



UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA
U NOVOM SADU




Михаела Осмајић

**Одговарање на питања са
визуелним контекстом у
области науке употребом
визуелно-језичких модела**

МАСТЕР РАД
- Мастер академске студије -

Нови Сад, 2024.

	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:	Мастер академске студије
Студијски програм:	Рачунарство и аутоматика
Руководилац студијског програма:	др Мирна Капетина (РА)

Студент:	Михаела Осмајић	Број	E2 13/2023
Област:	Електротехничко и рачунарско инжењерство		
Ментор:	Др Александар Ковачевић, редовни професор		
НА ОСНОВУ ПОДНЕТЕ ПРИЈАВЕ, ПРИЛОЖЕНЕ ДОКУМЕНТАЦИЈЕ И ОДРЕДБИ СТАТУТА ФАКУЛТЕТА ИЗДАЈЕ СЕ ЗАДАТАК ЗА ДИПЛОМСКИ РАД, СА СЛЕДЕЋИМ ЕЛЕМЕНТИМА: <ul style="list-style-type: none"> - проблем – тема рада; - начин решавања проблема и начин практичне провере резултата рада, ако је таква провера неопходна; - литература 			

НАСЛОВ МАСТЕР РАДА:

Одговарање на питања са визуелним контекстом у области науке употребом визуелно-језичких модела

ТЕКСТ ЗАДАТКА:

Описати систем за одговарање на питања са визуелним контекстом употребом визуелно-језичких модела. Анализирати сродна истраживања. Специфицирати и описати архитектуру решења. Анализирати и описати модуле коришћене у систему. Представити скуп података коришћен за обучавање и евалуацију. Анализирати експерименте и добијене резултате и дискутовати могуће правце унапређења. Документовати 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:	мастер рад
Аутор, АУ:	Михаела Осмајић
Ментор, МН:	др Александар Ковачевић, редовни професор
Наслов рада, НР:	Одговарање на питања са визуелним контекстом у области науке употребом визуелно-језичких модела
Језик публикације, ЈП:	српски
Језик извода, ЈИ:	српски / енглески
Земља публикавања, ЗП:	Србија
Уже географско подручје, УГП:	Војводина
Година, ГО:	2024
Издавач, ИЗ:	ауторски репринт
Место и адреса, МА:	Нови Сад, Факултет техничких наука, Трг Доситеја Обрадовића 6
Физички опис рада, ФО:	9/31/28/4/7/2/0
Научна област, НО:	Софтверско инжењерство и информационе технологије
Научна дисциплина, НД:	Софтверско инжењерство
Предметна одредница / кључне речи, ПО:	Одговарање на питања са визуелним контекстом, визуелно-језички модели, PaliGemma, ScienceQA
УДК	
Чува се, ЧУ:	Библиотека Факултета техничких наука, Трг Доситеја Обрадовића 6, Нови Сад
Важна напомена, ВН:	
Извод, ИЗ:	Одговарање на питања са визуелним контекстом захтева комбинацију визуелних и текстуалних информација, што може побољшати интеракцију између људи и машина. Фино су подешена два PaliGemma визуелно-језичка модела користећи ScienceQA скуп података. Остварени су резултати од 81% и 94% тачности на тест скупу. Резултати указују на могућност обучавања модела уз ограничене ресурсе, коришћењем техника оптимизације.
Датум прихватања теме, ДП:	
Датум одбране, ДО:	
Чланови комисије, КО:	
председник	др Јелена Сливка, ванредни професор
члан	др Лидија Крстановић, ванредни професор
ментор	др Александар Ковачевић, редовни 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 :	master thesis
Author, AU :	Mihaela Osmajić
Mentor, MN :	Aleksandar Kovačević, full professor, PhD
Title, TI :	Visual question answering in the field of science using vision-language models
Language of text, LT :	Serbian
Language of abstract, LA :	Serbian / English
Country of publication, CP :	Serbia
Locality of publication, LP :	Vojvodina
Publication year, PY :	2024
Publisher, PB :	author's reprint
Publication place, PP :	Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6
Physical description, PD :	9/31/28/4/7/2/0
Scientific field, SF :	Software Engineering and Information Technologies
Scientific discipline, SD :	Software Engineering
Subject / Keywords, S/KW :	Visual question answering, vision-language models, PaliGemma, ScienceQA
UDC	
Holding data, HD :	Library of the Faculty of Technical Sciences, Trg Dositeja Obradovića 6, Novi Sad
Note, N :	
Abstract, AB :	Visual question answering requires the integration of visual and textual information, enhancing human-machine interaction. Two PaliGemma vision-language models were fine-tuned using the ScienceQA dataset, achieving accuracies of 81% and 94% on the test set. These results demonstrate that effective model training is feasible with limited resources through the use of optimization techniques.
Accepted by sci. Board on, ASB :	
Defended on, DE :	
Defense board, DB :	
president	Jelena Slivka, associate professor, PhD
member	Lidija Krstanović, associate professor, PhD
mentor	Aleksandar Kovačević, full professor, PhD
Mentor's signature	

