Содержание и вклад

- MDETR: end-to-end архитектура для решения задач, связанных с детекцией объектов по тексту
- Две новые функции ошибки: soft token prediction и contrastive alignment
- Сбор данных и аннотаций
- Применение в различных downstream задачах сразу и после fine-tuning, проведение экспериментов

Сильные стороны

- Очень сильные результаты по сравнению с современными baseline-ами
- Исследование большого количества downstream задач
- Сравнение ResNet и EfficientNet для backbone модели
- Ablation study: использование только одной функции потерь и модификация модели для VQA

Слабые стороны

- Не совсем end-to-end: предобученные backbone и text encoder
- В качестве text encoder-а была попробована только одна архитектура(RoBERTa)
- Ablation study с функциями потерь был проведен только для modulated object detection

Качество текста

- Статья написана хорошо, но непросто читается, если не знаком с темой
- Описание DETR-а есть, но это скорее напоминание для тех, кто уже с ним знаком
- Нет описания метрик и некоторых специфичных понятий, например, 2D positional encoding
- Нет описания некоторых downstream задач: Referring Expression Segmentation, Visual Question Answering
- Прочтение appendix-а по сути обязательно для понимания

Воспроизводимость

- Весь код написан на pytorch и выложен на github
- Написан хорошо(на первый взгляд)
- Выложены данные и скрипт для обучения
- В статье указаны гиперпараметры обучения

Оценка

- Оценка(по 10-балльной шкале): 8
- Уверенность(по 5-балльной шкале): 4