LoRA: Low-Rank Adaptation of Large Language Models

Илья Пахалко БПМИ181, практик-исследователь

10 ноября 2021 г.

Статья подана в статусе препринта 17 июня 2021, её новая редакция - 16 октября 2021. На данный момент ещё нигде не опубликована.

Авторы статьи - сотрудники Microsoft Research. Два первых автора, помеченных как Equal Contribution - Edward Hu и Yelong Shen. Edward Hu - выпускник-бакалавр университета Джонса-Хопкинса 2019 года. Работает над GPT-3 и нейросетями бесконечной ширины. Публиковался ранее на крупных конференциях - ICLR Best Paper 2020 (Improved Image Wasserstein Attacks and Defenses), также есть публикация на ICML 2021. На данный момент статьи набрали 5 и 21 цитирований соответственно. Yelong Shen - PhD. Kent State University. Самая цитируемая работа (про эмбеддинги предложений в LSTM, 2016 год) набрала более 700 цитирований. Основной фокус статей - различные ответвления NLP: решение задач из бенчмарка GLUE, аугментация данных. Также среди соавторов статьи есть и другие сотрудники Microsoft AI, в том числе постдоки Принстона и Стенфорда.

Наибольшее влияние на данную работу, по словам самих авторов, оказали две статьи на схожие темы: Akhadzhanyan et al., 2020 [AZG20]; Li et al., 2018 [LFLY18] - Обе исследуют обучение моделей с ипользованием случайных проекций в пространстве параметров. Li et al. вводят понятие внутренней размерности оптимизационной задачи; Akhadzhanyan et al. с помощью данного понятия объясняют эффективность обучения языковых моделей как таковых. Соответственно, эти работы служат теоретическим обоснованием возможности поиска оптимальных параметров в пространствах меньшей размерности, а это и есть фокус рассматриваемой статьи.

Прямые конкуренты - в основном другие подходы к малозатратному файн-тюнингу - технологии обучения слоёв-адаптеров [HGJ+19] [LMF20] и префикс-слоёв [LL21]. С ними авторы проводят сравнение своего метода в 3-ей секции статьи.

Дополнительно в related works упоминается и другая вышедшая недавно статья [MHR21], которая для эффективного обучения адаптер-слоёв использует кронекерово произведение. Авторы отмечают, что схожий метод можно попробовать применить и к LoRA - и оставляют это как Future Work.

Цитирований у статьи пока немного, возможно, из-за статуса препринта. Среди них стоит отметить:

- 14.10.2021 [MMH+21] - комбинация различных PELT-методов (Parameter Eficient Lang Model Tuning), в т.ч. LoRA
- 12.10.2021 [LTLH21] - улучшение моделей, учитывающих приватность данных
- 28.10.2021 [JCR21] - построение полусиамских сетей на основе BERT для ранжирования

Очевидно, результатам статьи можно найти применение в индустрии. Во-первых открывается возможность быстрее и дешевле файн-тюнить огромные модели вроде GPT-3 под свои задачи, не обладая при этом суперкомпьютером - то есть значительно шире охват пользователей, которым доступны крупные модели. Например, если новостное издание захочет генерировать черновики статей на заданную тематику, то настройка под эту тематику обойдётся сильно дешевле, чем раньше. Во-вторых, в числе достоинств, отмечаемых авторами - отстутствие оверхеда по времени на этапе применения модели: нет дополнительных слоёв поверх базовой архитектуры, как в адаптерах. В случае индустрии это снижает загрузку железа и ускоряет ответ пользователю, что пригождается при маштабном использовании (читай - балабоба).

Список литературы

[AZG20] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *CoRR*, abs/2012.13255, 2020.

[HGJ+19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. *CoRR*, abs/1902.00751, 2019.

[JCR21] Euna Jung, Jaekeol Choi, and Wonjong Rhee. Semi-siamese bi-encoder neural ranking model using lightweight fine-tuning, 2021.

[LFLY18] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *CoRR*, abs/1804.08838, 2018.

[LL21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *CoRR*, abs/2101.00190, 2021.

[LMF20] Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 441–459, Online, November 2020. Association for Computational Linguistics.

[LTLH21] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners, 2021.

[MHR21] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *CoRR*, abs/2106.04647, 2021.

[MMH+21] Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Wen tau Yih, and Madian Khabsa. Unipelt: A unified framework for parameter-efficient language model tuning, 2021.