SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Primjena transformatora za automatsko generiranje opisa slika

Petra Ilić
Voditelj: *Zoran Kalafatić*

Zagreb, svibanj 2022.

SADRŽAJ

1. Uvod

Automatsko generiranje opisa slika kombinira računalni vid (engl. *Computer Vision*, CV) i obradu prirodnog jezika (engl. *Natural Language Processing*, NLP) pretvarajući slijed elemenata slike u slijed riječi. Osim prepoznavanja objekata i scene, ovaj proces zahtjeva i razumijevanje njihovog stanja i međusobnog odnosa te generiranje sintaktički i semantički ispravnog opisa. U početku je u rješavanju ovog problema većinom korištena koder-dekoder arhitektura s konvolucijskom neuronskom mrežom kao koderom (engl. *Convolutional Neural Networks*, CNN) i povratnom neuronskom mrežom kao dekoderom (engl. *Recurrent Neural Networks*, RNN). Pokazalo se da je ovakva arhitektura osjetljiva na duge ulazne sljedove jer se kod dugotrajne ovisnosti pojavljuje problem nestajućeg gradijenta. Također, kompresija cijele slike u statičku reprezentaciju može dovesti do gubitka bitnih informacija. Razvoj mehanizma pozornosti koji omogućuje istaknutim značajkama da dođu u prvi plan i kodira ulazni niz u slijed vektora, rezultirao je naprednijim generatorima opisa slika [1]. Google Brain je 2017. godine objavio rad *Attention is all you need* [2] u kojem je opisan model transformatora koji ostvaruje izvrsne rezultate u strojnom prevođenju. Od tada su razvijene brojne varijante transformatora primjenjive u multimodalnom kontekstu kao što je automatsko generiranje opisa slika. Primjeri su radovi *Captioning Transformer with Stacked Attention Modules* [3], *Image Caption Generation With Adaptive Transformer* [4] i *Entangled Transformer for Image Captioning* [5]. Jedna od najistaknutijih varijanti transformatora za rješavanje problema automatskog generiranja teksta je transformator mrežaste memorije ili M2 transformator [6]. M2 transformator ostvaruje *state-of-the-art* rezultate na skupu podataka COCO [7]. U nastavku je opisana arhitektura i način rada originalnog transformatora [2] te M2 transformatora [6].

2. Arhitektura i način rada transformatora

U nastavku slijedi opis transformatora predstavljenog u radu *Attention is all you need* [2]. Ovaj model se primjenjuje u neuralnom strojnom prevođenju gdje se na ulazu i izlazu nalazi tekst. Prilagodba transformatora za generiranje teksta iz slike bit će opisana kasnije u radu.

Transformator se sastoji od kodera i dekodera. Koder i dekoder tvore slojevi samopozornosti i potpuno povezani slojevi. Koder se sastoji od šest identičnih slojeva od kojih svaki ima dva podsloja. Prvi podsloj je mehanizam višestruke pozornosti (engl. *multi-head attention*), a drugi potpuno povezani sloj. Dekoder se također sastoji od šest slojeva. Uz dva navedena podsloja, ima i podsloj mehanizma maskirane višestruke pozornosti. U nastavku je detaljnije opisan sloj samopozornosti.

2.1. Samopozornost

Samopozornost je operacija slijed-u-slijed koja iz ulaznih vektora $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ generira odgovarajuće izlazne vektore $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$. Izlazni vektor $\boldsymbol{y}_i$ dobiva se težinskim prosjekom ulaznih vektora:

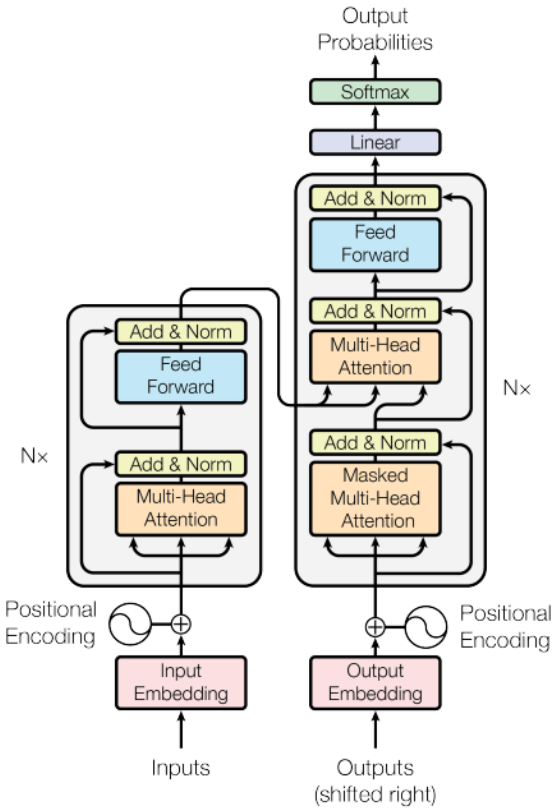
$$\boldsymbol{y}_i = \sum_j w_{ij} \boldsymbol{x}_j \tag{2.1}$$

Težine se dobivaju skalarnim produktom između ulaznih vektora:

$$w_{ij} = \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{2.2}$$

Skalarni produkt je mjera sličnosti dva ulazna vektora. Kako bi težine bile vrijednosti između 0 i 1, primjenjuje se funkcija softmax:

$$w_{ij} = \frac{\exp w_{ij}}{\sum_j \exp w_{ij}} \tag{2.3}$$



Slika 2.1: Arhitektura transformatora [2]

Tri važna elementa mehanizma samopozornosti su upiti (engl. *Queries*) Q, ključevi (engl. *Keys*) K i vrijednosti (engl. *Values*) V. Ova tri elementa predstavljaju tri uloge ulaznog vektora. Upiti se dobivaju uspoređivanjem svakog ulaznog vektora $\boldsymbol{x}_i$ s ostalim ulaznim vektorima kako bi se dobile težine w_{ij} za izlaz $\mathbf{y}_i$. Ključevi se dobivaju uspoređivanjem vektora $\boldsymbol{x}_i$ s drugim vektorima kako bi se dobile težine za izlaze drugih vektora. Vrijednosti se dobivaju množenjem odgovarajućih težina s vektorom $\mathbf{x}_i$. Uvode se matrice W_q, W_k i W_v. Matrice sadrže parametre koji se mogu optimizirati. Postupak se odvija prema sljedećim formulama:

$$\boldsymbol{q}_i = W_q \boldsymbol{x}_i \tag{2.4}$$

$$\boldsymbol{k}_i = W_k \boldsymbol{x}_i \tag{2.5}$$

$$\boldsymbol{v}_i = W_v \boldsymbol{x}_i \tag{2.6}$$

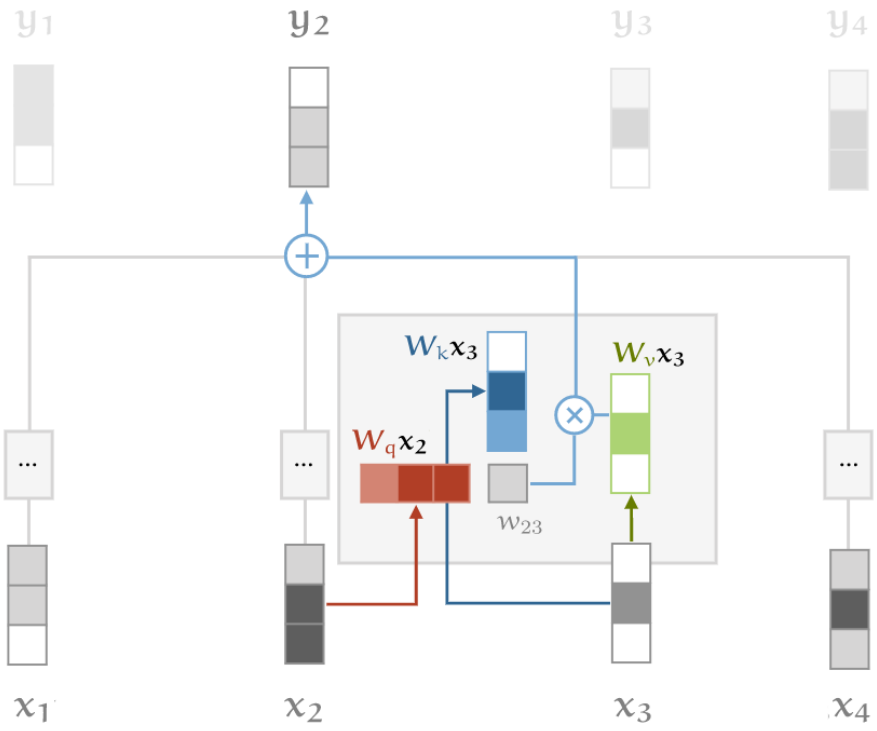
$$w_{ij} = \boldsymbol{q}_i^T \boldsymbol{k}_j \tag{2.7}$$

$$w_{ij} = softmax(w_{ij}) \tag{2.8}$$

$$\boldsymbol{y}_i = \sum_j w_{ij} \boldsymbol{v}_j \tag{2.9}$$

Funkcija softmax loše radi s velikim vrijednostima i rezultira nestajućim gradijentima. Stoga se produkt skalira dijeljenjem vrijednošću $\sqrt{d_k}$ gdje je d_k dimenzija matrice ključeva. Postupak se može sažeti sljedećom formulom:

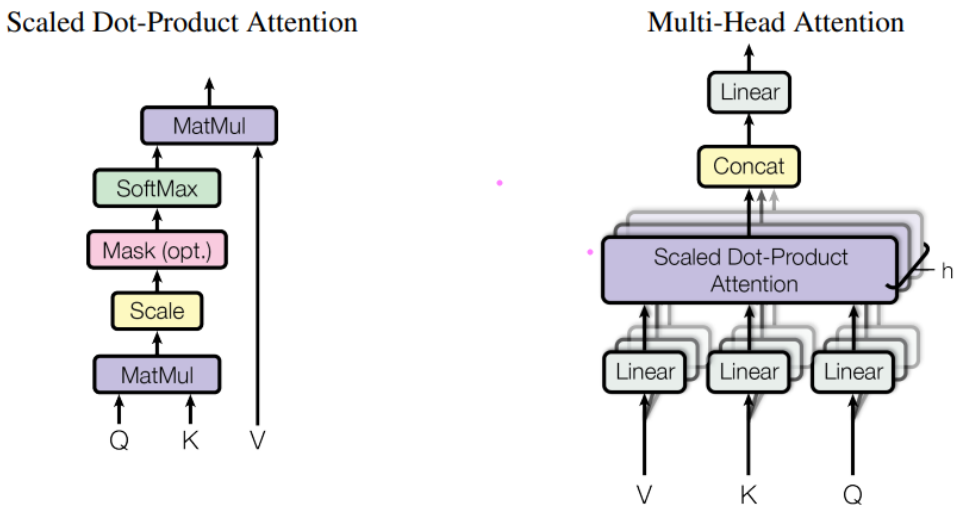
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.10}$$



Slika 2.2: Mehanizam samopozornosti uz upite, ključeve i vrijednosti [8]

2.2. Višestruka samopozornost

Jedna riječ može imati različito značenje u kombinaciji s drugim riječima. U prethodno opisanoj operaciji samopozornost sve informacije se sumiraju. U rečenicama *"John gave a present to Susan"* i *"Susan gave a present to Susan"* izlaz y_{gave} bi bio isti i ne bi se moglo razlikovati tko je kome dao dar. Rješenje je koristiti više matrica upita, ključeva i vrijednosti. Za svaki ulazni vektor $\boldsymbol{x}_i$ dobiva se više izlaznih vektora $\boldsymbol{y_i}$. Dobiveni vektori se konkateniraju i prolaze kroz linearnu transformaciju kako bi se smanjila dimenzionalnost.



