SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Vizualni transformeri

Kristo Palić

Voditelj: *Tomislav Petković*

Zagreb, svibanj, 2024

Sadržaj

1. Uvod

Jedan od važnijih razloga razvoja velikih jezičnih modela u prošlom desetljeću bile su povratne neuronske mreže (engl. Recurrent Neural Networks – RNN). Međutim, RNN-ovi imaju manu, njihovo treniranje se ne može paralelizirati. 2017. godine skupina Google-ovih inženjera objavljuje znanstveni rad [1]„Attention is all you need“ koji je do sada citiran preko 120 000 puta. U tom radu predlaže se nova vrsta arhitekture neuronske mreže – Transformer. Transformeri su neuronske mreže koje se zasnivaju na konceptu pažnje bez ikakve upotrebe RNN-ova. Nova arhitektura transformera omogućuje par magnituda efikasnije treniranje velikih jezičnih modela. Arhitektura transformera omogućila je mnogim pojedincima i znanstvenim institucijama koje nemaju resursa Google-a ili Amazona treniranje vlastitih velikih jezičnih modela. Nedugo nakon, objavljeno je mnogo znanstvenih članaka koji pokušavaju još više unaprijediti tehnologiju transformera i razvijaju se različite inačice modela od kojih svaka ima svoje prednosti i mane. Sve to kulminira 2023. godine kada na tržište dolazi planetarno poznati generativni predtrenirani transformer teksta ili chat-GPT tj. njegova treća verzija. Znanstvenici diljem svijeta pokušavaju prenamijeniti tehnologiju transformera na dvodimenzionalne ulaze kako bi mogli trenirati velike modele za klasifikaciju slika. Objavljeni su članci koji pokušavaju implementirati koncept pažnje u već postojeće konvolucijske i slične duboke modele. 2021. godine, skupina znanstvenika objavljuje članak [2]„An image is worth 16x16 words: Transformers for image recognition at scale“. U tom radu skupina znanstvenika pokazuje način na koji se treniraju vizualni transformeri. Njihov rad pokazuje kvalitetne rezultate s minimalnim podešavanjem parametara što pokazuje da je tehnologija transformera primjenjiva na dvodimenzionalne ulaze i da ima smisla nastaviti istraživati vizualne transformere.

U svom seminarskom radu detaljnije ću objasniti probleme koje vežemo uz povratne neuronske mreže, arhitekturu transformera, pokušati objasniti pojam pažnje i njegovu matematičku pozadinu, probleme koji postoje kada tehnologiju transformera pokušamo preslikati na vizualni input i objasniti dosadašnji tijek istraživanja i primjene tehnologije vizualnih transformera.

2. Transformeri

Kako bi shvatili što su vizualni transformeri, moramo shvatiti što su uopće transformeri, kako funkcioniraju i zašto su nastali. Njihova prvotna namjena bila je isključivo na području obrade prirodnog jezika (*engl. Natural Language Processing - NLP*).

2.1 Sekvencijalnost RNN-ova

Povratne neuronske mreže (RNN) su neuronske mreže koje koriste povratne veze. Za razliku od unaprijednih neuronskih mreža, RNN-ovi imaju petlje koje omogućuju prenošenje informacija iz jednog koraka sekvencijalnog unosa u sljedeći. Unutrašnja povratna veza omogućava RNN-ovima pamćenje prethodnih podataka u sekvenci. Glavne osobine modela RNN-ova su da ulazi mogu biti proizvoljne duljine, broj parametara ne ovisi o duljini slijeda i model je osjetljiv na redoslijed ulaznih podataka. Zbog toga je RNN arhitektura prikladna za zadatke prevođenja teksta, obrade prirodnog jezika, prepoznavanje govora i slično. Model se razlikuje od ostalih modela dubokog učenja po svom skrivenom stanju koje se ažurira nakon svakog novog ulaza. Generiranje novog stanja ovisi o stanju koraka prije njega, što ima smisla kad govorimo o obradi jezika. Trenutno stanje ovisi o svim riječima u rečenici koje su prethodile trenutnoj. Matematički gledano, generiranje novog stanja izgleda ovako:

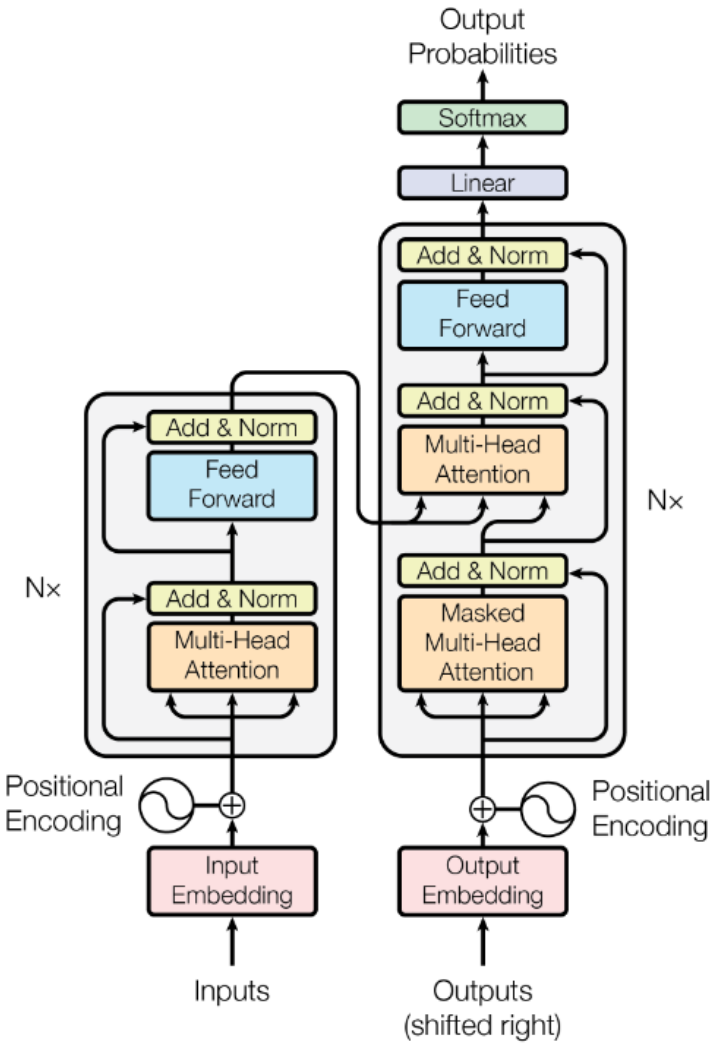
$$h^{(t)} = g\left(W_{hh}h^{(t-1)} + W_{xh}x^{(t)} + b_h\right) \quad (1)$$

Gdje su W_{hh}, W_{xh}, b_h parametri povratne afine transformacije, g je nelinearnost (sigmoida, tanh...), $h^{(t-1)}$ je skriveno stanje prijašnjeg koraka i $h^{(t)}$ je skriveno stanje trenutnog koraka.

Problem koji se javlja pri treniranju RNN-ova je sljedeći: Algoritam unazadne propagacije greške (engl. Backpropagation) koji koristimo pri treniranju RNN-ova sastoji se od izračuna diferencijala gubitka po svim parametrima koji su utjecali na izlaz modela. Osim parametara povratne afine transformacije, na izlaz modela utječu i sva prethodna stanja (na trenutno utječe prethodno, na prethodno utječe pretprethodno itd. do prvog skrivenog stanja). Takav proces je sekvencijalan i ne možemo ga paralelizirati. Upravo to je razlog traženja nove arhitekture koja će jednako dobro generalizirati nad podatcima, a istovremeno biti efikasnija za učenje.

