogu u ovom procesu, omogućujući prilagodbu modela za specifične zadatke s manjim resursima.

U zaključku, povijest i razvoj LLM-ova pokazuje kontinuirani napredak u razumijevanju i generiranju prirodnog jezika. Od ranih modela poput Word2Vec i GloVe do najnovijih transformera poput BERT-a, GPT-3, GPT-4 i Mistral, ovi modeli su transformirali način na koji komuniciramo s računalima i otvorili nove mogućnosti u različitim industrijama.

2.1.3 Primjene LLM-ova u Različitim Područjima

Veliki jezični modeli (LLM) imaju široku primjenu u različitim područjima zahvaljujući svojoj sposobnosti razumijevanja i generiranja prirodnog jezika. Ovdje ćemo obraditi neke od najvažnijih primjena koje su značajno unaprijedile razne industrije i znanstvena istraživanja.[5]

Prevođenje jezika

Jedna od prvih i najvažnijih primjena LLM-ova je automatsko prevođenje jezika. Modeli poput Google's Neural Machine Translation (GNMT) koriste duboke neuronske mreže za prevođenje teksta s jednog jezika na drugi. Ovi modeli koriste LLM-ove kako bi razumjeli kontekst rečenica i pružili precizne prijevode, često bolji od tradicionalnih metoda temeljenih na pravilima.

LLM-ovi su vrlo učinkoviti u zadacima sažimanja teksta, gdje je potrebno izvući ključne informacije iz duljih dokumenata. Modeli poput BERT-a i GPT-3 mogu generirati sažetke koji zadržavaju bitne informacije dok uklanjaju suvišne dijelove teksta. Ovo je posebno korisno za novinske članke, znanstvene radove i pravne dokumente.

Chatbotovi i virtualni asistenti

LLM-ovi se naširoko koriste u razvoju chatbotova i virtualnih asistenata. Primjeri uključuju asistente poput Amazone Alexe, Appleove Siri i Googleovog Asistenta. Ovi sustavi koriste LLM-ove za razumijevanje korisničkih upita i generiranje odgovora koji su koherentni i korisni. Ovaj oblik primjene omogućuje automatizaciju korisničke podrške i poboljšava korisničko iskustvo.

Analiza sentimenta

Analiza sentimenta je još jedna važna primjena LLM-ova, posebno u marketingu i istraživanju tržišta. Korištenjem LLM-ova, tvrtke mogu analizirati stavove i emocije izražene u recenzijama proizvoda, komentarima na društvenim mrežama i drugim tekstualnim izvorima. Ovo pomaže tvrtkama da razumiju percepciju svojih proizvoda i usluga te donesu informirane poslovne odluke.

Odgovaranje na pitanja

LLM-ovi su izuzetno učinkoviti u zadacima odgovaranja na pitanja, gdje se model trenira da pretražuje velike količine podataka i pronalazi relevantne odgovore na postavljena pitanja. Ova tehnologija se koristi u obrazovnim alatima, zdravstvenim aplikacijama i raznim drugim sustavima gdje je brz i precizan pristup informacijama od ključne važnosti.

Prepoznavanje imenovanih entiteta

LLM-ovi se koriste za prepoznavanje imenovanih entiteta (NER), što uključuje identifikaciju i klasifikaciju ključnih elemenata u tekstu, kao što su imena osoba, organizacija, mjesta i drugih entiteta. Ova primjena je korisna u raznim područjima, uključujući novinarstvo, pravne analize i istraživačke radove.

Generiranje kreativnih sadržaja

LLM-ovi se također koriste za generiranje kreativnih sadržaja kao što su pisanje pjesama, priča, scenarija i drugih oblika pisane umjetnosti. Ovi modeli mogu stvarati originalne tekstove koji su koherentni i estetski privlačni, otvarajući nove mogućnosti u umjetnosti i zabavi.

U zaključku, veliki jezični modeli imaju široku primjenu u različitim industrijama i znanstvenim disciplinama. Njihova sposobnost razumijevanja i generiranja prirodnog jezika omogućuje unapređenje mnogih procesa, povećanje efikasnosti i poboljšanje korisničkog iskustva u raznim kontekstima.

2.2 Uvod u Predikciju Događaja

Veliki jezični modeli (LLM-ovi) pokazali su izvanredne performanse u širokom spektru zadataka rezoniranja. U ovom radu istražujemo njihove sposobnosti u rezoniranju o stvarnim događajima i poboljšanju performansi modela sekvenci događaja.

Prompting jezičnih modela nedavno je postalo standardni pristup za obavljanje zadataka baziranih na tekstualnom rezoniranju. Konkretno, fokusiramo se na problem modeliranja sekvenci vremenski označenih događaja i predviđanja budućih događaja na temelju prošlih podataka. Na primjer, u ekonomskom području želimo modelirati sekvence vremenski označenih velikih financijskih događaja i predvidjeti njihove buduće promjene radi investicijskih prilika. Veliki jezični modeli mogu biti korisni

za unapređenje rješenja ovog problema jer su sekvence događaja često praćene bogatim tekstualnim informacijama koje jezični modeli izvrsno obrađuju[6].

- **Ekonomija**: Ekonomski događaji, poput promjena na financijskim tržištima ili poslovnih inovacija, često su popraćeni vijestima i analizama. Čitanjem takvih tekstualnih informacija, veliki jezični model može se prisjetiti ekonomskog znanja koje je stekao tijekom predtreniranja i zatim rezonirati o budućim ekonomskim trendovima i prilikama za ulaganje.

- **Politika**: Svaki politički događaj može generirati niz članaka koji opisuju političke agente uključene u događaj i raspravljaju o mogućim utjecajima. Jezični model koji čita te članke može se prisjetiti svog znanja stečenog tijekom predtreniranja o tim agentima, njihovim odnosima i osnovnim principima u politici, što mu omogućuje rezoniranje o budućim političkim događajima.

