Рецензия на статью “Vocabulary Learning via Optimal Transport for Neural Machine Translation”

Болотин Арсений

Содержание статьи.

В статье рассматривается новый подход к токенизации в задачах обработки естественного языка. Авторы предлагают подход VOLT с автоматическим подбором словаря токенов оптимального размера с помощью сведения к задаче линейного программирования - Optimal Transport. Предложенный метод значительно уменьшает размер используемого словаря и улучшает результаты на задачах машинного перевода.

Сильные стороны.

- Теоретическая обоснованность: в статье вводится, а затем оптимизируется понятие Marginal Utility of Vocabularization (MUV), для которого предварительно демонстрируется корреляция с метрикой BLEU. В статье приводятся необходимые математические выкладки и рассуждения для сведения к задаче Optimal Transport.

- Полнота эмпирического анализа: проведены эксперименты на задачах двуязычного и многоязычного перевода с использованием двух разных архитектур - Transformer-big и Convolutional Seq2Seq. Сравнение сделано на разных наборах данных и при разных размерах используемого словаря.

- Значимость: предложенный подход требует намного меньше ресурсов, чем подбор размера словаря для BPE как гиперпараметра. Также уменьшенный размер словаря несколько ускоряет дальнейшее обучение модели.

- Новизна: подбор оптимального словаря токенов действительно не рассматривался до данной работы. Стоит отметить, что в области изучаются используемые словари и их влияние на окончательный результат. Например, предлагается использовать разные разбиения на токены в виде некоторого способа регуляризации - https://arxiv.org/pdf/1804.10959.pdf.

Слабые стороны.

- В статье рассматривается только задача машинного перевода.
- Полученные улучшения метрики BLEU сравнительно небольшие, что ставит под вопрос практическую применимость работы. Результаты на WMT-14:https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-german

Насколько хорошо написана статья.

Статья достаточно простая для понимания, все идеи проиллюстрированы примерами, а математические выкладки сопровождены текстовым пояснением и подробным доказательством. Большинство терминов в статье пояснены в работе и не требуют самостоятельного изучения.

Воспроизводимость.

В работе приведён псевдокод алгоритма и указаны параметры используемых моделей. Также предоставлен репозиторий с имплементацией предложенного метода.

Дополнительные комментарии.

- Figure 4: в матрице расстояний не взят минус логарифм от аргументов.
- Conclusion, второе предложение: опечатка “informtaion-therotic” -> “information-theoretic”.

Оценка - 7: A good submission; an accept.

Уверенность - 3: You are fairly confident in your assessment.