SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Vizualni transformeri

Kristo Palić

Voditelj: *Tomislav Petković*

Zagreb, svibanj, 2024

Sadržaj

1. Uvod

Rekurentne neuronske mreže glavni su razlog značajnog razvoja velikih jezičnih modela prošlog desetljeća. Međutim, RNN-ovi imaju manu, sekvencijalni su. 2017. godine skupina Google-ovih inženjera objavljuje znanstveni rad „Attention is all you need“ koji je do sada citiran preko 120 000 puta. U tom radu predlaže se nova vrsta arhitekture neuronske mreže koja se zasniva na pažnji bez ikakve upotrebe RNN-ova. Taj rad je stvarno revolucionaran jer 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. Odjednom dolazi do mnoštva alternativnih 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 imaju vidljive rezultate, ali ništa od toga nije revolucionarno kao što je rad iz 2017. godine bio. Sve do 2021. godine, kada skupina znanstvenika objavljuje članak „An image is worth 16x16 words: Transformers for image recognition at scale“. U tom radu skupina znanstvenika pokazuje način na koji se mogu istrenirati vizualni transformeri uz što manje mijenjanje upravo onog istog revolucionarnog članka iz 2017. godine. Njihov rad pokazuje kvalitetne rezultate sa 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 teksta (*engl. Natural Language Processing - NLP*).

2.1 Sekvencijalnost RNN-ova

Povratne neuronske mreže *(engl. Recurrent Neural Networks – RNN)* su neuronske mreže koje koriste sekvencijalne podatke ili slijedne vremenske podatke (*engl. time series).* Glavne osobine modela RNN-ova je da ulazi mogu biti proizvoljne duljine, broj parametara ne ovisi o duljini slijeda i model je osjetljiv na redoslijed ulaznih podataka. Upravo zbog toga RNN arhitektura dubokih modela prikladna je 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. Tako ćemo na primjer, za rečenicu od sedam riječi imati sedam različitih skrivenih stanja. 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:

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

W_{hh}, W_{xh}, b_h → parametri povratne afine transformacije

g → nelinearnost (sigmoida, tanh...)

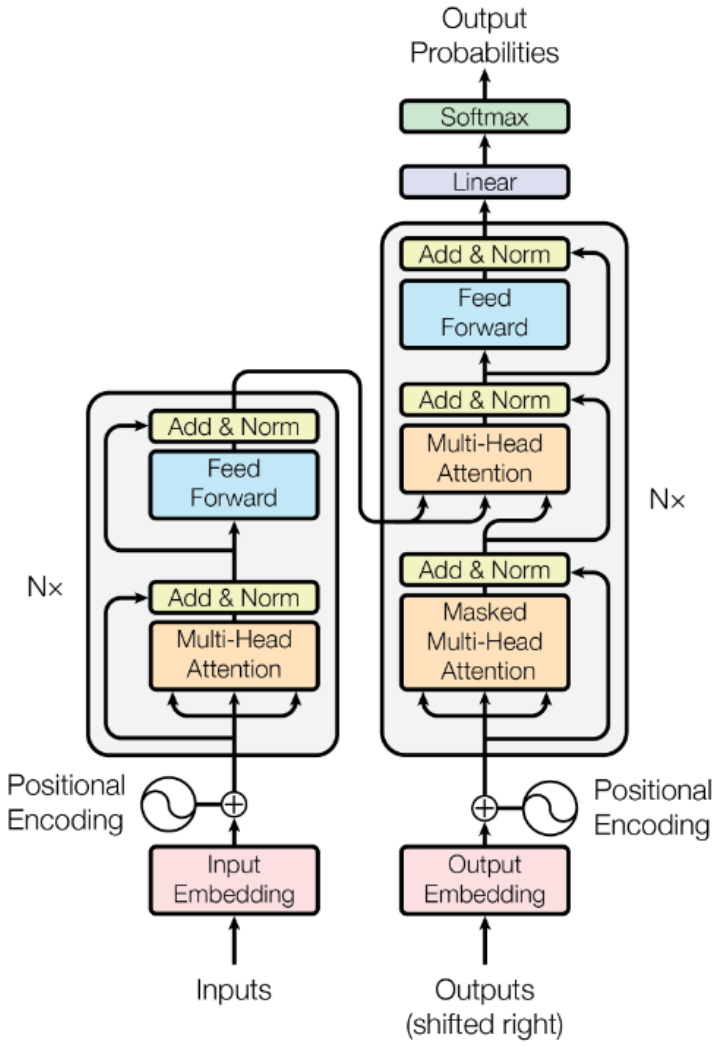
$h^{(t-1)}$ → skriveno stanje prijašnjeg koraka

