SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Primjena transformer modela za klasifikaciju slika

Rej Šafranko
Voditelj: *prof. dr. sc. Siniša Šegvić*

Zagreb, siječanj 2023.

SADRŽAJ

1. Uvod

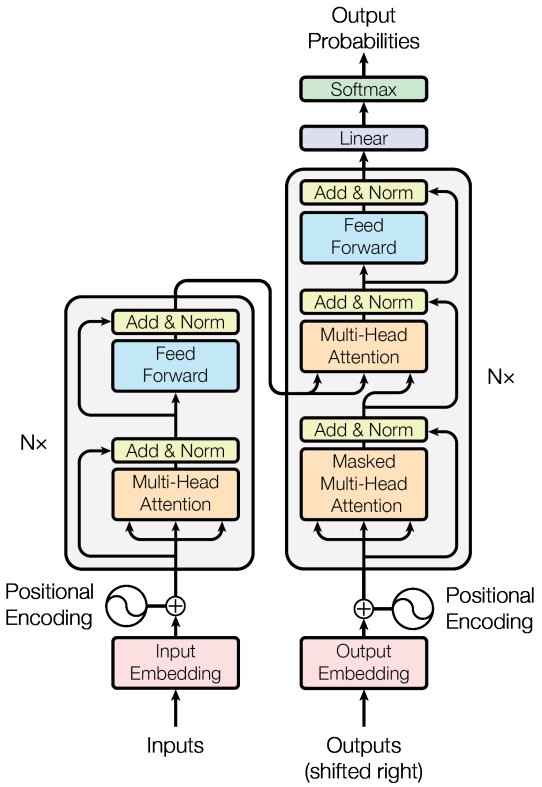
Vrhunski rezultati u obradi prirodnog jezika ostvaruju su uporabom transformer modela. Postavlja se pitanje mogu li se transformeri značajno primjeniti u računalnom vidu. Preciznije, je li moguće ostvariti rezultate u zadatku klasifikacije slika koji se mogu mjeriti sa rezultatima konvolucijskih neuronskih mreža, koje dominiraju područjem računalnog vida (1)?

Primjena računalnog vida u visokorizičnim situacijama zahtjeva razvoj sustava koji su garantirano robusni na nepredvidive promjene (ablacije) u ulaznim podacima. Sustavi računalnog vida griješe u klasifikaciji kad su izloženi kontradiktornim napadima (staklo, grafiti, odjeća, naljepnice). Obranu od kontradiktornih napada je teško ocjeniti jer se napade može prilagoditi i zakomplicirati. Ovo dovodi do potrebe za sustavima koji su robusni na napade bez empirijske evaluacije. Uporaba transformera za vid je jedno od mogućih rješenja ovog problema (2).

U ovom seminarskom radu istražit ću arhitekturu transformer modela, prilagodbu modela za zadatke računalnog vida te konkretan problem ostvarivanja garantirane robusne klasifikacije slika uporabom transformera za vid. Na kraju ću reproducirati rezultate eskperimenta iz (2) na skupu podataka CIFAR-10 (3).

2. Transformer model

Transformeri su trenutno dominantni modeli u području obrade prirodnog jezika koji modeliraju sekvence. Prije se taj zadatak obavljao konvolucijskim i povratnim modelima izvedenim kao koder-dekoder arhitekture. Transformeri se također zasnivaju na koder-dekoder arhitekturi, ali ne koriste ni konvolucijske slojeve ni povratne veze za modeliranje sekvenci. Koriste mehanizam pažnje za modeliranje ovisnosti dijelova sekvenci, neovisno o njihovoj udaljenosti unutar sekvence (globalno receptivno polje) (4).



Slika 2.1: Arhitektura transformer modela (4)

2.1. Koder-dekoder arhitektura

Koder modela se sastoji od 6 identičnih slojeva, a svaki sloj ima 2 podsloja. Prvi podsloj čini mehanizam pažnje s više glava, a drugi čini pozicijska potpuno povezana unaprijedna neuronska mreža. Svaki podsloj ima rezidualnu konekciju (5) koja, zajedno sa izlazom tog podsloja, ulazi u normalizacijski sloj (6).