Садржај

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА	4
KEY WORDS DOCUMENTATION.....	5
Увод.....	7
Сродна истраживања	9
Скуп података	11
Методологија.....	13
Претпроцесирање	13
Претпроцесирање скупа података.....	13
Форматирање улаза и излаза	14
Обучавање модела	15
Метод евалуације	18
Мере перформансе.....	18
Поступак евалуације.....	19
Резултати и дискусија.....	20
Резултати тачности модела	20
Насумичан избор одговора модела?	21
Ланац резоновања	23
Халуцинације	24
Просечна тачност по областима.....	24
Закључак	25
Литература.....	27
Биографија.....	31

Увод

Традиционална решења за одговарање на питања (*Question Answering*) су се ослањала искључиво на текстуалне податке [28]. Међутим, људска перцепција и разумевање света се не ослањају само на текст, већ и на визуелне информације [27]. Одговарање на питања са визуелним контекстом (*Visual Question Answering*, VQA) је задатак који захтева одговоре на питања која се односе на слике. Решења овог задатка имају потенцијал да унапреде интеракцију између људи и машина, омогућавајући апликације попут паметних помоћника или система за помоћ особама са оштећењем вида.

Проблем одговарања на питања са визуелним контекстом је посебно изазован јер захтева синтезу и интеграцију информација из две различите модалности: визуелне и текстуалне у облику природног језика. Решење изискује првобитно разумевање садржаја слике, препознавање релевантних објеката, сцена или радњи, а затим и повезивање тих информација са питањем како би се генерисао тачан одговор. На пример, питање „Која земља је најсевернија на слици?” захтева да модел препозна све земље на слици, разуме појам „север” и идентификује која од њих је најсеверније позиционирана. Овај задатак је изазован јер модел мора не само да идентификује објекте и земље, већ и да правилно интерпретира њихов релативни положај, што захтева комбинацију визуелне перцепције и просторне анализе, чинећи га сложенијим од основних задатака одговарања на питања.

У овом раду смо вршили fino подешавање (*fine-tuning*) два PaliGemma визуелно-језичка (*vision-language*) модела. Оптимизовали смо моделе да одговарају на питања која се односе на визуелни контекст. Модели су научили сложене везе између слика и текста. Први, базни модел, постиже тачност (ассигуру) од 81% на тест скупу, док други модел, који је већ fino подешаван, постиже тачност од 94% на истом тест скупу.

Решења су евалуирана кроз мерење перформанси модела на валидационом и тест скупу података, при чему је ChatGPT [2] модел процењивао семантичку исправност одговора у односу на коректне одговоре. На основу тих процена, коришћена је тачност као мера перформанси модела.

Структура рада је организована на следећи начин: У поглављу II представљена су сродна истраживања, пружајући преглед актуелног стања у релевантној области. Поглавље III садржи опис коришћеног скупа података. Поглавље IV описује методологију, укључујући

процесе претпроцесирања података и обучавања модела. Поглавље V је посвећено методама евалуације, које обухватају метрике перформанси и поступак евалуације. Након тога, у поглављу VI приказани су резултати уз пратећу дискусију. Поглавље VII закључује рад сумирањем главних доприноса и предлогом праваца за будућа истраживања.

Сродна истраживања

Један од првих радова у области VQA је рад [3] Антола и сарадника, који су представили VQA скуп података и поставили основне изазове за ову област. Како би их савладали, користили су конволутивну неуронску мрежу VGGNet [4] за екстракцију визуелних карактеристика и LSTM (*Long short-term memory*) [5] за обраду текстуалних питања. Овај рад је инспирисао даљи развој модела који користе самопажњу (*self-attention*) за боље повезивање текстуалних и визуелних информација.

Андерсон и сарадници су представили *Bottom-Up and Top-Down Attention* [6] модел, који користи механизме пажње за побољшање разумевања слика и генерисање прецизних одговора. Овај модел користи *bottom-up* механизам за генерисање региона од интереса (*Region Of Interest*) унутар слике, док *top-down* механизам омогућава моделу да усмерава пажњу на релевантне делове слике у контексту постављеног питања, значајно побољшавајући перформансе модела на VQA задацима. Појавом моћнијих трансформерских модела, попут BERT-а [7], долази до напретка у визуелно контекстним питањима и објаве ViLBERT-а [8]. Овај модел је показао супериорне перформансе у односу на претходне моделе.

Пионир PaLI серије визуелно-језичких модела [9], успешно је скалирао број параметара на 17 милијарди комбиновањем ViT [10] и mT5 [11] језичког модела. Његови наследници, PaLI-X [12] и PaLM-E [13], унапредили су ове резултате. PaLI-X је користио ViT-22B [14] и језички модел UL2 [15], док је PaLM-E користио исти визуелни трансформер и PaLM [16]. Ови модели су постигли још боље резултате на задацима визуелно-језичког разумевања, уз трошкове већих ресурса. PaLI-3 [17] је одржао корак са перформансама својих претходника на већини тестова, али са 10 пута мањим бројем параметара у односу на PaLI-X и 100 пута у односу на PaLM-E. Коришћењем SigLIP [18] енкодера за слике и UL2 језичког модела, скалирао је број параметара на 5 милијарди.

PaLI-Gemma [19] наставља овај тренд спајањем 400 милиона параметара SigLIP-а са 2 милијарде параметара Gemma модела [26], стварајући визуелно-језички модел са мање од 3 милијарде параметара, који и даље остварује перформансе упоредиве са PaLI-X, PaLM-E и PaLI-3.