Slika 2.3: Višestruka samopozornost [2]

Budući da transformator nema povratne veze niti konvolucije, redoslijed riječi ne utječe na izlaz. Permutirane rečenice na ulazu daju isti rezultat. Kako bi se iskoristio redoslijed, potrebno je dodati pozicijsko kodiranje u *embedding* matricu. Pozicijsko kodiranje svakom elementu ulaza određuje poziciju i može biti fiksirano ili naučeno.
Transformator koristi višestruku pozornost na tri načina:

- U koder-dekoder slojevima pozornosti, upiti dolaze iz prethodnog sloja u dekoderu, a ključevi i vrijednosti iz izlaza kodera. Ovako sve pozicije u dekoderu vide sve pozicije u ulaznom nizu.
- U sloju samopozornosti u koderu svi upiti, ključevi i vrijednosti dolaze iz izlaza kodera. Svaka pozicija u koderu vidi sve pozicije u prethodnom sloju kodera.
- Slično, sve pozicije u dekoderu vide sve pozicije u prethodnom sloju dekodera.

3. Transformator s mrežastom memorijom

Transformator s mrežastom memorijom ili M2 transformator [6] inspiriran je radom *Attention is all you need* [2]. Prilagođen je za rješavanje problema automatskog generiranja opisa slika. Ulaz kodera su značajke slike dobivene primjenom konvolucijske mreže. M2 transformator uvodi dvije ključne novosti: koder s proširenom memorijom i mrežasti dekoder. Koder procesira regije ulazne slike i konstruira veze između njih, a dekoder čita izlaze svih slojeva kodera i generira izlaz riječ po riječ.

3.1. Koder s proširenom memorijom

Operacija samopozornosti opisana izrazom:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.1}$$

može se koristiti kako bi se iz regija slike $\boldsymbol{X}$ dobilo kodiranje invarijantno na permutacije. Izlaz operacije je novi skup elemenata iste dimenzionalnosti kao $\boldsymbol{X}$. Svaki element u $\boldsymbol{X}$ zamijenjen je težinskom sumom vrijednosti. Operator samopozornosti kodira veze između regija slike, ali ne može modelirati apriorno znanje o vezama pojedinih regija slike. Primjerice, ako jedna regija predstavlja psa, a druga mačku, teško je zaključiti da pas lovi mačku bez apriornog znanja. Kako bi se riješio ovaj problem, autori M2 transformatora predlažu koder s proširenom memorijom. Skup ključeva i vrijednosti proširenim je dodatnom memorijom koja služi za kodiranje apriornog znanja. Budući da apriorno znanje ne ovisi o ulaznom skupu $\boldsymbol{X}$, dodatni ključevi i vrijednosti implementirani su kao vektori koji se mogu naučiti pomoću stohastičkog gradijentnog spusta. Novi operator može se prikazati izrazima:

$$M_{mem}(\boldsymbol{X}) = Attention(W_q\boldsymbol{X}, \boldsymbol{K}, \boldsymbol{V}) \tag{3.2}$$

$$\boldsymbol{K} = [W_k\boldsymbol{X}, \boldsymbol{M}_k] \tag{3.3}$$

$$\boldsymbol{V} = [W_v \boldsymbol{X}, \boldsymbol{M}_v], \tag{3.4}$$

gdje M_k i M_v predstavljaju dodatnu memoriju. Skup upita ostaje isti. Operator pozornosti s proširenom memorijom primjenjuje se h puta, svaki put s različitim matricama W_q, W_k i W_v, kao i M_k i M_v. Izlazi se konkateniraju i primjenjuje se linearna transformacija kako bi se smanjila dimenzionalnost. Izlaz operatora samopozornosti je ulaz potpuno povezanog sloja s dvije afine transformacije i jedne nelinearnosti:

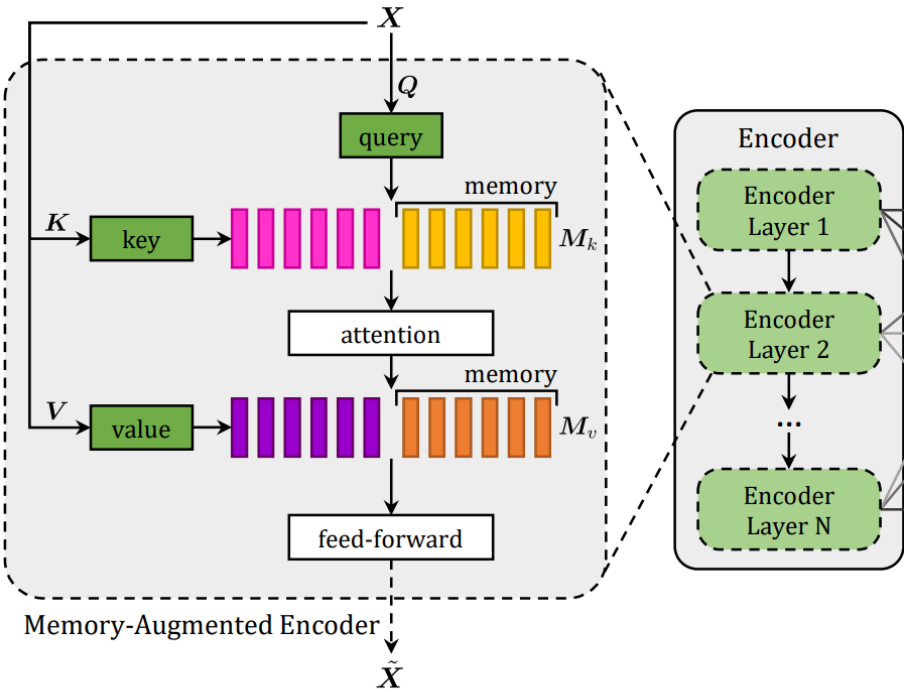
$$F(\boldsymbol{X})_i = U\sigma(V\boldsymbol{X}_i + b) + c, \tag{3.5}$$

gdje $\boldsymbol{X}_i$ predstavlja i-ti vektor ulaznog skupa, $F(\boldsymbol{X})_i$ i-ti vektor izlaza, $\sigma(\cdot)$ aktivacijsku funkciju ReLU, a V, U, b i c parametre. Izlazi oba podsloja se normaliziraju:

$$\boldsymbol{Z} = AddNorm(M_{mem}(\boldsymbol{X})) \tag{3.6}$$

$$\tilde{X} = AddNorm(F(\boldsymbol{Z})) \tag{3.7}$$

Slojevi kodiranja su složeni slijedno, tako da i-ti sloj prima izlaz sloja $i-1$. Ovako viši slojevi mogu rafinirati veze identificirane u prethodnim slojevima.



Slika 3.1: Koder s proširenom memorijom [6]

