Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Катедра „Компютърна информатика”

ДИПЛОМНА РАБОТА

на тема

„Автоматично отговаряне на многомодални въпроси”

Дипломант: **Симеон Емилов Христов**

Факултетен номер: **6MI3400191**

Специалност: **Информатика**

Магистърска програма: **Изкуствен интелект**

Научни ръководители:

Проф. Преслав Наков, катедра „Обработка на естествен език“,
Университет „Мохамед бин Зайед по Изкуствен интелект“ (MBZUAI)

Проф. д-р Иван Койчев, катедра „Софтуерни технологии“,
ФМИ, СУ „Св. Климент Охридски“

Консултант:

докт. Димитър Димитров, катедра „Софтуерни технологии“,
ФМИ, СУ „Св. Климент Охридски“

София, 2024 г.

Съдържание

1. Увод

Обобщен изкуствен интелект е възможността на машина да разбира широка гама съдържание (включ. текстово, визуално, аудио) и да взема решения и предприема действия, използвайки го. Концептуално тази възможност е връзката между текущите имплементации, които са силно специфицирани към решаване на конкретна и ясна задача, и системите с обобщен изкуствен интелект, често срещани в научно-популярните филми [1].

Необходимо е да се оцени модела цялостно в контекста както на прости разпознавателни умения в картинки, така и в комплексни дисциплини, изискващи задълбочено мислене и специфични за домейна знания. По този начин ще се подобри разбирането за прогреса към обобщен изкуствен интелект, който рефлектира типа експертиза и умения за мислене, очаквани от умели възрастни хора в най-разнообразни професионални области.

Важно в тази връзка е минимизиране на предубежденията при подбиране на моделите, наборите от данни, стратегиите за оценяване, тестовите множества, както и примерите, които стават основа за анализ на грешките на моделите впоследствие. В тази връзка поставянето на фокус за оценяване качеството на големи многомодални модели изцяло и само върху данни, събрани от образователната система може и да не е достатъчно за валидиране на съществуването на обобщен изкуствен интелект. Въпреки това постигането на добри резултати върху такива набори от данни със сигурност е признак на развита мисловна дейност, съответстваща на богата обща култура и задълбочено познаване на високо-специализирани области. Затова е и важно да се проверява и човешката точност. Това ще позволи по-добро сравнение между възможностите на моделите и експертните резултати, което от своя страна ще позволи по-ясно измерване на разстоянието от текущото състояние на изкуствения интелект до обобщения изкуствен интелект.

Многомодалните многоезични езикови модели заемат все по-важна роля в получаването на кратък и точен отговор на различен тип въпроси. От генериране на програмен код и анализиране на съдържанието на картинки до отговаряне на въпроси, свързани с управлението на човешки ресурси, тези модели демонстрират гъвкавост и адаптация към много области, независимо от езика. Те намират приложение и в контекста на образователната система с възможност за бързо даване на точен отговор на въпроси от затворен тип. Въпросите могат да се различават в три аспекта: могат да включват само текст, текст плюс картинка, или могат да се състоят само от текст, но отговорите им да съдържат картинки.

Настоящата дипломна работа има за цел да използва тези вариации и да извърши сравнителен анализ на съвременни многомодални многоезични езикови модели като Bard и GPT4 и да оцени тяхната точност чрез използване на въпроси, давани на изпити на ученици от различни държави и в различно ниво на обучение.

В настоящата дипломна работа са поставени и изпълнени следните задачи:

- Подробен обзор на изследванията в областта: проучване на използвани набори от данни, метрики и експерименти с езикови модели.
- Предварителна обработка на данните.
- Планиране и провеждане на експерименти за оценяване и сравняване на:

    - многомодални многоезични езикови модели;
    - големи езикови модели;
    - модели с автоматичен превод на въпроса на английски език.
- Анализ на грешките, допускани от моделите.

2. Преглед на областта

2.1. Основни понятия

Езиков модел (от англ. language model) - компютърна система, съпоставяща вероятност на последователност от думи.

Голям езиков модел (от англ. large language model) - езиков модел, предназначен за използване в големи мащаби и характеризиращ се с възможност за разбиране и генериране на език с общо предназначение.

Голям визуално-текстови модел (от англ. large vision-language model) - голям езиков модел, който може да получава на вход текстово и визуално съдържание.

Модалност на данни (от англ. data modality) - независим канал за протичане на информация при входно/изходни операции с човешко-машинен интерфейс. Най-популярните към момента модалности са текст, видео (в частност картинки) и аудио.

Чатбот (от англ. chatbot) - уеб интерфейс, използващ основен модел, за да имитира човешки разговор чрез обменяне на текстови или аудио съобщения.

Основен модел (от англ. foundation model) - модел, базиран на невронни мрежи, който има много параметри, стойностите на които са се получили в резултат на трениране с голяма част от данни с различни модалности, достъпни в Интернет. Характеризират се с възможност да симулират разсъждения в произволен контекст.

Голям многомодален модел (от англ. large многомодален model) - основен модел, който може да приема като вход данни от различни модалности и да създава резултати от същите или други модалности.

Образователни системи тип K-12 - обхваща годините на получаване на формално или задължително образование от детска градина до завършване на средно образование.

Затворен въпрос (от англ. close ended question) - въпрос, който има краен брой отговори, най-често от два до пет.

Отворен въпрос (от англ. open ended question) - въпрос, който провокира излагане на твърдение и поняка - негова защита или отрицание.

Анотация - множество от етикети, описващи прилежаща метаинформация, към единици данни.

2.2. Подобни разработки

2.2.1. Оценяване на набор от данни Visual QA

Задачата за даване на отговор в свободен стил след получена картинка като вход се въвежда за първи път в [6]. Тя цели да провери доколко големите

визуално-текстови модели могат да върнат правилен отговор в текстов вид по подадена картинка и въпрос за нея. Въпросът е отворен и в частност е насочен към конкретни части на картинката. Визуалното съдържание е подбрано от наборът данни MS COCO [14], който предоставя снимки от ежедневни ситуации в реалния свят, които са автентични и без допълнителна обработка. Обхващат широка гама сцени и имат разнообразен контекст от обекти. Поради тази строга специфичност обаче имат много детайли и шум. Авторите констатират, че това внася вид предубеденост и затова добавят втора част към формираното множество, която се състои от абстрактни сцени и обекти. Сцените и обектите могат да се групират по произволен начин и така да се създават нови сцени, които, въпреки че са нереалистични имат изчистен заден план и могат да се използват за създаване на картинки с цел фино оценяване на моделите.