Dekoder modela je građen kao i koder, no sadrži treći podsloj koji izvodi mehanizam pažnje s više glava nad izlazom kodera modela. Koriste se rezidualne konekcije za svaki podsloj na isti način kao i kod kodera. Mehanizam pažnje kod dekodera je prilagođen tako da na izlaz modela za neku poziciju sekvence mogu utjecati samo izlazi prijašnjih pozicija te sekvence (4).

2.2. Mehanizam pažnje

Pažnja je funckija koja mapira vektore upita Q, ključa K i vrijednosti V sa izlaznim vektorom.

U dekoderu vektor Q dolazi od prijašnjih slojeva dekodera, a vektori K i V dolaze od izlaza kodera. Ovo omogućuje da pozicija sekvence u dekoderu "obraća pažnju" na ostale (prijašnje) pozicije u sekvenci (4).

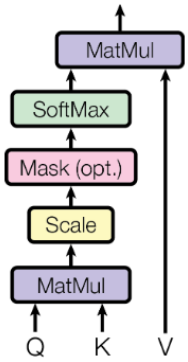
U koderu vektori Q, K i V dolaze iz prijašnjeg sloja kodera. Pozicija sekvence u koderu "obraća pažnju" na pozicije sekvence u svim prijašnjim slojevima kodera (4).

Ulaz u funkciju pažnje čine vektori Q i K dimenzije d_k te vektor V dimenzije d_v. Potrebno je izračunati softmax skalarnog produkta između Q i K podijeljenog sa $\sqrt{d_k}$. Na kraju se izračunaju težine kao skalarni produkt izlaza softmaxa i vektora V.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

Ovaj postupak se može provesti više puta. Vektori Q, K i V dimenzije d_{model} se preslikaju (sa različitim i naučenim linearnim preslikavanjima) h puta na dimenzije d_k, d_k i d_v. Sada se nad izlazom svakog preslikavanja paralelno provodi funkcija pažnje koja daje izlazne vektore (glave) dimenzije d_v. Glave se konkateniraju i ponovno preslikavaju što rezultira konačnim izlazom.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^0 \tag{2.2}$$



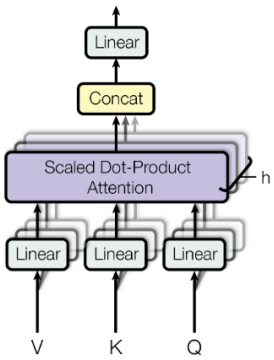
Slika 2.2: Funkcija pažnje kao skalarni produkt (4)

gdje je

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.3}$$

a preslikavanja su matrice parametara $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$ i $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ (4).

Opisani postupak se naziva mehanizam pažnje s više glava.



Slika 2.3: Mehanizam pažnje s više glava (4)

2.3. Pozicijsko kodiranje

Pošto model ne sadrži ni povratne veze ni konvolucijske slojeve, potrebno je unjeti informacije o relativnoj ili apsolutnoj udaljenosti pozicija u sekvenci kako bi model mogao iskoristiti poredak tokena u sekvenci (4). Prije ulaska u koder ili dekoder, ulazna kodiranja se sumiraju sa pozicijskim kodiranjima. Ulazna i pozicijska kodiranja su jednake dimenzije d_{model}.

Postoje razna pozicijska kodiranja koja mogu biti ili naučena ili fiksna (7). U radu (4) koji predstavlja transformer model, korištena su trigonometrijska kodiranja različitih frekvencija. Pretpostavka je da će odabirom ovih funkcija model lakše naučiti relativne odnose pozicija u sekvenci (4).

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{2.4}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{2.5}$$

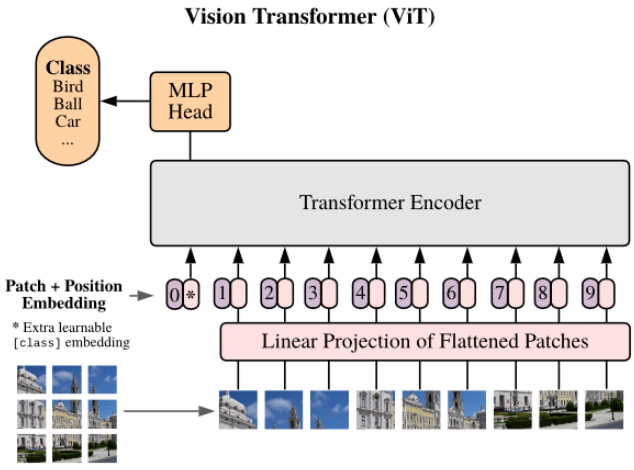
gdje pos predstavlja poziciju u sekvenci, a i dimenziju pozicijskog kodiranja.