- **Zdravstvo**: Svaki bolnički posjet ima liječničku bilješku koja sažima taj posjet, uključujući odjel koji je pacijent posjetio, klinička mjerenja i tretmane te buduće medicinske planove. Čitanjem takvih tekstualnih informacija, veliki jezični model može se prisjetiti medicinskog znanja koje je stekao tijekom predtreniranja i zatim rezonirati o budućim bolničkim posjetima, kao što su simptomi ili tretmani koje pacijent može imati.

- **Slični scenariji**: Slični scenariji javljaju se u komercijalnim, dijaloškim, financijskim i drugim kontekstima.

U ovom radu bavimo se primjenom velikih jezičnih modela za predikciju događaja s posebnim naglaskom na prepoznavanje investicijskih prilika iz novinskih članaka. Korištenjem tehnika finog podešavanja, kao što je LoRA, istražujemo kako ovi modeli mogu poboljšati točnost i efikasnost predikcija u stvarnom vremenu.

U arhitekturama rezidualnih mreža, kao što su transformeri, postoji slična ideja da želimo naučiti funkciju $h(X)$ od ulaznog X, ali može biti lakše naučiti rezidual $h(X)$ do identitetske mape, tj. $h(X) - X$. Ovdje, ako je funkcija $h(X)$ blizu identiteta, njezin rezidual će biti blizu nule, i stoga će biti manje za naučiti i učenje može biti učinkovitije. U oba slučaja cilj je iskoristiti strukturu kako bi se poboljšale predikcije: u slučaju financija, ideja je fokusirati se na predviđanje inovacija izvan onoga što implicira ukupno tržište, dok je za rezidualne mreže fokus na predviđanju inovacija prema identitetskoj mapi [6].

Ključni sastojak za impresivne performanse LLM-ova je njihova sposobnost razlučivanja afiniteta ili snaga između tokena preko dugih horizonata poznatih kao kontekstualni prozori. Na financijskim tržištima, sposobnost fokusiranja pozornosti kroz duge horizonte omogućuje analizu višestrukih fenomena, pri čemu se neki aspekti tržišnih promjena objašnjavaju kroz vrlo različite vremenske horizonte. Na primjer, u jednom ekstremu, fundamentalne informacije (npr. zarada) mogu biti uključene u cijene tijekom mjeseci, tehnički fenomeni (npr. moment) mogu se realizirati tijekom dana, dok, u drugom ekstremu, fenomeni mikrostrukture (npr. neravnoteža knjige narudžbi) mogu imati vremenski horizont od sekundi do minuta.

2.3 Tehnike Adaptacije Niskog Ranga (LoRA)

Tehnike adaptacije niskog ranga (Low-Rank Adaptation, LoRA) koriste se za fino podešavanje velikih jezičnih modela (LLM) kako bi se poboljšale njihove performanse u specifičnim zadacima.[7] Ove tehnike omogućuju prilagodbu modela uz smanjenje potrebnih računalnih resursa, što ih čini izuzetno korisnima u kontekstu primjene u industriji i istraživanju.

2.3.1 Osnove i Principi Rada LoRA Tehnika

Osnovni cilj LoRA tehnika je smanjenje broja parametara koji se trebaju prilagoditi tijekom procesa fino podešavanja velikih modela. Tradicionalne metode fino podešavanja često zahtijevaju značajne resurse jer uključuju ažuriranje svih parametara modela. LoRA tehnike, s druge strane, koriste niskorangirane matrice kako bi omogućile učinkovito podešavanje samo ključnih komponenti modela.[3]

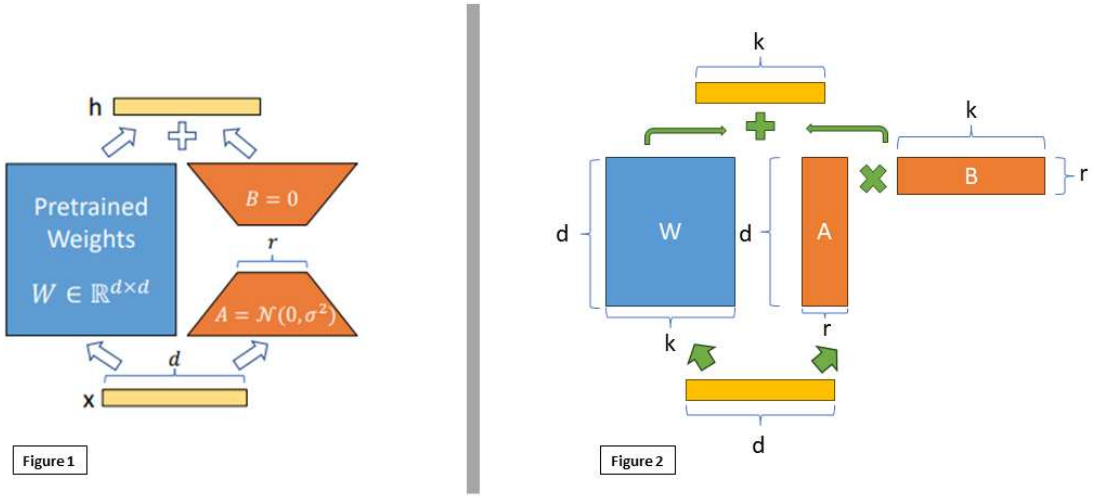
Matematička osnova LoRA tehnika

LoRA tehnike koriste dekompoziciju matrica kako bi reducirale složenost podešavanja. Konkretno, prilagodba se vrši putem niskorangiranih aproksimacija težinskih matrica modela. Neka W bude težinska matrica modela. Umjesto izravnog ažuriranja W, LoRA tehnika razlaže W na dva niskorangirana faktora A i B tako da je $W \approx A \cdot B$. Ovdje su A i B matrice niskog ranga koje imaju znatno manji broj parametara u usporedbi s originalnom matricom W.

$$W' = W + \Delta W \quad gdjeje \quad \Delta W = A \cdot B \tag{1}$$

Ova dekompozicija omogućuje učinkovitije prilagođavanje modela, jer je broj parametara koje treba optimizirati značajno manji. Osim toga, proces treniranja postaje brži i manje zahtjevan, jer se

ažuriraju samo niskorangirane komponente.



