Ижевская А.В., ИАД-4

Сравнительный обзор моделей, использованных для экспериментов по соревнованию Jigsaw Multilingual Toxic Comment Classification

В ходе решения соревнования были проведены эксперименты с 3 разными моделями, а именно Distilbert ('distilbert-base-multilingual-cased'), XLM-100 ('xlm-mlm-100-1280') и XLM-R (‘xlm-roberta-large'). Все три архитектуры основаны на одной базовой архитектуре Трансформер, впервые представленной в (Vaswani et al., 2017). Векторные представления, полученные с помощью данной модели (BERT) (Devlin et al., n.d.) на данный момент являются самыми популярными для решения задач с помощью transfer-learning в сфере обработки естественного языка. Предварительно обученные векторные представления легко можно дообучить под конкретную задачу и получить эталонное качество.

Так как модель Distilbert по сути является уменьшенной и оптимизированной версией Bert, то прежде всего необходимо описать архитектуру, на которой он основан. Существуют две важные стадии обучения модели – предварительное обучение и дообучение под конкретную задачу. Для предварительного обучения применяется модель Трансформер. Данная модель состоит из двух частей – энкодера и декодера Энкодер состоит из 6 слоев, которые, в свою очередь, включают подслои: первым является так называемый “multihead attention”, а вторым полносвязная нейронная сеть с применением positional encoding. Также, на каждом слое применяется нормализация и присутствуют residual connections. Декодер имеет примерно такую же архитектуру, но на каждом слое добавляется третий подслой для применения “multihead attention” на выходе энкодера. Кроме этого, применяется «masking», чтобы модель не обучалась еще и на последующих токенах – то есть, не могла заглянуть в будущее, ведь иначе она не сможет впоследствии качественно работать на новых данных. Отдельно стоит осветить 2 важных момента: “multihead attention” и positional encoding. Attention:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

где Q (query), K (key) and V (value) - матрицы. Скалярное произведение Q и K делится на d – размерность K - для численной устойчивости, затем применяется функция softmax и все умножается на V. Но обычно берут не одну такую функцию, а сразу много (их называют “heads”) – так получается “multihead attention”. Что касается positional encoding — это применение еще одной функции, в данном случае с синусами и косинусами, которая позволяет запоминать позицию токена в последовательности.

В случае Bert во время первичного обучения используются пары предложений (A и B), к которым добавляются специальные токены, обозначающие начало и конец предложения, а также паддинг; затем с помощью byte pairwise encoding создаются векторные представления и применяется masking. Последнее является основой обучения: в данных нет лейблов, вместо этого модель обучается предсказывать masked токены - такой подход называется MLM. Сами авторы отмечают, что хотя он и позволяет создать biderictionality, он также создает и некоторое несоответствие этапа обучения и дообучения, так как на второй стадии “masking” не применяется. Векторные представления BERT, полученные как выход Трансформера (его hidden layer), обладают большой универсальностью, однако модель является очень ресурсоемкой: Distilbert - более компактное решение практически без потери качества. Для его обучения используется дистилизация – специальная техника компрессии модели, в которой компактная модель учится воспроизводить поведение полной версии модели. Таким образом, Distilbert обладает преимуществами Bert:

запоминание длинных зависимостей и выделение наиболее важных кусков последовательности с помощью “multihead attention” на каждом слое. При этом он является облегченным и оптимизированным вариантом, что позволяет экономить ресурсы.

Хотя Distilbert уже является отличной моделью для работы с английским языком в случае с многоязычной задачей его multilingual версия не является самой лучшей моделью и уступает более сложным моделям, специально разработанным, чтобы работать с несколькими языками. XLM-100 (Lample & Conneau, 2019) является многоязычной моделью, которая обучается с использованием данных только на одном языке и затем применяется на нескольких языках. Существует несколько видов данной модели в зависимости от конкретной задачи; их процесс обучения также несколько отличается и используются разные подходы. В данном случае релевантен Masked Language Modeling (MLM). Эта модель использует byte pairwise encoding, как и Bert, и в целом имеет очень похожую архитектуру. Основное отличие: вместо пар предложений используется какое-то их количество. Также, чтобы преодолеть дисбаланс из-за редких токенов применяется сабсэмплирование частотных токенов: это происходит с помощью добавления весов, пропорциональных квадратному корню их обратных частот. Такой подход позволяет универсально кодировать текст на разных языках и бороться с проблемой редко встречающихся слов.

И, наконец, самой последней моделью является многоязычная Роберта (она же XLM-R) (Conneau et al., 2019). Создатели модели берут за основу оригинальную архитектуру Трансформер с MLM, обученную на моно лингвистических данных. Они используют сэмплирование из разных языков и masking: модель учится предсказывать masked токены. Токенизация с разбиением токенов на части применяется прямо на необработанные данные с использованием кусков предложений и униграмм. Батчи сэмплируются из разных языков и при этом не используются языковые векторные представления, что позволяет модели лучше обобщаться на разные языки. XLM-R обучается на 100 языках на огромном корпусе, превосходящем по размерам корпусы, на которых обучались предыдущие модели. Такая архитектура представляет собой дальнейший шаг в развитии многоязычных моделей и на данный момент является COTA: согласно авторам, она значительно превосходит по качеству, как Bert multilingual, так и XLM.

Эксперименты, проведенные для Jigsaw Multilingual Toxic Comment Classification, подтверждают эффективность данных многоязычных моделей, при этом Multilingual Distilbert показал самый худший результат, улучшенная модель XLM-100 оказалась несколько лучше, а самая современная XLM-R продемонстрировала наилучшее качество. Стоит отметить, что использовалась модель Multilingual Distilbert с hidden size 768, в то время как остальные модели имеют hidden size 1024 и 1028, что также играло существенную роль для качества обученной модели.

Таким образом, можно заключить, что все три рассмотренные архитектуры, по сути, достаточно похожи, так как имеют одну и ту же основу (архитектуру Трансформер) и обучаются с помощью MLM. Multilingual Distilbert является хорошей базовой моделью, но все-таки он недостаточно адаптирован под многоязычные задачи. Лучшим решением является XLM-100, также созданный на основе модели Трансформер, но специально для этой цели с учетом некоторых особенностей задачи. Его дальнейшим улучшением и COTA на настоящий момент является XLM-Roberta, основным преимуществом которой является то, что она была обучена на значительно большем объеме данных, что очень важно для подобной модели. Данные CommonCrawl, которые применялись для обучения, занимают целых 2,5 ТБ, что на несколько порядков больше, чем корпус Wiki-100, на котором был обучен XLM-10. Часть «Roberta» в названии означает тот факт, что ее обучение, так же, как

и в случае одноязычной модели Roberta, основано только на MLM. На данный момент это лучшая архитектура, которая может быть применена для решения многоязычных задач.

Библиография:

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. In *arxiv.org*. https://github.com/facebookresearch/cc

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (n.d.). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *arxiv.org*. Retrieved May 29, 2020, from https://github.com/tensorflow/tensor2tensor

Lample, G., & Conneau, A. (2019). *Cross-lingual Language Model Pretraining*. http://arxiv.org/abs/1901.07291

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*, 5999–6009.