Као део напретка у оптимизацији обучавања великих језичких модела, LoRA (*Low-Rank Adaptation of Large Language Models*) [20]

техника се истиче као иновативно решење за тренирање модела на специфичним задацима или доменима. Уместо да мења све параметре модела, LoRA уводи тренирајуће матрице ниског ранга у сваки слој трансформерске архитектуре, што омогућава смањење броја тренирајућих параметара док се унапред трениране тежине задржавају замрзнуте. Овај приступ може смањити број тренирајућих параметара и до 10.000 пута, док истовремено смањује потребу за меморијом графичке картице за 3 пута, а притом задржава перформансе модела у складу са конвенционалним финим подешавањем на различитим задацима [20]. QLoRA [21] је проширење LoRA технике које додатно оптимизује меморијске захтеве великих језичких модела. QLoRA квантизује тежинске параметре унапред тренираног модела на прецизност од 4 бита, смањујући потребу за меморијом, што омогућава фино подешавање модела на једној графичкој картици. Ова метода знатно смањује потребу за меморијом, омогућавајући покретање великих модела на слабијем хардверу, укључујући потрошачке графичке картице.

У последњих неколико година, истраживачи се нису фокусирали само на оптимизацију броја параметара модела и њиховог тренирања, већ су се такође бавили унапређењем самих захтева и података који се користе за обуку ових модела. *Prompt engineering* је представљен у раду [22] на GPT-3 моделу и показао је како промишљено креирање захтева може значајно утицати на перформансе модела у различитим задацима, наглашавајући важност дизајнирања захтева који омогућавају моделу да боље разуме и решава комплексне проблеме. Техника ланца резоновања (*Chain of Thought*, CoT) уводи секвенцијално резоновање приликом одговарања на питања, додатно побољшавајући перформансе модела одговарања на питања са визуелним контекстом [23]. Аутори су показали да модели могу ефикасније обрадити сложене упите када их води кроз низ логичких корака. Овај приступ омогућава моделима да боље разумеју међузависности између информација, што доводи до тачнијих и поузданијих одговора.

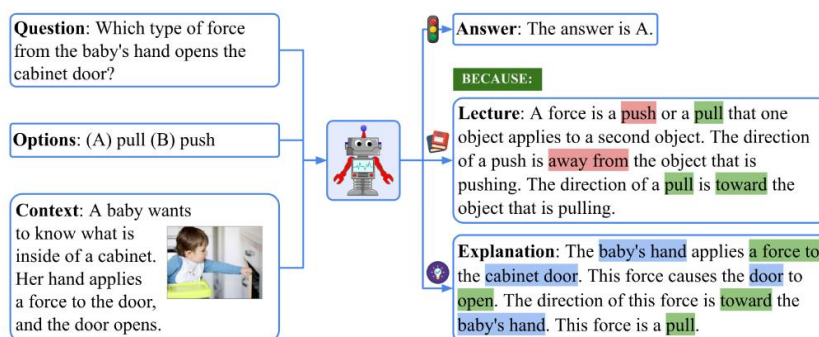
У овом раду се фино подешавање PaliGemma модела интегрише са QLoRA, као и оптимизација самих захтева, *prompt engineering*-ом, како би се уз што мање количине ресурса оствариле задовољавајуће перформансе одговарања на питања са визуелним контекстом.

Скуп података

ScienceQA [24] је јавно доступан скуп података развијен да унапреди способност вештачке интелигенције за решавање научних питања кроз мултимодални приступ. Скуп података обухвата 21.208 питања са вишеструким избором одговора и помоћи у њиховом решавању, покривајући широк спектар научних тема, те укључује објашњења и лекције везане за одговоре. Мултимодалност је кључна карактеристика овог скупа података, јер питања садрже различите модалитете као што су текст и слике, што омогућава сложеније и богатије разумевање проблема. Циљ *ScienceQA* је побољшање интерпретабилности и вишестепене логике (*multi-hop reasoning*) код вештачких интелигенција, омогућавајући им да генеришу лекције и објашњења као део ланца размишљања.

Питања обухватају области биологије, физике, хемије, математике, географије, историје, економије, језичке писмености и граматике. Од укупног броја инстанци, 10.332 (48.7%) имају контекст слике, 10.220 (48.2%) имају текстуални контекст, а 6.532 (30.8%) имају оба контекста. Већина питања је анотирана са лекцијама (83.9%) и детаљним објашњењима (90.5%). Лекције пружају опште знање, док објашњења дају специфичне разлоге за долазак до тачног одговора.