2.2 Prvi transformer – pažnja bez upotrebe RNN-ova

Potaknuti prethodno opisanim problemom, grupa Google-ovih znanstvenika je 2017. godine objavila rad pod nazivom Attention is all you need[1]. U njemu su predložili novi model arhitekture neuronske mreže koju su nazvali – Transformeri. Transformeri zaobilaze povratnu vezu koja postoji u RNN-ovima i sprječava paralelizaciju. Za izračunavanje semantičke povezanosti dviju riječi, bez obzira na njihovu poziciju u rečenici ili paragrafu, potreban je konstantan broj operacija. Prije svega bih htio pojasniti arhitekturu njihovog modela, kako bismo dobili općeniti dojam o modelu kroz koji podatci prolaze, a u zasebnom odjeljku posebnu pažnju posvetiti upravo pažnji; glavnom konceptu ovog rada, koji nam je bitan za naše vizualne transformere.



Slika 1. Arhitektura transformera (izvor [1])

Na slici 1 prikazana je arhitektura transformera. Transformer generira izlaz na način da se nakon prolaska cijele ulazne sekvence generiraju softmax vrijednosti izlaznog tokena. Token s najvećom vjerojatnošću se zatim odabire kao prvi izlazni token. Isti

taj izlazni token postaje ulaz izlaznog sloja ulaganja (engl. Output Embedding). Svaki sljedeći token biti će generiran uz pomoć ulazne sekvence i dotad generiranih izlaznih tokena.

Tokeni ulazne sekvence i generirani tokeni se u slojevima ulaganja enkodiraju u višedimenzionalni vektor. Nakon toga se dodatno upisuje njihova pozicija u rečenici (na slici 1 – pozicijsko enkodiranje, engl. positional encoding)

Slika 1 prikazuje model koji koristi koder-dekoder strukturu. Na ulaze kodera i dekodera dovodimo pozicijski enkodirane izlaze slojeva ulaganja. Takva arhitektura sastoji se od dvije povezane neuronske mreže: koder procesira ulazne podatke i transformira ih u neku drugu vrstu reprezentacije, dok dekoder s takvim ulazima i s prethodno generiranim izlazima generira novi izlazni token.

Koder se sastoji od 6 identičnih slojeva (na slici 1 prikazan je samo jedan). Svaki sloj ima dva podsloja, prvi je sloj pažnje s više glava (*engl. Multi-head attention)*, a drugi je potpuno povezana unaprijedna mreža. Nakon svakog podsloja dolazi sloj normalizacije. Dekoder se također sastoji 6 identičnih slojeva. Osim dva podsloja koji su identični koderu, dekoder sadrži treći podsloj, sloj pažnje s više glava čiji je ulaz kombinacija izlaza kodera i prethodno generiranih izlaznih tokena.

2.3 Pažnja

Ugrađujući vektori ulaznih tokena u višedimenzionalnom prostoru imaju i semantičko značenje. Tako će se na primjer, u okolini riječi „toranj“ nalaziti riječi sličnog značenja poput „kula“ ili „utvrda“. Sparivanjem tokena „Eiffelov“ i „toranj“ dobiti ćemo u potpunosti drugačiju vektorsku reprezentaciju u tom prostoru. Na taj način dobivamo novo, kompleksnije značenje koje nam omogućava analizu kompleksnijih i apstraktnijih rečenica. Za tu svrhu koristimo mehanizam pažnje.

Pažnja je funkcija čija je svrha odrediti odnose dvaju ili više tokena unutar modela. Jednostavno rečeno, želimo spariti riječi koje se odnose jedna na drugu. Kada se aktivira mehanizam pažnje, model analizira sve riječi i pridodaje važnost svakoj riječi ovisno o tome kako se ona odnosi na druge riječi. Mehanizam pažnje omogućuje modelu stvaranje preciznijih i kontekstualno relevantnijih izlaza jer razumije ne samo značenje pojedinih riječi, već i njihove međusobne odnose i utjecaje unutar rečenice. Kako je to izvedeno?

