**Описание задачи**

В крупных организациях создается большое количество разнообразных документов. В случае финтех-компаний это могут быть кредитные договоры, справки о доходах, выписки и т. д. Зачастую подобные документы предоставляются на бумажных носителях (распечатки электронных документов или сканов). Важной задачей для организаций является систематизация документов и управление ими.

Одним из возможных решений этой задачи является переход на электронный документооборот. Однако переход на него полностью может быть затруднительным, и даже если это будет сделано, как быть с существующими документами? Вам предстоит решить эту сложную задачу с помощью нейросетей!

**Цель кейса**

Необходимо реализовать нейронную сеть, которая сможет распознавать текст со сканов формата .jpg. Для обучения и оценки модели используются исключительно синтетические данные, специально сгенерированные для этого соревнования. Имитируются реалистичные сканы документов, с добавлением разнообразных искусственных шумов, вариаций освещения, искажений перспективы и деформаций текста.

**Входные данные**

* Длинный документ со сканами формата .jpg со словами, встречающихся в документах финтех-организаций.

**Ограничения при реализации решения**

* **Запрещено** использовать GPT или другие готовые модели генеративного ИИ в качестве конечного решения задачи OCR.
* **Не допускается** использование готовых коммерческих API для OCR (например, Google Vision, AWS Textract и т.д.).
* **Разрешено** использование библиотек для базовой обработки изображений (например, OpenCV) и открытых фреймворков для машинного обучения (PyTorch, TensorFlow и т.д.).

**Метрика качества**

• Основная метрика: Расстояние Левенштейна между распознанным текстом и оригинальным текстом (чем меньше расстояние, тем выше точность OCR).

**Результат**

Код решения загружайте в проект на github.

Принцип работы должен быть продемонстрирован на защите проектов.

**Защита проекта (что ожидается на презентации)**

Поиск лучшей ИИ-модели:

* Эксперименты с разными алгоритмами машинного обучения и/или нейросетевыми архитектурами.
* Оценка моделей с использованием метрик качества.
* Оптимизация гиперпараметров для достижения наилучших результатов.

Обоснование выбора модели:

* Сравнительный анализ производительности моделей.
* Интерпретация моделей и понимание принципов их работы.

Структура презентации:

* Чёткая структура: постановка задачи, выбор и обучение моделей, анализ результатов, выводы.
* Визуализация данных и результатов (графики, диаграммы, таблицы).
* Умение ответить на вопросы жюри и защитить свои решения.

**Не требуется и не будет оцениваться** дополнительно: представление решения в виде веб-сервиса, контейнера, приложения и т.д. и т.п. Цель соревнования – разработать наилучшую **модель** и продемонстрировать **сильные математические и ML-навыки**.

Удачи в решении задачи!