При подбора на моделите за оценяване, авторите избират:

- модел, който винаги дава като отговор най-често срещания отговор в тренировъчния набор от данни - “Да”;
- модел, който дава като отговор най-често срещания отговор за всеки тип на въпроса (има главно пет типа въпроси, започващи с петте най-популярни въпросителни думи - “какво”, “колко”, “защо”, “къде”, “кой”);
- модел, който дава отговор, който е най-често срещания отговор в K най-близки съседа на текущата двойка въпрос-картинка;
- модел, който дава отговор, след като кодира въпроса, използвайки предварително създаден bag-of-words речник. От най-често срещаните 1,000 думи във въпросите са извлечени най-често срещаните 10 първи, втори, и трети думи от въпросите. Така се създава влагане на думи с размерност 1,030;
- модел, използващ VGGNet, за да даде отговор на въпроса, използвайки единствено изображението;
- модел, базиран на еднопосочен LSTM.

Резултатите им показват, че моделът, разчитащ само на картинката, не се представя добре, при това постига по-слаби резултати от модела, който винаги отговаря с “Да”. В контраст с това, моделите, които разчитат само на текстовата част на въпроса, се представят доста добре, постигайки около 50% точност и задминавайки модела с най-близките съседи, който работи и с картинката. Авторите предполагат, че това се дължи на априорните вероятности, зададени от въпроса. Например, за отговаряне на въпроса “Какъв е цветът на бананите?” не е нужно да се гледа картинка.

Най-добрият модел постига точност около 60%. Той е комбинацията на двунивов LSTM за влагане на думите от въпроса и VGGNet с l2-нормализация за влагане на изображението. LSTM-ът се състои от два скрити слоя и води до получаването на влагане с размерност 2048. Влагането на текста е резултат от прилагането на хиперболичен тангенс върху конкатенацията на вектора от последната клетка и скритият вектор от последната клетка. Двете влагания са обединени чрез поелементно умножение, след което следва нелинейност и Софтмакс слой за получаване на крайния отговор. Наблюдава се, че моделът е добър в разпознаването на цветове и популярни обекти, но изпитва трудност при броене и в частност, когато верният отговор е стойност над 5. Спрямо трудността на въпросите, авторите показват, че моделът се представя като дете на 4,74 години. Като следствие, докато точността на модела е 61.07% във

възрастовата група от 3 до 4, тя спада до 47.83% при възрастите от 18 години нагоре.

2.2.2. Оценяване на набор от данни ScienceQA

Създаването на първия набор от данни фокусиран върху многомодални въпроси, взети от изпити, давани на ученици, е описано в [5]. ScienceQA е считан за голяма и важна стъпка към цялостното и систематичното оценяване на големи многомодални модели. Изследва се възможността на модели да създават верига от логически свързани мисли/твърдения, формирането на които позволява достигането до крайния отговор. Обхванати са три предмети, изучавани главно в класовете от първи до шести - Биология, Социология и Науката за езика. Отличителни черти на ScienceQA са големият брой многомодални въпроси и прилежащите към тях специални анотации - лекции и обяснения. Това позволява да се оцени качеството на *веригата от разсъждения*, които модела поражда при достигането до крайните отговори.

Оценяването се фокусира върху проучване на възможността за създаване на логически свързани обяснения с цел разобличаване на мисловия процес при отговаряне на въпросите в ScienceQA.

Авторите избират три типа модела за оценяване. Първият обхваща модели, базирани на евристиката: случаен избор и човешка точност. Случайният избор се изразява в избор на един от възможните отговори. Извършват се три обхождания на тестовото множество за пресмятане на средна точност. За пресмятане на човешката точност авторите използват Amazon Mechanical Turk. Amazon Mechanical Turk е платформа, в която регистрирани работници (реални хора) извършват услуги онлайн, които могат да варират от валидиране на данни до проучване на подходящи за използване спрямо спецификата на задачата методи за решаването ѝ. Вторият тип модели са големите езикови модели UnifiedQA [23] и GPT-3 [24]. В експериментите без подаване на предварително оценени примери, използваният формат за вход-изход е съответно QCM→A, където входът представлява конкатенация от текста на въпроса (Q), контекста му, който, в случая на многомодални въпроси, включва и обяснение за съдържанието на дадена картинка (C) и възможните отговори (M), а изходът е отговорът на модела. За получаване на обяснението на картинката се използват моделите ViT [25] и GPT-2 [26]. Провеждат се и експерименти с подаване на няколко предварително оценени примера с цел предоставяне на възможност за адаптация към специфичната задача на ScienceQA. Третият тип модели са фино настроени големи визуално-текстови модели (като VisualBERT и ViLT) и фино настроен UnifiedQA. Големите визуално-текстови модели приемат въпроса, тестовия му контекст (ако има такъв), възможните отговори и прилежащата картинка и създават точково разпределение над възможните отговори чрез линеен класификатор. Фината настройка на UnifiedQA запазва начина на използване при вторият тип експерименти - приема се текстовото съдържание, а визуалното се преобразува в текстово, използвайки обяснение на съдържанието.

Получените резултати показват, че от големите визуално текстови модели най-добре се представя VisualBERT, постигайки средна точност от 61.87%, следван от Patch-TRM, който се представя по-добре в предмети, свързани с наука за природата и в частност постига по-висока средна точност на въпроси, давани на ученици от седми до дванадесети клас (67.50% срещу съответно 59.92%).

Без фина настойка и без подаване на предварително оценени примери големият езиков модел UnifiedQA не успява да постигне по-добри резултати от кой да е визуално текстови модел, но все пак се представя по-добре от случайния избор. След фина настойка обаче моделът успява да постигне средна точност от 70.12%, която се увеличава с 4 пункта, когато изискваният резултат, включва и веригата от разсъждения. Тези резултати показват, че генерирането на веригата на разсъждения, заедно с отговора подобрява качеството на разсъжденията на езиковите модели.