Sloj pažnje s više glava razmatramo kao skup slojeva pažnje s jednom glavom. Matematički opisano, pažnja je funkcija:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

gdje Q označava matricu upita (engl. query), K matricu ključeva (engl. keys), a V matricu vrijednosti (engl. value). Način zapisa ove funkcije je izrazito kompaktan i zapravo označava izračun svih pojedinačnih upita q nad svim pojedinačnim odgovorima k, pomnožen sa svim vektorima v.

Sve opisane matrice Q, K i V su parametri našeg sloja pažnje i kao takvi se kalibriraju tijekom procesa učenja. Mehanizam pažnje je uveden kako bi se s lakoćom sparivale riječi koje se u ne nalaze jedna pored druge.

Mehanizam pažnje može se paralelizirati. Paralelno se može računati pažnja za svaki upit $q \in Q$. Matrični prikaz nam istovremeno daje numeričku vrijednost ovisnosti trenutne riječi o svakoj prijašnjoj riječi te za razliku od RNN-ova nema potrebe za odmotavanjem povratne veze tijekom procesa učenja. Broj upita, ključeva i vrijednosti ovisi o kontekstualnom prozoru našeg sloja pažnje s više glava. Što je veći kontekstualni prozor, tim više skrivenih uzoraka ili značenja možemo naučiti i samim time je naš model bolji.

3. Vizualni transformeri

Kada smo pričali o prvim transformerima i mehanizmu pažnje podrazumijevani ulaz je bio tekst. Osnovna gradivna jedinica teksta je slovo, ali tekst znamo rastaviti na riječi pa nismo imali problema oko određivanja što je ulazni token. Prvi problem nastaje kada kao ulaz u model transformera dovedemo sliku. Osnovna gradivna jedinica slike je piksel, ali što je onda token slike? Možemo li koristiti piksel kao token? Ako ne možemo, kako ih možemo grupirati? Navesti ćemo probleme s kojima su znanstvenici suočeni, rješenjem tih problema i primjenom vizualnih transformera.

3.1 Problem skaliranja vizualnih transformera

Kada smo opisivali mehanizam pažnje, rekli smo da je to sposobnost jednog tokena (riječ) da odredi ovisnost o drugom tokenu. Isti je slučaj i kod vizualnih transformera.

Možemo li koristiti piksele kao tokene? Mehanizam pažnje u vizualnim transformerima mora pratiti ovisnost svakog piksela o svakom pikselu. Tu nastaje problem. Pažnja kao kvadratna funkcija prati međuovisnosti parova tokena. Koliko ima parova tokena koje moramo pratiti? Idemo to izračunati jednostavnim primjerom.

Pretpostavimo da imamo sliku dimenzija 256 piksela.

$$Height = 256 \quad Width = 256$$

$$N = Height \cdot Width = 256^2 = 65536$$

$$Broj\ parova = N^2 = 4\,294\,967\,296$$

Korištenje piksela slike kao tokena znatno povećava broj parova koje mehanizam pažnje mora pratiti. Povećana računalna i memorijska složenost jedan je od glavnih izazova pri primjeni transformatora na vizualne podatke.

