



УНИВЕРЗИТЕТ У НОВОМ САДУ
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА
У НОВОМ САДУ




Петар Купусаревић

МУЛТИМОДАЛНО ГЕНЕРИСАЊЕ ОПИСА СЛИКА ИЗ АКАДЕМСКИХ РАДОВА

ДИПЛОМСКИ РАД
- Основне академске студије -

Нови Сад, 2023.

	УНИВЕРЗИТЕТ У НОВОМ САДУ ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА 21000 НОВИ САД, Трг Доситеја Обрадовића 6	Датум:
	ЗАДАТАК ЗА ИЗРАДУ МАСТЕР РАДА	Лист:
		1/1

(Податке уноси предметни наставник - ментор)

Врста студија:	Основне академске студије
Студијски програм:	Софтверско инжењерство и информационе технологије
Руководилац студијског програма:	проф. др Мирослав Зарић

Студент:	Петар Купусаревић	Број	SW 57/2019
Област:	Електротехничко и рачунарско инжењерство		
Ментор:	Др Јелена Сливка, ванредни професор		

НА ОСНОВУ ПОДНЕТЕ ПРИЈАВЕ, ПРИЛОЖЕНЕ ДОКУМЕНТАЦИЈЕ И ОДРЕДБИ СТАТУТА ФАКУЛТЕТА ИЗДАЈЕ СЕ ЗАДАТАК ЗА ДИПЛОМСКИ РАД, СА СЛЕДЕЋИМ ЕЛЕМЕНТИМА:

- проблем – тема рада;
- начин решавања проблема и начин практичне провере резултата рада, ако је таква провера неопходна;
- литература

НАСЛОВ МАСТЕР РАДА:

МУЛТИМОДАЛНО ГЕНЕРИСАЊЕ ОПИСА СЛИКА ИЗ АКАДЕМСКИХ РАДОВА

ТЕКСТ ЗАДАТКА:

Направити систем за аутоматско генерисање описа слике из академског рада:

1. Анализирати стање у области.
2. Израдити спецификацију захтева софтверског решења.
3. Израдити спецификацију дизајна софтверског решења.
4. Имплементирати софтверско решење према израђеној спецификацији.
5. Тестирати имплементирано софтверско решење.

Руководилац студијског програма:	Ментор рада:

Примерак за: ☐ - Студента; ☐ - Ментора

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, РБР:	
Идентификациони број, ИБР:	
Тип документације, ТД:	монографска публикација
Тип записа, ТЗ:	текстуални штампани документ
Врста рада, ВР:	дипломски рад
Аутор, АУ:	Петар Купусаревић
Ментор, МН:	др Јелена Сливка, ванредни професор
Наслов рада, НР:	Мултимодално генерисање описа слика из академских радова
Језик публикације, ЈП:	српски
Језик извода, ЈИ:	српски / енглески
Земља публикавања, ЗП:	Србија
Уже географско подручје, УГП:	Војводина
Година, ГО:	2023
Издавач, ИЗ:	ауторски репринт
Место и адреса, МА:	Нови Сад, Факултет техничких наука, Трг Доситеја Обрадовића 6
Физички опис рада, ФО:	10 / 35 / 44 / 5 / 1 / 0 / 0
Научна област, НО:	Софтверско инжењерство и информационе технологије
Научна дисциплина, НД:	Машинско учење
Предметна одредница / кључне речи, ПО:	Вештачка Интелигенција, Машинско Учење, Генерисање Описа Слика, Мултимодалност
УДК	
Чува се, ЧУ:	Библиотека Факултета техничких наука, Трг Доситеја Обрадовића 6, Нови Сад
Важна напомена, ВН:	
Извод, ИЗ:	Да би се остварила генерална вештачка интелигенција, модели ће требати да буду мултимодални и да разумеју научни текст. Овај рад ће представити једну потенцијалну методу за решавање ових проблема која се заснива на тренирању модела на задатку генерисања описа слика из академских радова. Представљена метода остварује 14% боље перформансе од претходног најбољег решења.
Датум прихватања теме, ДП:	
Датум одбране, ДО:	
Чланови комисије, КО:	
председник	др Никола Лубурић, доцент
члан	др Александар Ковачевић, редовни професор
ментор	др Јелена Сливка, ванредни професор
Потпис ментора	

KEY WORDS DOCUMENTATION

Accession number, ANO :	
Identification number, INO :	
Document type, DT :	monographic publication
Type of record, TR :	textual material
Contents code, CC :	bachelor thesis
Author, AU :	Petar Kupusarević
Mentor, MN :	Jelena Slivka, associate professor, PhD
Title, TI :	Multimodal Image Captioning from Academic Papers
Language of text, LT :	Serbian
Language of abstract, LA :	Serbian / English
Country of publication, CP :	Serbia
Locality of publication, LP :	Vojvodina
Publication year, PY :	2023
Publisher, PB :	author's reprint
Publication place, PP :	Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6
Physical description, PD :	10 / 35 / 44 / 5 / 1 / 0 / 0
Scientific field, SF :	Software Engineering and Information Technologies
Scientific discipline, SD :	Machine Learning
Subject / Keywords, S/KW :	Artificial Intelligence, Machine Learning, Image Captioning, Multimodality
UDC	
Holding data, HD :	Library of the Faculty of Technical Sciences, Trg Dositeja Obradovića 6, Novi Sad
Note, N :	
Abstract, AB :	In order to achieve General Artificial Intelligence (AGI), future models will have to be multimodal as well as understand scientific text. This work presents one possible method for solving these problems, which is based on training a model to generate captions from images in academic papers. The presented method achieves 14% better performance than the previous best solution.
Accepted by sci. Board on, ASB :	
Defended on, DE :	
Defense board, DB :	
president	Nikola Luburić, assistant professor, PhD
member	Aleksandar Kovačević, full professor, PhD
mentor	Jelena Slivka, associate professor, PhD
Mentor's signature	

