Charformer: Fast Character Transformers via Gradient-based Subword Tokenization

О статье

CHARFORMER

Публикация

- Первая версия: Июнь 2021
- Итоговая версия: Февраль 2022, конференция ICLR 2022

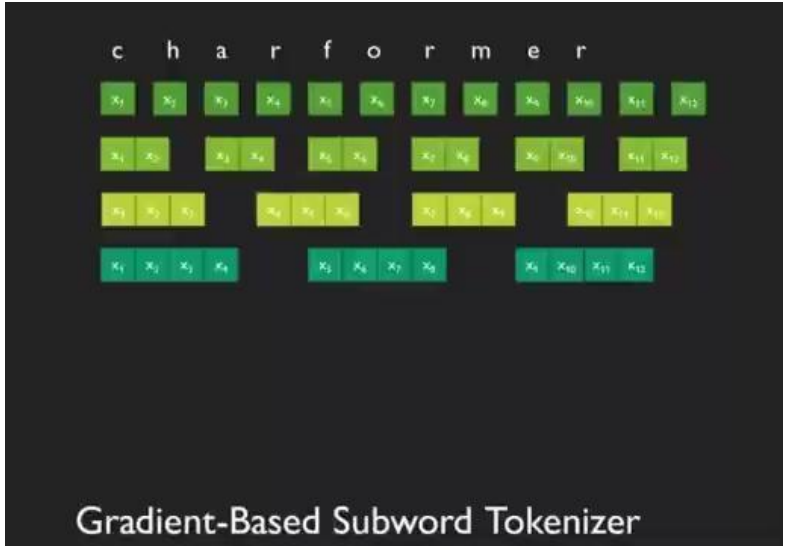
Авторы

- Google Research and DeepMind
- 8 авторов: Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri Zhen Qin, Simon Baumgartner, Cong Yu, Donald Metzler
- Авторы занимаются NLP

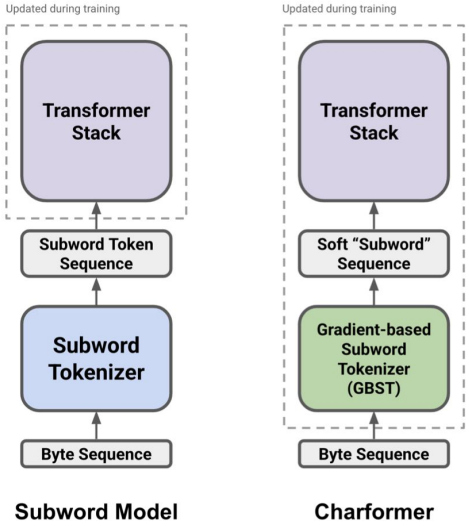
О статье

Цель: устранить зависимость моделей NLP от внешних токенизаторов

Способ: GBST



Результаты: производительность конкурирует с Byte-level T5 и часто аналогична subword моделям, но при этом более эффективна в FLOPS



Достоинства

~~**Предобработка**~~	Отсутствие необходимости предобработки входного текста
↓ Размер словаря	Во много раз снижается размер словаря для хранения представлений (фиксирован 256 для любого языка)
Производительность	Эксперименты показывают конкурентные результаты и фокусируются как на аспектах производительности, так и на скорости
Интерпретируемость	Подход к взвешиванию n-граммов позволяет весам суб-токенов оставаться в некоторой степени интерпретируемыми
Расширение возможностей	Возможность работать с многоязычными моделями и зашумленными данными

Недостатки

Получение представлений

Для получения значимых представлений нельзя просто усреднить векторы токенов, нужно запускать отдельную часть модели

Время

Затрачивается время на получение эмбеддингов

Влияние статей

Subword tokenization

- модели, изученные с жесткой маркировкой, плохо справляются с вариациями в языке (BPE)
- методы требуют больших вычислительных затрат из-за необходимости выполнения нескольких прямых проходов для каждой сегментации примера (DPE)

https://aclanthology.org/2020.acl-main.275/

https://aclanthology.org/P16-1162/

ByT5

- Байты подаются на вход
- Изменяется баланс encoder/decoder частей
- Вход ~ числу байт (нет группировки)

ByT5: Towards a token-free future with pre-trained byte-to-byte models 2021

https://arxiv.org/abs/2105.13626

CANINE

- Группировка векторов
- Нет предобработки
- На вход символы (необходимо хэширование)

Canine: Pre-training an efficient tokenization-free encoder for language representation 2021

https://arxiv.org/abs/2103.06874

Дальнейшая работа

كتب	k-t-b	“write” (root form)
كَتَبَ	**k**a**t**a**b**a	“he wrote”
كَتَّبَ	**k**a**tt**a**b**a	“he made (someone) write”
اِكْتَتَبَ	i**k**ta**t**a**b**a	“he signed up”

Table 1: Non-concatenative morphology in Arabic.[4]

References

- Charformer: Fast Character Transformers via Gradient-based Subword Tokenization https://arxiv.org/abs/2106.12672
- ByT5: Towards a token-free future with pre-trained byte-to-byte models https://arxiv.org/abs/2105.13626
- CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation https://arxiv.org/abs/2103.06874
- https://aclanthology.org/2020.acl-main.275/
- https://aclanthology.org/P16-1162/