

# Решавање проблема машинског превођења помоћу архитектуре трансформера

Драгана Нинковић, ученица мастер студија Електротехничког факултета у Београду, смер Системи и сигнали

**Абстракт**—Иако је постојало доста модела који су успешно решавали проблем превођења секвенце у секвенцу попут рекурентних неуралних мрежа са разним механизмима пажње, ови модели били су доста сложени и захтевали су много времена за тренирање. Такође, њихова секвенцијална архитектура спречава било коју врсту паралелизације у току тренирања. Трансформери су веома проста архитектура заснована искључиво на механизму пажње и обичним линеарним слојевима без коришћења рекурентних мрежа и осталих сложених слојева. Показује се да дају јако висок BLUE скор у односу на остале моделе, да су јако добри у генерализацији и подржавају паралелизацију. У овом пројекту приказани су резултати добијени коришћењем трансформера за превођење са немачког на енглески језик.

**Главни појмови**—трансформер, машинско превођење, Multi30k скуп података

## I. УВОД У ТРАНСФОРМЕРЕ

Архитектура трансформера је једна од најчешће коришћених архитектура у проблемима који уче да секвенцу на улазу претворе у секвенцу на излазу, а најчешћа примена им је машинско превођење. Након што су се показали као један од најбољих модела за овај проблем, људи су почели да испробавају разне адаптације овог модела за друге области попут препознавања говора, предикције временских серија, и слично. Овај модел је чак допуњен за проблеме машинске визије под именом визијски трансформер где је такође дао добре резултате. Овај модел је коришћен за обраду језика од стране неких од највећих компанија које се баве тим проблемом попут OpenAI-а и DeepMind-а.

## II. ОПИС АРХИТЕКТУРЕ ТРАНСФОРМЕРА

Трансформер користи енкодер-декодер архитектуру при чему оригинални рад користи 6 енкодера и 6 декодера (тј.  $N = 6$ ). Ова архитектура подразумева да енкодер прима секвенцу и од ње прави скривену секвенцу која се шаље на улаз декодера који на излазу даје предикцију излазне секвенце. Граф целе архитектуре приказан је на слици 1. Унутар енкодера прва компонента је ембединг слој који се користи да секвенцу речи претворимо у бројеве јер компјутер не може да ради са текстом. Најчешће се прво свака реч представи вектором димензије речника који се састоји од нула и једне јединице на позицији те речи у речнику а на излазу даје вектор бројева тако да семантички сличне речи буду близу једна другој у векторском простору. Следећа компонента је позициони енкодер који

служи да дода информацију о позицији речи у реченици јер у неким језицима у зависности од позиција речи реченица може да има различито значење. Позициони енкодер мора да испуни три услова: да даје различити излаз за сваки временски тренутак, да даје константну дистанцу између два еквидистантна временска корака, да може да се генерализује за различите дужине реченица и да је у питању детерминистичка функција. Најчешће се користи следећа функција:

$$PE_{(pos, 2i)} \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (1)$$

$$PE_{(pos, 2i+1)} \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (2)$$

која испуњава све наведене услове, и позицију представља вектором а не само једним бројем. Наравно, да би се сабрала са излазом из ембединг слоја мора бити истих димензија да бисмо могли да их саберемо. Следећа компонента је пажња са више глава која уноси у модел информацију о повезаности тренутне речи са осталим речима у реченици тако што за сваку реч имамо три матрице кључ, вредност и упит. За једну реч формирамо самопажњу тако што њен кључ скаларно помножимо са сваки од упита и нађемо softmax функцију ових вредности помножимо вредности свих речи њима одговарајућом softmax вредности и поделимо са кореном величине вектора. Када овај поступак поновимо за сваку од речи добијамо пажњу са више глава причему је неопходно да их објединимо у излаз одговарајућих димензија множењем једном великом матрицом.

$$Attention(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Последњи део енкодера је најобичнији потпуно повезани слој при чему пре сваког од ових делова имамо резидуалне везе као и слој за додавање и нормализовање како не би била заборављена нека од оригиналних информација из претходних слојева и како би било спречено експлодирање или нестајање градијената и учење текло стабилно. Декодер се састоји из истих слојева као енкодер и на самом крају му се додаје линеарни слој и softmax функција како би се добили крајњи излази. Енкодеру поред излаза енкодера који му даје кључ и вредности користи и оригинални очекивани излаз који му даје упите при чему треба да замаскирамо речи од те па надаље да модел не би варао већ да би користио оригиналне преводе речи пре те у току тренирања. На крају, цео модел се састоји од стека више енкодера и декодера при чему се сваки излаз енкодера шаље на улаз свих декодера.

