Računarski fakultet

DIPLOMSKI RAD

Slika je vredna 16x16 reči: Vision Transformeri

Vanja Kovinić
RN 42/2020

Mentor:
dr Nemanja Ilić

Komisija:
dr Nemanja Ilić
dr Nevena Marić

Beograd, septembar 2024.

Sadržaj

Apstrakt

Ovaj diplomski rad istražuje **Vision Transformere** (**ViT**), nov pristup u oblasti računarskog vida koji koristi arhitekturu transformera prvobitno razvijenu za obradu prirodnog jezika. Prvi deo rada pruža detaljan pregled arhitekture transformera, uključujući ključne komponente kao što su ***self-attention*** **mehanizam** i **poziciono enkodovanje**, i diskutuje njihove svrhe i funkcionalnosti. Nakon toga, fokus se prebacuje na Vision Transformere, objašnjavajući kako se slike transformišu u **tokene** i obrađuju kroz **enkoder transformera** kako bi se primenili na rešavanje vizuelnih zadataka.

Rad zatim ulazi u praktične aspekte implementacije Vision Transformera, uključujući izbor i podešavanje **hiperparametara** za poboljšanje performansi. Izvršeno je i poređenje sa referentnim implementacijama, i predložen pristup za poboljšanje performansi. Prikazani su različiti eksperimenti, zajedno sa diskusijom njihovih rezultata, pružajući uvid u efikasnost i izazove povezane sa Vision Transformerima.

Na kraju, rad naglašava značaj Vision Transformera u oblasti računarskog vida, prikazujući njihov potencijal i ograničenja, kao i njihove praktične primene.

1 Uvod

"Pre otprilike 540 miliona godina, Zemlja je bila obavijena tamom. Ovo nije bilo zbog nedostatka svetlosti, već zato što organizmi još uvek nisu razvili sposobnost da vide. Iako je sunčeva svetlost mogla da prodre u okeane do dubine od 1.000 metara i hidrotermalni izvori na dnu mora isijavali svetlost u kojoj je život cvetao, nijedno oko nije se moglo naći u tim drevnim okeanima, nijedna retina, rožnjača ili sočivo. Sva svetlost i život nikada nisu viđeni. Koncept gledanja nije ni postojao tada i ova sposobnost nije ostvarena sve dok nije stvorena.

Iz nama nepoznatih razloga, trilobiti su se pojavili kao prva bića sposobna da spoznaju svetlost. Oni su prvi prepoznali da postoji nešto izvan njih samih, svet okružen višestrukim jedinkama. Rađenje vida se smatra da je pokrenulo kambrijsku eksploziju, period u kojem se veliki broj vrsta životinja pojavljuje u fosilnom zapisu. Vid je započeo kao pasivno iskustvo, jednostavno propuštanje svetlosti, ali je ubrzo postao aktivniji. Nervni sistem je počeo da evoluira, vid je prešao u uvid, gledanje je postalo razumevanje, a razumevanje je dovelo do akcije, a sve to je dovelo do nastanka inteligencije.

Danas nismo više zadovoljni vizuelnom spoznajom koju nam je priroda dala. Radoznalost nas je navela da stvorimo mašine koje mogu da "vide" kao mi, pa čak i inteligentnije." - Li Fei-Fei [1]

1.1 Istorija i motivacija

1.1.1 Rani Razvoj u Računarskom Vidu

Koreni računarskog vida potiču iz ranih pokušaja da se razume i interpretira vizuelni podatak korišćenjem matematičkih modela i računara. U početku, istraživanja u oblasti računarskog vida fokusirala su se na jednostavne zadatke kao što su detekcija ivica, prepoznavanje objekata i osnovna obrada slika. Rane metode su se u velikoj meri oslanjale na ručnu izradu karakteristika (engl. ***features***) slike i algoritme dizajnirane

da imitiraju osnovne aspekte ljudskog vida.

1.1.2 Uspon Dubokog Učenja

Značajan preokret u računarskom vidu dogodio se sa pojavom **dubokog učenja**. **Konvolucione Neuronske Mreže** (**CNNs**), koje su predstavili Yann LeCun i drugi [4] krajem 1980-ih i početkom 1990-ih, revolucionisale su ovu oblast uvođenjem automatskog ekstraktovanja karakteristika kroz slojeve koji se uče (engl. ***learnable features***). **CNN**-ovi su pokazali izuzetne performanse u različitim zadacima klasifikacije slika, omogućavajući računarima da nauče složene reprezentacije vizuelnih podataka. Ovo otkriće je kasnije propaćeno uspehom modela kao što su **AlexNet** [3], **VGGNet** [5] i **ResNet** [2], koji su postavili nove standarde u izazovima prepoznavanja slika.

1.1.3 Ograničenja CNN-ova

Uprkos svom uspehu, **CNN**-ovi imaju inherentna ograničenja koja su motivisala potragu za novim pristupima. Jedan od značajnih nedostataka je njihova poteškoća u povezivanju udaljenih zavisnosti i globalnog konteksta unutar slike. **CNN**-ovi obično obrađuju slike kroz seriju lokalizovanih konvolucionih operacija, što može ograničiti njihovu sposobnost da razumeju odnose između udaljenih elemenata na slici.

1.1.4 Motivacija za uvođenje Vision Transformera

Pojava **Vision Transformera** (**ViT**) predstavlja odgovor na ova ograničenja. Inspirisani uspehom modela transformera u obradi prirodnog jezika (**NLP**), istraživači su pokušali da primene iste principe u računarskom vidu. Transformeri koriste ***self-attention*** mehanizam za povezivanje globalnih zavisnosti, što ih čini pogodnim za zadatke koji zahtevaju razumevanje složenih odnosa unutar vizuelnih podataka.

Vision Transformeri rešavaju nekoliko izazova sa kojima se suočavaju **CNN**-ovi. Pretvaranjem slika u sekvence parčića (engl. ***image patches***) i primenom ***self-attention*** mehanizma preuzetim iz **transformera**, ViTs mogu efikasnije modelovati globalni kontekst. Ovaj pristup omogućava **ViT**-ovima da postignu vrhunske performanse na različitim testovima klasifikacije slika i pokazuje njihov potencijal da unaprede oblast računarskog vida.

References

[1] Li Fei-Fei. *With spatial intelligence, AI will understand the real world.* URL: https://www.ted.com/talks/fei_fei_li_with_spatial_intelligence_ai_will_understand_the_real_world.

[2] Kaiming He et al. "Deep residual learning for image recognition". In: (2016), pp. 770–778.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: NIPS'12 (2012), pp. 1097–1105.

[4] Yan LeCun. "Backpropagation applied to handwritten zip code recognition". In: *Neural Computation* 1 (1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541.

[5] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: (2015).