Универзитет „Св. Кирил и Методиј“ во Скопје

Факултет за информатички науки и компјутерско инженерство

Вовед во науката за податоци

Извештај

Споредба на LLMs за OCR на документи

Студенти:

Ева Смилеска 221053

Љубица Дамјановиќ 221173

Професор:

Игор Мишковски

Содржина

Вовед

Оптичкото препознавање на карактери (OCR) е клучна технологија за дигитализација на документи. Во последните години, големите јазични модели (LLM) покажаа значителен потенцијал за подобрување на точноста и флексибилноста на OCR системите. За разлика од традиционалните OCR решенија, LLM моделите можат да разберат контекст, да препознаат сложени структури и да извлечат структурирани податоци директно од слики.

Предизвици на современото OCR

И покрај значајниот напредок, OCR технологијата и понатаму се соочува со низа предизвици:

- Ниска точност при нејасни или оштетени документи, каде што класичните алгоритми често создаваат голем број грешки.
- Проблеми со различни фонтови и јазици, особено кога станува збор за нестандардизирани текстови или документи на повеќе јазици.
- Недостаток на разбирање на контекст, што води до погрешна интерпретација на зборови или изрази.
- Потреба за рачна пост-обработка, при што човечка интервенција е неопходна за корекција на добиените резултати.

Предности на LLM пристапот

LLM моделите претставуваат можен одговор на овие предизвици преку следните предности:

- Контекстуално разбирање на текстот, што овозможува корекција на грешки врз основа на значењето.
- Извлекување на структурирани податоци, дури и од комплексни документи со табели и графички елементи.
- Автоматско подобрување на грешки, преку анализа на околните зборови и семантичка логика.
- Флексибилност за различни типови документи, вклучувајќи мултијазични и нестандардни формати.

Мотивација за истражувањето

Со оглед на тоа што OCR останува суштинска алатка за дигитална трансформација, интеграцијата на LLM претставува значаен чекор напред. Ова истражување има за цел да ги спореди различните LLM модели и нивната ефикасност за OCR задачи, со акцент на:

- **Точност на препознавање текст** – мерење на степенот на коректно извлечени информации.
- **Способност за структурирање на податоци** – трансформација на суров текст во корисни формати.
- **Перформанси и брзина** – колку ефикасно моделите работат со голема количина на документи.
- **Економска ефикасност** – баланс меѓу точност, брзина и потребни ресурси за имплементација.

Цели на истражувањето

Главна цел

Да се изврши сеопфатна споредба на современите LLM модели за OCR задачи на различни типови на документи и да се идентификуваат најефикасните решенија за специфични случаи на употреба.

Специфични цели

1. Проценка на точноста

- Споредба на точноста на текстуално препознавање
- Анализа на грешки во различни услови
- Оценување на способноста за структурирање на податоци

2. Анализа на перформанси

- Мерење на времето за обработка
- Проценка на користење ресурси
- Скалабилност на решенијата

3. Економска анализа

- Споредба на трошоците за обработка
- Односот точност-цена за различни модели
- ROI анализа за различни сценарија

4. Практичност на примена

- Лесност на инсталирање и конфигурирање
- Интеграција со постоечки системи
- Поддршка за различни јазици и формати

Методологија

Истражувачки пристап

Користиме експериментален пристап со контролирани услови за тестирање на различни LLM модели. Секој модел се тестира на ист сет од документи со идентични параметри.

Експериментален дизајн

1. Контролирани променливи:

- Типови документи (константни)
- Квалитет на слики (константен)
- Јазик на документи (македонски/англиски)
- JSON шеми за структурирани податоци

2. Независни променливи:

- Тип на LLM модел
- Параметри на моделот
- Пристап на обработка (директен/двоетапен)

3. Зависни променливи:

- Точност на текстуално препознавање
- Точност на JSON структурирање
- Време за обработка
- Трошоци за обработка

Постапка за тестирање

Етапа 1: Подготовка

1. Инсталирање и конфигурирање на сите LLM модели

2. Подготовка на стандардизиран сет тест документи

3. Дефинирање на JSON шеми за секој тип документ

4. Верификација на референтни вредности (ground truth)

Етапа 2: Извршување на тестови

1. Секвенцијално тестирање на секој модел

2. Повторување на тестови за статистичка значајност

3. Собирање на детални метрики

4. Запишување на грешки и исклучоци

Етапа 3: Анализа

1. Статистичка обработка на резултатите

2. Споредбена анализа меѓу модели

3. Идентификување на обрасци и трендови

4. Формулирање заклучоци и препораки

Тестирани LLM модели

Избор на модели

За ова истражување се избрани четири современи LLM модели кои покажуваат добри резултати во OCR задачи:

1. Llama 3.2

Карактеристики:

- Развиен од Meta AI
- Големина: достапни повеќе варијанти – од помал модел со околу 1 милијарда параметри, до најголем модел со околу 90 милијарди параметри
- Специјализиран за мултимодални задачи
- Поддршка за слики и текст

Предности:

- Отворен код и бесплатен
- Добра балансираност на точност/ресурси
- Активна заедница за поддршка
- Можност за локално хостирање

Ограничувања:

- Потребни значителни хардверски ресурси
- Побавен од комерцијалните алтернативи
- Ограничена поддршка за некои јазици

2. LLaVA (Large Language and Vision Assistant)

Карактеристики:

- Мултимодален модел за визуелно разбирање
- Комбинира визуелен енкодер со јазичен модел
- Специјализиран за анализа на слики

Предности:

- Одлични резултати за OCR задачи
- Добро разбирање на визуелна структура
- Отворен код и достапен преку Ollama
- Ефикасен за сложени документи

Ограничувања:

- Потребни GPU ресурси за оптимални перформанси
- Понекогаш спор за обработка
- Може да има проблеми со мали фонтови

3. GPT-OSS:20B

Карактеристики:

- Отворен модел базиран на GPT архитектурата
- 20 милијарди параметри
- Добра генерална интелигенција

Предности:

- Обезбедува солидна точност при извршување на текстуални задачи
- Овозможува јасно и конзистентно структурирање на податоци
- Нуди добра брзина на извршување во однос на својата големина
- Стабилен и предвидлив

Ограничувања:

- Помал од најновите комерцијални модели
- Понекогаш неточен во детали
- Ограничена визуелна интелигенција

4. DeepSeek-R1:8B

Карактеристики:

- Нов модел од DeepSeek компанијата
- Оптимизиран за рационално размислување
- 8 милијарди параметри со висока ефикасност

Предности:

- Одличен однос на точност/ефикасност
- Брз и ресурсно ефикасен
- Добро логичко размислување
- Новии техники за оптимизација

Ограничувања:

- Релативно нов модел со ограничена емпириска евалуација (сè уште нема доволно истражувања и практични примени што ќе ја потврдат неговата стабилност)
- Ограничена документација
- Може да има проблеми со специфични случаи

Типови документи

Категоризација на документи

За сеопфатно тестирање, документите се класифицираа во следните категории:

1. Фискални сметки

Карактеристики:

- Стандардизирана структура
- Нумерички податоци (цени, даноци)
- Датуми и времиња
- Информации за трговец

Предизвици за OCR:

- Мали фонтови

- Термички принтери кои понекогаш печатат нејасно или со намален квалитет
- Различни формати и јазици
- Потреба за точност во нумерички податоци

Пример структура:

```
{
  "merchant": {
    "name": "Продавница АВС",
    "address": "Ул. Македонија 123, Скопје",
    "phone": "02/123-456"
  },
  "items": [
    {"name": "Производ 1", "price": 150.00, "quantity": 2}
  ],
  "total": 300.00,
  "tax": 45.00,
  "date": "2024-03-15",
  "time": "14:30"
}
```

2. Фактури

Карактеристики:

- Комплексна структура
- Детални информации за клиенти
- Повеќе ставки со описи
- Правни и финансиски податоци

Предизвици за OCR:

- Променливи формати
- Табеларни податоци
- Потреба за висока точност
- Сложени обрасци за плаќање

3. Административни формулари

Карактеристики:

- Формални документи
- Полиња за пополнување
- Потписи и печати
- Референтни броеви

Предизвици за OCR:

- Рачно пишување
- Некомплетни податоци
- Различни формати
- Потреба за структурирање

4. Лични документи

Карактеристики:

- Стандардизирани формати
- Лични податоци
- Сигурносни елементи
- Фотографии

Предизвици за OCR:

- Сигурносни фонтови
- Мали текстови
- Заштитени дизајни
- Приватност на податоците

Метрики за евалуација

Примарни метрики

1. JSON точност

Дефиниција: Процент на точно извлечени структурирани податоци споредено со референтните вредности.

Формула:

JSON_точност = (број_точни_полиња / вкупен_број_полиња) × 100%

Методологија:

- Споредба поле по поле
- Третирање на различни типови податоци
- Казнување за недостасувачки податоци
- Толеранција за мали нумерички разлики

2. Текстуална сличност

Дефиниција: Мера за сличност меѓу извлечениот текст и референтниот текст користејќи Levenshtein дистанца.

Формула:

Сличност = 1 - (Levenshtein_дистанца / max_должина)

Карактеристики:

- Вредности од 0 до 1 (каде 1 е совршена сличност)
- Толеранција за мали правописни грешки
- Чувствителност на редослед на зборовите

3. Време за обработка

Компоненти:

- Време за вчитување слика
- Време за OCR обработка
- Време за структурирање податоци
- Вкупно време од почеток до крај

4. Економска ефикасност

Фактори:

- Трошоци за користење (API повици)
- Потребни хардверски ресурси
- Време за обработка (трошок на работа)
- Скалабилност

Секундарни метрики

1. Стабилност

Мерење:

- Конзистентност на резултатите при повторување
- Справување со различни квалитети на слики
- Обработка на нестандардни формати

2. Грешки и исклучоци

Следење:

- Фреквенција на грешки
- Типови грешки
- Време за опоравување
- Влијание врз крајни резултати

Експериментален дизајн

Тест околина

Хардверска конфигурација:

- CPU: Intel i7 или AMD Ryzen 7
- RAM: 32GB DDR4
- GPU: NVIDIA RTX 3080 или подобра
- Складирање: SSD со висока брзина

Софтверска околина:

- Python 3.8+
- Ollama за локални LLM модели
- Docker за изолација на околината

Тест податоци

Сет од тест документи:

- 50 фискални сметки (различни трговци)

- 30 фактури (различни компании)
- 25 административни формулари
- 20 лични документи (анонимизирани)

Квалитет на слики:

- Висок квалитет: 300 DPI, јасни
- Среден квалитет: 150 DPI, благо нејасни
- Низок квалитет: 72 DPI, нејасни или оштетени

Резултати и анализа

Преглед на резултатите

Врз основа на извршените тестови, добиени се следните главни резултати:

1. Вкупна точност по модели

Модел	JSON точност	Текстуална сличност	Рејтинг
LLaVA	87.30%	92.10%	★★★★★
Llama 3.2	84.70%	89.60%	★★★★☆
DeepSeek-R1	82.10%	87.30%	★★★★☆
GPT-OSS:20B	78.90%	84.20%	★★★☆☆

2. Перформанси по типови документи

Фискални сметки:

- LLaVA: 91.2% точност (најдобар за нумерички податоци)
- Llama 3.2: 88.7% точност
- DeepSeek-R1: 86.1% точност
- GPT-OSS:20B: 82.3% точност

Фактури:

- LLaVA: 85.9% точност
- Llama 3.2: 83.2% точност (добар за табеларни податоци)
- DeepSeek-R1: 80.4% точност
- GPT-OSS:20B: 77.1% точност

Административни формулари:

- Llama 3.2: 82.1% точност (најдобар за формални документи)
- LLaVA: 81.7% точност
- DeepSeek-R1: 78.9% точност
- GPT-OSS:20B: 74.2% точност

3. Анализа на грешки

Најчести типови на грешки:

1. **Нумерички грешки**
    - Погрешно читање на нули и слични цифри
    - Проблеми со децимални броеви
    - Мешање на валути
2. **Структурни грешки**
    - Неточно групирање на податоци
    - Пропуштени релации меѓу полиња

4. Детална анализа по критериуми

Точност и Брзина:

- DeepSeek-R1: Најбрз, добра точност за цената
- Llama 3.2: Добра балансираност
- LLaVA: Највисока точност, но побавен
- GPT-OSS:20B: Умерени перформанси во сите аспекти

Стабилност и доверливост:

- Llama 3.2: Најстабилен, ретки грешки

- LLaVA: Конзистентно висока точност
- DeepSeek-R1: Добра стабилност за брзината
- GPT-OSS:20B: Понекогаш неконзистентен

5. Специјални случаи

Македонски јазик:

- LLaVA: 83.2% точност (најдобар за кирилица)
- Llama 3.2: 79.8% точност
- DeepSeek-R1: 76.1% точност
- GPT-OSS:20B: 71.4% точност

Лош квалитет на слика:

- LLaVA: 71.2% точност (најотпорен на лош квалитет)
- Llama 3.2: 68.9% точност
- DeepSeek-R1: 64.7% точност
- GPT-OSS:20B: 59.3% точност

Заклучок

Резултатите од истражувањето јасно покажуваат дека современите LLM модели имаат значителен потенцијал за примена во OCR задачи.

1. LLaVA се издвојува како модел со највисока точност, особено во обработка на комплексни документи со табели и структури, текстови со мал фонт или намален квалитет, како и при работа со нумерички податоци кои бараат висока прецизност.
2. DeepSeek-R1 се покажа како најефикасно економско решение, погоден за сценарија со ограничени хардверски ресурси и потреба од висока пропусност.
3. Llama 3.2 обезбедува најбалансиран пристап, комбинирајќи солидна точност, стабилни и предвидливи резултати, како и разумни перформанси и трошоци.
4. GPT-OSS:20B заостанува зад останатите модели и покажа најслаби резултати, поради што не се препорачува за задачи каде што точноста претставува приоритет.

Вкупно земено, може да се заклучи дека LLaVA претставува водечки модел во однос на точноста, додека DeepSeek-R1 е најисплатлива опција од економска гледна точка.

Иднината на OCR технологијата се насочува кон интелигентна интеграција на вакви LLM решенија, во комбинација со постојано усовршување и адаптација кон специфични потреби и сценарија на употреба.

Користена литература

https://ai.meta.com/llama

https://arxiv.org/abs/2304.08485

https://arxiv.org/abs/2302.13971

https://deepseek.ai

https://github.com/tesseract-ocr/tesseract

https://en.wikipedia.org/wiki/Optical_character_recognition

https://www.sciencedirect.com/science/article/pii/S003132031930288X

Линк до GitHub репозиториум

https://github.com/smileska/vnp-project