Slika 2: Shematski prikaz LoRA tehnike[1]

Primjena LoRA tehnika u različitim kontekstima

LoRA tehnike se mogu primijeniti na različite slojeve unutar neuronskih mreža, uključujući potpuno povezane slojeve, konvolucijske slojeve i slojeve pozornosti u transformerima. Na primjer, u kontekstu transformera, LoRA se može koristiti za optimizaciju slojeva pozornosti, omogućujući modelu da se učinkovitije prilagodi specifičnim zadacima obrade prirodnog jezika.

Primjenom LoRA tehnika, istraživači i inženjeri mogu postići visoku razinu prilagodbe i optimizacije modela bez potrebe za opsežnim resursima, što ih čini idealnim za primjenu u stvarnim sustavima i aplikacijama. U sljedećoj sekciji istražujemo razlike između osnovnih LoRA tehnika i naprednijih verzija poznatih kao LoRA+.

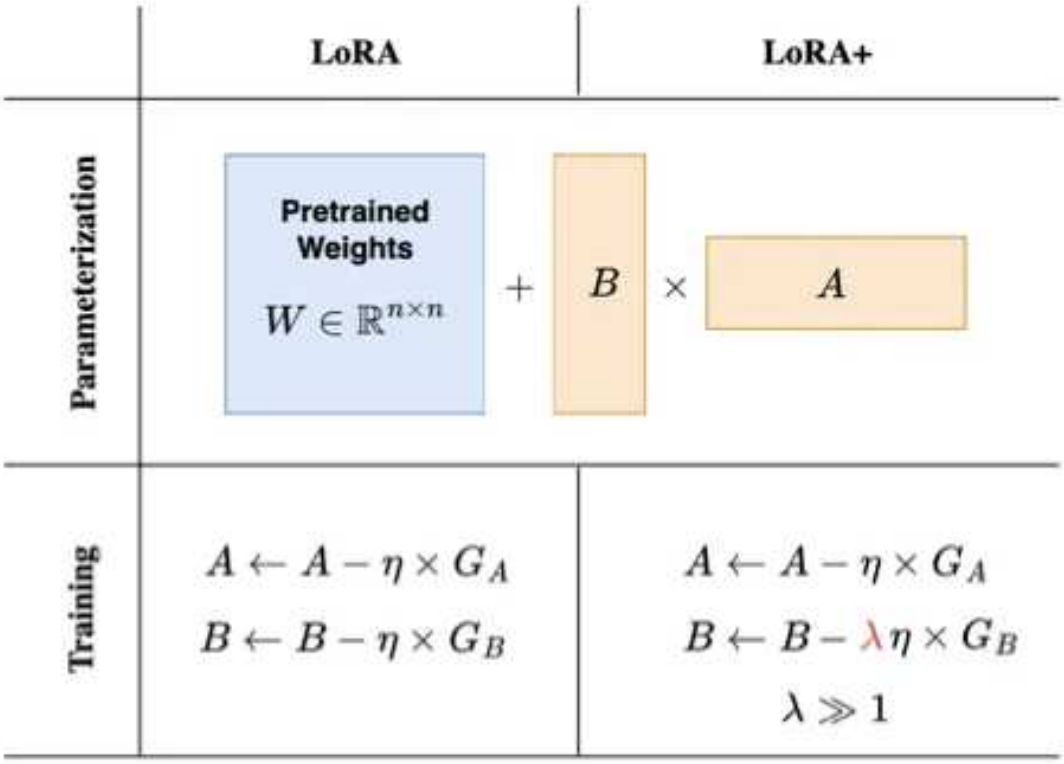
2.3.2 Razlike između LoRA i LoRA+

LoRA+ tehnike predstavljaju unaprijeđenu verziju osnovnih LoRA tehnika, omogućujući još bolju prilagodbu modela i postizanje viših performansi u specifičnim zadacima. Osnovna razlika između LoRA i LoRA+ leži u dodatnim optimizacijama koje se primjenjuju na niskorangirane faktore tijekom procesa treniranja.

Dodatne optimizacije u LoRA+

LoRA+ uključuje dodatne korake kao što su dinamička prilagodba ranga matrica, adaptivno učenje stopa i sofisticiranije metode regularizacije. Ove optimizacije omogućuju modelu da učinkovitije nauči relevantne značajke iz podataka bez prekomjerne prilagodbe, što je čest problem u osnovnim LoRA tehnikama.

$$W' = W + \Delta W \quad gdjeje \quad \Delta W = A \cdot B + \gamma \tag{2}$$



Slika 3: Usporedba između LoRA i LoRA+[4]

U ovom slučaju, γ predstavlja dodatni faktor regularizacije koji pomaže u kontroli prekomjerne prilagodbe i poboljšava generalizaciju modela. Dinamička prilagodba ranga omogućuje fleksibilnije podešavanje modela tijekom treninga, što rezultira boljim performansama u specifičnim zadacima.

Jedna od ključnih optimizacija u LoRA+ tehnici je uvođenje različitih stopa učenja za matrice A i B. Obično, kod treniranja neuronskih mreža, koristi se jedna stopa učenja koja se jednako primjenjuje na sve težinske matrice. Međutim, kod adaptera korištenih u LoRA tehnici, pokazalo se da je suboptimalno koristiti jedinstvenu stopu učenja. Trening postaje učinkovitiji postavljanjem stope učenja za matricu B znatno višom od one za matricu A.

Postoji teorijsko opravdanje za ovaj pristup, koje se temelji na numeričkim ograničenjima inicijalizacije neuronskih mreža ako model postane vrlo širok u smislu broja neurona. Intuitivno, možete zamisliti da matrica B, koja je inicijalizirana s nulama, može koristiti veće korake ažuriranja od nasumično inicijalizirane matrice A. Osim toga, empirijski dokazi također pokazuju poboljšanje ovim pristupom. Postavljanjem stope učenja matrice B 16 puta većom od one za matricu A, trening postaje učinkovitiji i rezultira boljim performansama modela.