Без фина настойка и без подаване на предварително оценени примери големият езиков модел GPT-3 постига много добра точност - 74.04%. Положителният ефект на извличането на веригата на разсъждения заедно с отговора може да се види и при GPT-3. Моделът постига 74.17% точност с нейното добавяне, с което се превръща и в най-добрия модел във всички експерименти. Точността му не е толкова далеч от тази на човек (88.40%), което показва, че дори и да обработва картинките косвено (вместо пряко), голямото количество информация, с която е претрениран, помага значително за обобщаване на разсъжденията към разнообразни области.

2.2.3. Оценяване на набор от данни Textbook QA

Наборът от данни TQA [10] цели да оцени възможността на модел да даде отговор на краен брой многомодални въпроси, използвайки като контекст реални уроци от учебни материали с прилежащи диаграми и картинки.

Авторите създават свой собствени модели и ги оцянават върху TQA.

Базовият модел използва LSTM и приема на вход текстовото съдържание на въпроса и урока. Анализ на данните в TQA показва, че в повечето случаи разполагане с текстовата информация е достатъчно за даване на отговор на въпросите, съдържащи само текст. Тази информация обаче не е достатъчна за отговаряне на въпросите, съдържащи диаграми и картинки. На входа базовият модел приема урока като контекст, единствен въпрос и възможни отговори, достигащи най-много 7 възможности. Очаква се като изход правилния отговор. Размерът на урок средно е над 1,000 думи, което не позволява съхраняването му на една графична карта. За да се справят с този проблем, авторите използват подход, базиран на извличане на информация: за всеки параграф те пресмятат сходство (чрез скаларно произведение) до въпроса, сумирайки tf-idf стойностите на всички думи и избират параграфа с най-голямо сходство. За получаването на влагания поотделно за всяко изречение от параграфа, въпроса и отговорите се използва LSTM. След това чрез механизма внимание се избира влагането на думата с най-голяма добавена стойност, която се смята за отговор на въпроса. Това влагане се сравнява по сходство с всеки от отговорите и за верен отговор се избира този с най-голямо сходство.

Визуалните модели са разширение на базовия текстови модел. Разликата между двата типа модели е добавеният визуален контекст. Авторите сравняват два визуални модела: един, базиран на методи, използвани в големи визуално-текстови модели, и модел, разширение на DSDP-NET [27]. В първия вид визуален модел картиката преминава през VGG невронна мрежа и стойностите на параметрите от последния конволюционен слой се считат за съответния визуален контекст, който поради спецификата на VGG се изразява като 49 вектора, всеки с по 512 елемента. Получените вектори се трансформират към размерността, използвана за влагане на думите, използвайки два скрити

слоя с активационна функция тангенс хиперболичен. Резултатните вектори са конкатенирани към изходните такива от LSTM-а и по този начин моделът може да прилага внимание и към картинката. Вторият тип визуален модел включва използване на структурната информация в картинка за получаване на свързан ориентиран граф от свързани обекти, който в последствие може да бъде трансформиран към изречения, описващи началото и края на всяка дъга. Задачата на модела е да постави семантично значение върху връзката между двата обекта. Следвайки този подход, всяко изображение се трансформира към няколко изречения, които се добавят в параграфа, който реферира към изображението. След това се прилага изчисляването на сходство и процесът продължава, както е в текстовия модел.

Резултатите от експериментите сочат, че текстовия модел не се представя добре на въпросите от тип “истина” - “лъжа”, постигайки едва 50.2% точност. Това е възможно да се дължи на голямата трудност на тези въпроси - за правилен отговор се изисква повече създаване на верига от разсъждения, а не толкова търсене в текста, в което моделите, базирани на памет и внимание не са добри. Моделът се представя по-добре на въпроси с множество отговори, задминавайки случаен избор с близо 10% пункта точност. Въпреки това точността остава ниска - 32.9%, което отново се държи на високата трудност на въпросите. На въпросите, включващи картинка, първият тип визуален модел не постига по-добри резултати от текстовия модел, но вторият тип визуален модел, използващ структурен граф за описание на картинката, постига подобрение от почти пункт и половина над останалите. Това е най-добрият тестван модел. Точността остава ниска - 31.3%, което се дължи на богатото съдържание на повечето диаграми, изискващо развита възможност за изграждане на причинно-следствени връзки, формирането на които изисква информация от целия урок.

2.2.4. Оценяване на набор от данни SciBench

Наборът от данни SciBench се фокусира върху изследването на представянето на модели върху въпроси от предмети, свързани с Математика [7]. Целта е систематично да се изследват нужните уменията за разсъждаване над и решаване на комплексни математически задачи. Състои се от две множества данни, съдържащи въпроси от трудност в университет: с отворен характер и със затворен характер. SciBench не съдържа многомодални въпроси.

Отворените въпроси са 695 и са събрани от популярни университетски курсове: Физика, Термодинамика, Класическа механика, Квантова химия, Диференциално и интегрално смятане и Статистика. Второто множество се състои от 104 затворени въпроса взети от семестриални контролни и изпити. С цел намаляне на вероятността за отгатване на верния отговор затворените въпроси също са представени като отворени, т.е. при оценяването на моделите не се използват възможни отговори.

Сравнени са моделите GPT-3.5-Turbo (основа на чатбота ChatGPT) и GPT-4 с температура 0, за да се наблегне на точността и на детерминистичните разсъждения. Разгледани са два вида стратегии за инструктиране на моделите - чрез извличане на верига от разсъждения и преобразуване на част от разсъжденията към програмен код, написан на езика Python или Wolfram. Това се прави с цел постигането на по-точни резултати при смятане. Наред с това в експериментите, изключващи предоставянето на примери, се проверява

добавената стойност на инструкциите, описващи типа и категориите на въпросите и какво се очаква от модела. По този начин авторите формират седем експеримента за оценяване:

1. без предоставяне на примери и без инструкции;
2. без предоставяне на примери и с инструкции;
3. с предоставяне на няколко примера;
4. без предоставяне на примери и с извличане на верига от разсъждения;
5. с предоставяне на примери и с извличане на верига от разсъждения;
6. с предоставяне на примери и конвертиране на части от разсъжденията към Python;
7. с предоставяне на примери и конвертиране на части от разсъжденията към Wolfram.