3. Transformer model za vid

U području računalnog vida dominantni modeli su konvolucijske neuronske mreže (1). Nakon uspjeha transformer modela u obradi prirodnog jezika, nekoliko radova je pokušalo povezati konvolucijske arhitekture i mehanizam pažnje (8; 9). To je inspiriralo razvoj transformer modela za vid koji ima minimalne promjene u odnosu na standardni transformer model (1).

Arhitektura transformera za vid je u skladu sa arhitekturom standardnog transformer modela. Razlike su u tome što je transformer za vid prilagođen za rad sa slikama. S toga se najveća razlika očekuje na ulazu modela, a minimalna razlika u koder-dekoder arhitekturi.

Slika se podijeli u isječke fiksne veličine. Isječci se linearno preslikaju u kodiranja. Kodirani isječci se sumiraju sa pozicijskim kodiranjima. Dobivena sekvenca vektora čini ulaz u model. Da bi radili klasifikaciju, potrebno je dodati klasifikacijski token u sekvencu.



Slika 3.1: Transformer model za vid (1)

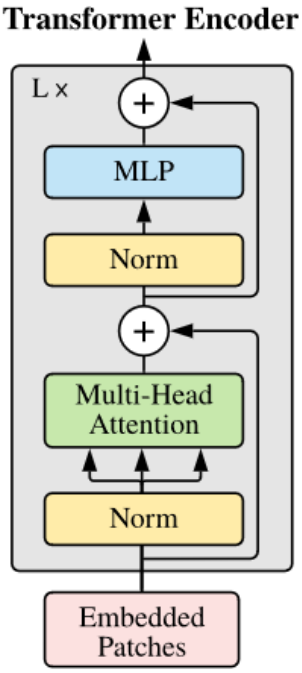
3.1. Ulaz u model

Standardni transformer model prima ulaz kao 1D sekvencu tokena (ulazno kodiranje). Kako bi model mogao raditi sa slikama (2D ulazima), potrebno je promijeniti dimenzije slika. Slika $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ se pretvori u sekvencu $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$koja sadrži 2D isječke slike. (H, W) predstavlja rezoluciju originalne slike, a C je broj kanala. $(P, P$ je rezolucija svakog 2D isječka originalne slike, a $N = HW/P^2$ je broj isječaka. N je ujedno i duljina ulazne sekvence.

Model radi sa vektorima dimenzije D kroz sve slojeve pa se isječci izravnaju i mapiraju na na D dimenzija sa naučenim linearnim preslikavanjem. Izlazi preslikavanja zovu se kodirani isječci.

Pozicijska kodiranja sumiraju se sa kodiranim isječcima radi održavanja prostorne informacije. Također, u sekvencu se ubacuje klasifikacijski token x_{xlass} koji se može naučiti (1).

3.2. Koder modela



Slika 3.2: Koder transformer modela za vid (1)

Koder modela sadrži alternirajuće slojeve mehanizma pažnje s više glava i MLP blokova. Normalizacija sloja se primjenjuje prije svakog bloka, a rezidualne konekcije se

dodaju izlazu svakog bloka (1).

3.3. Induktivna pristranost modela

Bitno je napomenuti da je induktivna pristranost specifična za slike manja kod transformer modela za vid nego kod konvolucijskih neuronskih mreža. Kod konvolucijskih mreža lokalnost i struktura 2D susdjedstva su prisutni kroz sve slojeve modela. Kod transformer modela za vid, lokalnost je prisutna samo kod MLP slojeva, dok su slojevi mehanizma pažnje globalni. Struktura 2D susjedstva je prisutna samo na početku kod podjele slike na isječke. Inicijalna pozicijska kodiranja ne nose nikakvu informaciju od 2D susjedstvu, već se ona mora naučiti (1).

Jedno alternativno formiranje ulazne sekvence je korištenje mapi značajki konvolucijskih neuronskih mreža umjesto isječaka originalne slike. Ovakav hibridni model bi mogao imati induktivnu pristranost više specifičnu za slike.