САДРЖАЈ

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА.....	3
KEY WORDS DOCUMENTATION.....	4
1. УВОД.....	6
2. ПРЕГЛЕД СТАЊА ОБЛАСТИ.....	8
3. ТЕОРИЈСКИ ПОЈМОВИ И ДЕФИНИЦИЈЕ.....	11
4. МЕТОДОЛОГИЈА.....	12
4.1 Скуп Података.....	12
4.2 Тренирање Модела.....	13
4.3 Коришћени Алати и Хардвер.....	13
5. ЕКСПЕРИМЕНТИ.....	15
5.1 <i>BLIP-2 Q-Former</i> модел.....	15
5.2 <i>BLIP</i> модел.....	16
5.3 <i>BLIP-2 QLoRA</i> модел.....	16
6. РЕЗУЛТАТИ.....	18
7. ДИСКУСИЈА.....	20
8. ЗАКЉУЧАК.....	23
9. ЛИТЕРАТУРА.....	25
10. БИОГРАФИЈА.....	29

1 УВОД

Модерни велики језички модели (енгл. *Large Language Models*) показују веома обећавајуће перформансе на проблемима који укључују разумевања природног говора. Модели попут *GPT4* [1][2] и *Llama 2* [3] представљају данашњи *state-of-the-art* што се тиче генерисања текста. Међутим, да би се направио следећи корак ка општој генералној интелигенцији, модели ће морати да постану мултимодални, то јест да имају способност да раде са другим врстама података попут слике, видеа и звука. Поред тога, још један кључни аспект оваквих модела ће бити њихова способност разумевања научних појмова.

Овај рад ће представити једну потенцијалну методу постизања ових циљева, која се заснива на тренирању модела на задатку генерисања описа слика из академских радова. Пошто је акценат на мултимодалности, за генерисање описа ће се користити додатне текстуалне информације из академског рада. Комбинација слике и текста чине овај модел мултимодалним.

Улаз у модел ће бити слика од које треба да се генерише опис. Поред слике ће се додатно убацити текстуалне информације, у виду наслова и апстракта рада, у циљу давања додатног контекста моделу.

Решење овог проблема се заснива на до-тренирању пре-тренираног *BLIP-2* [4] модела над подацима извучених из академских радова. *BLIP-2* је врста модела која користи "замрзнут" визијски енкодер и говорни декодер модел за генерисање описа слика. До-трениравање пре-тренираног *BLIP-2* модела се постиже тако што се тренира један међу-модел, што су аутори *BLIP-2* рада назвали *Querying Transformer (Q-Former)*, који стоји између та два замрзнута модела. *Q-Former* пружа могућност додавања префикса уносу, који метода представљена у овом раду користи као механизам за постизање мултимодалности. У поглављу 4 ће се више говорити о методологији и техничким детаљима представљеног решења. Потребно је исто нагласити да је тренутни *state-of-the-art* модел за генерисање описа слика управо *BLIP-2* [5].

Тренирана су два модела, један трениран на скупу података који се састоји од 10 хиљада радова из области рачунарске науке, и један трениран на скупу података који се састоји од 15 хиљада радова из разних научних области попут физике, економије, математике, биологије, козмологије и пуно других. Сваки скуп података садржи отприлике 20 хиљада парова слика и њихових описа. У поглављу 4.1 ће се спецификати како су подаци прикупљени. Крајње решење ће моћи да се користи од стране чланова академске заједнице као помоћ при писању академских радова.

Евалуација је вршена тако што су се поредиле перформансе обичног пре-тренираног *BLIP-2* модела са перформансама представљеног до-тренираног *BLIP-2* модела. Метрика која је коришћена је косинусна сличност [6] између генерисаног текста и правог текста. Кроз евалуацију закључујемо да представљена метода постиже 14% боље перформансе од тренутно најбољег решења. Тачне метрике су представљене у поглављу 6, док се детаљна анализа дискусија метрика налази у поглављу 7.

У следећем поглављу се налази преглед стања области, док се у поглављу 3 налазе теоријски појмови и дефиниције потребне за разумевање рада. У поглављу 5 ће бити представљени сви експерименти који су били вршени током имплементације решења, а поглавље 8 ће садржати закључак и сумаризацију главних поенти рада и неке друге идеје за решавање датог проблема као и могућих побољшања представљеног решења.

2 ПРЕГЛЕД СТАЊА ОБЛАСТИ

Област генерисања описа слика (енгл. *Image Captioning*) је релативно нова област. Циљ области јесте да се генерише дескриптиван опис на основу улазне слике. Методе генерисања описа слике се могу поделити у две категорије: у темплејт методе и у енкодер-декодер методе.