Primjenom LoRA+ tehnika, modeli mogu postići bolje rezultate u različitim NLP zadacima, uključujući prepoznavanje entiteta, sažimanje teksta i prevođenje jezika. Dinamička prilagodba stope učenja omogućuje finiju kontrolu nad procesom treniranja, što vodi ka učinkovitijem učenju i bolje optimiziranim modelima.

2.3.3 Prednosti Korištenja LoRA za Fino Podešavanje Modela

Korištenje LoRA tehnika za fino podešavanje velikih jezičnih modela donosi brojne prednosti koje su ključne za primjenu u stvarnim sustavima i industriji.

1. Smanjenje računalnih resursa

Jedna od najznačajnijih prednosti LoRA tehnika je smanjenje broja parametara koje je potrebno prilagoditi tijekom treniranja. Ovo rezultira manjim zahtjevima za memoriju i računalne resurse, što je ključno za velike podatkovne centre i aplikacije u oblaku.

2. Brže fino podešavanje

Smanjenjem broja parametara, proces fino podešavanja postaje brži[4], što omogućuje bržu iteraciju i optimizaciju modela. To znači da se modeli mogu prilagoditi specifičnim zadacima u kraćem vremenu, što je posebno korisno u dinamičnim okruženjima gdje su brze prilagodbe ključne.

3. Manja potrošnja energije

Optimizacija modela uz manji broj parametara smanjuje ukupnu potrošnju energije tijekom treniranja, što je važno za održivost i ekonomičnost podatkovnih centara. Manja potrošnja energije također znači manji ekološki otisak, što je sve važnije u današnjem svijetu.

Korištenjem LoRA tehnika, istraživači i inženjeri mogu postići visoku razinu prilagodbe i optimizacije modela uz smanjenje potrebnih resursa i vremena treniranja.

2.3.4 Izazovi i Ograničenja LoRA Tehnika

Iako LoRA tehnike donose brojne prednosti, postoje izazovi i ograničenja u njihovoj primjeni koje je važno razmotriti.

1. Specifične prilagodbe

LoRA tehnike često zahtijevaju specifične prilagodbe za različite zadatke i tipove modela. To znači da inženjeri moraju pažljivo dizajnirati i testirati različite konfiguracije kako bi postigli optimalne performanse. Ovaj proces može biti dugotrajan i zahtijeva duboko razumijevanje modela i zadatka.

2. Ograničenja u generalizaciji

Iako LoRA tehnike pomažu u smanjenju prekomjerne prilagodbe, mogu se pojaviti problemi s generalizacijom kada se model koristi na potpuno novim domenama ili zadacima. To znači da model možda neće postići iste razine performansi kada se primijeni na različite tipove podataka od onih korištenih tijekom treniranja.

3. Složenost implementacije

Implementacija LoRA tehnika može biti složena, posebno za inženjere koji nisu upoznati s matematičkim osnovama i algoritmima dekompozicije matrica. To može ograničiti primjenu LoRA tehnika u manjim timovima ili organizacijama koje nemaju potrebne stručnjake.

Unatoč ovim izazovima, LoRA tehnike predstavljaju značajan korak naprijed u fino podešavanju velikih jezičnih modela, omogućujući učinkovitiju prilagodbu modela uz smanjenje potrebnih resursa i vremena treniranja. Kroz daljnja istraživanja i razvoj, moguće je prevladati ova ograničenja i unaprijediti primjenu LoRA tehnika u različitim kontekstima.

3 Predikcija Događaja Pomoću Jezičnih Modela

Predikcija budućih događaja ključna je za donošenje politika i strateških odluka u raznim područjima, uključujući ekonomiju, politiku i javno zdravstvo. U ovom radu istražujemo mogu li jezični modeli (LM-ovi) predviđati na razini kompetitivnih ljudskih prediktora. U tu svrhu razvili smo sustav jezičnih modela s poboljšanim dohvatom informacija dizajniran za automatsko pretraživanje relevantnih informacija, generiranje predikcija i agregiranje rezultata.

3.1 Uvod u Predikciju Događaja

Predikcija događaja uključuje predviđanje vjerojatnosti budućih događaja na temelju dostupnih podataka i stručne procjene. Tradicionalno, postoje dva glavna pristupa predikciji:

- **Statistička Predikcija:** Koristi alate vremenskih serija i druge statističke metode za predikciju budućih događaja na temelju povijesnih podataka. Ova metoda je učinkovita kada postoji velika količina podataka i minimalne distribucijske promjene.
- **Prosudbena Predikcija:** Oslanja se na ljudske prediktore koji koriste svoje znanje, intuiciju i širok raspon izvora informacija za dodjelu vjerojatnosti budućim događajima. Ovaj pristup je posebno vrijedan kada su podaci oskudni ili postoje značajne distribucijske promjene.

S obzirom na ograničenja ljudske predikcije, kao što su visoki troškovi i ograničena skalabilnost, sve veći interes postoji za automatizacijom procesa pomoću LM-ova.

3.2 Automatizirana Predikcija Događaja s Jezičnim Modelima

Nedavni napredak u obradi prirodnog jezika (NLP) omogućio je LM-ovima izvršavanje složenih zadataka, uključujući predikciju događaja. LM-ovi mogu brzo analizirati i generirati tekst, što ih čini pogodnima za generiranje pravovremenih predikcija. Dodatno, mogu koristiti ogromne količine prethodno obrađenog međudomenskog znanja za poboljšanje svojih predikcija.

3.2.1 Jezični Modeli s Poboljšanim Dohvatom Informacija

Kako bi se poboljšala točnost predikcija, istraživači su razvili LM sustave s poboljšanim dohvatom informacija. Ovi sustavi integriraju tri ključne komponente tradicionalne predikcije:

- **Dohvat:** Automatski pretražuje relevantne informacije iz izvora vijesti i drugih baza podataka.

- **Razmišljanje:** Analizira dohvaćene informacije i generira predikcije.
- **Agregacija:** Kombinira više predikcija kako bi proizveo agregiranu predikciju.