3.2. Mrežasti dekoder

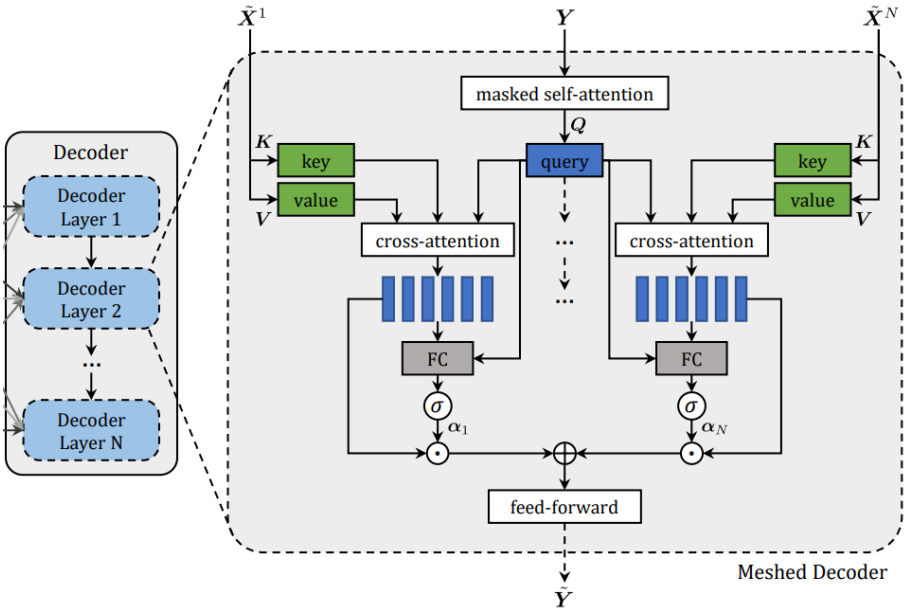
Dekoder originalnog transformatora povezuje ulazni niz vektora $\boldsymbol{Y}$ s izlazom posljednjeg sloja kodera. Dekoder M2 transformatora povezuje ulazni niz vektora sa svim izlazima kodera što omogućuje iskorištavanje značajki nižih i viših razina. Operator pozornosti se primjenjuje na sve izlaze kodera i rezultati se sumiraju:

$$M_{mesh}(\tilde{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \alpha_i \odot C(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y}) \tag{3.8}$$

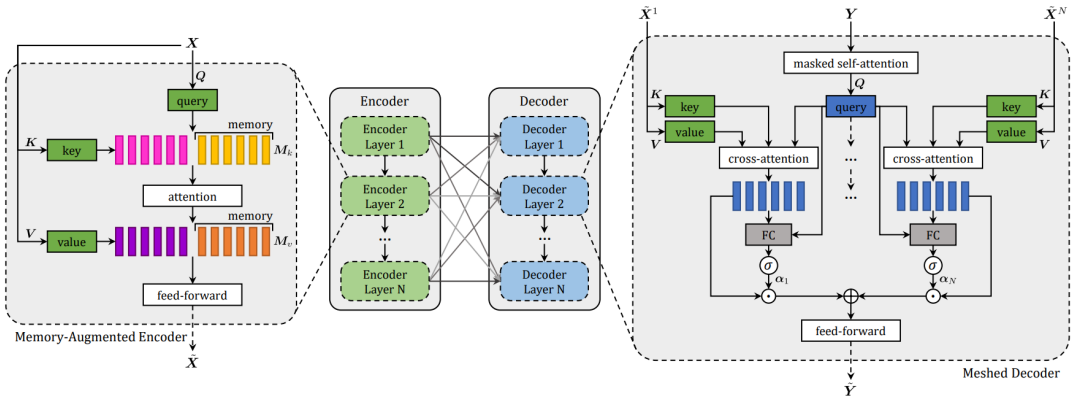
Operacija C predstavlja operaciju pozornosti između izlaza slojeva kodera $\tilde{\boldsymbol{X}}^i$ i ulaza dekodera $\boldsymbol{Y}$ u kojoj se koriste upiti kodera, a ključevi i vrijednosti dekodera. Matrica težina α_i modelira doprinos svakog izlaznog sloja kodera i relativnu povezanost različitih slojeva. Mjeri se značajnošću veze izlaza operacije pozornosti C i ulaznog upita. Može se prikazati izrazom:

$$\alpha_i = \sigma(W_i[\boldsymbol{Y}, C(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y})] + b_i), \tag{3.9}$$