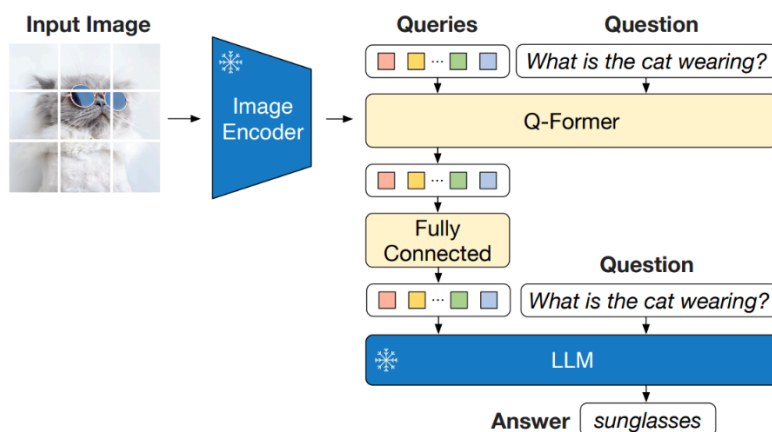
Прве методе су засноване на темплејт методи, где се од улазне слике екстрахују кључне карактеристике попут објеката који се налазе у слици, па се ти подаци проследе неком лингвистичком темплејту који склопи те податке у реченицу. *Farhadi* и сарадници [7] користе информације о објекту који се налази на слици, радњи објекта на слици и околине у коме се објекат налази. Користећи ове информације и њихове релације, може се генерисати реченица. *Li* и сарадници [8] користе темплејт у формату <придев1, објекат1> <препозиција> <придев2, објекат2> да конструишу реченицу. Рад [9] ради слично као рад [8], уз додатак да се може детектовати више објеката и њихови атрибути. Због овога се могу генерисати дуже реченице које садрже информације о вишеструком броју објеката. С обзиром на то да се темплејт методе свode на попуњавање предефинисаних темплејта за генерисање описа, крајњи формати излаза су репетитивни.

Појава рекурентних и конволуционих неуронских мрежа је увела велику промену у област генерисања описа слика. Методе за генерисање описа слике засноване на дубоким неуронским мрежама се свode на то да енкодер (конволуциона неуронска мрежа) екстрахује информације из слике у вектор, на основу кога декодер (рекурентна неуронска мрежа) ауто-регресивно генерише описну реченицу. *Karpathy* и сарадници [10] и *Vinyals* и сарадници [11] су представили прве примере оваквих система.

Недостатак енкодер-декодер метода је што се информације о слици користе само на првом кораку генерисања описа, због чега, при генерисању дужих реченица, модел почиње да губи информацију о слици. *Xi* и сарадници [12] решавају овај проблем помоћу механизма пажње (енгл. *Attention Mechanism*), који даје могућност декодеру да уз сваки корак генерисања "обраћа пажњу" на одређени део улазне слике. Механизам пажње постаје кључан за нову врсту енкодер-декодер језичких модела под називом Трансформер [13].

Појава Трансформер модела представља још један корак унапред у области генерисања описа слика. Као што је поменуто у поглављу 1, данашњи *state-of-the-art* приступ је *BLIP-2* [4], који користи "замрзнут"

визијски енкодер (*ViT*) [14] и "замрзнут" говорни декодер (*OPT*) [15] да би генерисао опис слике. Ово се постиже тако што се тренира један међу-модел, што су аутори *BLIP-2* рада назвали *Querying Transformer* (*Q-Former*), који стоји између та два замрзнута модела. *Q-Former* је исте архитектуре као *BERT* [16] модел, само што је још додат *Cross-Attention* [13] механизам који је повезан са излазом визијског енкодера. Улаз у *Q-Former* су 32 научена "кверија" (енгл. *query*) и, опциони, префикс текст. Сваки вектор излаза *Q-Former*-а је линеарно пројектован на сакривену (енгл. *hidden*) димензионалност декодер модела. Ови пројектовани вектори служе као улаз у декодер који наставља ауто-регресивно да генерише излазни текст. На слици 1 је илустрована архитектура *BLIP-2* модела. Пошто се током тренирања *BLIP-2* модела једино мењају параметри *Q-Former* модела, који се састоји од само 188 милиона параметара, то чини процес тренирања релативно јефтиним. У поглављу 6 ће бити представљени резултати *BLIP-2* модела са до-тренираном верзијом истог над сакупљеним скуповима података.



Слика 1. Илустрација архитектуре *BLIP-2* модела.

Улазна слика пролази кроз визијски енкодер, док су улази у *Q-Former* 32 научена кверија и питање на које треба да се одговори.

Информације са улазне слике интерагују са кверијима *Q-Former*-а помоћу *Cross-Attention* механизма. Затим су излазни кверији линеарно пројектовани на сакривену димензионалност декодер модела уз поновљено питање. Излаз из модела је генерисан одговор на питање.

Слика преузета из рада [4]

Још једна модерна метода која се може користити за генерисање описа слика је *CLIP* [17]. *CLIP* је метода која користи два једно-модална енкодер модела, један визиони енкодер за слику и један говорни енкодер за текст. Циљ *CLIP*-а јесте да репрезентација слике и репрезентација описа буду што сличније. Користећи једноставне функције за израчунавање сличности два вектора, може се изабрати реченица која најбоље описује слику. *Mokady* и сарадници [18] користе комбинацију робусне репрезентације *CLIP* модела и генеративне способности *GPT2* [19] модела да генеришу описе слика. Поред овога, користе малу Трансформер мрежу да мапирају излаз из *CLIP* модела на улаз *GPT2* модела, веома слично као *Q-Former* у *BLIP-2* моделу.