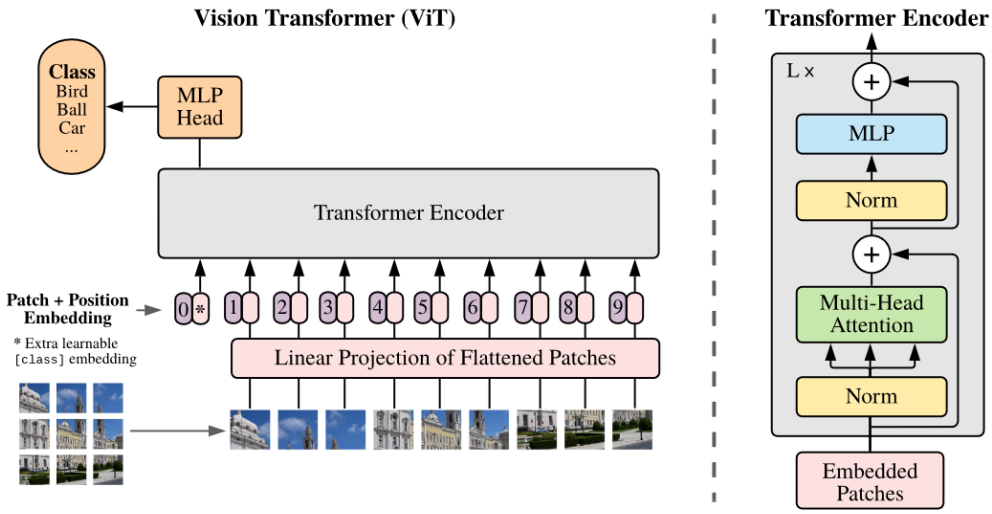
Očito je da ne možemo koristiti piksele kao tokene, ali kako onda podijeliti sliku? Znanstvenici su pokušali riješiti ovaj problem spremajući samo lokalne ovisnosti u obliku kvadratnog susjedstva piksela, što je naravno teorijski temelj već postojećih konvolucijskih dubokih modela. Osim toga, napravljene su inačice transformera koje kombiniraju neke druge duboke modele, što također nije donijelo značajan uspjeh.

3.2 An image is worth 16x16 words

Četiri godine nakon već spomenutog rada „Attention is all you need“ grupa znanstvenika Google-a objavljuje rad: [2] „An image is worth 16x16 words: Transformers for image recognition at scale“. Njihov rad temelji se na rješavanju gore opisanog problema uz što manju promjenu parametara i slojeva prvog ikad transformera iz 2017. godine. Prođimo prvo kroz arhitekturu vizualnog transformera pa zatim detaljno kroz eksperimentalne rezultate koji su takvim modelom postignuti.

3.2.1 Arhitektura vizualnog transformera

Navedeni problem nepoznavanja što su tokeni u vizualnom transformeru, znanstvenici su riješili seciranjem slike na komade veličine 16x16 piksela. Ti isti komadi su zatim, u pravilnom redoslijedu, enkodirani skupa s pozicijskim brojem. Numerirani komad slike je zatim, transformiran u višedimenzionalni vektor na jednak način kao što je riječ enkodirana u višedimenzionalni vektor tekstualnog transformera.



Slika 2 – Arhitektura vizualnog transformera

Na slici 2 preuzetoj iz [2] vidimo detaljniju ilustraciju njihovog Vision Transformera (ViT). Komade slike možemo predstaviti kao tenzor trećeg reda dimenzija 16x16x3 (za RGB sliku), gdje vrijednost svakog elementa tenzora predstavlja vrijednost subpiksela unutar tog komada. Taj isti tenzor je zatim „spljošten“ u matricu s 256x3 dimenzija (po jedan vektor za vrijednosti svake boje u RGB slici). Kako smo na ulaze tekstualnih transformera transformirali riječi u višedimenzionalne vektore, isto moramo napraviti i u vizualnim transformerima. Tome služi sloj na slici nazvan Linear Projection of Flattened Patches. Sloj se sastoji od matrice ulaganja (embedding matrix) koja svaki ulaz transformira u vektor istih dimenzija. Nakon linearne projekcije, u vektor se ugrađuje njegova pozicija u slici.

Nakon ovih transformacija slijedi koder transformatora identičan onome iz 2017. godine. Jedina razlika je u dodanom „specijalnom“ ulazu (0*) koji je „naučljiv“. Izlaz kodera nultog ulaza dovodimo na ulaz višeslojnog perceptrona i koristimo ga za klasifikaciju. Ostale izlaze kodera odbacujemo i oni nam nisu potrebni.