Резултатите показват, че моделът GPT-4 се представя по-добре от модела GPT-3.5-Turbo във всички 7 експеримента. Най-големи подобрения са наблюдават при предоставяне на примери и с извличане на верига от разсъждения и при предоставяне на примери и конвертиране на части от разсъжденията към Python - съответно с 16.36% и с 15.89%. Също така в общия случай не се наблюдава голяма добавена стойност на включването на примери. Осреднената точност на моделите с и без предоставени примери е съответно 12.17% и 11.99% за GPT-3.5 и 28.52% и 28.35% за GPT-4. Във високоспециализирани предмети като Квантова химия добавянето на примери все пак води до по-висока точност с почти 3 пунка за GPT-4, макар че в предмети като Физика добавянето на примери води до намалена точност с 6.99%. Тази вариации могат да се дължат на степента на представителност на примерите, които се използват - очаква се, че примери, цялостно представящи разнообразието на множеството, към което принадлежат, ще водят до постигането на по-висока точност. Наблюдава се също и тенденцията за по-висока точност при извличане на веригата от разсъждения. Експериментите, насочени към изследването на зависимостта между генериране на код на Python и Wolfram, показват, че използването на Python увеличава средно точността съответно с 7.92% за GPT-3.5-Turbo и 7.45% за GPT-4 спрямо използването на верига от разсъждения, но използването на Wolfram намалява точността с 4.12% за GPT-3.5-Turbo и с 12.79% за GPT-4.

2.2.5. Оценяване на набор от данни MathVista

MathVista [4] е множество от данни с фокус върху Математиката и разсъждаването върху фигури и математически обекти. Въпросите са събрани от седем дисциплини: Алгебра, Аритметика, Геометрия, Логическо мислене, Прости числови операции, Научно мислене и Статистическо мислене. Обхваща широка гама визуално съдържание - естествени картини, геометрични фигури, графики и изкуствени сцени и диаграми.

Авторите отбелязват, че досега анализът почти винаги е бил фокусиран върху свеждането на развитието на модели до качествена оценка. За разлика от това, те провеждат количествен и качествен систематизиран анализ на резултатите на основни модели, за да проверят доколко развита възможност имат те в разсъждаването при визуален контекст.

Те създават нов начин за оценяване на отговорите, давани от големи многомодални модели. Нуждата за това е продиктувана от тенденцията големи езикови и големи многомодални модели да се тренират по начин, който изисква

получаването на дълъг отговор в разговорен стил, вместо кратък отговор. Авторите предлагат автоматизиран начин за задаване на въпрос и извличане на отговора му, следвайки процедура от три стъпки: създаване на отговора, извличане на отговора и пресмятане на точност. Първата стъпка включва подаване на входа инструкция заедно с въпроса и получаването на отговор. Инструкцията включва описание на формата, в който трябва да бъде върнат отговорът, въпроса, възможните отговори и метаинформация за въпроса. Следващата стъпка цели да извличе буквата или цифрата, съответстваща на верния отговор. Авторите предлагат използване на система, за извършване на тази задача, базирана на голям езиков модел. Те използват GPT-4 и посочват усъвършенстваното му умение за извличане на информация от текст. Моделът успява да извлече правилно отговора от текста в 99.5% от случаите. Последната стъпка включва преобразуване на извлечения отговор в предварително дефиниран формат, който след това позволява пресмятане на точността спрямо очаквания отговор.

Авторите разпределят моделите, които използват, в три направления. Първото включва големите езикови модели GPT-3.5-Turbo, GPT-4 и Claude-2 в контекста без и с предварително подаване на два примера и с извличане на верига от разсъждения и преобразуване на част от веригата, включваща изчисления, към код. Второто направление включва същите големи езикови модели, но в този случай освен текстовото съдържание на въпроса към входа се добавя и генерираното от чатбота Bard обяснение за съдържанието на картинката и разпознатият текст в картинката от EasyOCR. Третото направление включва големи многомодални модели, част от които са GPT-4V, Bard (not a model?), LLaMA-Adapter-V2-7B, Vicuna и miniGPT-4. За пресмятане на човешката точност авторите използват Amazon Mechanical Turk.

Резултатите показват, че всички модели се справят по-добре от случаен избор на отговор, но в същото време не успяват да се постигат по-добри резултати от човек (60.3%). Най-добрият модел в първото направление е GPT-4 с два предварително предоставени примера и извличане на верига от разсъждения. Той постига точност от 29.2%. Това показва, че наборът от данни изисква разсъждения, базирани на визуален контекст. След добавяне на описание на съдържанието в картинката и намереният текст в нея се наблюдава цялостно подобрение на точността. Най-добрият модел е GPT-4, постигащ точност от 33.9% с два предварително предоставени два примера и преобразуване на част от разсъжденията към програмен код на Python.

При големите многомодални модели се наблюдава голяма разлика между GPT-4V и всички останали. GPT-4V постига точност от 49.9%, което е с 15.1% по-високо от втория най-добър модел - Bard + Google Lens (34.8%). Моделите с отворен програмен код не се представят задоволително. Моделът LLaVA постига най-висока средна точност от 26.1%, което показва, че в тези модели се наблюдава липса на възможност за добри разсъждения в математически контекст, разпознаване на текст и фигури и разбиране на графики.

Качествения анализ на резултатите показва, че чатботът Bard и моделите GPT-4 и GPT-4V често връщат правилен отговор, но с грешки в обясненията. Bard изпада в тази ситуация с частично грешни разсъждения в 6.8% от случайте и с изцяло грешни разсъждения, но правилен отговор в 8.1% от случайте.

2.2.6. Оценяване на набор от данни MMMU

Авторите на [9] представят MMMU като набор от данни, създаден, за да оценява многомодални модели на масивно многодисциплинарни задачи, решаването на които изисква умения, придобити в университетски курсове, и съзнателни разсъждения. Цели да измери успеваемостта на моделите в три основни направления: възприятие, знания и разсъждения.

Експериментите са фокусирани в оценяването на големи многоезикови и големи многомодални модели без предварително предоставени примери. За проверка дали разпознаването и добавянето на текста от картинките води до по-добри резултати, се използва MMOCR, а за описание на съдържанието на картинките се използва LLaVA-1.5. За извличане на отговорите от дълги параграфи, върнати от моделите, се използват регулярни изрази и процедури за последваща обработка. При липса на валиден отговор за въпрос с няколко възможности, се избира случаен отговор, а при отворени въпроси отговорът се смята за грешен.

