CHARFORMER: FAST CHARACTER TRANSFORMERS VIA GRADIENT-BASED SUBWORD TOKENIZATION

Краткий обзор статьи

Цель работы: устранить зависимость моделей NLP от внешних токенизаторов. Авторы предлагают основанный на градиенте модуль токенизации подслова, который может быть использован в любой нейронной модели.

Подход: сначала авторы “смешали” символы друг с другом по соседству, затем снова “смешивали”, но на этот раз ортогонально первому смешиванию и применили mean pooling. При этом существующие модели достаточно просто перестроить под работу с символами (но могут возникнуть проблемы с производительностью).

Результат: Полученная модель была протестирована на GLUE и нескольких межъязыковых задачах. Производительность конкурирует с Byte-level T5 и часто аналогична subword моделям, но при этом более эффективна в FLOPS.

Авторы

Первая версия статьи появилась в июне 2021, но опубликована только в феврале 2022 на конференции ICLR 2022. Над статьей работали 8 авторов (Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri Zhen Qin, Simon Baumgartner, Cong Yu, Donald Metzler) из Google Research и DeepMind. Авторы занимаются NLP (ранее публиковали статьи по NLP), а также публиковали статьи, где полезно использовать Charformer (в том числе и в мультиязычных моделях).

Связанные работы

Subword tokenization

Стандартными алгоритмами для детерминированной токенизации подслова являются кодирование пар байтов (BPE Wordpiece SentencePiece. Предыдущая работа выявила проблемы с некоторыми из этих алгоритмов и в целом отметила, что модели, изученные с такой жесткой маркировкой, плохо справляются с вариациями в языке. Необходимо сделать модель более устойчивой к морфологическому и композиционному обобщению, были предложены алгоритмы вероятностной сегментации, такие как регуляризация подслова (Kudo, 2018) и BPE-dropout (Provilkov et al., 2020), которые отбирают различные сегментации во время обучения. Однако такие методы требуют больших вычислительных затрат что делает их непригодными для предобучения.

Модели символьного уровня

Недавно ByT5 (Xue et al., 2021) установил новые передовые результаты для моделей без токенизации, работая на байтовом уровне. Эта работа выполняется наравне с ByT5 или превосходит ее, значительно повышая скорость и эффективность вычислений. Идеи (байт = символ, предобучение на предсказании нескольких последовательных байт) у статей схожи, но авторы этой статьи используют идею группировки входных данных в последовательность меньшей длины.
Идея группировки также использовалась в статье CANINE (Clark et al., 2021), но Charformer все еще более эффективный. Не смотря на то, что CANINE тоже является моделью без токенизации, эти модели достаточно сильно отличаются. Например, входом модели CANINE являются Unicode символы, а не байты, это означает, что необходимо дополнительное хэширование для сопоставления векторов.

Продолжение

Прямых продолжений статьи пока нет, т.к. она достаточно новая. Лучший метод из Charformer группирует символы для решения проблемы увеличения длины текстовых последовательностей, но допускает утечку информации при применении к декодеру Transformer. Авторы статьи [5] решают эту проблему утечки информации, тем самым позволяя группировать символы в декодере. Они показывают, что понижающая дискретизация символов не имеет очевидных преимуществ в NMT по сравнению с предыдущими методами понижающей дискретизации с точки зрения качества перевода, однако ее можно обучить примерно на 30% быстрее. Многообещающие результаты при переводе с английского на турецкий указывают на потенциал моделей символьного уровня для морфологически богатых языков.

Достоинства и недостатки

Авторам удалось отказаться от предобработки входного текста, что является достаточно сложной задачей и усложняется при работе с несколькими языками.

Авторы смогли зафиксировать размер словаря для хранения векторов представлений - 256 (для любого языка), т.е. уменьшить его размеры во много раз. Однако параметры, которые удалось сэкономить в этом месте перенесены в энкодер. А также необходимо запускать часть модели для получения значимых представлений, что может быть очень затратной задачей.

Также авторы утверждают, что достигли лучшего качества, чем аналогичные модели без токенов, а также превзошли качество стандартных языковых моделей в некоторых задачах, однако эти утверждения

недостаточно подтверждены в статье. Хотелось бы увидеть больше экспериментов по сравнению качества моделей в конкретных задачах.

References

[1] https://arxiv.org/pdf/2106.12672
[2] https://arxiv.org/abs/2202.11176
[3] https://arxiv.org/abs/2103.06874
[4] https://arxiv.org/abs/2105.13626
[5] https://arxiv.org/abs/2205.14086