About ScienceQA



Слика 3.1 Илустрован приказ обележја скупа података, при чему су са леве стране приказани елементи питања, а са десне стране елементи одговора

У овом раду користили смо подскуп података који укључује наведене четири области: хемија, физика, биологија и

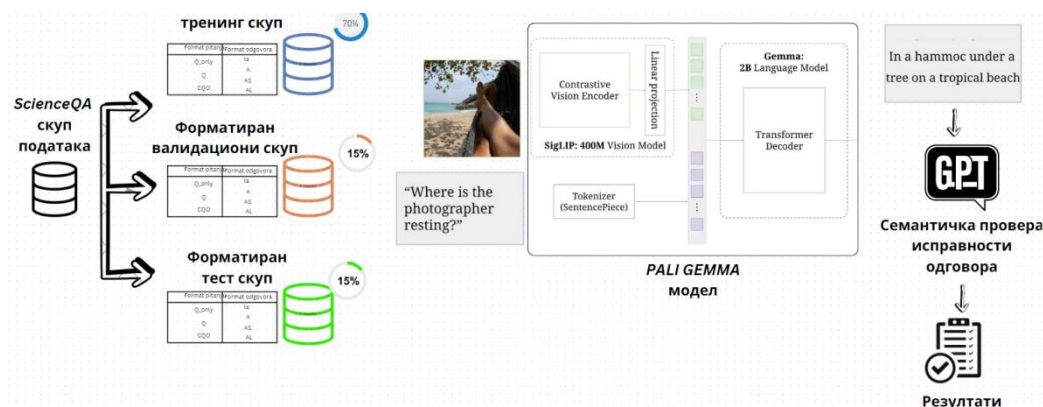
географија. Тиме смо редуковали скуп података на 560 инстанци без недостајућих вредности, 140 у свакој области. Након претпроцесирања, коначан скуп података се састоји од следећих атрибута:

1. Питање
2. Помоћ
3. Вишеструки избор одговора
4. Слика
5. Одговор
6. Објашњење
7. Лекција

Питање и слика представљају улаз у модел, док сам одговор представља излаз.

Методологија

У овом поглављу је представљена методологија имплементације система за фино подешавање PaliGemma модела за потребе одговарања на питања са визуелним контекстом. Овај систем користи напредне технике за оптимизацију, попут LoRA и QLoRA, како би се смањила потреба за меморијом и убрзао процес тренирања. Улаз у систем чини слика са припадајућим питањем, док је очекивани излаз текстуални одговор на постављено питање. Методологија обухвата све кораке од обраде података до обуке модела. На слици 4.1. налази се дијаграм решења.



Слика 4.1 Дијаграм решења

Претпроцесирање

Претпроцесирање података представља битан корак у припреми за обучавање модела, омогућавајући побољшање квалитета података и ефикасности процеса учења.

Претпроцесирање скупа података

У оквиру овог рада, иницијални скуп података је филтриран како би се уклониле колоне које нису релевантне за

тренирање модела, смањујући комплексност и фокусирајући анализу на кључне информације.

Након почетног чишћења, подаци су филтрирани на нивоу области, задржане су само четири релевантне области: географија, биологија, физика и хемија. Елминисани су редови са недостајућим вредностима. Подаци су додатно избалансирали по областима, што је обезбедило равномерну заступљеност сваке области.

Како је формат тачног одговора, заправо индекс тачног одговора у вишеструком избору, последњи корак је укључивао креирање нове колоне са одговорима у текстуалном облику. Након тога, скуп података је спреман за обучавање модела. Важно је напоменути да PaliGemma при прослеђивању слика сама врши поједине трансформације попут промене резолуције и скалирања, из тог разлога, није било потребно вршити такве трансформације. Једина неопходна трансформација била је конверзија слика у RGB формат.

Форматирање улаза и излаза

Податке за обуку: префикс, суфикс и слику, потребно је форматирали за свако питање. Формати префикса су обухватили различите комбинације три елемента, као што су (1) питање, (2) понуђени одговору и (3) помоћ, док су формати суфикса могли да садрже 4 елемента: (1) само одговор, (2) одговор са префиксом који упућује на његову природу одговора, (3) лекцију и (4) објашњење. У Табели 4.1 се налазе сви наведени формати. Након форматирања текста, врши се токенизација текста и слика.

	Симбол формата	Формат
Формати питања	Q_only	{question}.
	Q	{question}. Options: {options}
	CQO	{hint}. {question}. Options: {options}
Формати одговора	Ia	{answer}
	A	The answer is {answer}
	AL	The answer is {answer}. {solution}
	AS	The answer is {answer}. {lecture}

Табела 4.1 Формати питања и одговора

У наставку налази пример једне инстанце са форматом CQO-A и кореспондирајућом сликом (слика 4.2).

Питање: *The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material. Think about the magnetic force between the magnets in each pair. Which of the following statements is true? Options: ["The strength of the magnetic force is the same in both pairs.", "The magnetic force is stronger in Pair 2.", "The magnetic force is stronger in Pair 1."]*

Одговор: *The answer is the magnetic force is stronger in Pair 1.*



Слика 4.2 Улазна слика за пример инстанце CQO-A формата

Обучавање модела

У овом раду су фино подешавана два модела PaliGemma. Први модел, `pali-gemma-3b-pt-224` (у даљем тексту модел А), представља базни модел са 3 милијарде параметара, који ради са сликама резолуције 224 x 224. Други модел, `pali-gemma-3b-mix-224` (у даљем тексту модел Б), има исту резолуцију и број параметара, али је додатно подешаван за неколико задатака, попут оптичког препознавања карактера. Одабрали смо резолуцију 224 x 224 јер веће резолуције захтевају значајно више меморије због дужих улазних секвенци.

Три милијарде параметара овог модела изискују значајну количину меморије. У овом случају, за фино подешавање свих параметара модела потребне су или две L4 графичке картице или једна A100/H100 [25].