$h^{(t)}$ → skriveno stanje trenutnog koraka

Bez nepotrebnog ulaska u detalje, jer RNN-ovi su tema za sebe, želio bih predstaviti problem koji se javlja pri treniranju RNN-ova. Backpropagation algoritam 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). Pogledajmo rečenicu: „Lijepa je na Dunavu kuća". „Lijepa" je riječ koja opisuje riječ „kuća", ali nisu jedna pored druge. Ovakvih rečenica ima bezbroj i ne možemo jednostavno zaustaviti računanje gradijenata unatrag po volji, već moramo uvijek računati gradijente skrivenih stanja sve do gradijenta 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 revolucionarni rad pod nazivom Attention is all you need. 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. Konstantan broj operacija potrebnih za izračunavanje semantičke povezanosti dviju riječi, bez obzira na njihovu poziciju u rečenici ili paragrafu, čini ovu arhitekturu stvarno revolucionarnom. 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.



Na slici je prikazana arhitektura transformera. Sastoji se od Embedding slojeva koji pretvaraju ulazni token (riječ/slog) u višedimenzionalni vektor. Svaka riječ u rječniku može se na taj način, mapirati u višedimenzionalni vektor. Chat-GPT-ev embedding vektor sastoji se od preko 12000 dimenzija, od kojih svaka nosi nekakvo semantičko

značenje. Ulaz Input Embedding sloja je cijela ulazna sekvenca (tokeni), a ulazi Output Embeddinga su prethodno generirani izlazni tokeni.

Model koristi koder-dekoder strukturu. 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 sa prethodno generiranim izlazima generira novi izlazni token.

Koder se sastoji od 6 identičnih slojeva (na slici lijevo je prikazan 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

Ključan dio nove arhitekture transformera zasniva se na pažnji bez upotrebe RNN-ova, stoga smatram da trebamo detaljno objasniti o čemu se tu radi.

Pažnja je funkcija čija je svrha odrediti odnose dvaju ili više tokena unutar modela. Jednostavno rečeno, želimo kombinirati riječi koje se odnose jedna na drugu. Na taj način možemo doći do semantički kompliciranijeg značenja. Na primjer: „Jabuka nije bila samo ukusna već i prelijepa.“ Prijevodi ove rečenice na drugi jezik zahtijevaju razumijevanje kako se značenja riječi „ukusna“ i „prelijepa“ odnose na „jabuka“, a ne na nešto drugo u rečenici. U početku se svaka riječ u rečenici transformira u vektor u embedding prostoru, koji sadrži informacije o njenom značenju i upotrebi. Kada se aktivira mehanizam pažnje, model analizira sve riječi i pridaje važnost svakoj riječi ovisno o tome kako se ona odnosi na druge riječi. Na primjer, u gornjoj rečenici, pažnja bi se mogla više usmjeriti na povezanost između „jabuka“ i „ukusna“ te „jabuka" i „prelijepa". Tijekom ovog procesa, vektori za „ukusna“ i „prelijepa“ pomaknut će se bliže vektoru za "jabuka" u embedding prostoru, što znači da model prepoznaje i pojačava njihovu međusobnu povezanost. Na kraju ćemo umjesto vektora „jabuka“ dobiti vektor „prekrasna ukusna jabuka“. Dakle, s vremenom i s više informacija koje pažnja obradi, embedding vektori za te riječi postaju sve usklađeniji s kontekstom u kojem se koriste. 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:

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

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. Pokušati ću ilustrirati ovu funkciju nad

rečenicom: „Marljivi studenti pišu seminar prije krajnjeg roka“. Umjesto matrica svih upita, ključeva i vrijednosti uzmimo samo jedan upit i jedan ključ.

Na upit (query) q možemo gledati kao pitanje: „Jesi li pridjev?“. Taj upit upućen je riječi u svakom retku. ako je riječ iz retka pridjev koji se u rečenici nalazi neposredno prije imenice iz stupca, odgovor k će biti pozitivan vektor i rezultat množenja $q \cdot k^T$ biti će pozitivan broj. Opet laički rečeno, kažemo da imenica mora „obratiti pažnju“ na pridjev.

Ako riječ iz retka dolazi nakon riječi iz stupca, njihov umnožak pitanja i odgovora „maskiramo“ na negativnu beskonačnost jer želimo gledati samo podatke koji su došli prije odabrane riječi

	marljivi	studenti	pišu	seminar	prije	krajnjeg	roka
marljivi	$q_i \cdot k_j^T$	75.3	0.23	1	-1.25	-5.56	-2.4
studenti	$-\infty$	$-\infty$	-5	0	1.7	-4.5	-1.4
pišu	$-\infty$	$-\infty$	$-\infty$	-4	0.4	1.2	-0.4
seminar	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-7	1.3	0.4
prije	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-2.1	0.5
krajnjeg	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	87.89
roka	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$

Ovakve izlaze podjelimo s dimenzionalnosti ključeva $\sqrt{d_k}$ zbog numeričke stabilnosti i ubacimo u softmax funkciju po stupcima kako bi dobili vjerojatnosti da riječi trebamo gledati kao cjelinu i obraćati pažnju na njih.

	marljivi	studenti	pišu	seminar	prije	krajnjeg	roka
marljivi	$softmax(q_i \cdot k_j^T)$	1	0.99	0.75	0.75	0.75	0.001
studenti	0	0	0.01	0.20	0.23	0.22	0.001
pišu	0	0	0	0.05	0.01	0.01	0.001
seminar	0	0	0	0	0.01	0.01	0.001
prije	0	0	0	0	0	0.01	0.001
krajnjeg	0	0	0	0	0	0	0.995
roka	0	0	0	0	0	0	0

Za kraj, dobivene izlaze softmaxa pomnožimo sa vrijednošću parametara v. Vektor v je vrijednost za koju se moramo pomaknuti u višedimenzionalnom prostoru embeddanih vektora da bi došli na poziciju koja označava našu novu cjelinu. U našem primjeru trebamo translatirati ulazni vektor sa pozicije „studenti“ na poziciju „marljivi studenti“. Tom transformacijom vektora dobili smo potpuno novo značenje. Embeddani vektori u višedimenzionalnom prostoru su „bliski“ ako im je značenje slično. Ovako prikazana transformacija dovodi do promjene značenja riječi ili sintakse te omogućava još složenije ili apstraktnije izraze. Još jedan primjer je „Eiffelov toranj“. Riječ toranj u višedimenzionalnom prostoru u svojoj neposrednoj blizini ima riječi poput „kula“, „utvrda“ i slično jer su značenjem bliski, ali dodavanjem riječi „Eiffelov“ dobit ćemo potpuno drugu interpretaciju i samim time, drugačiji vektor u tom višedimenzionalnom prostoru.