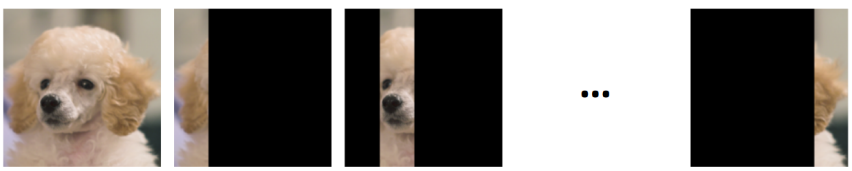
4. Certificirana robustnost klasifikacije slika

Sustavi računalnog vida koriste se u situacijama koje su visokog sigurnosnog rizika (npr. cestovni i zračni promet) te moraju biti pouzdani u svakom mogućem scenariju (2). Jedan od glavnih problema u računalnom vidu su smetnje u podacima. To može biti neka klasa koju model još nije vidio (izvandistribucijska klasa) ili neprijateljski primjeri. Neprijateljski primjeri su skup pojava koje zavaraju sustav računalnog vida (staklo, grafiti, naljepnice, odjeća) te sustav čini pogrešku u klasifikaciji. Jednostavan primjer neprijateljskog primjera je čovjek na cesti u odjeći koja se uklapa sa okolinom ceste.

Aktivna tema istraživanja u računalnom vidu je razvoj sustava koji su garantirano otporni, tj. robusni, na neprijateljske primjere. Za takve sustave se kaže da su certificirano robusni. Jedan način postizanja certificirano robusne klasifikacije je korištenje transformera za vid kao osnovicu za obranu zaglađivanjem. Promatrana obrana zaglađivanjem se zove derandomizirano zaglađivanje i temelji se na predikcijama klasifikatora na ablacijama ulazne slike (2).

4.1. Ablacije slike

Ablacije slike su varijacije slike kod kojih je većinski dio slike maskiran. Postoje stupčaste ablacije i ablacije u blokovima. Dalje razmatram samo stupčaste ablacije.



Slika 4.1: Stupčaste ablacije (2)

Kod stupčastih ablacija, nemaskirani dio slike je fiksne veličine. Za sliku **x** rezolucije $h \times w$, skup svih mogućih stupčastih ablacija širine b je $S_b(\mathbf{x})$. Stupčasta ablacija može početi na bilo kojoj poziciji od njih w i može biti omotana oko slike što znači da je w ukupni broj ablacija u $S_b(\mathbf{x})$ (2).

4.2. Derandomizirano zaglađivanje

Derandomizirano zaglađivanje je jedan od načina obrane zaglađivanjem koji konstruira zaglađeni klasifikator. Taj klasifikator se sastoji od dva glavna dijela: baznog klasifikatora i skupa ablacija slike $S_b(\mathbf{x})$. Bazni klasifikator se zaglađuje tim skupom ablacija. Rezultirajući zaglađeni klasifikator vraća najčešću predikciju baznog klasifikatora na skupu ablacija $S_b(\mathbf{x})$ (2).

4.2.1. Zaglađeni klasifikator

Za ulaznu sliku x, skup ablacija $S_b(\mathbf{x})$ te bazni klasifikator f, zaglađeni klasifikator g se definira kao:

$$g(\mathbf{x}) = \operatorname*{argmax}_c n_c(\mathbf{x}) \tag{4.1}$$

gdje je

$$n_c(\mathbf{x}) = \sum_{\mathbf{x}' \in S_b(\mathbf{x})} \mathbb{I} f(\mathbf{x}') = c \tag{4.2}$$

broj ablacija slike koje su klasificirane kao klasa c. Skup slika koje je zaglađeni klasifikator točno klasificirao zove se standardna preciznost (2).

4.2.2. Certificirana robusnost zaglađenog klasifikatora

Zaglađeni klasifikator je certificirano robusan za ulaznu sliku ako je broj točno klasificiranih ablacija najčešće klase veći od broja točno klasificiranih ablacija druge najčešće klase za dovoljno veliku marginu. Velika margina onemogućava neprijateljskom primjeru da promjeni predikciju zaglađenog klasifikatora jer jedan neprijateljski primjer može biti pristuan samo na ograničenom broju ablacija (2).