Овај проблем је превазиђен употребом методе LoRA, помоћу Hugging Face PEFT (*Parameter-Efficient Fine-Tuning*) библиотеке¹. Ова метода омогућава замрзавање (*freezing*) постојећих тежина, што значи да се обучава само неколико слојева модела, неведених у

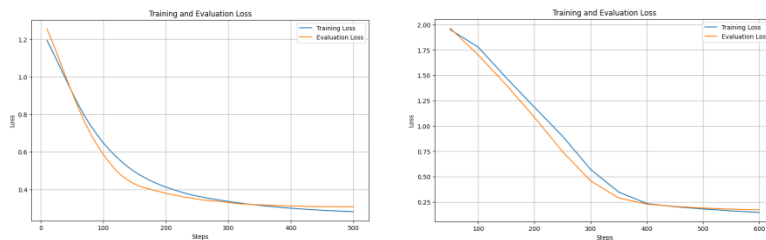
¹ PEFT: Parameter-Efficient Fine-Tuning, <https://github.com/huggingface/peft>

подешавањима. Како бисмо додатно редуковали употребу ресурса, осим што се обучавају параметри само неких слојева, такође су претходно и квантизовани у четворобитну вредност, уместо уобичајне 32-битне или 16-битне, тачније коришћена је QLoRA. Постоје многе форме квантизације, а у раду је коришћена *BitsAndBytes*² интеграција. У случају финог подешавања ових модела, обучавањем је само језички декодер, док су замрзнути визуелни енкодер и мултимодални пројектор, што би значило да је обучавањем 13 милиона параметара од 3 милијарде, односно приближно 45% параметара модела у четворобитној вредности. Ова стратегија је смањила потребу за меморијом и омогућила ефикасније прилагођавање језичких компоненти модела.

Како фино подешавамо модел са 400 тренинг инстанци, ефективна величина *batch*-а за базни модел А била је 8, док је за други модел, модел Б, била 32. Међутим, због ограничења меморије користили смо технику акумулације градијената (*gradient accumulation*) са корацима од 4 и 8 у другом случају, чиме смо постигли еквивалентну величину жељених величина *batch*-а без прекорачења меморијских капацитета. Обучавање је изведено кроз 10 епоха у случају базног модела А и 20 епоха у случају модела Б, са стопом учења (*learning rate*) од $1e-5$. Није коришћено насумично искључивање (*dropout*), док је опадање тежина (*weight decay*) од $1e-3$ било присутно за обучавање модела Б, а оптимизатор *paged AdamW* је био специфично прилагођен за 4-битну квантизацију, омогућавајући ефикасно тренирање модела уз минималну употребу меморије.

Пратили смо *Cross-Entropy* функцију губитка (*loss*) над тренинг и валидационим скупом на сваких 10 корака оптимизације. Графикони функција губитка (Слика 4.3) показују сталан пад на оба скупа, валидационом и тренинг скупом. Такође, имплементирана је техника раног заустављања како би се зауставило обучавање када нема значајног побољшања на валидационом скупу, чиме је спречен један начин претренирања.

² BitsAndBytes:8-bit optimizers and quantization routines,
<https://github.com/TimDettmers/bitsandbytes>



Слика 4.2 Графיקони функција губитака модела А и Б на тренинг и валидационом скупу

Обучавање модела је изведено на Google Colab платформи са Т4 графичком картицом. Иако, због ограничених ресурса, није било могуће користити сложеније методе оптимизације хиперпараметара, постигнути су задовољавајући резултати коришћењем одабраних хиперпараметара и техника за ефикасно коришћење меморије. Обучавање је показало како се може постићи задовољавајући ниво перформанси чак и са ограниченим ресурсима.

Метод евалуације

Метод евалуације обухвата поступак оцењивања перформанси модела. Подразумева одабир одговарајућег поступка на основу којег ће се евалуација извршити и одговарајућих мера перформанси за које ће се евалуација извршити.

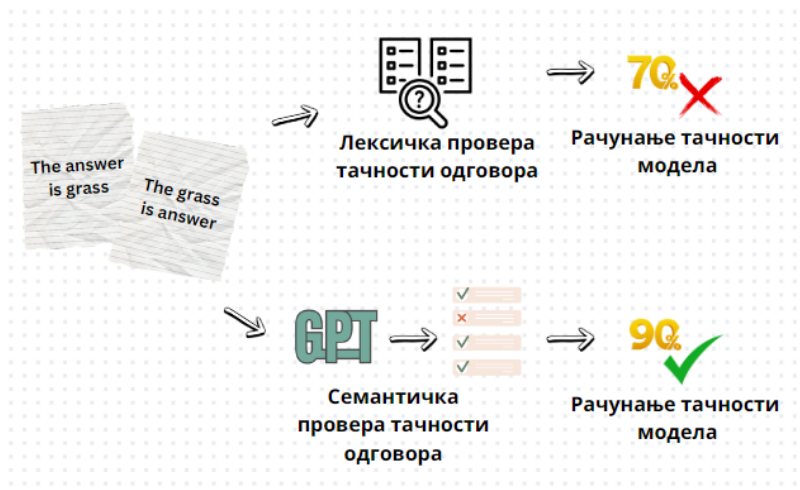
Мере перформансе

За меру перформанси користи се тачност, као и у већини радова који се баве овим проблемом. Међутим, традиционалне методе процене тачности одговора не узимају у обзир семантичку сличност између одговора модела и тачног одговора. То значи да чак и ако модел пружи одговор који је суштински тачан, али се не подудара лексички са очекиваним одговором, тачност ће бити смањена. Проблем може да се јави и код најпростијих формата одговора, попут формата А из Табеле 4.1, где је уочено 3% грешке евалуације помоћу традиционалних метода. Нарочито је изражен код сложенијих формата одговора, где разлике у формулацији могу довести до значајних разлика у оцени тачности, иако су оба одговора семантички исправна.

