Dobar dan,

Ja sam Kristo Palić, moj mentor je Tomislav Petković. Tema mog seminarskog rada bili su vizualni transformeri.

Prije nego što krenemo pričati što su vizualni transformeri moramo prvo objasniti što su uopće transformeri i kako funkcioniraju.

2017. godine skupina Google-ovih znanstvenika objavljuje rad „Attention is all you need“ u kojem predlažu novu vrstu arhitekture neuronske mreže s mehanizmom pažnje. Transformer arhitektura se u potpunosti oslanja na mehanizam pažnje, za razliku od tradicionalnih povratnih neuronskih mreža (RNN) i konvolucijskih neuronskih mreža (CNN), koje koriste sekvencijalne ili lokalne operacije. Na slici vidimo prikaz te arhitekture.

Ovaj transformer ima 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 s prethodno generiranim izlazima generira novi izlazni token.

Prvo objasnimo ulaz u mrežu. Pošto sada govorimo o tekstualnim transformerima, ulaz je naravno tekstualna sekvenca u kojoj svaka riječ predstavlja jedan token. U prvom sloju matrica ulaganja pretvara riječ u višedimenzionalni vektor. Svaka riječ koja postoji u rječniku ima svoju vektorsku reprezentaciju. Te dimenzije imaju i semantičko značenje pa će se na primjer, riječi poput kula, toranj ili utvrda u prostoru nalaziti blizu jedna drugoj. Dodavanjem pridjeva „Eiffelov“ ispred riječi toranj dobit ćemo vektor koji u tom istom prostoru ima potpuno drugačiju vrijednost. Na taj način dolazimo do kompleksnijih i apstraktnijih izraza.

U sljedećem sloju takvu vektorsku reprezentaciju enkodiramo s pozicijskim brojem. To nam je naravno bitno kako bismo znali redosljed riječi u sekvenci.

Nakon njega imamo sloj pažnje kojem ćemo se posebno posvetiti. Mehanizam pažnje je ključna komponenta transformera koja omogućuje modelu da stvara povezanosti s relevantnim tokenima ulazne sekvence prilikom generiranja izlaza. Evo kako to funkcionira:

 **Upiti, Ključevi i Vrijednosti (Q, K, V)**:

* Svaka riječ u ulaznoj sekvenci se predstavlja s tri različita vektora:
  + **Upit (Query, Q)**: Što tražimo.
  + **Ključ (Key, K)**: Što imamo.
  + **Vrijednost (Value, V)**: Informacija koju koristimo.

 **Izračunavanje Q, K i V**:

* Q, K i V se dobivaju tako da se vektori ulaganja pomnože s težinskim matricama koje se treniraju tijekom učenja modela:
* Ovdje je X matrica ulaznih vektora, a su težinske matrice.

Slojevi normalizacije i zbrajanja rezidualnih veza U praksi, svaki "Add & Norm" sloj u transformatoru radi sljedeće:

Dodavanje: Zbraja izlaz prethodnog sloja sa svojim originalnim ulazom (rezidualna veza).

Normalizacija: Primjenjuje normalizaciju sloja na rezultat dodavanja kako bi se osigurala stabilnost i konzistentnost tijekom treniranja.

Na kraju imamo softmax sloj koji je prikladan kad imamo zadatke klasifikacije kako bi logite pretvorili u vjerojatnosti.

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?

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

Odgovor je ne. Ne možemo. Zašto? Zato što smo rekli da je mehanizam pažnje koja prati ovisnost između parova tokena

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.