УДК 004.85

ИССЛЕДОВАНИЕ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ TRANSFORMER ЛИНГВИСТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПРАКТИЧЕСКИХ ЗАДАЧ КЛАССИФИКАЦИИ

Посохов П.А. Пятигорский государственный университет

Аннотация: Доклад включает в себя исследование конкурирующих моделей Transformer архитектур нейронных сетей для выявления наиболее подходящих условий использования каждой из них в сфере анализа информационного пространства, а также предложения по практическому применению.

Введение. Последние несколько лет глубокое обучение активно развивается в направлении анализа естественного языка. Основная причина такого роста интереса к NLP в машинном обучении — появление transformer моделей. Именно такой тип архитектуры стал наиболее эффективным в данной сфере и привел к появлению различных его модификаций. Выделяют два основных типа архитектур: Transformer Decoder, представленный базовой моделью GPT-2, Transformer Encoder, представленный BERT. Практическое применение архитектур предполагает выявление наиболее релевантной модели. Сейчас в этой области имеются сравнения BERT и GPT-2 для английского языка, но в русском языке эти модели начали применяться недавно, что обуславливает малое количество и несистематизированность исследований в этом направлении. Целью нашей работы является исследование конкурирующих моделей, выбор и обоснование критериев сравнения, проведение сравнительного анализа и формирование списка рекомендаций для применения на практике.

Основная часть. Эмпирической основой исследования стали задачи классификации различных типов с целью демонстрации практических различий моделей. В рамках выполняемого анализа была проведена работа с корпусом новостных текстов. Такой выбор обусловлен рядом причин. Во-первых, это позволяет рассмотреть язык в диахронии, поскольку корпус составляют новости с 2000 г. по 2020 г. Во-вторых, публицистический стиль наиболее разнообразно отражает строй языка, включая в себя как художественные приемы, так и разговорную речь. В-третьих, собранный корпус семантически разнообразен, поскольку новости несут в себе самые различные смысловые концепты, описывающие политику, спорт, культуру и т.д.

В рамках исследования на основе данных текстов модели решали следующие задачи классификации:

- Бинарная классификация fake news. Dataset был собран самостоятельно и включил в себя новости из различных источников, размеченные на два класса fake и true. Так как классы мало сбалансированы, в качестве метрики точности используется F1-мера в совокупности с balanced accuracy. Объем выборки составляет 16221 новостная статья.
- Multilabel multiclass классификация включает в себя общедоступный корпус новостей размером более 800 тыс. примеров, который был переразмечен на 20 взаимопересекающихся классов, в которые входят: политика, спорт, культура, экономика и т.д. Эти классы так же слабо сбалансированы. В качестве метрик для этой задачи используется Jaccard index и F1-мера.

В ходе работы были построены модели классификаторов на основе моделей BERT и GPT-2 (предобученных для русского языка), реализованы несколько десятков экземпляров каждой и проведено их обучение.

Выводы. Исходя из полученных в ходе обучения данных можно сказать, что скорость оптимизации параметров decoder base моделей, представленных GPT-2, гораздо выше, в сравнении с encoder моделями, однако конечные результаты точности ниже в обеих задачах. Кроме того, анализ roc curve обеих моделей позволяет сказать, что систематические ошибки моделей различны, а значит — нет возможности сделать вывод о превосходстве одной из

них. Причина этого — в коренном различии decoder и encoder моделей, а также в двунаправленной природе BERT (что особенно важно для языков без строгого порядка слов). В ходе обучения был собран перечень рекомендаций по оптимизации гиперпараметров моделей классификатора на основе transformer.

Посохов П.А. (автор)