Поред нових архитектура говорних модела и бржег хардвера, велики део успеха најновијих великих говорних модела јесте експоненцијално повећавајућа количина податка која се може користити за тренирање модела. Два позната и популарна скупа података за област генерисања описа слика су *MSCOCO* [20] и *Flickr30k* [21].

Поред области генерисања описа слика, овај рад се исто делом бави са мултимодалношћу. Због обимности и велике брзине напредовања ове области, овај рад ће само набројати пар истакнутих модела. За обимнији преглед стања области, *Yin* и сарадници [22] су направили преглед модерних метода.

Тренутна два најпознатија модела, који могу "нативно" да представе слике уз текст као њихов улаз, су Мајкрософтов *Kosmos-1* [23] и *OpenAI*-ев *GPT4* [1]. Поред ова два модела је веома недавно представљен *NExT-GPT* [24] који је такозвани *any-to-any* мултимодални систем. Користећи различите улазне енкодере и излазне моделе на бази дифузије (енгл. *Diffusion*) [25], *NExT-GPT* може да ради са било којом комбинацијом улазних и излазних модалности попут текста, слике, видео и аудија.

3 ТЕОРИЈСКИ ПОЈМОВИ И ДЕФИНИЦИЈЕ

Већина појмова кроз овај рад су адекватно објашњени кроз њихово увођење или су превише познати или једноставни да није потребно додатно објашњење. Поред овога додата је одговарајућа литература у поглављу 9 за сваки уведен значајан појам. У овом поглављу ће једино бити објашњена *QLoRA* [26] метода пошто је она играла кључну улогу у постизању најбољег резултата у поглављу 6.

QLoRA је комбинација метода квантизације [27] и *LoRA* [28]. Квантизација (енгл. *quantization*) је техника која смањује рачунску и меморијску цену система дубоког учења тако што се параметри модела представљају са мањом прецизношћу (нпр. *Int8* или *Int4*). *LoRA* додатно смањује потребну количину меморије тако што уместо да се сви параметри модела тренирају, *LoRA* замрзне параметре модела и ињектује матрице декомпозиције ранга у сваки слој архитектуре Трансформера. Ово смањује потребан број параметара који треба да се тренирају за фактор од 10 хиљада, док се још одржава могућност ефективног до-тренирања. Ова метода омогућује до-тренирања великих модела са три пута мање ресурса него што би било потребно без такозваних "адаптера".

Поред овога, финални производ процеса до-тренирања користећи *QLoRA* технику је адаптер који је много мањи од првобитног модела. Ово омогућује брзу и лаку дистрибуцију до-тренираних солуција пошто је само адаптер трениран и само он треба да се дистрибуира. Резултат постигнут представљеним решењем је био постигнут користећи *QLoRA* методологију, крајњи адаптер је само 70MB велик наспрам 16GB *BLIP-2* [4] модела.

4 МЕТОДОЛОГИЈА

У овом поглављу ће бити објашњена методологија која се користила за израду решења. У поглављу 4.1 ће бити описан процес прикупљања и процесирања скупова података. У поглављу 4.2 ће се представити методологија за тренирање модела. У поглављу 4.3 ће бити набројане које технологије и библиотеке су се користиле за израду решења, као и опис окружења за развој решења и тренирање модела.

4.1 Скуп Података

За решавање проблема су прикупљена два скупа података користећи исту методологију. Први скуп података се састоји од 10 хиљада научних радова из области рачунарске науке, који ће се у остатку рада звати *CS10K*. Други скуп података се састоји од 15 хиљада научних радова из разних области попут физике, економије, математике, биологије, космологије и много других. У остатку рада ће се овај скуп података звати *Mix*. За прикупљање *CS10K* скупа је било потребно око 8 сати, а за *Mix* око 12 сати. Сви подаци су преузети са *ArXiv* [29] архиве.

Сви преузети радови су у *PDF* формату, што их чини незгодним за процесирање и екстракцију података, у односу на друге формате, попут *XML* или *HTML*, који садрже хијерархијску репрезентацију документа. Помоћу *PyMuPDF* [30] *Python* [31] библиотеке могуће је екстраховати све слике и блокове текста једне странице. Користећи ову библиотеку, стратегија за процесирање података је била да се прво идентификују блокови текста који се односе на слику – сваки опис слике обично почиње са "*Figure*" или "*Fig.*", па се помоћу регуларних израза може адекватно идентификовати који блок је опис. Након тога се пореди број слика са бројем блокова текста који се односе на слике. Уколико је овај број исти, претпоставка је да су ово парови слика-опис и спојени су по редоследу на страници.

Описана метода ће промаштити велики број парова, али ће гарантовати високу прецизност изабраних парова. Због изабране стратегије процесирања радова, из око 5 хиљада радова из *CS10K* се није успело издвојити ни један пар слике и описа, а из *Mix* скупа око 10 хиљада радова није резултовало ни једним екстрахованим паром слика-опис. Крајњи број парова је на крају био око 20 хиљада, по скуп података.

Коришћен је још и *arXiv Metadata* [32] скуп података, из ког су издвојене информације о наслову и апстракт у сваког рада.