gdje $[\cdot, \cdot]$ označava konkatenaciju, σ sigmoidnu aktivacijsku funkciju, a W_i i b_i težine. Operator pozornosti se kao i kod kodera primjenjuje višestruko. Budući da predikcija riječi mora ovisiti samo o prethodno predviđenim riječima, dekoder koristi maskiranu operaciju pozornosti koja povezuje upite t-tog elementa ulaznog niza $\boldsymbol{Y}$ s ključevima i vrijednostima prethodnih podnizova $\boldsymbol{Y}_{\leq t}$. Nakon sloja pozornosti slijedi potpuno povezani sloj. Izlazi slojeva pozornosti i potpuno povezanih slojeva su normalizirani. Dekoder na ulazu prima vektore koji predstavljaju riječi. Izlazni element u koraku t predstavlja predikciju riječi za korak $t + 1$. Nakon linearne projekcije i operacije softmax, dobivaju se vjerojatnosti riječi u rječniku.



Slika 3.2: Mrežasti dekoder [6]



Slika 3.3: Arhitektura M2 transformatora [6]

4. Zaključak

Razvoj modela transformatora donio je znatan napredak u rješavanju različitih problema, od strojnog prevođenja do automatskog generiranja opisa slika. Transformator se oslanja na mehanizam pozornosti i nema konvolucije i povratne veze kao klasični koder-dekoder modeli za generiranje opisa slika. Razvijene su brojne varijante transformatora za generiranje opisa slika, a jedna od najistaknutijih je M2 transformator. Operator samopozornosti korišten u klasičnom transformatoru kodira veze između regija slike, ali ne može modelirati apriorno znanje. M2 transformator uvodi dvije glavne promjene. Prva je koder s proširenom memorijom koji pomoću dodatnih vektora memorije modelira apriorno znanje. Druga promjena je mrežasti dekoder koji povezuje ulazni niz s izlazima svih slojeva kodera, za razliku od originalnog transformatora gdje je ulazni niz dekodera povezan samo s posljednjim izlazom kodera. M2 transformator ostvaruje *state-of-the-art* rezultate na skupu podataka COCO [7].

5. Literatura

[1] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

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

[3] Xinxin Zhu, Jing Liu, Peng Haipeng, and Xinxin Niu. Captioning transformer with stacked attention modules. *Applied Sciences*, 8:739, 05 2018. doi: 10.3390/app8050739.

[4] Wei Zhang, Wenbo Nie, Xinle Li, and Yao Yu. Image caption generation with adaptive transformer. In *2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 521–526, 2019. doi: 10.1109/YAC.2019.8787715.

[5] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8927–8936, 2019. doi: 10.1109/ICCV.2019.00902.

[6] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning, 2020.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[8] Transformers From Scratch, 2018. URL `http://peterbloem.nl/blog/transformers`.

6. Sažetak

U radu je opisana primjena transformatora u automatskom generiranju opisa slika. U rješavanju ovog problema često je korišten koder-dekoder model s konvolucijskom neuronskom mrežom kao koderom i povratnom neuronskom vezom kao dekoderom. Pokazalo se da ovakav model ima velike nedostatke kada je u pitanju obrada sljedova velike duljine i lako dolazi do gubitka informacija. Transformatori koriste mehanizam pozornosti koji rješava probleme klasične koder-dekoder arhitekture te omogućuje paralelnu obradu svih elemenata ulaznog slijeda što značajno ubrzava cijeli proces. Transformator s mrežastom memorijom ili M2 transformator jedna je od najznačajnijih varijanti transformatora za automatsko generiranje opisa slika. Ostvaruje *state-of-the-art* rezultate na skupu podataka COCO [7]. M2 transformator uvodi koder s dodatnom memorijom za modeliranje apriornog znanja o vezama između regija ulazne slike i dekoder čiji je ulaz povezan sa svim izlazima kodera što omogućuje iskorištavanje značajki nižih i viših razina.