3.2.2 Eksperimentalni Postav i Rezultati

Recentni napredak u korištenju jezičnih modela (LM-ova) omogućio je njihovu primjenu u predikciji događaja, često približavajući njihove performanse onima ljudskih prediktora. Sustavi temeljeni na LM-ovima dizajnirani su za predviđanje binarnih ishoda i fino su podešeni korištenjem samonadziranog pristupa. Eksperimenti se obično fokusiraju na tri glavna aspekta:

- **Generiranje Upita za Pretraživanje:** LM generira upite za pretraživanje relevantnih članaka.
- **Sažimanje Članaka:** LM sažima najrelevantnije članke.
- **Generiranje Predikcija:** LM generira predikcije na temelju sažetih informacija.

Sustavi temeljeni na LM-ovima često nadmašuju ljudske prediktore u mnogim scenarijima, posebno u onima gdje ljudski prediktori pokazuju visoku nesigurnost. Ovo je bilo posebno evidentno u scenarijima s visokim stupnjem složenosti i promjenjivosti.

3.2.3 Primjena na Predikciju Investicijskih Signala

Primjena jezičnih modela na predikciju budućih investicijskih prilika koristi tehnike slične onima opisanim gore. Sustav LM-a s poboljšanim dohvatom informacija koristi se za identifikaciju relevantnih članaka, analizu njihovog sadržaja i generiranje investicijskih signala. Ovaj pristup omogućuje prepoznavanje potencijalnih prilika za ulaganje kroz analizu tekstualnih informacija iz različitih izvora.

4 Eksperimentalni Rad

U ovom poglavlju detaljno opisujemo eksperimentalni rad proveden tijekom istraživanja, uključujući prikupljanje i obradu podataka, te trening modela.

4.1 Prikupljanje i Izrada Dataseta

Prikupljanje podataka za trening modela bio je ključni korak u ovom istraživanju. Podaci su prikupljeni iz različitih izvora, uključujući BBC, The Guardian, TechCrunch, TechRadar i Wccftech, uz manji broj članaka iz Reutersa. Ukupno je prikupljeno nešto više od 35,000 članaka koji su zadovoljili kriterije za uključivanje u dataset.

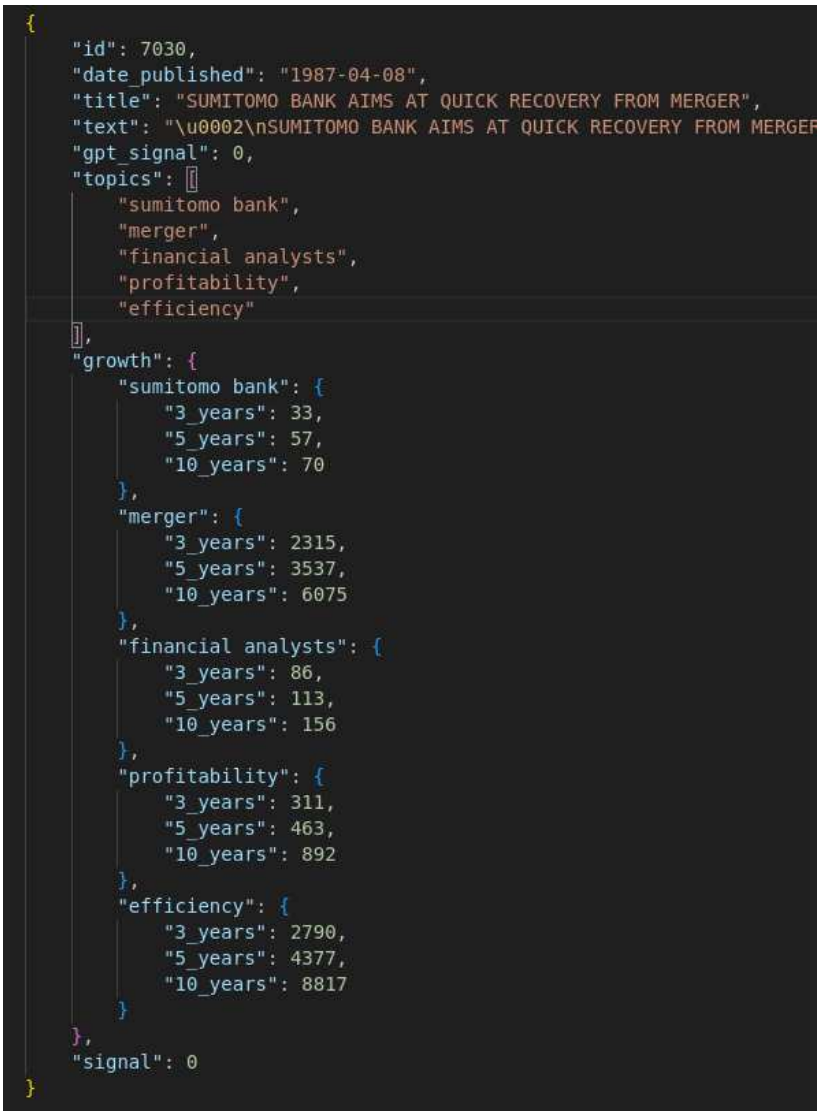
Dataset se sastojao od sljedećih polja:

- **ID:** Jedinstveni identifikator članka.
- **Date Published:** Datum objave članka, ključan za praćenje vremenskog okvira interesa.
- **Title:** Naslov članka.
- **Text:** Sadržaj članka.
- **GPT Signal:** Ocjena GPT-3.5 modela o tome treba li članak biti signal za investicijske prilike.
- **Topics:** Popis tema obrađenih u članku.
- **Growth:** Interes za teme kroz tri vremenska razdoblja (3 godine, 5 godina, 10 godina) na temelju Google Trends podataka.
- **Signal:** Kombinacija GPT signala i rasta interesa za teme.

Proces prikupljanja i izrade dataseta uključivao je nekoliko koraka:

- Filtriranje i čišćenje podataka: Uklanjanje nepotpunih ili nerelevantnih članaka. Proces filtriranja podataka uključivao je procjenu relevantnosti članaka na temelju njihovih embeddinga. Koristili smo tehnike učenja temeljene na embeddinzima kako bismo odredili sličnost između članka i skupa trening podataka. Ako je embedding članka pokazivao visoku sličnost s postojećim relevantnim člancima, taj je članak uključen u dataset. Na taj način smo osigurali da samo relevantne vijesti, koje su značajne za trening modela, budu uključene u konačni dataset.