4.2 Тренирање Модела

Укупно су вршена 3 експеримента за решавање овог проблема и сви ће бити истакнути у поглављу 5. Експерименти се разликују један од другог по мањим детаљима, док генерална методологија за тренирање остаје иста кроз експерименте. Сви експерименти укључују неки вид до-тренирања *BLIP-2* [4] модела који је био представљен у поглављу 1 и 2. У овом поглављу ће бити презентована процедура тренирања.

Сваки експеримент је подразумевао тренирање модела кроз две целе епохе података, са истим начином претпроцесирања података. За сваки податак је насумично одабран формат префикса који је могао да буде један од три. Прва врста префикса садржи само инструкцију да се генерише опис слике, друга врста, поред ове инструкције, убацује информацију о наслову научног рада, док трећа врста додатно убацује и апстракт рада. Било би корисније да се тренира по један модел за сваку врсту префикса, пошто би ово омогућило поређење перформанси између различитих нивоа додатог контекста, али у интересу времена овај корак је замењен горепоменутом хибридном методом.

Сви модели су тренирани користећи или *Adam* [33] или *AdamW* [34] оптимизатор, са коефицијентима учења (енгл. *learning rate*) између $5e-5$ и $5e-4$. Сам одабир ових хипер-параметара није значајно мењао експерименте. Највећу разлику је направила стратегија бирања дела модела чији ће параметри бити оптимизовани током тренирања, као што ће бити наведено у поглављу 5. У циљу уштеде меморије, подаци нису били груписани и паралелно процесирани кроз мрежу, него су пролазили један по један кроз мрежу са 8 или 16 корака акумулације градијената. Пре сваког ажурирања параметара, градијенти су лимитирани у вредност опсега између -1 и 1.

Израз из модела је вероватноћа следећег токен-а на основу свих претходних токена и унете слике. Израчунате вероватноће се упоређују са вероватноћом тачног следећег токена користећи *Cross Entropy* [35] функцију грешке. Ова функција се током тренирања оптимизује.

4.3 Коришћени Алати и Хардвер

Најпопуларнији програмски језик за задатке везане за област вештачке интелигенције је *Python* [31], због интеграције са разним библиотекама за дубоко учење, као и брзина и лакоћа развоја и имплементације. Због ових разлога је представљено решење имплементирано у *Python-u* [36].

За прикупљање података је коришћена *BeautifulSoup* [37] библиотека, а за процесирање и читање преузетих радова је коришћен *PyMuPDF* [30]. За руковање, аквизицију и тренирање модела су коришћене *transformers* [38] и *PyTorch* [39] библиотеке, за квантизацију [27] су коришћене *bitsandbytes* [40] и *accelerate* [41] библиотеке. Последња битна библиотека која је коришћена је *peft* [42], која је омогућила *LoRA* [28] тренирање.

За тренирање модела и вршење експеримената се користило *Google Colaboratory* [43] окружење са *NVIDIA A100* графичком картицом која садржи *40GB VRAM* меморије. Ово је уједно и највећа количина *VRAM* меморије која се може добити користећи *Google Colaboratory* сервис, па су сви експерименти били конструисани око овог ограничења.

5 ЕКСПЕРИМЕНТИ

У овом поглављу ће бити описани експерименти који су били вршени као и недостаци сваког решења. У поглављу 4.2 је дефинисана методологија тренирања, па због тога ово поглавље неће улазити превише у те детаље. Одабир хипер-параметара се исто може наћи у поглављу 4.2, па зато те информације неће бити наведене у овом поглављу, осим ако су сматране битним. У поглављу 4.3 је наведено хардверско ограничење које је често било лимитирајући фактор у првим експериментима и које је одређивало следеће потенцијалне кораке за даље експерименте.

Укупно су вршена три експеримента, где је примарна разлика између сваког била стратегија која се користила за бирање параметара за тренирање. Ови експерименти су наведени хронолошки кроз следећа три подпоглавља, где резултат сваког претходног директно одређује методологију следећег.

Сви експерименти укључују неки вид до-тренирања *BLIP-2* [4] модела који је био представљен у поглављу 1 и 2.

5.1 *BLIP-2 Q-Former* модел

Методологија за тренирање би била веома једноставна уколико не би било хардверских ограничења – у овом случају би се једноставно тренирао цео модел. Као што је наведено у поглављу 4.3, имплементација свих експеримената је била ограничена меморијским капацитетом *NVIDIA A100* графичке картице која садржи *40GB VRAM* меморије. Ово је и више него довољно за покретање *BLIP-2* модела чија је величина *16GB*, али није довољно за тренирање целог модела.

Због овога је први експеримент ограничен искључиво на тренирање међу-модела, то јест *Q-Former*-а, који се састоји од само 188 милиона параметара што је око *0,7 GB*.

Кроз овај експеримент је било утврђено да је прва реченица описа коју генерише овај до-трениран модел обећавајућа. Међутим, реченице које јој следе су наизглед насумичне и неvezане за претходно генерисан текст. Ово се може објаснити тиме што, уз дуже реченице, говорни декодер "преузме" процес генерисања текста и престаје да обраћа пажњу на улазну слику и претходно генерисан текст. Ово је директан резултат тренирања искључиво међу-модела, а остављања декодера

замрзнутим. Стога су извршени и додатни експерименти где су се тренирали и параметри последњих пар блокова декодера, као и почетних пар блокова декодера, али крајњи резултат је остао исти. Кроз овај експеримент је било очигледно да се, поред *Q-Former* модела, треба и тренирати цео декодер модел, али је меморијска рестрикција чинила ово немогућим.