Най-добрият голям многомодален модел е Gemini Ulta, следван от GPT-4V, постигащи съответно 59% и 56% точност (те не тестват Gemini Ulta на test, а на val, т.ч. това трябва да се промени). Отново се наблюдава голям спад в качеството на получените отговорите при модели с отворен код като BLIP2-FLAN-T5XXL и LLaVA-1.5, които достигат едва 34% точност.

Авторите наблюдават, че добавянето на разпознат текст в и описание на картинката не води до статистически значимо подобрение в резултатите, което означава, че модел, който се справя добре би следвало да има възможност за добра интеграция на визуална и текстова информация.

Спрямо различните предмети средната точност варира. Наблюдава се, че при дисциплини като Изкуство и Социология средната точност на моделите е по-висока, което се дължи на сравнително по-лесните и естествени въпроси. Дори и моделите с отворен код се представят сравнително добре в категории като Снимки и Рисунки най-вероятно, защото са често срещани при трениране. В дисциплини като Наука, Медицина и Компютърни науки обаче точността е по-ниска, т.к. въпросите изискват по-задълбочено мислене. Това е част от наблюдавана тенденция на постепенно намаляване на добавената стойност на големи и комплексни модели като GPT-4V при увеличаване на трудността на въпросите. Например, докато разликата между моделите InstructBLIP-T5-XXL (40.3%) и GPT-4V (76.1%) е 34.8% на въпроси, класифицирани като лесни, тя бързо намалява при въпроси, класифицирани като трудни - моделите постигат съответно 29.4% и 31.2% точност.

Авторите провеждат анализ на грешките, които GPT-4V допуска. За целта те избират на случаен принцип 150 примера, за които са получени грешни резултати. Те са ръчно обходени за идентифициране на причината за грешките. Анализът показва, че 35% от грешките се дължат на грешки при възприятието. Тези грешки могат да се разбият на два типа: повърхностни и специфични за домейна. Повърхностни грешки моделът допуска при неразбиране на посоки. Грешки, специфични за домейна, моделът допуска, когато няма нужните знания. Също така се наблюдава тенденция моделът да поставя по-голямо внимание и тежест на текстовото съдържание. Грешките от незнание са причина за грешен отговор в 29% от разгледните случаи. Пример за това е неразбирането на крайни състояния в крайни детерминистични автомати. В контекста на медицината се наблюдава липса на достатъчно обширен контекст за правилното разпознаване на болест по подадено описание и таблица. В 26% от случаите се наблюдават грешки в разсъжденията. Това си личи най-добре в контекста на математиката,

където моделът разпознава правилно необходимата информация, но липсата му на развити умения за боравене с математически инструменти резултира в грешен отговор.

2.2.7. Оценяване на набор от данни M3Exam

Наборът от данни M3Exam се характеризира с многомодалност, многоезичност и въпроси, които са взети от всички нива на обучение до достигане на средно образование [8]. Обхванате са езиците Английски, Китайски, Италиански, Португалски, Виетнамски, Тайски, Суахили, Африкаанс и Явански (Индонезия). Изборът е продиктуван главно от желанието да се създаде набор от данни, който има не само лингвистични, но и културни разлики. Така се покриват различни езикови групи и езици с неравномерна популярност, за които няма много информация в Интернет.

Авторите подбират големи езикови модели и големи многомодални модели. Сред големите езикови модели са GPT-3.5-Turbo, GPT-4, Claude, BLOOM и Vicuna. Сред големите многомодални модели са BLIP-2, InstructBLIP, Fromage и OpenFlamingo. Оценяват се моделите без предварително подаване на примери, главно поради това, че репликират най-добре реалния свят, трудно е да се подадат няколко картинки за отделни примери и при трениране моделите преминават през инструкционно нагласяне, т.ч. са готови да отговорят на въпросите. Авторите включват името на предмета, за който се отнася въпросът, в тяхната инструкция и инструктират модела да не поражда разсъждения, а само да представи крайния отговор, който смята за верен. Във връзка с моделите бъдейки генеративни инструкцията след като предостави въпроса, завършва с *Answer:*, подтиквайки модела към запълване с отговора. За всеки език инструктиращият текст се превежда.

Всички въпроси са затворени и съдържат краен набор от отговори. Авторите наблюдават, че в някои случаи моделите връщат дълги отговори, съдържащи не само идентификатора на вярната опция, но и аргументация и допълнителни примери, както и аргументи защо другите опции са грешни. По този начин в целия отговор се появяват всички опции. Авторите вземат първата срещната опция и считат нея за верен отговор. Това крие своите рискове, т.к. е възможно първо моделът да аргументира защо грешни опции са грешни и след това да напише вярната опция по метода на изключването.

Резултатите показват, че докато средната точност, нужна за минаване на изпитите е около 54%, единствено моделите GPT-3.5-Turbo и GPT-4 успяват да постигнат по-високи средни резултати - съответно 57.57% и 72.92%. Наблюдава се, че GPT-4 се представя най-добре в различните езици. Моделът BLOOM се представя по-лошо от случайния избор. Средните точностите на моделите GPT-3.5-Turbo, Claude и Vicuna варират спрямо езика. Докато GPT-3.5-Turbo и Claude имат сходна точност за английски (75.98% и 74.25%), GPT-3.5-Turbo постига по-високи резултати в останалите езици, демонстрирайки по-развита многоезична способност. На ниво език, средно моделите имат намалена точност при нелатинските езици като Китайски (въпреки, че за него има доста данни, т.е. това е неочаквано) и езици, за които няма много данни като Явански, въпреки че в него главно се използва латинската азбука (което е очаквано, но и неочаквано?).

Авторите експериментират с две стратегии за инструктиране на GPT-3.5-Turbo в контекста на различни езици. Първата се изразява в превод на

иструкцията на Английски език, но запазвайки съдържанието на въпроса на изходния език. Втората се изразява в превеждането и на инструкцията и на съдържанието на Английски език. Резултатите показват, че превеждането само на инструкцията на Английски език не подобрява консистентно резултатите. Възможно е това да се дължи на това, че изходният език е оригинален и не е получен чрез превод от Английски. Също така е възможно използването на Английски език да не подсказва правилно на модела какви умения и знания са нужни за решаването на задачата. Превеждането на инструкцията и данните също не подобрява консистентно резултатите. От една страна, повечето въпроси са тясно свързани с езика, от който са написани, и превеждането им може да доведе до загубена ключова информация, което резултира в намалена точност. От друга страна, използването на преведен вариант може да премахне бариерите към разбирането на някои езици особено тези, за които GPT-3.5-Turbo вижда трудности като Тайски и Явански. За тях се наблюдава увеличаване на точността с по 20 пункта.