Да би се овај проблем заобишао и тачност проценила за сваки формат праведно, користи се помоћни језички модел, као што је ChatGPT верзије gpt-3.5-turbo. Њему се шаље захтев који садржи тачан одговор, питање и предвиђени одговор модела (модел који смо тренирали) за процену њихове семантичке тачности. Захтев се формулише тако да се од помоћног модела захтева процена да ли је предвиђени одговор тачан у односу на тачан одговор за дата питања. Захтев је приказан следећим форматом:

For the question “{question}”, the expected answer is “{correct}”. In relation to the expected answer, would “{response}” be correct for the given question? Please answer with yes/no.

На основу ове процене, свако предвиђање модела се класификује као тачно или нетачно. Коначна тачност се израчунава као однос броја тачних предвиђања према укупном броју питања, чиме се добија мерило које су перформансе модела за све формате. На слици 5.1 је приказ разлике између традиционалног начина евалуације тачности одговора и примененог у овом раду.



Слика 5.1 Приказ евалуације перформанси традиционалним и начином коришћеним у раду на једном примеру

Оваква проверка омогућава да одговори модела не морају бити потпуно исти као у скири података, већ да су суштински тачни у односу на очекиване одговоре. Комбиновањем ове две методе, тачност као квантитативну меру и семантичку сличност као квалитативну процену, постиже се свеобухватан увид у перформансе модела.

Поступак евалуације

Поступак евалуације модела укључује поделу скупа података на тренинг, валидациони и тест скуп у односу 70:15:15. Тренинг скуп се користи за обучавање модела, омогућавајући му да научи обрасце и структуре из података. Валидациони скуп служи за фино подешавање хиперпараметара и спречавање претренирања, док се тест скуп користи за коначну евалуацију перформанси модела. Ова подела осигурава да се модел евалуира на невиђеним подацима, пружајући реалистичну процену његове способности генерализације нових примера.

Резултати и дискусија

Евалуацијом више експеримената, анализирали смо утицај формата питања и одговора на перформансе модела. У овој секцији приказани су резултати експеримената и дискусија на њихову тему. Детаљи о форматима питања и одговора приказани су у Табели 4.1 у одељку методологије.

Резултати тачности модела

Најосновнији формат одговарања укључује постављање питања уз понуђене опције одговора и једноставан одговор без додатног образложења (формат Q-A). У поређењу су коришћена два модела обучавана током истраживања, Модел А и Модел Б, заједно са референтним *paligemma-3b-mix-224* моделом (*baseline model*), фино подешеним за различите задатке. У Табели 6.1. приказани су резултати.

Формат питања	Формат одговора	Модел	Тачност валидационог скупа	Тачност тест скупа
Q	A	Модел А	85%	81%
Q	A	Модел Б	95%	94%
Q	Ia	paligemma-mix-3b-224	95%	89%

Табела 6.1 Резултати тачности модела на валидационом и тест скупу за формат Q-A

Модел А постигао је тачност од 85% на валидционом скупу и 81% на тест скупу, док је модел Б остварио тачност од 95% на валидционом и 94% на тест скупу. Референтни модел постигао је 95% тачности на валидционом и 89% на тест скупу. Како је Модел Б претходно био фино подешен за разноврсне проблеме, а потом и додатно обучаван за *ScienceQA* скуп података, очекивано је остварио најбоље резултате од 95% на валидционом скупу и 94% на тест скупу. Модел А је остварио најмању тачност међу анализираним моделима, с разликом што је он базни модел, који је фино подешен само на основу скромног скупа података и ограничених ресурса у односу на остале моделе. Процена људске тачности при одговарању на питања износила је 88%, при чему им је постављено питања, дата помоћ и опције

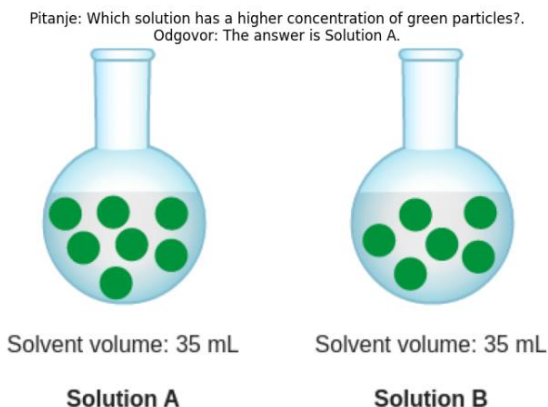
одговора и нису морали да дају образложење већ искључиво да одговоре на питања [24].

Резултати су показали да наш приступ омогућава задовољавајуће перформансе за решавање проблема одговарања на питања са визуелним контекстом. Референтни модел, у свом раду наводи тачност од 95% на тест скупу [19], док је у нашој евалуацији тачност износила 89%. Претпостављамо да је ова разлика узрокована случајним одабиром тежих примера у нашем тест скупу. Такође, сви модели су постигли боље резултате на валиднионом у односу на тест скуп, што такође сугерише на могући дисбаланс у тежини питања између ових скупова.