3.2.2 Eksperimentalni rezultati

Promotrimo detaljnije eksperimentalne rezultate vizualnog transformera iz rada [2]

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

tablica 3 – parametri modela ViT – preuzeto iz [2]

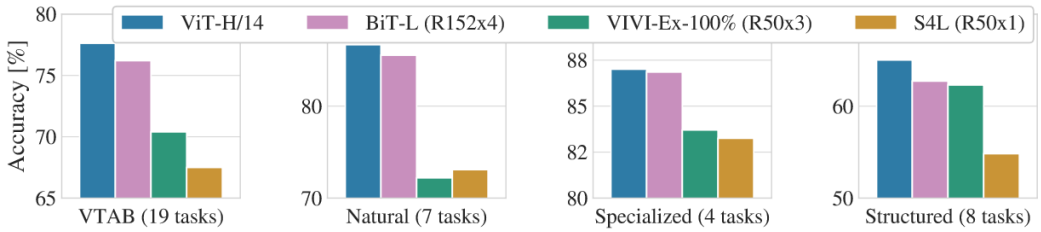
Za potrebe testiranja napravljena su tri različita modela čije parametre možemo vidjeti na slici 3. Ideja testiranja je bila da se različite inačice vizualnih transformera i konvolucijskih modela trenira na istom skupu podataka za učenje, a zatim testira na drugim dostupnim velikim skupovima podataka. Rezultate možemo vidjeti na sljedećoj slici:

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Tablica 4 – eksperimentalni rezultati – preuzeto iz [2]

U tablici 4 u prvom stupcu nalaze se imena skupova podataka za testiranje. U drugom, trećem i četvrtom stupcu nalaze se različite inačice modela transformera. Četvrti i peti stupac su konvolucijski modeli koji su do tad smatrani najboljima.

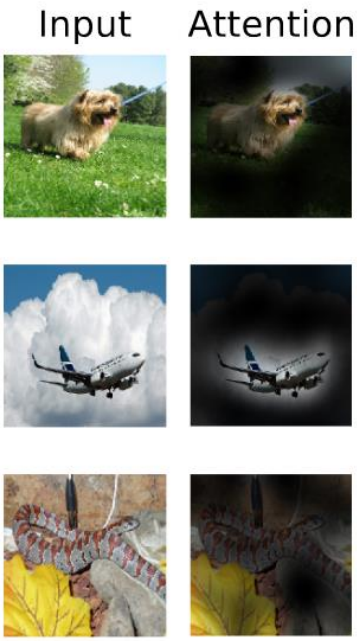
Iz prikazanih rezultata vidimo da vizualni transformer ViT-H/14 u kojem su dimenzije komada slike 14x14 premašuje rezultate dotad najboljeg ResNet-a uz 4 puta jeftiniju cijenu treniranja (zadnji red – TPUv3-core-days). Iz rezultata također vidimo da se samo s promjenom arhitekture s konvolucijskog modela na model vizualnog transformera (ViT-L/16) mogu dobiti jednaki ili jako bliski rezultati uz čak 15 puta jeftinije treniranje.



Slika 5 – rezultati različitih modela – preuzeto iz 2

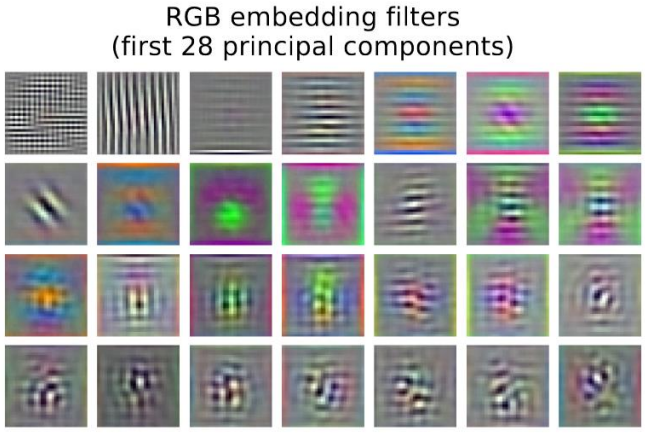
3.2.3 Analiza modela

Zahvaljujući detaljnoj analizi koja je u radu [2] predstavljena, možemo kvalitetno analizirati unutarnje parametre vizualnih transformera.