И в тази статия авторите провеждат експерименти, свързани с предварително подаване на примери. Те дават инструкцията след това няколко примера и след това тестовият въпрос. И те наблюдават, че предоставянето им не води до консистентно покачване на точността в общия случай. В частни случай като Португалски и Виетнамски има подобрение, в други езици като Китайски и Суахили се наблюдава деградация. Това отново показва колко важно е подбраните примери правилно да представят цялото, за което се водят представителни. Също така е възможно съществуващите големи езикови модели да са наясно/свикнали с формата на въпросите и добавянето на примери да не носи добавена стойност. Допълнителни фактори, които влияят на ефективността на примерите са трудността на езика, знанията на модела, начина на избор и други.

Авторите експериментират и с многомодални модели. По време на създаване на статията обаче не съществуват многомодални модели, които да имат многоезиково разпознаване, затова използваният език е английски. Използва се Flan-T5 въпреки че е само текстови модел. Той постига добри резултати. Към него подават само текстовата част на въпроса, а към многомодалните модели - BLIP-2 и InstructBLIP, за които се използва Flan-T5 като енкодер, подават първото изображение, ако има повече от един. Неочаквано е, че многомодалните модели не се представят значително по-добре от Flan-T5. Всъщност единствено BLIP-2 постига по-висока точност, но едва с 0.76%. Анализ на грешките показва, че многомодалните модели се затрудняват при разбирането на сложни детайли като детайли за осите в математически въпроси и детайли за картата в географски въпроси. Въпреки че моделите Fromage и OpenFlamingo са специално тренирани, за да могат да работят с множество картинки, те не се представят по-добре и дори се представят по-зле от модели като BLIP-2 и InstructBLIP, които работят само с една картинка. Анализът показва, че разбиране дори и на една картика е трудно за тези модели (Fromage и OpenFlamingo). Това показва, че претрениране с няколко изображение не води непосредствено до по-добро разбиране на многомодален вход и правене на крос-референции между няколко снимки.

Разпределението на точността на моделите спрямо нивата на учениците е неравномерно. Учудващо, с увеличаване на нивото, авторите не наблюдават тенденции за спад на точността. Най-добрите резултати се наблюдават за второ ниво, а най-ниските за трето (най-сложното).

2.2.8. Оценяване на набор от данни Exams

Наборът от данни е създаден от Хардалов и колеги [2]. Фокусира се върху оценяване на големи езикови модели, в частност, сравнява каква точност постигат, когато отговарят на еднакви въпроси, преведени на различни езици. Експериментите включват фина настройка на многоезичния BERT модел и модела XLM-RoBERTa. Най-добрият модел е XLM-RoBERTa, който е фино настроен чрез RACE, AI2 English science datasets и най-накрая на тренировъчното множество на Exams, постигайки средна точност от 42%. Авторите наблюдават, че въпросите от Наука за природата затрудняват най-много моделите. Това се дължи главно на предметите Химия и Физика. Предметът Информатика също затруднява моделите, т.к. изисква разбиране и създаване на програмен код, както и познаването и работенето с различни бройни системи.

2.2.9. Сравнителен анализ и изводи

Таблица X показва сравнителен анализ на експериментите и получените от тях резултати.

Име	Публикуван (Година/Месец)	Многомодалност	Най-добър модел	Втори най-добър модел	Предоставяне на примери помага	Текст + Картинка по-добро от текст	OCR & Описание на картинка помага	Верига на разсъжденията помага
VQA	2016/10	да	deeper LSTM Q + l2 norm VGGNet	LSTM Q + VGGNet	-	да	-	-
ScienceQA	2022/10	да	GPT-3	UnifiedQA	до 2 примера	да	-	да
TQA	2017/07	да	Text + DPG	Text + Image | Text Only	-	да	-	-
SciBench	2023/07	не	GPT-4	GPT-3.5	не	-	-	да
MathVista	2023/10	да	GPT-4V	Bard + Google Lens	-	да	да	-
MMMU	2023/12	да	GPT-4V	Qwen-VL-PLUS	не	да	не	-
M3Exam	2023/11	да	GPT-3.5-Turbo	BLIP-2	не	не	не	-
Exams	2020/11	не	XLM-R (RACE + SciENs + Exams)	XLM-R (RACE + SciENs)	-	-	-	-
Exams2	2024/01	да	Gemini Ultra	GPT-4V	-	-	-	-

Таблица X: Сравнителен анализ

Обзорът на статиите показва, че:

1. Въпроси, успешното отговаряне на които изисква използването на голям контекст, са по-трудни от специфични за домейна въпроси.
2. Конкатенирането на репрезентации на многомодални данни в една структура допринася за по-голяма трудност на модела при разсъждаване.
3. За цялостното оценяване на модел е необходимо разнообразие от типове въпроси, които варират в количеството разсъждения, нужни за достигането до правилен отговор, и количеството контекст, необходимо за формирането на тези разсъждения.

4. Тенденцията е модели, тренирани с по-голямо количество данни, да се представят по-добре от модели, тренирани с по-малко количество данни, независимо от контекста и специфичността.
5. Използването на стратегии за инструктиране, които включват извличане на веригата от разсъждения, създадена от модела водят до по-добри резултати.
6. Наблюдава се тенденция GPT-4V да постига по-добри резултати от Bard (there is a problem here, right - gpt is a model, while bard is a chatbot - clear this up)?
7. Докато средно точността на големи многомодални модели като GPT-4V не надмива тази на човек, тези модели се справят по-добре от човек в задачи, свързани с детайлно познаване на специфичен домейн.
8. Няма много статии, използващи Bard (Gemini), но има много с GPT-3.5-Turbo и GPT-4(V).
9. Няма много статии, анализиращи грешките, които Bard допуска.