5.2 BLIP модел

Идеја другог експеримента је била да, ако није могуће тренирање целог *BLIP-2* модела, можда се могу постићи задовољавајуће перформансе тренирањем претходника тог модела, то јест *BLIP* модела [44]. *BLIP* модел има исту архитектуру као *BLIP-2*, али значајно мањи број параметара (око 500 милиона, што је око 2GB).

Мања величина модела је омогућила тренирање целог модела, али исто тако негативно утицала на перформансе. Крајњи до-трениран модел је био превише примитиван за дати проблем. Поред овога, *BLIP* нема способност генерисања великих слова, што значи да генерисан опис неће бити капитализован.

5.3 BLIP-2 QLoRA модел

Трећи експеримент представља последицу прва два експеримента. Закључак првог експеримента је био да је потребно тренирати више него само параметре међу-модела, а закључак другог експеримента је био да су ипак потребне напредне способности *BLIP-2* модела за решење дефинисаног проблема. Да би се оба описана услова постигла, вршено је до-тренирање *BLIP-2* модела коришћењем *QLoRA* [26] технике, која је представљена у поглављу 3. Ово је омогућило тренирање целог модела без прекорачења меморијских рестрикција дефинисаних у поглављу 4.3.

За модел трениран на *CS10K* скупу података, међу-репрезентација (ранг) *LoRA* [28] адаптера је била 16 док је за *Mix* скуп података била 32. Повећање овог хипер-параметра није пуно значило за крајње перформансе, једино је адаптер за *Mix* модел дупло већи наспрам адаптера *CS10K* модела. Сви под-модели *BLIP-2* су били до-тренирани са *LoRA* адаптерима. Ово укључује визијски енкодер, *Q-Former*, говорни декодер и још пројекција излаза из *Q-Former* модела на улазну димензију говорног декодер модела.

У поглављу 6 ће бити представљени резултати евалуације до-тренираних модела из трећег експеримента и стандардним пре-тренираним *BLIP-2* моделом, као и њихово међусобно поређење.

6 РЕЗУЛТАТИ

У овом поглављу ће бити поређене перформансе обичног пре-тренираног *BLIP-2* [4] модела са представљеним до-тренираним моделом из поглавља 5.3. Као што је поменуто у поглављу 1, метрика која је коришћена је косинусна сличност [6] која представља сличност између две реченице као вредност између 0 и 1. За добијање вектора реченица се користио *BERT* [16] модел. Евалуација и тренирање је вршено кроз два скупа података: *CS10K* и *Mix*. Евалуциони скуп података се састојао од 15% броја укупних података из сваког скупа (око 3 хиљаде парова слике и описа).

Конкретно су тестирана три модела:

- 1) *BLIP-2* до-трениран – До-трениран модел који је представљен у поглављу 5.3
- 2) *BLIP-2* са префиксом – Пре-трениран *BLIP-2* модел где су током евалуације у префикс додате информације о наслову и апстракту рада, на исти начин као што је представљено у поглављу 4.2
- 3) *BLIP-2* без префикса – Пре-трениран *BLIP-2* модел где се током евалуације није користио префикс, него је улаз била само слика

У следећим табелама су представљене метрике евалуације:

<i>CS10K</i> Скуп Података	
Модел	Косинусна Сличност
<i>BLIP-2</i> до-трениран	0.81922
<i>BLIP-2</i> без префикса	0.71623
<i>BLIP-2</i> са префиксом	0.55020

Табела 1. Метрике тестираних модела над *CS10K* скупом података. Са леве стране се налазе називи тестираних модела, а са десне средње вредности израчунатих косинусних сличности.

Mix Скуп Података	
Модел	Косинусна Сличност
BLIP-2 до-трениран	0.81957
BLIP-2 без префикса	0.72126
BLIP-2 са префиксом	0.52231

Табела 2. Метрике тестираних модела над *Mix* скупом података. Са леве стране се налазе називи тестираних модела, а са десне средње вредности израчунатих косинусних сличности.

Из горњих табела се види да до-трениран модел постиже најбоље резултате за оба скупа података, превазилазећи перформансе *BLIP-2* модела без префикса за 14%. *BLIP-2* модел са префиксом постиже најгоре резултате, из овога се може закључити да обичан *BLIP-2* модел не подноси добро префиксе који служе да додају контекст одговору. Ово се потврђује у раду [4] где је представљен *BLIP-2* модел, где се префикс само користио да дефинише кратко питање на које треба да се одговори са информацијом која се налази на улазној слици.

Следеће поглавље садржи дискусију добијених решења. Додатно ће бити показани и упоређени поједини генерисани описи из горепомнутих модела.

7 ДИСКУСИЈА

Ово поглавље садржи дискусију резултата који су представљени у поглављу 6. Резултати су показали да модел представљен у овом раду постиже 14% боље перформансе од претходних најбољег модела. За боље разумевање генеративних способности модела поменутих у претходном поглављу, у овом поглављу ће бити представљени примери генерисаних описа свих модела.

Генерисан Опис	Тачан Опис
<i>Energy calibration for spectra generated from data taken in the NUCLEUS experiment</i>	<i>Figure 5.21. Energy spectrum for the Run 1 NUCLEUS prototype data.</i>
<i>расследование фактора 3 луч</i>	<i>Figure 1. Research framework</i>
<i>The caption reads "Identifying and Classifying Issues in Issue Reports."</i>	<i>Figure 3: An overview process of identifying and classifying privacy requirements in issue reports</i>
-----	<i>Figure 1: one point function</i>
<i>ইউজ</i>	<i>Figure 1: Server-aided PSI.</i>

Табела 3. Лево: Примери генерисаних описа *BLIP-2* модела са префиксом. Десно: Тачни описи улазних слика.