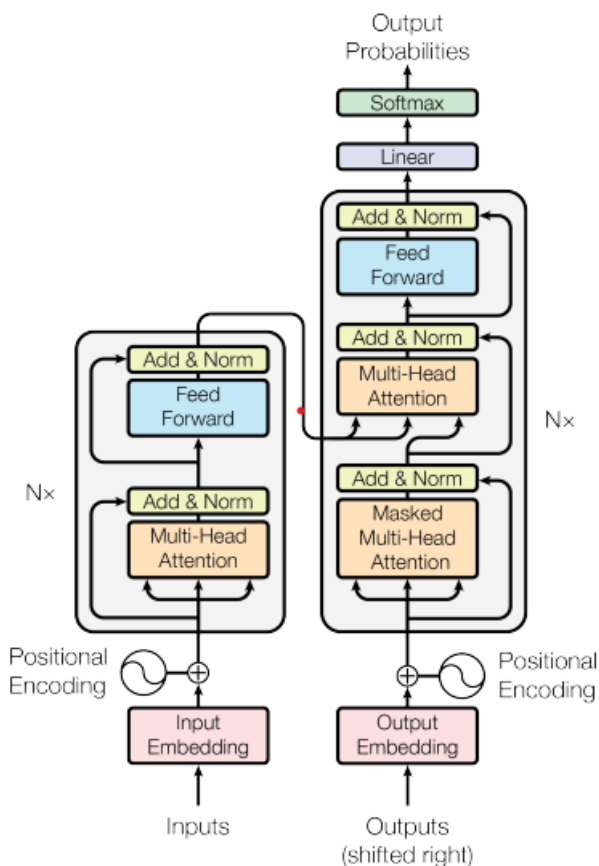


Fig. 1. Архитектура трансформера

### III. РЕЗУЛТАТИ

За овај пројекат коришћен је скуп података Multi30k који је направљен од стране стручних преводиоца и садржи преводе реченица са немачког на енглески и са енглеског на немачки за моделе машинског превођења које представљају описе слика. Постоји додаток на овај скуп података који чине слике чији су то описи за тренирање модела да генерише описе слика али оне нису коришћене у склопу овог пројекта.

Због великог времена тренирања и велике сложености модела коришћена су само три енкодера и декодера. Коришћени параметри модела су: величина ембединг слоја 512, број глава у слоју пажње са више глава је 8, димензија потпуно повезаног слоја 512. За хиперпараметре модела је узета величина пакета 128 и стопа учења 0.0001. Као оптимизатор је коришћен Адам. Модел је трениран укупно 30 епоха, а просечно време потребно за тренирање једне епохе је око 40 миинута.

Приказ валидационе и обучавајуће криве приказан је на слици 2. Можемо приметити да након одређеног броја епоха губитак модела на валидационом скупу креће да расте и да се удаљава од губитка на обучавајућем скупу. Из тог разлога изабран је модел добијен у шеснаестој епохи јер је у њој грешка на валидационом

скупу минимална а након те епохе грешка почиње монотонно да расте.

Губитак овог модела на обучавајућем скупу је 1.904, на валидационом скупу је 1.907, док је на тест скупу 1.86. Тачност као метрика у случају машинског превођења не говори много јер било се изостави једна мање битна реч у преводу или нека која значајно мења превод тачност ће бити иста. Неинтуитивност ове метрике потврђује и чињеница да је тачност на валидационом скупу око 32.7 процената док је на тест скупу 67.6 процената што је дупло више иако валидациони скуп користимо за бирање модела а тест скуп модел никада није видео.

Примери превода на случајно изабраних пет примера из валидационог скупа:

Оригинал: Zwei Hunde beißen spielerisch einen dritten Hund, dessen Zunge heraushängt.

Превод модела: Two dogs are wrestling a third third, with his tongue out.

Превод професионалног преводиоца: Two dogs playfully bite a third dog, which has its tongue sticking out.

Оригинал: Zwei Männer rauchen außerhalb eines Parks.

Превод модела: Two men smoking outside a park.

Превод професионалног преводиоца: Two men smoking outside of a park.

Оригинал: Ein Mann in einem schwarzen T-Shirt und braunen Cargohosen hält eine Schaufel über seinem Kopf.

Превод модела: A man in a black t-shirt and brown shorts is holding a shovel.

Превод професионалног преводиоца: A male in a black shirt and brown cargo pants holding a shovel above his head.

