MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, Zaid Harchaoui

Автор рецензии: Аюпов Шамиль

Содержание и вклад

В статье предложен метод оценивания качества открытой генерации текста. Метод основан на идее сравнения распределений над текстами с помощью дивергенции Кульбака-Лейблера в низкоразмерном квантизованном пространстве с использованием внешней языковой модели. Предложенный метод позволяет оценивать качество открытой генерации текстов, не прибегая к ручной разметке.

Сильные стороны:

- предложенный метод прост идейно и в реализации
- метод актуален, поскольку в задаче открытой генерации нет общепринятой метрики качества, которая при этом хорошо коррелирует с человеческим восприятием
- проведено достаточно логичных экспериментов, которые показывают, что метод:
    - очень хорошо коррелирует с человеческим восприятием
    - согласуется с известными свойствами сгенерированных текстов
    - устойчив к изменению внутренних составных частей (почти всех)

Слабые стороны:

- метод использует внешнюю языковую модель, из этого вытекают некоторые проблемы (похожие на проблемы FID), например, устаревание модели, необходимость строгого протокола оценивания, и, как следствие, неудобство для сравнения с другими статьями.
- есть некоторая неразбериха с данными, один из используемых наборов данных (WebText) использовался в обучении внешней

языковой модели, другие наборы нет. Кажется неявной зависимости от этого нет, но комментарий нужен
- в качестве языковой модели в рамках одной задачи используется одна модель (GPT-2 для WebText и Stories, Grover для News), это не очень хорошо для model-agnostic метода, можно добавить нетрансформерную языковую модель для сравнения.

Воспроизводимость и качество текста
Текст супер хорошего качества: последователен, отлично структурирован и логичен; при этом очень прост для чтения.
Метод выложен как библиотека для Python. Также есть репозиторий с экспериментами. Использованные гиперпараметры также описаны в самой статье (Appendix).

Оценка: 9 (Top 15% of accepted NeurIPS papers. An excellent submission; a strong accept)
Уверенность: 4 (You are confident in your assessment, but not absolutely certain)