Као што је дискутовано у поглављу 6, *BLIP-2* модел не подноси добро дугачке префиксе чији је циљ да дају контекст. Ово се потврђује у табели 3. Када је модел кориштен са префиксом, генерисани описи често изгледају као грешке. Није ретко да генерисан текст буде на другом језику или да садржи невезане и необичне симболе. У табели 3 су показана два примера која донекле добро описују слику, али чак и добри примери често имају проблеме око форматирања, као, на пример, велики број размака између речи (у табели 3 су избрисани ови размаци ради лакшег прегледања). Ово није велики проблем пошто се крајњи излаз може процесирати да не садржи лоше форматирање, али могуће је да ти проблематични генерисани токени утичу на квалитет остатка генерисаног текста.

Генерисан Опис	Тачнан Опис
<i>the pie chart shows the average % of the energy budget of the universe</i>	<i>Figure 1.1: The mass-energy budget of the Universe determined by Wilkinson Microwave Anisotropy Probe (WMAP) [1]</i>
<i>a greek coin with two heads with a crown in the head of one</i>	<i>Figure 1.1: The picture of a Roman coin with a depiction of the God 'Janus'. Picture taken from Wikipedia</i>
<i>a diagram of the lens array, and how it works</i>	<i>Figure 1. Sketch of the geometry for a SIS lens.</i>
<i>the logo for the flag 2019 conference</i>	<i>Figure 1: FLAG 2019 compilation of results for $f\pi$ and fK. Meaning of the colors is explained the the text.</i>
<i>conclusion construction rule modify each internal line by one</i>	<i>Figure 8: Construction rule 1.</i>

Табела 4. Лево: Примери генерисаних описа *BLIP-2* модела без префикса. Десно: Тачни описи улазних слика.

Из табеле 4 се види да *BLIP-2* модел без префикса може солидно да опише слику, али генерисане реченице још увек не личе на тачне описе и не садрже довољно детаља. Поред овога, генерисани описи не садрже капитализоване речи, нити су конструисани као реченице. До-трениран модел представљен у овом раду покушава да поправи ове грешке.

Генерисан Опис	Тачнан Опис
<i>Fig. 1. A schematic diagram of the EHC-OFDM-IM scheme.</i>	<i>Fig. 1: System block diagram of the proposed EHC-OFDM-IM system.</i>
<i>Figure 2: Average of the G5.900-0.430 signal.</i>	<i>Fig. 1: Averaged spectrum of the 6.7 GHz methanol maser associated with G5.900–0.430. All 437</i>
<i>Figure 1: Snapshot of the daily tweets in the 2017 Hurricane Andrew dataset.</i>	<i>Figure 1: Wordcloud of disasters in our dataset.</i>
<i>Fig. 3. An example of an Euler diagram from Figure 2.1.</i>	<i>Figure 1: Plot of the real 2-soliton solution in Example 3.9.</i>
<i>Figure 6. Model of the SWGO detector.</i>	<i>Figure 7: 3D model of the cross holder.</i>

Табела 5. Лево: Примери генерисаних описа до-тренираног BLIP-2 модела представљеног у овом раду.
Десно: Тачни описи улазних слика.

Из табеле 5 се види да до-трениран модел генерише описе сличне тачним и да прати конвенције и формат за описе слика у академским радовима.

Једна мана процеса евалуације јесте што се још не може мерити доменско знање горепоменутих модела. Као што је поменуто у поглављу 1, једна мотивација овог рада јесте да ће модели морати да разумеју научан текст, али без постојања скупа података за визуелно одговарање на питања (енгл. *Visual Question Answering*) из научних области, мерење доменског знања оваквих модела је веома тежак посао.

Остале мане и побољшања ће бити представљена у следећем поглављу.

8 ЗАКЉУЧАК

Циљ овог рада је био да се направи систем који ће да генерише описе слика из академских радова уз додатно коришћење текстуалног контекста у виду наслова и апстракта рада. Мотивација за решавање овог проблема јесте што ће у будућности модели бити мултимодални, то јест мораће да разумеју и генеришу више модалности попут слике, видеа, аудија и наравно текста.

Решавање овог проблема се сводило на до-тренирању пре-тренираног *BLIP-2* [4] модела над паровима слика и описа из академских радова. Сакупљена су два скупа података, *CS10K* који садржи 10 хиљада радова из области рачунарске науке, и *Mix* скуп података који садржи 15 хиљада радова из разних научних области попут физике, математике, биологије, космологије, економије и много других. До-тренирање је било вршено у *Google Colaboratory* [43] окружењу са *Nvidia A100* графичком картицом са *40GB VRAM* меморије. Додатно се користила *QLoRa* [26] метода за меморијски ефикасно до-тренирање.

Представљен модел је успео да оствари косинусну сличност [6] од 0.81922 над *CS10K* скупом података, и 0.81957 над *Mix* скупом података, остваревши 14% боље резултате од претходног најбољег модела, што је *BLIP-2*.