Sve opisane matrice Q, K i V su parametri našeg sloja pažnje i kao takvi se kalibriraju tijekom procesa učenja. Zbog jednostavnosti sam kao uvjet uzeo neposredno predhođenje imenici, ali u stvarnosti to nije slučaj. Mehanizam pažnje se bez problema može aktivirati nad udaljenim riječima i to je često i slučaj. Moram naglasiti kako je ovo izuzetno teško vizualizirati na papiru te preporučujem dodatnu pretragu na internetu.

Mehanizam pažnje može se paralelizirati. Paralelno se može računati pažnja za svaki upit $q \in Q$. Matrični prikaz, koji smo maloprije ilustrirali, nam istovremeno daje numeričku vrijednost ovisnosti trenutne riječi o svakoj prijašnjoj riječi te za razliku od RNN-ova nema potrebe za dubljim iterativnim računanjem 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.

Svakome koga ovo gradivo zanima preporučio bih samostalnu analizu znanstvenog rada iz poglavlja 2.2 te radove koji se nadovezuju na temu transformera. Sada kada smo objasnili mane RNN-ova, tehnologiju transformera i sloj pažnje u njima, vrijeme je da se krenemo baviti vizualnim transformerima i njihovom primjenom.

3. Vizualni transformeri

Kada smo pričali o prvim transformerima i mehanizmu pažnje podrazumijevani ulaz je bio tekst. Ulazni podatci u enkoder su zapravo bili embedded vektori ulaznih tokena. U ovom poglavlju bavit ćemo se transformerima kojima je ulaz slika, stoga ulaz ima dvije dimenzije. 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 jedne riječi da odredi ovisnost o drugoj riječi. Isti je slučaj i kod vizualnih transformera, samo ulazi nisu riječi, već slike. Točnije, tokeni dvodimenzionalnog transformera su pikseli. 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 tekstualni paragraf s 512 tokena.

$$N = 512$$

$$Broj\ parova = N^2 = 262{,}144$$

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 umjesto tekstualnih tokena eksponencijalno povećava broj parova koje mehanizam pažnje mora pratiti. Povećana računalna i memorijska složenost jedan je od glavnih izazova pri primjeni transformatora na vizualne podatke.

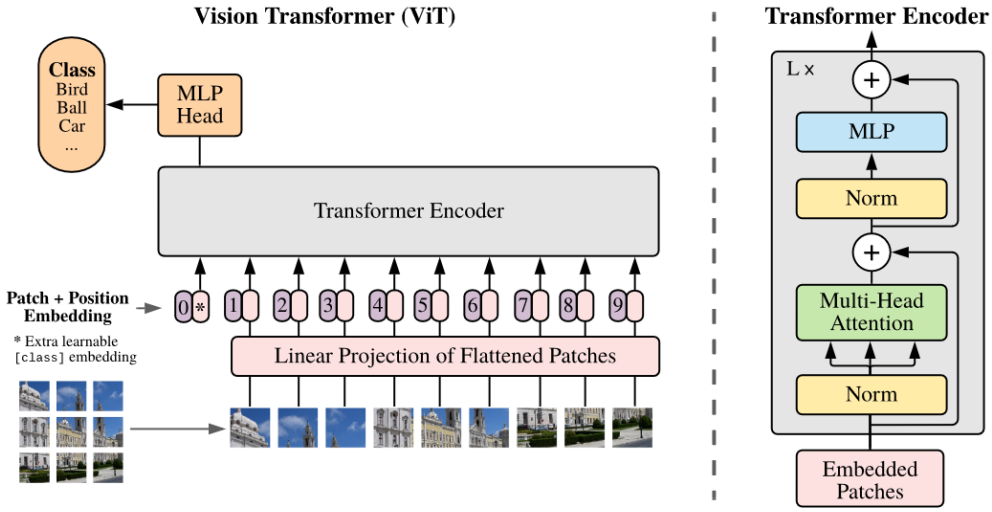
Znanstvenici su pokušali riješiti ovaj problem spremajući samo lokalne ovisnosti u obliku kvadratnog susjedstva piksela, što je naravno teorijski temelj već postojećih konvolucijskih dubokih modela. Osim toga, napravljene su inačice transformera koje kombiniraju neke druge duboke modele, što također nije donijelo značajan uspjeh.