Оригинал: Eine Person sitzt vor einer Menschenmenge auf einem Stuhl.

Превод модела: A person is sitting in a chair in front of a crowd.

Превод професионалног преводиоца: A person sitting on chair in front of a crowd.

Оригинал: Ein Junge steht auf alberne Weise bei einem geselligen Anlass herum.

Превод модела: A boy is standing around a potato sack race at a formal event.

Превод професионалног преводиоца: A boy stands goofily in the middle of a social gathering.

Примери превода на пет случајно изабраних примера из тест скупа:

Оригинал: Zwei Hunde beißen spielerisch einen dritten Hund, dessen Zunge heraushängt.

Превод модела: Two dogs are wrestling a third third , with his tongue out .

Превод професионалног преводиоца: Two dogs playfully bite a third dog, which has its tongue sticking out.

Оригинал: Zwei Arbeiter streichen Zement auf einen Ziegelbau.

Превод модела: Two workers are painting cement on a brick brick wall .

Превод професионалног преводиоца: Two workers spread cement onto a brick building.

Оригинал: Ein kleiner Junge hängt von einer Kleiderstange.

Превод модела: A little boy hangs off a ramp .

Превод професионалног преводиоца: A young boy is hanging onto a clothing rack.

Оригинал: Ein Mann in einem Karateanzug macht eine Vorführung vor zwei Punktrichtern.

Превод модела: A man in a karate uniform is performing a trick in front of two metal pipe .

Превод професионалног преводиоца: A man in a karate is performing in front of two judges.

Оригинал: Zwei Fußballmannschaften auf dem Feld.

Превод модела: Two soccer teams are on the field .

Превод професионалног преводиоца: Two soccer teams are on the field.

Можемо приметити да су реченице на валидационом скупу доста боље преведене као и да је модел доста добро научио сам превод речи али да доста греша граматику и по негде семантику унутар реченице

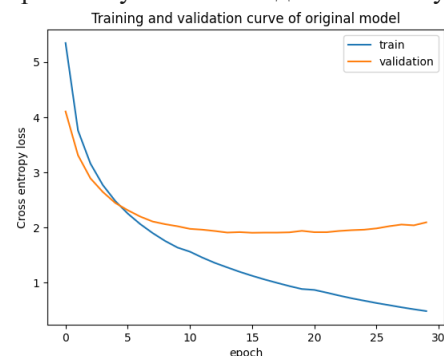


Fig. 2. Обучавајућа и валидациона крива са хиперпараметрима из рада

ТАБЕЛА I

ГРЕШКЕ И СКОРОВИ СА ОРИГИНАЛНИ МОДЕЛ

	скуп	вредност
Губитак	тренинг	0.828
Губитак	валидациони	1.907
Губитак	тест	1.856
BLEU скор	валидациони	0.2885
BLEU скор	тест	0.2848

Након што је успешни обучен модел из рада мењане су вредности хиперпараметара како би се проверио њихов утицај на перформансе модела.

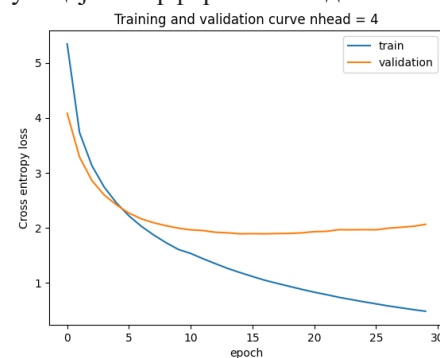


Fig. 3. Обучавајућа и валидациона крива за број глава 4

ТАБЕЛА II

ГРЕШКЕ И СКОРОВИ СА БРОЈ ГЛАВА 4

	скуп	вредност
Губитак	тренинг	0.936
Губитак	валидациони	1.893
Губитак	тест	1.841
BLEU скор	валидациони	0.294
BLEU скор	тест	0.284

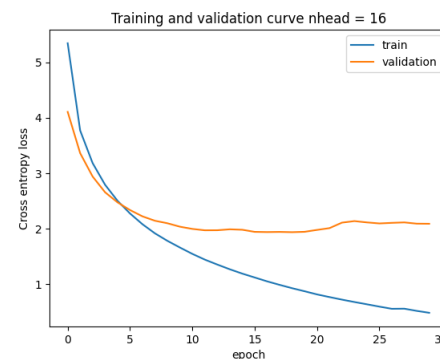


Fig. 4. Обучавајућа и валидациона крива за број глава 16

ТАБЕЛА III

ГРЕШКЕ И СКОРОВИ СА БРОЈ ГЛАВА 16

	скуп	вредност
Губитак	тренинг	0.747
Губитак	валидациони	1.943
Губитак	тест	1.904
BLEU скор	валидациони	0.2846
BLEU скор	тест	0.2766

Из добијених резултата можемо закључити да посматрајући валидациони скуп, број глава 4 даје боље резултате него друга два модела, док са бројем глава 16 долази до највећег преобучавања. Тест скуп јасно потврђује ово тиме што је најмањи скор за последњи модел.