2.3. Многомодални модели

2.3.1. Чатбот Bard

Bard е многомодален и многоезичен чатбот, поддържан от софтуерната компания Гугъл. Той се предоставя за свободно използване през март 2023 и е широко приеман да е пряк конкурент на създадения по-рано и постигнал много успехи чатбот, създаден от компанията OpenAI - ChatGPT. Докато ChatGPT е базиран на архитектурата GPT 3.5 Turbo, Bard използва първоначално LaMDA (Language Models for Dialog Applications), след това PaLM 2 (Pathways Language Model 2), а наскоро и най-новия голям многомодален модел на Google - Gemini. Първоначално Bard, използвайки LaMDA, приема инструкции само под формата на текст и извършва разнообразни задачи, свързани с отговор на въпроси, резюмиране на голямо съдържание и създаване на различни видове текстово съдържание [15]. На 13 юли 2023 г. Bard се превръща в първия свободно достъпен многоезичен и многомодален модел, който може приема картинки и текст като вход. По това време съществуват и други модели, за които се твърди, че могат да работят с картинки, но те все още не са публични. Такъв модел е GPT4. Той наследява GPT 3.5 Turbo - големият езиков модел, използван от ChatGPT. Новите модели стъпват върху възможностите на предишните, но ги и надграждат и подобряват - моделът LaMDA се характеризира с възможност за провеждането на дълги разговори, които силно наподобяват реални. Моделът PaLM 2 надгражда и подобрява възможностите за генериране на нов текст и превод между различни езици. Моделът Gemini разширява тези възможности и поставя фокус върху генерирането и обясняването на програмен код, както и работа с нови видове вход - аудио и видео. Като резултат Google Bard може да превключва между различни начини на комуникация и да адаптира своите отговори въз основа на контекста на разговора. В допълнение чатботът има достъп до Интернет и не е ограничен във времето, т.е. може да бъде използван, за да резюмира и анализира нова информация. Основният конкурент на Bard - ChatGPT е ограничен да работи с информация само до определен момент във времето. Към момента това е януари 2023 г.

2.3.2. Чатбот LLaVa

todo

2.3.3. Чатбот ChatGPT

todo

2.3.4. Модел GPT4

todo

2.3.5. Модел Quen-VL

todo

3. Набор от данни Exams2

3.1. Същност

todo

3.2. Създаване на множества за тестване

train / test, test-mini

todo

4. Експерименти

4.1. Големи многомодални модели

4.1.1. Същност и цели

Целта е по подадена картинка да се върне правилния отговор. Към картинките не се прилага предварителна обработка. По този начин се проверява доколко може моделът да идентифицира и разбере въпроса в картинката, както и доколко може да идентифицира и разбере изображения или диаграми, прилежащи към въпроса.

4.1.2. Планиране и подготовка на среда за провеждане на експерименти

Използваната инструкция е: *В изображението има въпрос с от 2 до 5 възможни отговора. Изведи правилния отговор във формат JSON, както*

следва: {'answer': 'xxx'}. Замести „xxx“ със съответната буква: „A“, ако първият отговор е правилен, „B“, ако вторият отговор е правилен, „C“, ако третият отговор е правилен, „D“, ако четвъртият отговор е правилен, или „E“, ако петият отговор е правилен.

Въпреки че инструкцията е подробна и в частност изисква получаване на отговора във формат JSON, получените резултати от Bard рядко спазват посочения формат. В повечето случай се представя даден отговор и последващи го обяснения. Това налага нуждата от последваща обработка върху получените отговори от Bard за изваждане на буквата, съответстваща на отговора. Част от последващата обработка се извършва чрез код, използващ регулярни изрази. Понеже възможностите имат голяма вариативност и са неконсистентни е нужна и ръчна обработка чрез подходящ софтуер за визуализиране и обработка на таблици - в случая бе използван LibreOffice Calc.

За комуникация с програмния интерфейс на моделите бе използван Python пакетът requests. Той предоставя възможност за изпращане и получаване на заявки във формата на HTTP/1.1. От гледна точка на потребителя, библиотеката е написана на високо ниво и позволява автоматично изграждане и поддържане на множество сесии към HTTP сървъри. Примери за това са автоматичното добавяне на параметри към GET заявки, автоматичното прехвърляне на данните като параметри на заявката вместо на URL адреса при заявки от тип POST, SLL верификация, автоматично декодиране на върнатите отговори и други.

Чатботът Bard е свободно достъпен чрез уеб приложение в браузъра, използвайки адреса https://bard.google.com/chat . Към момента няма свободно достъпен програмен интерфейс за комуникация чрез HTTP заявки с Bard, който официално да се поддържа от Гугъл. Това е недостатък спрямо конкурентите на Bard и в частност чатботът ChatGPT - компанията OpenAI предоставя такъв пакет, който може да се използва, след закупуване на ключ за достъп. Въпреки това в [16] може да бъде намерена неофициална и поддържана от обществото версия на програмен интерфейс. Функционалностите имат технически лимитации. Честите обновявания на функционалностите на Bard понякога водят до неочаквано поведение при използване на методи.

Начинът на взаимодействие според [16] се изразява в създаване на сесия чрез requests и задаване на стойности на определени бисквитки, които позволяват изпращането на заявки и получаване на отговори. При изтичането на валидността на поне една от тях, се появява грешка с текст *Response Error*. Трите бисквитки имат временна валидност и изтичат нееднакво. При изтичането на коя да е от тях се нарушава възможността за комуникация чрез Bard. Това води до непредвидимо количество заявки, които могат да се подават към модела. Оказва се, че заявките изтичат различно в зависимост от държавата и региона, от които се правят заявките, и в зависимост от времето на деня, когато се правят те. Освен това влияние оказва и неравномерното натоварване на платформата, поддържаща Bard.

След множество експерименти се оказа, че максималният брой заявки за клиент, разположен в България, е 89. В допълнение след една успешна сесия с 89 въпроса, трябва да се изчака поне 10 часа, за да се извърши отново успешна сесия с 89 въпроса.

4.1.3. Резултати

todo

https://github.com/Vision-CAIR/MiniGPT-4
https://huggingface.co/spaces/Vision-CAIR/minigpt4
https://huggingface.co/spaces/Vision-CAIR/MiniGPT-v2

5. Големи езикови модели

5.1. Същност и цели

todo

5.2. Планиране и подготовка на среда за провеждане на експерименти

todo

5.3. Резултати

Автоматично отговаряне на въпроси след разпознаване на текста от въпроса и описване на прилежащите изображения

https://huggingface.co/google/flan-t5-xxl
https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

todo
a table with the results;
statistics for each model

6. Големи езикови модели с превод на въпроса на английски език

6.1. Същност и цели

todo

6.2. Планиране и подготовка на среда за провеждане на експерименти

Автоматично отговаряне на въпроси след разпознаване на текста от въпроса и описване на прилежащите изображения.