3.2 An image is worth 16x16 words

Četiri godine nakon već spomenutog, revolucionarnog rada „Attention is all you need“ grupa znanstvenika Google-a objavljuje rad: „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 eksponencijalnog povećanja parametara pažnje znanstvenici su riješili seciranjem slike na komade. Ti isti komadi su zatim, u pravilnom redoslijedu, enkodirani skupa sa pozicijskim brojem. Numerirani komad slike je zatim, transformiran u višedimenzionalni vektor na jednak način kao što je riječ enkodirana u višedimenzionalni vektor tekstualnog transformera.



Na gornjoj slici 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 piksela unutar tog komada. Taj isti tenzor je zatim „spljošten“ u matricu sa 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 ugradnje (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

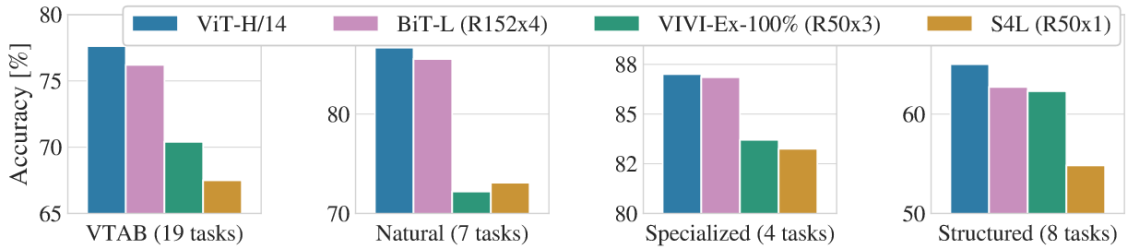
Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Za potrebe testiranja napravljena su tri različita modela čije parametre možemo vidjeti u tablici iznad. Ideja 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 tablici:

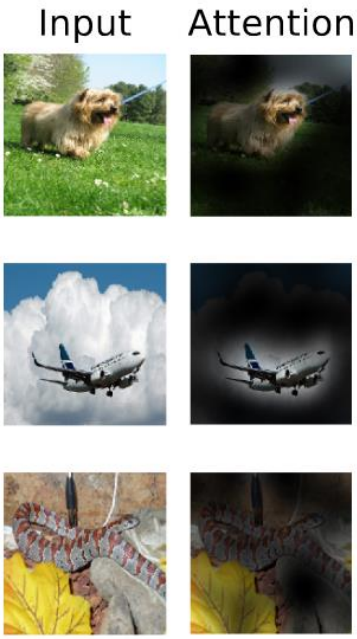
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 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

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 sa 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.

