Рецензия на MDETR

Никита Степанов

24 ноября 2021 г.

В рассматриваемой статье авторы предложили новую архитектуру MDETR для задачи modulated detection. Так как она довольно сильно отличается от других моделей, использующихся в этой области, авторам также потребовалось описать в статье метод ее обучения. Для этого они предложили 2 новые функции ошибки, а также описали процесс сбора данных и аннотаций для обучения. Также они провели достаточно много экспериментов и применили MDETR в различных downstream задачах.

Начну с сильных сторон. Архитектура MDETR показывает очень сильные результаты по сравнению с современными baseline-ами. Авторы также предложили, как можно модифицировать модель, чтобы сделать fine-tuning для разных downstream задач. Во многих рассмотренных задачах у MDETR удалось обойти state of the art методы, причем в некоторых задачах улучшение очень существенное. Если учесть, что в каждой рассмотренной задаче MDETR сравнивали с разными, специализированными под задачу архитектурами, то результат очень впечатляет. Был проведен ablation study, который показал важность совместного использования обеих функций ошибки. Были проведены эксперименты с разными архитектурами для backbone сети: авторы сравнили ResNet и EfficientNet. Авторы также исследовали возможность применения MDETR для классической задачи детектирования объектов с помощью fine-tuning-а на небольшой части обучающей выборки. Хоть превзойти state of the art методы не получилось, результат оказался достойным, даже лучше, чем у DETR-а.

В статье было проведено много экспериментов, но некоторые модули архитектуры использовались в качестве черного ящика, и эксперименты с ними не проводились, например, в качестве text encoder-а была попробована только одна архитектура(RoBERTa) с предобученными весами. Интересно было бы посмотреть, насколько сильно изменяется качество в downstream задачах в зависимости от изменений в text encoder. Авторы статьи характеризуют свою модель как end-to-end, однако это не совсем так, ведь сначала нужно предобучить text encoder и backbone модель. Авторы могли бы поэкспериментировать с этим, например, попробовать случайную инициализацию вместо предобучения, или можно было попробовать заморозить веса предобученных моделей, а не переобучать их. Возможно, от этого бы увеличилась обобщающая способность модели. Когда авторы проводили ablation study о важности использования обеих функций потерь, они указывали только результаты для modulated object detection, хотя интересно было бы также посмотреть, как меняется результат в visual question answering, ведь эта задача напрямую не связана с object detection.

Статья, на мой взгляд, написана довольно хорошо, но она рассчитана на подготовленного читателя. Описания downstream задач написаны не очень подробно, описания используемых метрик вообще нет, поэтому если читатель не сильно разбирается в multimodal object detection, то ему придется в этом разбираться самостоятельно. Почему-то авторы описали постановки задач phrase grounding и referring expression comprehension, но не сделали того же для задач referring expression segmentation и visual question answering. Мне не понравилось, что некоторые важные куски изложения почему-то находятся в appendix-е. Во время первого прочтения мне показалось, что авторы очень размыто объясняют, как они модифицируют MDETR для применения в downstream задачах, а потом я обнаружил, что это более подробно написано в appendix-е. Непонятно, почему нельзя было сразу описать эти процедуры подробно, ведь это сильно затрудняет понимание статьи. Тем не менее, в итоге у меня не осталось впечатления, что авторы опустили какие-то детали, важные для реализации.

Авторы статьи прилагают код, написанный на pytorch. Он довольно хорошо написан и структурирован, по-крайней мере на первый взгляд, в описании репозитория есть инструкция, как им пользоваться. Также в статье приведены гиперпараметры, с которыми проводились эксперименты. Хочу отметить, что авторы выложили данные, на которых они обучались, и это важно, потому что они собирались нетривиальным образом и воспроизвести результат без наличия этих данных было бы крайне тяжело. Авторы также выложили скрипт для обучения модели и описали как им пользоваться, поэтому проблем с вопроизводимостью быть не должно.

Я ставлю статье оценку 8, свою уверенность я оценю на 4 балла. Я довольно внимательно читал статью, но все же я мало знаком с областью modulated detection, не знаком со стандартными методами решения задач, рассмотренных в статье, поэтому мне тяжело в полной мере оценить, насколько революционны методы, предложенные авторами, и их результаты.