- Normalizacija i formatiranje podataka: Osiguravanje kompatibilnosti podataka s modelom.
- Prikupljanje podataka iz Google Trends-a: Korištenje Python skripte bazirane na *pytrends* library-ju za dobivanje podataka o trendovima interesa.



Slika 4: Prikaz strukture obrađenoga dataset-a

S obzirom na velike zahtjeve za broj zahtjeva prema Google Trends API-ju, proces prikupljanja podataka predstavljao je značajan izazov. Skripta je kontinuirano slala zahtjeve i spremala rezultate u dataset.

4.2 Trening Modela

Trening modela proveden je korištenjem tehnika LoRA i LoRA+ kako bi se usporedile njihove performanse. Trening je obavljen korištenjem PyTorch biblioteke i Hugging Face Transformers frameworka, uz praćenje napretka i evaluacije putem Weights Biases alata.

Korištene su sljedeće postavke za trening:

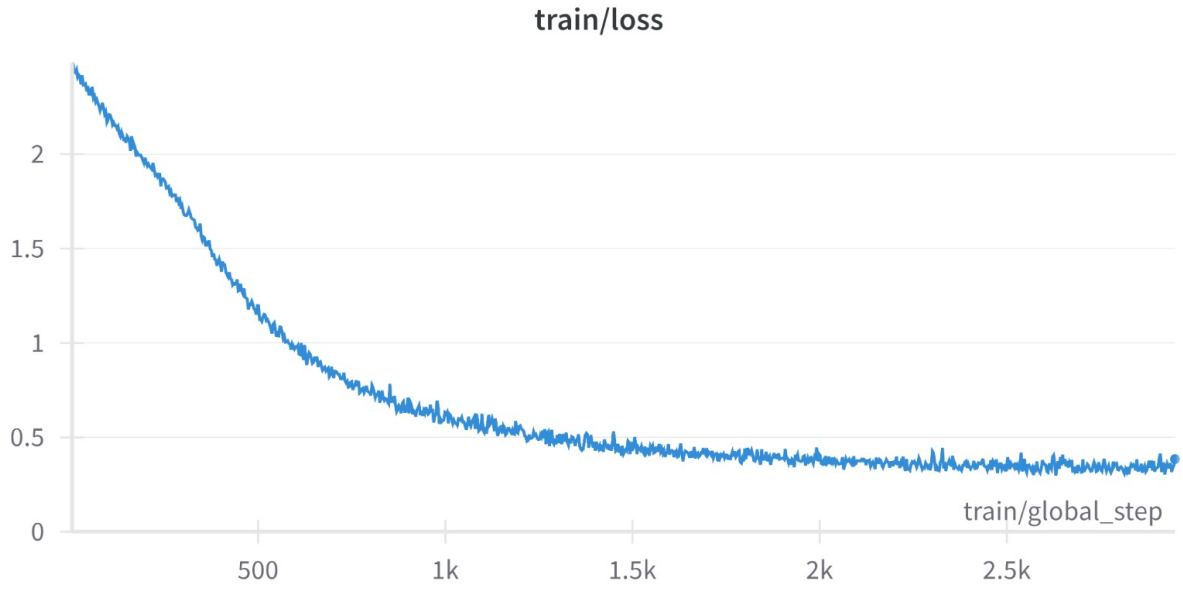
- **Learning Rate:** 0.001 - Brzina učenja označava stopu kojom se parametri modela ažuriraju u smjeru suprotnom od gradijenta funkcije gubitka. Manja vrijednost learning rate-a omogućuje finiju prilagodbu modela, dok veća vrijednost ubrzava trening ali može uzrokovati oscilacije u procesu optimizacije.
- **Batch Size:** 5 - Broj primjera iz skupa podataka koji se obrađuju prije nego što model ažurira svoje parametre. Manji batch size omogućuje češća ažuriranja, ali može biti manje stabilan, dok veći batch size nudi stabilnije ali rjeđe ažuriranje parametara.
- **Broj Epoha:** 1 - Broj prolaza kroz cijeli skup podataka za trening. Veći broj epoha omogućuje modelu da bolje nauči uzorke iz podataka, ali može dovesti do prekomjerne prilagodbe (overfitting) ako je previše epoha.
- **LoRA Parametri:**
  - **r:** 16 - Rang niskorangirane aproksimacije težinske matrice.
  - **lora_alpha:** 32 - Multiplikativni faktor za r.
  - **lora_dropout:** 0.1 - Postotak neurona koji se isključuju tijekom svakog koraka treninga kako bi se spriječilo prekomjerno prilagođavanje modela.

Eksperimentirali smo s različitim parametrima treninga kako bismo identificirali optimalne postavke. Najbolje/Najoptimalnije rezultate postigli smo s learning rate-om od 0.001, batch size-om od 5 i brojem epoha od 1, većim brojem epoha nismo dobivali znatno bolje rezultate.

Trening je obavljen u iterativnim ciklusima, pri čemu su se modeli periodično evaluirali kako bi se pratile promjene u performansama. Eksperimentiranjem s različitim konfiguracijama i parametrima uspjeli smo značajno smanjiti ”loss” modela, što je vidljivo na grafikonima performansi.

Eksperimentirali smo s LoRA i LoRA+ tehnikama. Rezultati loss funkcije bili su slični za obje tehnike, ali LoRA+ tehnika značajno je pridonijela brzini treninga, smanjujući vrijeme potrebno za trening za faktor od 1.4 (3500/2500) u usporedbi s LoRA tehnikom koristeći iste računalne resurse.

Na grafikonu za LoRA tehniku (slika 3) vidljivo je postupno smanjenje vrijednosti loss funkcije tijekom treninga, pri čemu je model dosegao stabilnu vrijednost nakon približno 2500 globalnih koraka.