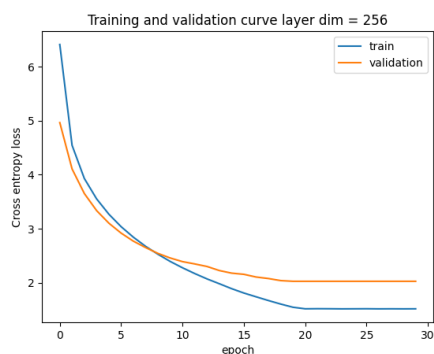


Fig. 5. Обучавајућа и валидациона крива за величину слоја 256

ТАБЕЛА IV  
ГРЕШКЕ И СКОРОВИ ЗА ВЕЛИЧИНУ СЛОЈА 256

	скуп	вредност
Губитак	тренинг	1.275
Губитак	валидациони	2.025
Губитак	тест	1.978
BLEU скор	валидациони	0.2661
BLEU скор	тест	0.2544

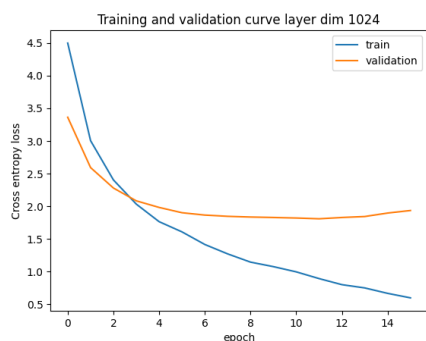


Fig. 6. Обучавајућа и валидациона крива за величину слоја 1024

ТАБЕЛА V  
ГРЕШКЕ И СКОРОВИ ЗА ВЕЛИЧИНУ СЛОЈА 1024

	скуп	вредност
Губитак	тренинг	
Губитак	валидациони	
Губитак	тест	
BLEU скор	валидациони	

BLEU скор	тест	
-----------	------	--

Посматрајући промену величине слојева модела можемо видети да ни повећање ни смањење овог хиперпараметра не даје боље резултате. Смањење доводи до подобравања, а повећавање до преобучавања и из тог разлога је скор на тест и на валидационом скупу у оба случаја мањи.

#### IV. CONCLUSION

Иако можда делује да је проблем машинског преводјења и генерално модела који се баве обрадом језика данас решен јер је доживео до тада незамислив напредак у последњих пар година, постоји још доста простора да се модели попут овог унапреде уз помоћ стручњака из ове области како би модел могао поред генерисања превода који је довољно добар за споразумевање да омогући и семантичку и граматичку коректност.

Добијени резултати су добри узевши у обзир да овај скуп података има само тридесет хиљада реченица док већина скупова података коришћена у ове сврхе има милионе реченица које крећу од знатно једноставнијих реченичних склопова и модел може много лакше и много више из њих да научи. Такође, овај BLEU скор се поклапа са јавно доступним резултатима где је репродукован овај рад над истим скупом података [6] као и над познатим скупом података WMT 2014 [7] са добијеним BLEU скором 0.29.

#### РЕФЕРЕНЦЕ

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," Adv. in Neural Inf. Process. Syst., vol. 30, 2017.
- [2] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German Image Descriptions," Proc. of the 5th Workshop on Vision and Language, Berlin, Germany, pp. 70-74, Aug. 2016.
- [3] Pytorch, "Language translation with nn.transformer and torchtext", PyTorch Tutorials 2.2.0+cu121 documentation
- [4] J. Alamm, The Illustrated Transformer – Jay Alamm – Visualizing machine learning one concept at a time. (jalammar.github.io), 27. Jun 2018.
- [5] A. Kazemnejad, Transformer Architecture: The Positional Encoding - Amirhossein Kazemnejad's Blog, 20. Sep. 2019.
- [6] Multi30K Benchmark (Multimodal Machine Translation) | Papers With Code, 2017
- [7] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. M. Rush, The Annotated Transformer (harvard.edu)