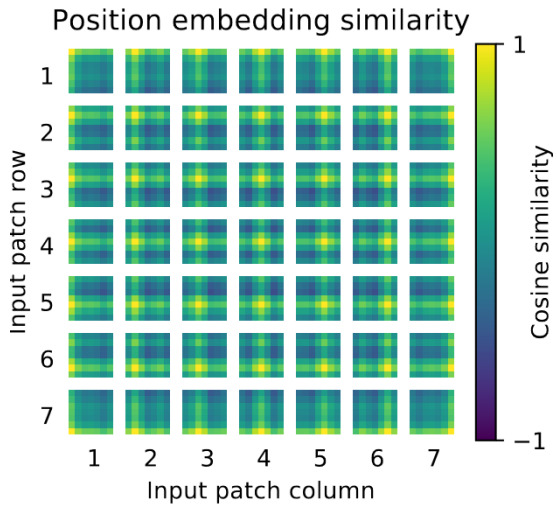
Slika 6 – Pažnja u ViT-u – preuzeto iz [2]

Slika 6 demonstrira sposobnost vizualnih transformera da usmjere svoju pažnju na relevantne dijelove slike, ignorirajući nebitne informacije. U svakom od primjera, model je uspješno identificirao glavni objekt (pas, avion, zmija) i koncentrirao pažnju na njega, što je ključno za točne klasifikacijske zadatke. Vizualizacija može pomoći u razumijevanju kako model donosi odluke i koji dijelovi slike najviše doprinose konačnoj odluci.



Slika 7 – filteri linearnog ulaganja – preuzeto iz [2]

Također, na slici 7 vidimo filtere linearnih ulaganja (engl. embedding filters) koji izrazito podsjećaju na filtere konvolucijskih mreža. Svaki kvadrat na slici predstavlja jedan filter koji se primjenjuje na ulazne podatke. Filteri izvlače različite značajke iz ulaznih podataka, kao što su rubovi, teksture, i boje.



Slika 8 – pozicijska sličnost – preuzeto iz [2]

Iz slike 8 sličnosti pozicijskih ulaganja vidimo sličnost komada slike sa svim ostalim komadima. Model je bez da smo mu išta eksplicitno rekli, sam pravilno razumio da je ulaz dvodimenzionalan i napravio je pravilnu rekonstrukciju svoje i svih ostalih pozicija.

3.3 Primjene vizualnih transformera

3.3.1 Prepoznavanje objekata

Jedna od najznačajnijih primjena vizualnih transformera je detekcija objekata. Detekcija objekata uključuje identificiranje i lokalizaciju objekata unutar slike. Klasične metode, poput Faster R-CNN, koriste složene cjevovode i ručno dizajnirane mehanizme za predikciju središta okvira (bounding boxes). S druge strane, model DEtection TRansformer (DETR) [3] predstavlja inovativan pristup koji koristi transformere za direktno predviđanje setova objekata, eliminirajući potrebu za nekim od tradicionalnih koraka poput maksimalne supresije (non-maximum suppression). DETR koristi bipartitni mehanizam podudaranja za jedinstvenu predikciju objekata, što rezultira značajno boljim performansama na velikim objektima.

Deformable DETR [3] rješava probleme sporog konvergiranja i ograničene prostorne rezolucije karakteristika transformera u obradi slika, uvodeći deformirani modul

pažnje koji se prirodno proširuje na agregaciju višeslojnih značajki. Ovi napredni pristupi omogućuju precizniju detekciju objekata i bržu obuku modela.

3.3.2 Praćenje objekata

Transformeri su također značajno unaprijedili praćenje objekata. TrackFormer [3] je model temeljen na transformerima koji koristi koder-dekoder arhitekturu za praćenje i segmentaciju više objekata. Pristup uvodi ugrađene upite za praćenje koji prate objekte kroz video sekvence autoregresivno, eliminirajući potrebu za dodatnim mehanizmima podudaranja, optimizacijom ili modeliranjem pokreta i izgleda.