Насумичан избор одговора модела?

Да бисмо проценили да ли и колико модели насумично бирају одговоре, без стварног разумевања тачног решења, спровели смо експеримент у којем су уклоњене опције одговора. Моделима је прослеђено само питање, а од њих се очекивало да самостално генеришу одговор. На жалост, овај поступак није било могуће евалуирати због слободe одговарања коју модели имају. Примера ради:

1. На слици 6.1. приказана је слика и питање које се прослеђује моделу, као и тачан одговор. Модел је одговорио да слика са леве стране има већу концентрацију. Иако је одговор тачан, није га могуће евалуирати без помоћног визуелно-језичког модела који би имао улогу проценитеља тачности.



Слика 6.1 Пример инстанце тест скупа

Pitanje: Which bird's beak is also adapted to crack large, hard nuts?
 Odgovor: The answer is scarlet macaw.



Слика 6.2 Пример инстанце тест скупа

2. На питање са слике 6.2, које птице имају адаптиран кљун за ломљење ораха, модел је одговорио "папагај", док је тачан одговор био "*scarlet macaw*", врста папагаја. (Не)тачност може да се сагледа на различите начине. Можемо рећи да је одговор тачан зато што је *scarlet macaw* папагај, али такође одговор може да се посматра као нетачан зато што није сваки папагај адаптиран на ломљење ораха.

3. На слици 6.3. приказана је слика која је улаз у модел, уз њу такође улази и питање шта је заједничко свим стварима на слици. Модел је одговорио "да су од злата", док је тачан одговор био "да сијају". Обравивши пажњу на слику, модел није дао нетачан одговор у ширем контексту, али је то нетачан одговор за стварни контекст овог питања.

Pitanje: Which property do these three objects have in common?
 Odgovor: The answer is shiny.



Слика 6.3 Пример инстанце тест скупа

Ланац резоновања

Још један начин провере насумичног избора одговора из понуђених опција је генерисање тока размишљања (CoT) за дат одговор [23]. Експеримент који смо спровели укључује генерисање објашњења или лекција за одређено питање. Због ограничених ресурса, нисмо могли обучавати моделе за генерисање и објашњења и лекција истовремено, па је то учињено појединачно, користећи формате одговора AL и AS.

Формат питања	Формат одговора	Модел	Тачност валидационог скупа	Тачност тест скупа
Q	AS	Модел А	82%	75%
		Модел Б	93%	91%
Q	AL	Модел А	84%	78%
		Модел Б	94%	91%

Табела 6.2 Приказ резултата тачности AS и AL на валидационом и тест скупу

Резултати из Табеле 6.2. су, у односу на друге формате, показали слабије перформансе при генерисању објашњења, што је очекивано с обзиром на тежину оваквог задатка. Генерисање лекција уз одговор показало се ефикаснијим од генерисања објашњења, верујемо да је то због понављања истих лекција кроз слична питања.

Занимљиво је да је укључивањем објашњења, тачност модела Б за сам избор одговора заправо повећала и достигла 96% на валидационом и 95% на тест скупу, али су објашњења била лоша и у односу на истинито објашњење, семантички нетачна. Што би значило да чак 3% одговора на валидационом и 4% на тест скупу имају тачан одговор, али погрешно објашњење. Овај проценат повећања перформанси може бити појава насумичног избора одговора, али такође може бити и појава учења тока размишљања. Приметили смо да су модели за поједина питања изгенерисали тачан одговор, али су дошли до њега потпуно другим, и даље тачним, начином размишљања у односу на објашњење које се налази у тестном скупу.

Претпоставивши да модели боље уче уколико им се укаже на начин тока размишљања, испитали смо како додатна помоћ у прослеђеном одговору утиче на перформансе модела користећи формат CQO из Табеле 5.1. За моделе које смо ми обучавали, ова помоћ је повећала тачност за 1-2,5% за генерисање формата одговора А, али уз већу потрошњу ресурса, због чега је било могуће утврдити

како утиче помоћ на одговоре у којима је потребно објаснити решење или написати лекцију. За референтни модел, додатна помоћ је статистички негативно утицала на перформансе. Верујемо да је скромно повећање перформанси резултат неодговарајуће природе пружене помоћи за одређена питања. Такође, остаје отворено питање каквог би утицаја на перформансе имало обучавање модела са пруженом помоћи у питањима, као и са објашњењима и лекцијама у одговорима у виду тока размишљања.

Халуцинације

Потребно је напоменути да, као и код свих језичких модела, долази до халуцинација. Модел пружи тачан одговор и добро објашњење, али у наставку придода одређени текст који није потребан или релевантан. Примећени су поједини случајеви где у одговору модел понавља неколико пута одређену реченицу. Како се овај рад не бави нужно креирањем модела за евалуацију перформанси, омакве одговоре нисмо сматрали грешком, али ово запажање оставља простора за нека нова истраживања и унапређења.

