Распознавание документов на русском языке на основе единой нейросетевой архитектуры с использованием дистилляции знаний

Иртуганов Мансур Наилевич

Студент

Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия

E-mail: irtuganovmn@gmail.com

Научный руководитель *— Грабовой А. В.*

В задачах автоматической обработки документов и построения вопросно-ответных систем (RAG) критически важным является сохранение семантической структуры: иерархии заголовков, топологии таблиц и связей между иллюстрациями и текстом. Классические каскадные подходы (OCR + Layout Analysis) часто нарушают целостность данных, а современные SOTA-модели (GPT-5) обладают высокой задержкой и стоимостью. В данной работе предлагается метод создания компактной модели структурного анализа документов (IDP-OCR) на базе архитектуры Qwen2.5-VL посредством дистилляции знаний.

Для обучения модели использовался метод *Sequence-Level Knowledge Distillation* [1]. В качестве учителя выступала модель Gemini 2.5 Pro, с помощью которой был размечен датасет из 1 млн изображений. Особенностью работы является разработанный формат *Extended Markdown*, включающий: 1) HTML-представление для сложных таблиц; 2) специальные токены для локализации визуальных объектов ($< figure >$).

В качестве базовой модели выбрана Qwen2.5-VL (3B параметров) [2]. Для адаптации модели к задаче структурного анализа применялся метод LoRA [3]. Экспериментально установлено, что для корректного восприятия пространственной структуры документа (Layout Understanding) необходим экстремально высокий ранг адаптеров ($r = 512$), в отличие от стандартных текстовых задач ($r = 8 \dots 16$).

Сравнительный анализ на отложенной выборке (200 сложных документов) показал превосходство дистиллированной модели над индустриальным бейзлайном и базовой версией модели .

Результаты демонстрируют, что компактная модель после обучения достигает качества, сопоставимого с моделями, чьи размеры на

Таблица 1: Сравнение качества распознавания по типам документов (BoW-F1)

Модель	**Чеки**	**Фин. таблицы**	**Платежки**	**Среднее**
Baseline	0,8373	0,9530	0,9263	0,9055
ChatGPT-5	0,7937	0,9488	0,9527	0,8984
Qwen3-VL 235B	**0,9211**	0,9659	0,9649	0,9506
Обученная модель 3B	0,9171	**0,9724**	**0,9838**	**0,9578**

порядок превышают текущую, и превосходящего закрытые проприетарные решения. Это подтверждает эффективность подхода дистилляции знаний для создания специализированных решений в домене обработки документов.

Литература

1. Kim Y., Rush A. M. Sequence-Level Knowledge Distillation // arXiv:1606.07947. 2016.
2. Wang, Q. et al. Qwen2-VL: To See the World More Clearly // arXiv:2409.12191, 2024.
3. Hu E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models // arXiv:2106.09685. 2021.