TransTrack [3] koristi mehanizam query-key za praćenje objekata, što omogućuje jednostavnu paradigmu zajedničke detekcije i praćenja. Koristi naučene upite objekata za detekciju novih objekata u svakom okviru, osiguravajući visoku točnost i jednostavniju implementaciju.

3.3.3 Klasifikacija akcija

Vizualni transformeri također imaju primjenu u klasifikaciji akcija, gdje se prepoznaju radnje osoba u videozapisima. ActionTransformer koristi transformere za analizu interakcija između ljudi u sceni, omogućujući modelu da prepozna složene akcije na temelju interakcija s okolinom.

TimeSformer [3] je još jedan značajan model koji koristi "Divided Space-Time Attention" za klasifikaciju akcija u videozapisima. Ovaj pristup prvo primjenjuje samopažnju na sve dijelove unutar istog vremenskog okvira, a zatim na prostorne dijelove, omogućujući modelu da bolje razumije dugoročne odnose u videozapisima.

4. Zaključak

U proteklih nekoliko godina, transformeri su postali značajna arhitektura u području dubokog učenja, značajno unapređujući performanse u različitim zadacima obrade prirodnog jezika i računalnog vida. Od svoje primjene u NLP-u, transformeri su se proširili na područje računalnog vida, gdje su donijeli niz poboljšanja i omogućili nove metode obrade vizualnih podataka. Vizualni transformeri (ViT) posebno su se istaknuli u različitim primjenama, uključujući prepoznavanje objekata, praćenje objekata i klasifikacije akcija.

Vizualni transformeri pokazali su izvanredne rezultate u zadacima restauracije slika, gdje su se pokazali superiornima u odnosu na CNN-ove u zadacima kao što su super-rezolucija, uklanjanje šuma, opće poboljšanje slike, smanjenje artefakata JPEG kompresije i uklanjanja zamućenja. Svi ovi zadatci zahtijevaju detaljno razumijevanje i obradu finih detalja unutar slike, a transformeri su se pokazali izuzetno sposobnima u učenju i rekonstrukciji ovih detalja.

Unatoč svojim prednostima, vizualni transformeri također se suočavaju s izazovima, poput povećane računalne složenosti i potrebe za velikim količinama podataka za obuku. Međutim, kontinuirani napredak u optimizaciji algoritama i arhitektura obećava daljnje poboljšanje efikasnosti i šire prihvaćanje ove tehnologije.

Zaključno, transformeri su se etablirali kao ključna tehnologija u modernom dubokom učenju. Njihova sposobnost da efikasno obrađuju složene uzorke podataka i pružaju vrhunske performanse čini ih nezamjenjivim alatom u današnjem svijetu. Budućnost vizualnih transformera obećava daljnje inovacije i poboljšanja koja će dodatno unaprijediti sposobnosti obrade i analize vizualnih podataka, čineći ih ključnim elementom u razvoju naprednih sustava umjetne inteligencije.

5. Sažetak

U posljednjem desetljeću, konvolucijske neuronske mreže (CNN) dominirale su područjem računalnog vida. Međutim, nedavno su transformerske mreže, prvotno razvijene za obradu prirodnog jezika, prilagođene i pokazale jako dobre rezultate u računalnom vidu. Ovaj rad fokusira se na nastanak modela Vision Transformer (ViT), prvu arhitekturu koja je efikasno primijenila model transformera na slike. Detaljno ćemo istražiti kako i zašto su nastali transformeri, kako ViT radi, kako se uspoređuje s tradicionalnim CNN-ima, i koje su njegove prednosti i ograničenja. Također ćemo razmotriti primjene ViT-a u različitim zadatcima računalnog vida.

6. Literatura

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

[3] Jiarui Bi, Qinglong Meng, Zengliang Zhu: Transformer in Computer Vision. CEI 2021