Slika 5: Prikaz loss-a za LoRA tehniku

Na grafikonu za LoRA+ tehniku (slika 4) primjećujemo inicijalnu nestabilnost u vrijednostima loss funkcije tijekom prvih 500 globalnih koraka, nakon čega slijedi stabilizacija i značajno smanjenje vrijednosti loss funkcije. Ova nestabilnost može biti posljedica agresivnijeg prilagođavanja parametara koje omogućava brže konvergiranje modela. Konačno, LoRA+ tehnika omogućila je postizanje stabilne vrijednosti loss funkcije brže nego LoRA tehnika.

Slika 6: Prikaz loss-a za LoRA+ tehniku za isto vrijeme kao na prikazu loss-a za LoRA tehniku

Ovi rezultati ukazuju na to da, iako LoRA+ tehnika nije nužno poboljšala krajnju vrijednost loss funkcije u usporedbi s LoRA tehnikom, značajno je smanjila vrijeme potrebno za trening, što je ključno za primjene u stvarnim sustavima gdje je vrijeme treninga kritično.

5 Evaluacija Performansi

Evaluacija performansi modela ključan je dio istraživanja. U ovom poglavlju koristimo matricu konfuzije kako bismo detaljno analizirali uspješnost modela u prepoznavanju signala. Matrica konfuzije omogućuje vizualizaciju performansi klasifikacijskog modela prikazivanjem broja ispravnih i pogrešnih predikcija za svaku klasu.

5.1 Matrica Konfuzije

Matrica konfuzije je tablica koja se koristi za procjenu performansi algoritma za klasifikaciju. Sastoji se od četiri glavne komponente:

- **TP (True Positive):** Ispravno predviđeni pozitivni primjeri.
- **FP (False Positive):** Pogrešno predviđeni pozitivni primjeri.
- **TN (True Negative):** Ispravno predviđeni negativni primjeri.
- **FN (False Negative):** Pogrešno predviđeni negativni primjeri.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Tablica 1: Struktura matrice konfuzije

5.2 Evaluacija Početnog Modela

Početni model testiran je na skupu podataka koji se sastoji od 8452 članka, od kojih je 1762 bilo signal (21%), a 6690 nije signal (79%). Rezultati matrice konfuzije za početni model prikazani su u tablici 2.

	Predicted Signal	**Predicted No Signal**
Actual Signal	1521	241
Actual No Signal	1640	5050

Tablica 2: Matrica konfuzije za početni model

Početni model je dao 1521 ispravan pozitivan signal, 1640 pogrešnih pozitivnih signala, 5050 ispravnih negativnih signala i 241 pogrešnih negativnih signala. Vidljivo je da je početni model imao tendenciju precijenjivanja pozitivnih signala, što je rezultiralo visokim brojem lažno pozitivnih predikcija.

5.3 Evaluacija Istreniranog Modela

Nakon uspješnog treniranja, model s najboljim loss-om testiran je na istom skupu podataka. Rezultati matrice konfuzije za istrenirani model prikazani su u tablici 3.

	Predicted Signal	**Predicted No Signal**
Actual Signal	1633	279
Actual No Signal	1521	5119

Tablica 3: Matrica konfuzije za istrenirani model

Istrenirani model je dao 1633 ispravnih pozitivnih signala, 1521 pogrešnih pozitivnih signala, 5119 ispravnih negativnih signala i 279 pogrešnih negativnih signala. Usporedbom s početnim modelom, istrenirani model je:

- Smanjio broj lažno pozitivnih predikcija za 119 (1640 na 1521).
- Povećao broj ispravnih pozitivnih predikcija za 112 (1521 na 1633).
- Povećao broj ispravnih negativnih predikcija za 69 (5050 na 5119).
- Povećao broj pogrešnih negativnih predikcija za 38 (241 na 279).

Iako je došlo do povećanja broja pogrešnih negativnih predikcija, značajno smanjenje lažno pozitivnih predikcija i povećanje ispravnih pozitivnih predikcija pokazuje značajno poboljšanje modela u prepoznavanju relevantnih signala, čime je smanjena tendencija ka pogrešnom prepoznavanju signala.

```
You are an assistant for signal search and keyword extraction. Given the provided article (in JSON
format with multiple fields), you need to determine if it is a signal. An article is considered a
signal if it presents significant news for its time and has subsequently demonstrated lasting impact,
successful implementation, or recognized advancement in the respective field. Signals are defined as
unusual or exceptional instances that deviate from the norm or expected patterns. Your task is to
analyze the provided text and respond with ONLY a "1" if it exhibits positive anomalies meeting the
specified criteria, or ONLY a "0" if it does not. If you encounter the same or a very similar article
again, you should give an IDENTICAL result as you gave when you first saw the article. The
information should be groundbreaking for its time period and have shown sustained impact or
advancement beyond the initial publication. Ensure that the news is of crucial matter and impactful
to the entire sector/industry or the world.
```

Slika 7: Prikaz prompt-a korištenoga za testiranje modela i obrađivanje polja gpt-signal u datase-u

Ove rezultate smo dobili korištenjem Python skripte koja je koristila početni prompt (slika 7) za dobivanje *gpt_signal* polja i ispitivala oba modela, početni GPT-2 model i istrenirani GPT-2 model, kako bi ocijenila jesu li vijesti signal ili ne.

Ovi rezultati potvrđuju da je primjena LoRA i LoRA+ tehnika uspješno poboljšala performanse modela u prepoznavanju investicijskih prilika, posebno smanjujući broj lažno pozitivnih signala, što je bila glavna slabost početnog modela.

6 Zaključak

U ovom radu istražili smo primjenu tehnika adaptacije niskog ranga (LoRA) za optimizaciju velikih jezičnih modela (LLM) s ciljem prepoznavanja investicijskih prilika. Kroz nekoliko ključnih faza, od odabira modela, prikupljanja i obrade podataka, do treninga i evaluacije performansi, analizirali smo učinkovitost LoRA tehnika u prilagodbi modela za specifične zadatke uz ograničene računalne resurse.

Ključni nalazi uključuju:

Odabir Modela: Kao početni model odabrali smo GPT-2 zbog njegove široke dostupnosti i razumnih zahtjeva za računalne resurse. GPT-2 je pokazao solidne performanse u raznim zadacima obrade prirodnog jezika, čineći ga pogodnim za eksperimentiranje s tehnikama adaptacije kao što su LoRA i LoRA+.