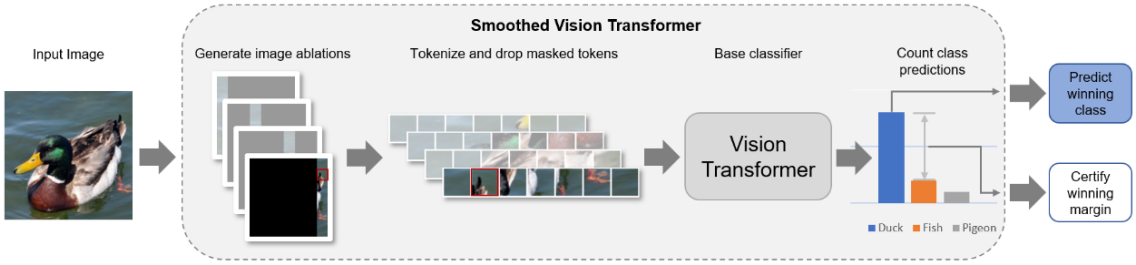
Formalno, neka je $\triangle$ maksimalni broj ablacija u skupu ablacija $S_b(\mathbf{x})$ koje jedan neprijateljski primjer može obuhvatiti istovremeno (za stupčaste ablacije širine b, neprijateljski primjer rezolucije $m \times n$ može obuhvatiti najviše $\triangle = m + b - 1$ ablacija).

Onda je zaglađeni klasifikator certificirano robusan na ulazu x ako za predviđenu klasu c vrijedi:

$$n_c(\mathbf{x}) > \max_{c' \neq c} n_{c'}(\mathbf{x}) + 2\triangle \tag{4.3}$$

Ako se zadovolji ovaj uvjet, najčešća klasa će garantirano ostati nepromijenjiva čak i ako neprijateljski primjer kompromitira svaku ablaciju koju obuhvati (2).

4.3. Zaglađeni transformer za vid



Slika 4.2: Zaglađeni transformer za vid (2)

Transformer za vid posjeduje 2 svojstva koja ga čine pogodnim za derandomizirano zaglađivanje:

1. Obrađuje sliku kao sekvencu isječaka te slike. Iz toga slijedi da može odbaciti nepotrebne isječke iz sekvence i ignorirati veće dijelove slike.
2. Mehanizam pažnje u modelu dijeli informacije globalno kroz sve slojeve. To je pogodno za klasifikaciju ablacija jer će model pridati više pažnje manjem nemaskiranom dijelu slike.

Korištenjem transformer modela za vid kao bazni klasifikator te zaglađivanjem istog ablacijama, dobiva se zaglađeni transformer za vid.

Korisno je povući poveznicu sa konvolucijskim neuronskim mrežama kako bi se opravdala uporaba transformera. Razlika je što konvolucijske mreže ne obrađuju sliku kao sekvencu isječaka pa se receptivno polje, koje je lokalno, mora postepeno izgraditi. To povlači da konvolucijske mreže moraju obraditi i maskirane dijelove slike koje transformer ignorira. Sada je intuitivno jasnije da je transformer model za vid pogodniji odabir za postizanje certificirane robusnosti klasifikacije slika.

5. Reproduciranje rezultata

U ovom poglavlju prikazujem ishod reprodukcije rezultata rada (2) koji predstavlja zaglađeni transformer za vid za postizanje certificirane robusnost klasifikacije.

5.1. Programska implementacija

Kako bih reproducirao rezultate rada (2), prilagodio sam originalni Python kod autora kao Jupyter bilježnicu. Korišteni radni okviri su Pytorch i MadryLab Robustness. Jupyter bilježnicu pokrećem na platformi Google Colaboratory koja nudi besplatan pristup grafičkoj procesnoj jedinici NVIDIA Tesla T4.

Treniram ViT-T (2) model na skupu podataka CIFAR-10 (3). Parametri treniranja su sljedeći:

- Broj epoha: 30
- Stopa učenja: 0.01
- Propadanje težina: $5 \cdot 10^{-4}$
- Korak učenja: 10
- Veličina grupe: 128
- Vrsta ablacija: stupčasta
- Širina ablacija: 4