6.3. Резултати

todo

7. Анализ на допусканите грешки

Моделите успяват да постигнат добри резултати и показват голям потенциал за успешно бъдещо развитие. Все пак се наблюдават и затруднения в даването на отговор за широка гама въпроси. Грешните отговори могат да се класифицират в два типа: (a) моделът не успява да разбере правилно и цялостно многомодалния вход и няма достатъчно опит в приложната област, за да достигне до правилния отговор; (b) моделът генерира грешни разсъждения с ненужна, невярна или непълна информация.

8. Заключение

Настоящата дипломна работа използва наборът от данни Exams2 за предоставяне на множество отправни точки за бъдещи сравнения с големи многомодални модели, включително Gemini Pro и GPT-4. Допълнителни експерименти са проведени с цел изследване на добавената стойност на визуалното съдържание и допринесената трудност на езици, различни от Английски. Експериментите показват, че най-добрият модел върху Exams2 е ??? и че спрямо различните езици точността може да варира много ??? малко.

С цел по-нататъшно развитие може да се добавят още въпроси от тези модалности; могат да се добавят и изследват различни нови модалности; могат да се използват по-нови и по-добри големи многомодални модели.

Обобщение на изпълнението на началните цели
Насоки за бъдещо развитие и усъвършенстване

- добавяне на човешки baseline
- експериментиране с различни стратегии за извличане на разсъжденията - chain of thought, program of thought
- може да се направи за всеки изпит и оценка на това дали моделът получава над 2, което би означавало взет изпит (M3Exam)
- повече експерименти - few shot vs zero shot, cot vs without, pot vs without, picture only vs text only and the combinations from those

9. Използвана литература

1. Goertzel, Ben. "Artificial general intelligence: concept, state of the art, and future prospects." *Journal of Artificial General Intelligence* 5.1 (2014): 1.
2. Hardalov, Momchil, et al. “Exams: A Multi-subject high school examinations dataset for cross-lingual and multilingual question answering.” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5 Nov. 2020, https://doi.org/10.18653/v1/2020.emnlp-main.438.
3. Hardalov, Momchil, Code and Data for Exams: https://github.com/mhardalov/exams-qa
4. Lu, Pan, et al. ‘MathVista: Evaluating Math Reasoning in Visual Contexts with GPT-4V, Bard, and Other Large Multimodal Models’. arXiv [Cs.CV], 2023, http://arxiv.org/abs/2310.02255. arXiv.
5. Lu, Pan, et al. ‘Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering’. arXiv [Cs.CL], 2022, http://arxiv.org/abs/2209.09513. arXiv.
6. Antol, Stanislaw, et al. ‘VQA: Visual Question Answering’. CoRR, vol. abs/1505.00468, 2015, http://arxiv.org/abs/1505.00468.
7. Wang, Xiaoxuan, et al. ‘SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models’. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2307.10635. arXiv.
8. Zhang, Wenxuan, et al. ‘M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models’. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2306.05179. arXiv.
9. Yue, Xiang, et al. ‘MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI’. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2311.16502. arXiv.
10. Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4999–5007, 2017.
11. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In ICCV, 2015
12. J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In EMNLP, 2013
13. Q. Cai and A. Yates. Large-scale semantic parsing via schema matching and lexicon extension. In ACL, 2013
14. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ ar, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, 2014
15. Qin, Haotong, et al. “How Good Is Google Bard’s Visual Understanding? An Empirical Study on Open Challenges.” *Machine Intelligence Research*, vol. 20, no. 5, Oct. 2023, pp. 605–13. *arXiv.org*, https://doi.org/10.1007/s11633-023-1469-x.
16. Bard-API. *Daniel Park*, Minwoo Park 2023. *GitHub*, https://github.com/dsdanielpark/Bard-API.
17. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6904–6913, 2017.

18. Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
19. Huang, Lei, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232, arXiv, 9 Nov. 2023. arXiv.org, http://arxiv.org/abs/2311.05232.
20. Edouard Belval, A wrapper around the pdftoppm and pdftocairo command line tools to convert PDF to a PIL Image list: https://pypi.org/project/pdf2image .
21. Ivan Goncharov, A modified version of https://github.com/Cartucho/OpenLabeling OpenLabelling tool: https://github.com/ivangrov/ModifiedOpenLabelling .
22. Jeffrey A. Clark (Alex), The Python Imaging Library adds image processing capabilities to your Python interpreter: https://pypi.org/project/Pillow .
23. Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. In Findings of the Association for Computational Linguistics (EMNLP), pages 1896–1907, 2020.
24. Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big selfsupervised models are strong semi-supervised learners. Advances in neural information processing systems (NeurIPS), 33:22243–22255, 2020.
25. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In The International Conference on Learning Representations (ICLR), 2021.
26. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.
27. A. Kembhavi, M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In ECCV, 2016.
28. Bai, Jinze, et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966, arXiv, 12 Oct. 2023. arXiv.org, http://arxiv.org/abs/2308.12966.

1. Това е препоръчителен шаблон, в зависимост от конкретното задание.

2. Йерархията на структуриране на съдържанието да не бъде повече от 3 нива, номерирани с арабски цифри – напр. 1.2.3.

3. Чуждестранните термини да бъдат преведени, а където това не е възможно – цитирани в курсив и нечленувани.

4. Страниците да бъдат номерирани с арабски цифри, в долния десен ъгъл.

5. Използваният шрифт за основния текст на описанието да бъде Times 12 или Arial 10, и Courier 9 за кода, с междуредие 16pt.

6. Да се избягват пренасянията на нова страница на заглавия на секции, фигури и таблици.

7. Да се избягват празни участъци на страници вследствие пренасянето на фигури на нова страница.

8. Всички фигури и таблици да бъдат номерирани и именовани (непосредствено след фигурата или таблицата).

9. Всички фигури и таблици да бъдат цитирани в текста.

10. Използваните фигури от други източници да бъдат цитирани.

11. Всички цитати да бъдат отразени в списъка на използваната литература.

12. Всички източници от списъка на използваната литература да бъдат цитирани в текста.

13. Използваната литература да се цитира съгласно MLA Style - http://www.library.mun.ca/guides/howto/mla.php