Просечна тачност по областима

Када упоредимо просечну тачност свих модела по областима (Табела 6.3.), уочавамо да географија има највећи проценат тачности. Претпостављамо да је то зато што су питања из географије јасна, без уланчавања мисли, захтевају чињенице на којима су језички модели добро обучени, а додатни контекст слике олакшава проналажење тачног одговора. Супротно томе, питања из биологије, физике и хемије захтевају разумевање дешавања на слици, комбинацију чињеница и њихову синтезу, што отежава тачно одговарање.

Области	Просечна тачност на валидационом скупу	Просечна тачност на тест скупу
Биологија	76%	90%
Хемија	86%	80%
Географија	95%	90%
Физика	86%	73%

Табела 6.3 Просечна тачност свих модела по областима на валидационом и тест скупу

Закључак

Овај рад истражује примену фино подешених, визуелно-језичких PaliGemma модела за одговарање на питања која укључују визуелни контекст. Комбиновањем текстуалних и визуелних информација, можемо значајно проширити могућности вештачке интелигенције у разумевању и одговарању на комплексна питања везана за стварни свет.

Решење је подразумевало претпроцесирање *ScienceQA* скупа података, након чега је уследила обука модела оптимизована напредним техникама редуковања меморијских захтева, без којих не би била могућа. Перформансе модела су оцењиване користећи тачност семантичке сличности, док је подела скупа података на тренинг, валидациони и тест скуп омогућила бољу генерализацију модела.

Фино подешен paligemma-mix-3b-224 је постигао најбоље резултате, са тачношћу од 95% на валидационом и 94% на тест скупу, што га чини најуспешнијим међу анализираним моделима. Његова супериорност је очекивана, јер је модел фино подешен за разноврсне проблеме, укључујући и *ScienceQA* скуп података. Употреба LoRA и QLoRA техника је оптимизовала обуку, значајно смањујући потребу за меморијом и омогућавајући ефикасније коришћење ресурса, што је резултовало смањењем трошкова обуке великих модела.

Такође, истраживање је показало да различити формати питања и одговора утичу на перформансе модела: најједноставнији формат, питање уз опције са одговором, се показао као најефикаснији, што је очекивано с обзиром на то да су сложенији формати, попут формата са објашњењем, захтевали већу софистицираност модела. Иако је генерисање објашњења смањило перформансе због тежине свог задатка, уочено је повећање тачности одговора без објашњења. Ова појава може бити последица случајности насумичног избора, али и учења модела вишестепеним размишљањем, које захтева више примера за прецизно објашњење одговора.

Коришћење ChatGPT модела за процену семантичке сличности омогућило је за 3% прецизнију евалуацију тачности одговора за најједноставније формате одговора, док је за сложеније формате ово био једини начин евалуације, јер је фокусирало оцену на суштинску, а не само дословну тачност одговора, уз напомену да у овом истраживању није придата пажња на понављање неких реченица више пута или додавање непотребних халуцинација на крају одговора.

Ограничења студије укључују немогућност аутоматске евалуације модела у слободном формату одговарања, као и потребу за

већим ресурсима за сложеније формате и оптимизације параметара. У будућем раду, проширење ресурса за обуку модела могло би значајно побољшати перформансе и тачност одговора. Такође, истраживање примене визуелно-језичких модела за евалуацију семантичке сличности одговора представља занимљиву и корисну тему за будућа истраживања.

Литература

- [1] Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- [2] OpenAI. (2024). ChatGPT (verzija gpt-turbo-3.5). Dostupno na: <https://www.openai.com/>
- [3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425-2433).
- [4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- [5] Hochreiter, S. (1997). Long Short-term Memory. *Neural Computation MIT-Press*.
- [6] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6077-6086).
- [7] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [8] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- [9] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A. J., Padlewski, P., Salz, D., ... & Soricut, R. (2022). Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.
- [10] Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022). Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12104-12113).
- [11] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

- [12] Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., ... & Soricut, R. (2023). Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.
- [13] Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... & Florence, P. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.
- [14] Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., ... & Houlsby, N. (2023, July). Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning* (pp. 7480-7512). PMLR.
- [15] Team, C. (2024). Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.
- [16] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240), 1-113.
- [17] Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., ... & Soricut, R. (2023). Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.
- [18] Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 11975-11986).
- [19] Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., ... & Zhai, X. (2024). PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.
- [20] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- [21] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- [22] Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- [23] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.
- [24] Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., & Bhattacharyya, P. (2022). Scienceqa: A novel resource for question answering on

scholarly articles. *International Journal on Digital Libraries*, 23(3), 289-301.

- [25] Rogge, N. (2023). Fine-tune PaliGemma for image > JSON [Jupyter Notebook]. GitHub.
https://github.com/NielsRogge/Transformers-Tutorials/blob/master/PaliGemma/Fine_tune_PaliGemma_for_image_%3EJSON.ipynb
- [26] Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., ... & Kenealy, K. (2024). Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.0829
- [27] Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational psychology review*, 3, 149-210.
- [28] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163, 21-40.

Биографија

Михаела Осмајић, рођена је 6.4.2000. године у Новом Саду. Факултет техничких наука у Новом Саду, смер Рачунарство и аутоматика уписује 2019. године. Наставља мастер студије на истом смеру и ступа у радни однос сарадника у настави, све испите полаже и студије завршава у року, 2024. године. Исте године издаје свој први научни рад на тему *Utilizing Large Language Models for Automated Grading of Free-Form Test Questions*.