1 Аннотация

Размеры современных state of the art моделей машинного обучения растут с каждым годом, что является препятствием к их использованию в условиях высоких требований к производитлеьности и размеру модели, например, на мобильных устройствах. Для решения подобной проблемы может помочь метод называемый дистилляцией знаний (KD, knowledge distillation), нацеленный на то, чтобы обучить малую модель имитиовать поведение большой более точной предобученной модели. В данной работе мы исследуем возможность приблизиться к качеству модели построенной на основе BERT-base, используя несколько рекуррентных архитектур нейронных сетей и подоход KD на задаче распознавания именованных сущностей в текстах экономической тематики. Было показано, что KD позволяет значительно улучшить качество прогнозирования по сравнению с обычным подходом к обучению модели с нуля. Более того, в нашей задаче рекуррентым малым моделям обученным с использованием KD подхода удалось достичь качества не сильно хуже исходной модели на основе BERT-base.

2 Тема работы

Дистилляция BERT-based модели моделями рекуррентных нейронных сетей в задаче распознавания именованных сущностей на корпусе экономических текстов.

3 Состав коллектива

Малахов Илья Павлович - ЭФ НГУ, направление бизнес-информатика, кафедра применения математических методов в экономике, i.malakhov1@g.nsu.ru

4 Научное содержание работы

4.1 Постановка задачи

Цель работы заключалась в том, чтобы на примере задачи распознавания именованных сущностей показать возможность существенного снижения количества параметров BERT-based модели без значительной потери качества прогнозирования посредством применения подхода дистилляции знаний с использованием моделей рекуррентных нейронных сетей. Дистиляция знаний – подход к снижению размеров модели машинного обучения – модели-учителя – при котором малая модель-ученик стремится сымитировать поведение большой, более точной модели-учителя, посредством аппроксимации выходов некоторых слоев модели-учителя в процессе обучения.

4.2 Современное состояние проблемы

Перед началом исследования были изучены существующие работы по теме дистилляции знаний и распознавания именованных сущностей. Ключевые ближайшие по теме работы в основном используют одинаковые архитектуры моделей учителя и ученика в подходе дистилляции знаний, сокращая количество слоев или их размерности у модели-ученика (Buciluundefined, Caruana и Niculescu-Mizil 2006; Hinton,

Vinyals и Dean 2015; Sanh и др. 2019; Turc и др. 2019; Chen и др. 2017). Наш же подход заключается в использовании концептуально различных архитектур моделей уителя (BERT) и ученика (рекуррентные нейронные сети).

4.3 Подробное описание работы, включая используемые алгоритмы

В рамках данного исследования было выполнено 2 работы: статья на студенческой сессии конференции Dialog по компьютерной лингвистики и студенческая курсовая работа на ЭФ НГУ. Работы несколько отличались, в первую очередь использованными наборами данных для обучения В статье на конференции Dialog использовался датасет RuNNE (Artemova и др. 2022), в курсовой работе – датасет RuREBus (Ivanin и др. 2020). В обеих работах использовались одинаковые архитектуры моделей нейронных сетей. Моделю-учителем выступала модель на основе BERT. В качестве моделей-учеников были предложены несколько архитектур рекуррентных нейронных сетей на основе LSTM (Hochreiter и Schmidhuber 1997), SRU (Lei и др. 2017) и SRU++ (Lei 2021). Дистиляция знаний выполнялась с помощью минимизации среднеквадратичной ошибки (MSE) между выходами модели-учителся и ученика (Hinton, Vinyals и Dean 2015) алгоритмом AdamW (Loshchilov и Hutter 2017). Одноврменно с минимизацией MSE в задаче дистилляции, производилось дообучение на задаче распознавания именованных сущностей (NER), посредством минимизации Cross Entropy Loss (CE) при формализации NER в виде sequence tagging задачи. Итоговая функция потерь – сумма функций потерь дистилляции и CE с некоторыми эвристическими коэффициентами. Подробнее со статьей можно ознакомиться по ссылке `https://www.dialog-21.ru/media/5727/malakhovi130.pdf`.

4.4 Полученные результаты

В результате в работе на конференцию Dialog удалось получить рекуррентную модель в 25 раз меньше по объему и в 24 раза быстрее на CPU чем модель учитель. В курсовой же работе модели-ученики были немного учеличины в целях улучшения качества прогноза, в результате чего размеры итоговых моделей составили в 12-20 раз меньше модели-учителя при несущественном снижении качетва прогнозирования.

Список литературы

[1] Cristian Buciluundefined, Rich Caruana и Alexandru Niculescu-Mizil. «Model Compression». В: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, с. 535—541. ISBN: 1595933395. DOI: 10.1145/1150402.1150464. URL: https://doi.org/10.1145/1150402.1150464.

[2] Geoffrey Hinton, Oriol Vinyals и Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. DOI: 10.48550/ARXIV.1503.02531. URL: https://arxiv.org/abs/1503.02531.

[3] Victor Sanh и др. «DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter». В: *CoRR* abs/1910.01108 (2019). arXiv: 1910.01108. URL: http://arxiv.org/abs/1910.01108.

[4] Iulia Turc и др. «Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation». В: *CoRR* abs/1908.08962 (2019). arXiv: 1908.08962. URL: http://arxiv.org/abs/1908.08962.

[5] Guobin Chen и др. «Learning Efficient Object Detection Models with Knowledge Distillation». В: *NIPS*. 2017.

[6] Ekaterina Artemova и др. «RuNNE-2022 Shared Task: Recognizing Nested Named Entities». В: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog"* (2022).

[7] Vitaly Ivanin и др. «RuREBus-2020 Shared Task: Russian Relation Extraction for Business». В: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*. Moscow, Russia, 2020.

[8] Sepp Hochreiter и Jürgen Schmidhuber. «Long short-term memory». В: *Neural computation* 9.8 (1997), с. 1735—1780.

[9] Tao Lei и др. *Simple Recurrent Units for Highly Parallelizable Recurrence*. 2017. DOI: 10.48550/ARXIV.1709.02755. URL: https://arxiv.org/abs/1709.02755.

[10] Tao Lei. «When Attention Meets Fast Recurrence: Training Language Models with Reduced Compute». В: *CoRR* abs/2102.12459 (2021). arXiv: 2102.12459. URL: https://arxiv.org/abs/2102.12459.

[11] Ilya Loshchilov и Frank Hutter. «Fixing Weight Decay Regularization in Adam». В: *CoRR* abs/1711.05101 (2017). arXiv: 1711.05101. URL: http://arxiv.org/abs/1711.05101.

5 Эффект от использования кластера в достижении целей работы

С использованием графических ускорителей (GPU) вычислительного кластера ИВЦ НГУ были обучены глубокие нейронные сети рассмотренные в работе. Без использования GPU обучение нейронных сетей подобных размеров не представляется возможным.

6 Перечень публикаций, содержащих результаты работы

- Compressing Bert 25 Times by RNN in Named Entity Recognition Task, Ilya Malakhov, Novosibirsk State University, 2022 - https://www.dialog-21.ru/media/5727/malakhovi130.pdf (статьи студенческой сессии не индексируется)