Prikupljanje i Obrada Podataka: Prikupili smo značajnu količinu članaka iz različitih izvora kako bismo izradili dataset koji uključuje relevantna polja za analizu investicijskih prilika. Proces prikupljanja podataka predstavljao je izazov zbog velikog broja zahtjeva prema različitim API-jevima, a podaci su se prikupljali pomoću Python skripte.

Trening Modela: Eksperimentirali smo s različitim parametrima treninga kako bismo identificirali optimalne postavke. Parametri su uključivali različite vrijednosti za learning rate, batch size i broj epoha, kao i specifične postavke za LoRA tehniku. Najbolji rezultati postignuti su s određenim kombinacijama ovih parametara.

Evaluacija Performansi: Evaluacija performansi modela provedena je korištenjem matrice konfuzije. Početni model imao je tendenciju davanja lažno pozitivnih signala, što je rezultiralo visokim brojem pogrešno predviđenih pozitivnih primjera. Nakon primjene LoRA i LoRA+ tehnika, istrenirani model je značajno smanjio broj lažno pozitivnih predikcija i povećao broj ispravnih pozitivnih i negativnih predikcija, što pokazuje poboljšanje u prepoznavanju relevantnih signala.

Na temelju dobivenih rezultata, možemo zaključiti da su tehnike adaptacije niskog ranga (LoRA) učinkovite za prilagodbu velikih jezičnih modela u specifične zadatke, posebno kada su računalni resursi ograničeni. LoRA tehnike omogućuju smanjenje broja parametara potrebnih za fino podešavanje modela, čineći proces treniranja bržim i manje zahtjevnim. Primjena LoRA+ tehnike dodatno poboljšava učinkovitost treninga smanjujući vrijeme potrebno za optimizaciju modela bez značajnog gubitka performansi.

Praktične implikacije primjene LoRA tehnika uključuju povećanu efikasnost resursa, brže fino

podešavanje modela i poboljšanu točnost u prepoznavanju relevantnih signala. Organizacije s ograničenim računalnim resursima mogu koristiti ove tehnike za učinkovitu prilagodbu velikih jezičnih modela, omogućujući im postizanje visokih performansi bez potrebe za opsežnim hardverskim ulaganjima.

U zaključku, ovaj rad pokazuje da su tehnike adaptacije niskog ranga, kao što su LoRA i LoRA+, vrlo korisne za fino podešavanje velikih jezičnih modela u specifične zadatke, posebno kada su računalni resursi ograničeni. Poboljšanja u točnosti i brzini treninga čine ove tehnike ključnim alatima za buduća istraživanja i primjene u različitim industrijama i domenama.

7 Literatura

Literatura

[1] Hugging Face Documentation. *LoRA*. Dostupno na: https://huggingface.co/docs/peft/main/en/conceptual$_guide$

[2] ArXiv. *Paper on LoRA*. Dostupno na: https://arxiv.org/html/2402.18563v1.

[3] Databricks. *Efficient Fine-Tuning: A LoRA Guide for LLMs*. Dostupno na: https://www.databricks.com/blog/efficient-fine-tuning-lora-guide-llms.

[4] Towards Data Science. *An Overview of the LoRA Family*. Dostupno na: https://towardsdatascience.com/an-overview-of-the-lora-family-515d81134725.

[5] IBM. *Large Language Models*. Dostupno na: https://www.ibm.com/topics/large-language-models.

[6] The Gradient. *Financial Market Applications of LLMs*. Dostupno na: https://thegradient.pub/financial-market-applications-of-llms/.

[7] Prasun Mishra. *Transforming AI Landscape with LoRA's Technical LLM Training*. Dostupno na: https://www.linkedin.com/pulse/transforming-ai-landscape-loras-technical-llm-training-prasun-mishra/.

[8] Toloka Team *The history, timeline, and future of LLMs*. Dostupno na: https://toloka.ai/blog/history-of-llms/.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar... *Attention Is All You Need*. Dostupno na: https://proceedings.neurips.cc/paper$_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf$.

[10] Kaggle. *Links to All Articles from Big Tech News Sites*. Dostupno na: https://www.kaggle.com/datasets/itsayushk/links-to-all-articles-from-big-tech-news-sites.

Optimizacija velikih jezičnih modela korištenjem LoRA tehnike za identifikaciju investicijskih prilika

Sažetak

Raspoznavanje investicijskih prilika u financijskim tržištima zahtijeva napredne metode analize i obrade velikih količina podataka. U ovom radu istražit će se primjena tehnike niskorangirane adaptacije (engl. Low-Rank Adaptation, LoRA) za fino podešavanje velikih jezičnih modela (engl. Large Language Models, LLM) s ciljem poboljšanja procesa prepoznavanja investicijskih prilika. Korištenjem LoRA tehnike, model će biti treniran na podacima dobivenim iz financijskih e-biltena i vijesti, kako bi se identificirali relevantni događaji koji utječu na tržište dionica. Ova metoda omogućava učinkovitije i brže prilagođavanje modela novim informacijama, poboljšavajući njihovu točnost i performanse u stvarnom vremenu.

Ključne riječi: veliki jezični modeli, LoRA/LoRA+, investicijske prilike, financijska tržišta, fino podešavanje, adaptacija niskog ranga, prepoznavanje obrazaca

Optimization of Large Language Models Using LoRA Technique for Identifying Investment Opportunities

Abstract

Recognizing investment opportunities in financial markets requires advanced methods for analyzing and processing large amounts of data. This paper explores the application of Low-Rank Adaptation (LoRA) technique for fine-tuning large language models (LLMs) with the aim of improving the process of identifying investment opportunities. Using the LoRA technique, the model will be trained on data obtained from financial newsletters and news to identify relevant events that impact the stock market. This method enables more efficient and faster adaptation of the model to new information, enhancing their accuracy and performance in real-time scenarios.

Keywords: large language models, LoRA/LoRA+, investment opportunities, financial markets, fine-tuning, low-rank adaptation, pattern recognition