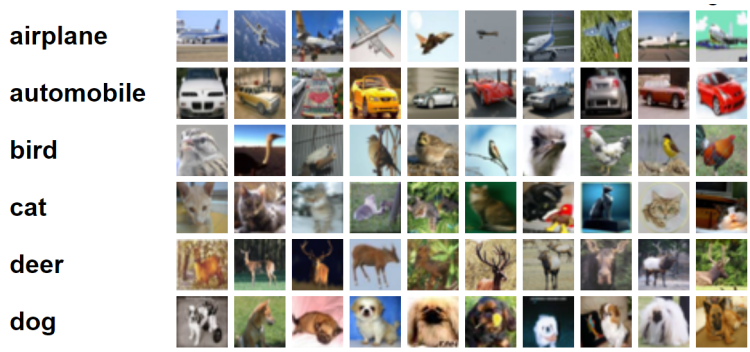
Nakon treniranja treba certificirati robusnost modela. Parametri certifikacije su:

- Veličina grupe: 128
- Vrsta ablacija: stupčasta
- Širina ablacija: 4
- Veličina neprijateljskih primjera: 5

5.2. Skup podataka CIFAR-10

CIFAR-10 sadrži 60 000 RGB slika rezolucije 32×32. Slike su podijeljene u 10 klasa te svaka klasa ima 6000 slika. CIFAR-10 se dijeli na skup za učenje od 50 000 slika i skup za ispitivanje od 10 000 slika. Skup za ispitivanje se sastoji od po 1000 slučajno odabranih slika za svaku od 10 klasa (3).

Klase u skupu podataka su sljedeće: avion, automobil, ptica, mačka, pas, žaba, konj, brod, kamion.



Slika 5.1: 10 slučajno odabranih slika za svaku klasu iz CIFAR-10 (3)

5.3. Rezultati

Rezultati prikazani u radu (2) predstavljaju uprosječene vrijednosti metrika. Eksperiment je ponovljen 50 puta. Metrike kojima se evaluira zaglađeni model su: standardna točnost, zaglađena točnost i certificirana točnost.

Standardna točnost	Zaglađena točnost	Certificirana točnost
85.53	85.15	58.5

Rezultati prikazani u radu (2)

Nakon reprodukcije ostvario sam sljedeće rezultate. Metrike nisu uprosječene pa rezultati predstavljaju ishod jednog eksperimenta.

Standardna točnost	Zaglađena točnost	Certificirana točnost
85.36	85.43	58.2

Rezultati postignuti reprodukcijom

6. Zaključak

Konvolucijske i povratne neuronske mreže dominiraju područjem računalnog vida, dok transformer model dominira područjem obrade prirodnog jezika. Uporaba transformer modela u računalnom vidu je i dalje ograničena. Međutim, postoje problemi u kojima je transformer model primjenjiv te parira dominantim modelima u računalnom vidu. U prikazanom problemu certifikacije robusnosti modela transformer model se intuitivno pokazuje kao dobro rješenje za ablacije slika. Arhitektura transformera pojednostavljuje učenje na ablacijama ignoriranjem maskirnaih tokena. Također, model je brži u predikcijama od dominantnih modela računalnog vida. Očekujem da će razvoj transformera za vid pridonijeti boljim rezultatima certificirane robusnosti u budućnosti. Transformeri definitivno imaju mjesto u računalnom vidu, no zasad su uglavnom ograničeni na klasifikaciju slika. Zanimljivo je promatrati njihovu primjenu u npr. detekciji objekata i segmentaciji slike.

7. Literatura

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

[2] Hadi Salman, Saachi Jain, Eric Wong, and Aleksander Madry. Certified patch robustness via smoothed vision transformers. 2021.

[3] Alex Krizhevsky. CIFAR-10.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[7] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning, 2017.

[8] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks, 2017.

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

8. Sažetak

Kroz ovaj seminarski rad opisan je transformer model kao dominantan model u obradi prirodnog jezika. Detaljno je opisana arhitektura modela. Opisan je i transformer model za vid, njegove prilagodbe za rad sa slikama i razlike u odnosu osnovni transformer model. Nadalje, prikazana je primjena transformer modela za vid u konkretnom problemu računalnog vida. Opisan je problem ceritificirane robusnosti modela za klasifikaciju slika kao rješenje za visokorizične situacije u praksi. Opisan je zaglađeni transformer za vid koji se koristi za rješavanje spomenutog problema. Prikazani su rezultati zaglađenog transformera za vid na skupu podataka CIFAR-10 koji su predstavljeni u radu (2). Na kraju je prikazana i reprodukcija tih rezultata.