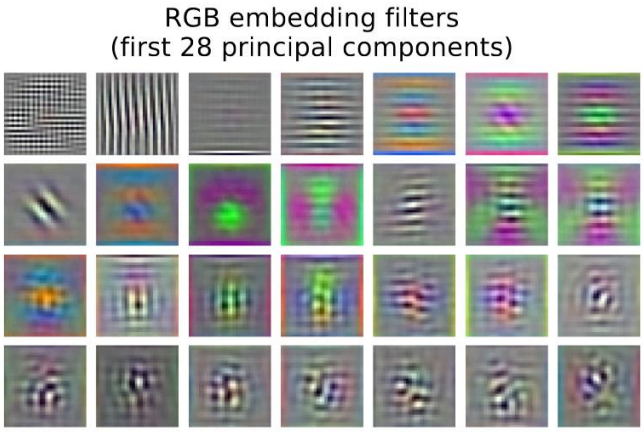


3.2.3 Analiza modela

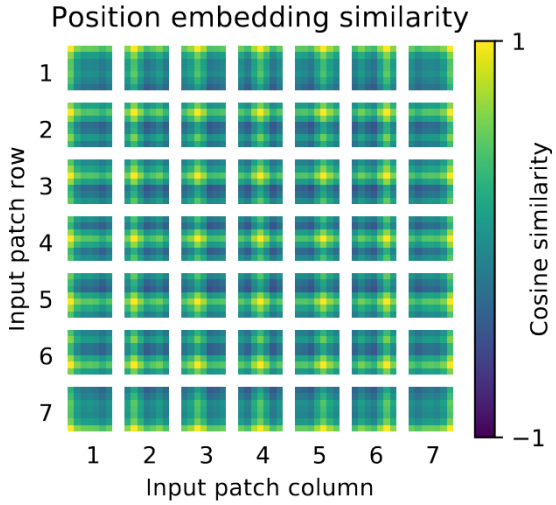


Zahvaljujući detaljnoj analizi koja je u radu predstavljena, možemo kvalitetno analizirati unutarnje parametre vizualnih transformera.

Slika desno 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.



Također, na slici desno vidimo filtere linearnih embeddinga 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.



Iz slike sličnosti pozicijskih embeddinga 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) 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 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 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 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 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.

3.3.4 Restauracija slika

Vizualni transformeri također se uspješno koriste u restauraciji slika, području koje uključuje zadatke poput super-rezolucije, uklanjanja šuma, općeg poboljšanja slike, smanjenja artefakata JPEG kompresije, uklanjanja zamućenja, uklanjanja nepovoljnih vremenskih uvjeta i uklanjanja zamućenja slike.

Vizualni transformeri pokazuju značajne prednosti u ovim zadacima zahvaljujući svojoj sposobnosti da efikasno uče značajke iz velikih količina podataka i boljoj robusnosti u ekstrakciji značajki. ViT modeli često nadmašuju konvolucijske neuronske mreže (CNN) u zadacima restauracije slike, posebno kada je dostupno mnogo podataka za obuku.

4. Zaključak

U proteklih nekoliko godina, transformeri su postali dominantna arhitektura u području dubokog učenja, značajno unapređujući performanse u različitim zadacima obrade prirodnog jezika i računalne vizije. Od svoje revolucionarne 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, klasifikaciju akcija i restauraciju slika.

Vizualni transformeri nadmašili su tradicionalne konvolucijske neuronske mreže (CNN) u mnogim aspektima, zahvaljujući svojoj sposobnosti paralelizacije obrade i boljeg razumijevanja globalnih odnosa unutar podataka. Dok su CNN-ovi ograničeni u svojoj sposobnosti da obuhvate dugačke domete odnosa zbog svoje lokalne prirode, transformeri koriste mehanizam pažnje kako bi efikasno identificirali i obradili ključne značajke bez obzira na njihovu poziciju.

Vizualni transformeri su pokazali 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, nadmašujući tradicionalne metode i otvarajući nove mogućnosti za istraživanje i primjenu u raznim domenama. 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 izvanredne rezultate u računalnom vidu. Ovaj rad fokusira se na nastanak modela Vision Transformer (ViT), prvu arhitekturu koja je efikasno primjenila 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

- 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.

- 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

- Jamil, S.; Jalil Piran, M.; Kwon, O.-J. A Comprehensive Survey of Transformers for Computer Vision. Drones 2023, 7, 287. https://doi.org/10.3390/drones7050287

- Jiarui Bi, Zengliang Zhu, Qinglong Meng; Transformer in Computer Vision. 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)

- Ali, A.M.; Benjdira, B.;Koubaa, A.; El-Shafai, W.; Khan, Z.;Boulila, W. Vision Transformers in Image Restoration: A Survey. Sensors 2023, 23, 2385. https://doi.org/10.3390/s23052385