Један недостатак јесте што се још не може дискутовати стечено доменско знање представљених модела. Да би ово било могуће, потребно је постојање скупа података за визуално одговарање на питања из области науке (енгл. *Visual Question Answering*). Један могући начин да се направи овакав скуп података је да се користи велики језички модел (енгл. *Large Language Model*) за генерисање парова питања и одговора на основу тачног описа слика из академских радова. Затим би се могао поново тренирати модел, стим што би сада префикс био генерисано питање, а тачан излаз из модела би био одговор.

Тренирање ових модела је скуп и дуготрајан процес, па би зато било корисно да се овај експеримент додатно понови са различитим хипер-параметрима, попут коефицијента учења (енгл. *learning rate*) и другачијег оптимизатора. Исто је важно знати да ли контекст у виду наслова и апстракта академских радова помаже код генерисања описа, што у овом тренутку није експлицитно доказано. За ово би требало да се тренирају три модела са различитим нивоима контекста и да се упореде метрике. Добра подела би била да први модел за улаз има само

слику, други додатно има контекст наслова рада, док трећи добије и наслов и апстракт рада. Додатно би било корисно да се повећа количина парова слика и описа, и да се побољша стратегија за екстраховање ових парова која је била представљена у поглављу 4.1.

Последња мана представљеног решења јесто што се величина улазне слике претвори у димензије 224×224 јер је ова димензијалност потребан услов за визијски енкодер. Ова димензијалност понекад није довољна за препознавање ситнијег текста у слици, као и генералних ситнијих детаља. Експеримент би могао да се понови уз додатак визијског енкодера који може да функционише са већом димензијом улазне слике.

9 ЛИТЕРАТУРА

- [1] GPT4 <https://openai.com/research/gpt-4>
- [2] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*
- [3] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- [4] Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*
- [5] Papers With Code Image Captioning Benchmark
<https://paperswithcode.com/sota/image-captioning-on-coco-captions>
- [6] Cosine Similarity https://en.wikipedia.org/wiki/Cosine_similarity
- [7] Ali Farhadi, Mohsen Hejrati, Mohammed Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier and David Forsyth. 2010. Every Picture Tells a Story: Generating Sentences from Images.
Computer Vision – ECCV 2010, Volume 6314
- [8] Li, S., Kulkarni, G., Berg, T., Berg, A. and Choi, Y., 2011, June. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the fifteenth conference on computational natural language learning* (pp. 220-228).
- [9] Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL. 2013. Babytalk: understanding and generating simple image descriptions. *IEEE Trans Pattern Anal Mach Intell*
- [10] Andrej Karpathy and Fei-Fei Li. 2014. Deep Visual-Semantic Alignments for Generating Image Descriptions. *arXiv preprint arXiv:1412.2306*
- [11] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitry Erhan. 2015. Show and Tell: A Neural Image Caption Generator. *arXiv preprint arXiv:1411.4555*

- [12] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015, June. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057). PMLR.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA
- [14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- [15] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V. and Mihaylov, T., 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training for Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*
- [17] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.
- [18] Ron Mokady, Amir Hertz, Amit H. Bermano. 2021. ClipCap: Clip Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*
- [19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *Paperswithcode Preprint 2019*
- [20] COCO <https://paperswithcode.com/dataset/coco>
- [21] Flickr30k <https://paperswithcode.com/dataset/flickr30k>
- [22] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, Enhong Chen. 2023. A survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*
- [23] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei. 2023. Language Is Not All

You Need: Aligning Perception with Language Models. *arXiv preprint arXiv:2302.14045*

[24] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, Tat-Seng Chua. 2023. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv preprint arXiv:2309.05519v2*

[25] Jonathan Ho, Ajay Jain, Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2006.11239*

[26] Tim Dettmers, Artidoro Pagoni, Ari Holtzman, Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314*

[27] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang,rew Howard, Hartwig Adam, Dmitry Kalenichenko. 2017. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. *arXiv preprint arXiv:1712.05877*

[28] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*

[29] arXiv <https://arxiv.org/>

[30] PyMuPDF <https://pymupdf.readthedocs.io/en/latest/>

[31] Python <https://www.python.org/>

[32] arXiv Metadata dataset
<https://www.kaggle.com/datasets/Cornell-University/arxiv>

[33] Diedrik P. Kingma, Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*

[34] Ilya Loshchilov, Frank Hutter. 2017. Decoupled Decay Regularization. *arXiv preprint arXiv:1711.05101*

[35] Cross-entropy <https://en.wikipedia.org/wiki/Cross-entropy>

[36] Multimodal-Image-Captioning Repository
<https://github.com/Cabbagito/Multimodal-Image-Captioning>

[37] BeautifulSoup4
<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

[38] transformers <https://huggingface.co/docs/transformers>

[39] PyTorch <https://pytorch.org/>

- [40] bitsandbytes <https://github.com/TimDettmers/bitsandbytes>
- [41] accelerate <https://huggingface.co/docs/accelerate>
- [42] peft <https://github.com/huggingface/peft>
- [43] Google Colaboratory <https://colab.google/>
- [44] Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*

10 БИОГРАФИЈА

Петар Купусаревић је рођен 15.03.2000. у Суботици, где је завршио основну школу "Јован Јовановић Змај" и техничку средњу школу "Иван Сарић", смер "Информационе Технологије".

Школске 2019/20 године уписује Факултет Техничких Наука на студијски програм "Софтверско Инжињерство и Информационе Технологије". Положио је све испите предвиђене планом и програмом и стекао услов за одбрану завршног рада.