RoseLoRA

Команда RoseLora: Абасов Э.Э., Очиров О.В., Шематович П.В

Введение

Для дообучения моделей существует популярный метод - LoRA (Low Rank Adaptation), основанный на разложении матриц весов в малоранговом приближении. Однако, в данном случае обновляются все веса модели, тем самым теряется структура, существовавшая внутри предобученных весов, что может быть существенным для некоторых задач.

Для уменьшения влияния на начальную структуру, мы используем новый метод - RoseLoRA. За счет введения разреживания в матрицы LoRA, итоговое число ненулевых параметров значительно уменьшается, что позволяет достигать близких с LoRA результатов ценой меньших изменений весов.

Основная задача:

- Реализовать алгоритм RoseLoRA и сравнить со стандартными LoRA методами на различных задачах
  - Дает ли алгоритм преимущество в точности, как заявляют авторы статьи?
  - Если да, то для каких задач?

LoRA - Low-rank Adaptation

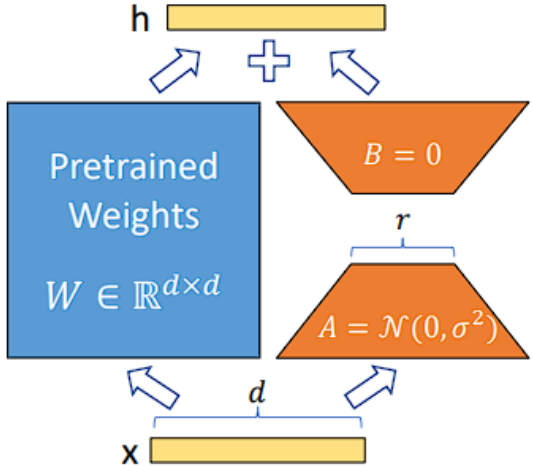
$$\boldsymbol{W} = \boldsymbol{W}^o + \Delta = \boldsymbol{W}^o + \boldsymbol{BA}$$

y = $W_0 x + \Delta W x = W_0 x + BAx$

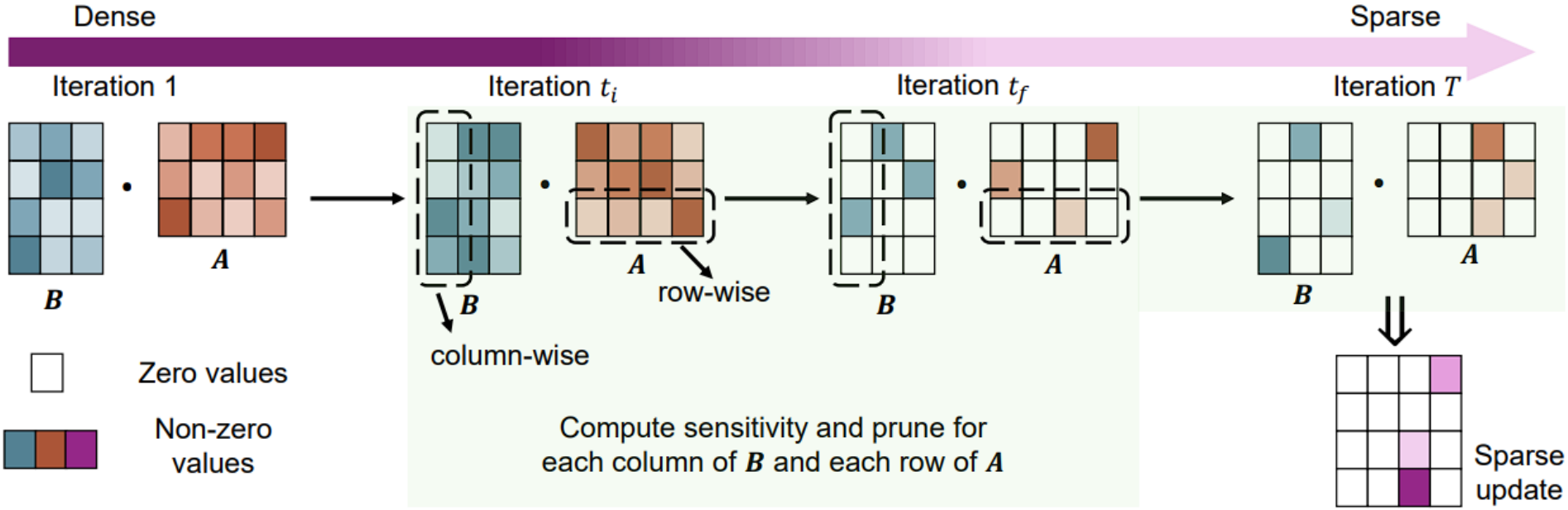
$B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$

W_0 - предобученные веса исходной модели

- Ускоряет дообучение
- Меняет все веса слоев, к которым применена



Row and c**o**lumn-wi**se** sparse **lo**w-**r**ank **a**daptation



RoseLoRA: чувствительность

Ключевая идея:

1. Считаем ***чувствительность*** элементов матрицы A и B.
2. Оставляем только топ самых чувствительных элементов (у A по строкам, у B по столбцам)

$$I(\boldsymbol{W}_{ij}) = |\boldsymbol{W}_{ij} \cdot \nabla_{W_{ij}} \mathcal{L}|.$$

$$\bar{I}^{(t)}(\boldsymbol{W}_{ij}) = \beta \bar{I}^{(t-1)}(\boldsymbol{W}_{ij}) + (1-\beta) I^{(t)}$$

Обновление весов

$$\tilde{\boldsymbol{A}}^{(t)} = \boldsymbol{A}^{(t)} - \nabla_{\boldsymbol{A}^{(t)}} \mathcal{L},$$
$$\tilde{\boldsymbol{B}}^{(t)} = \boldsymbol{B}^{(t)} - \nabla_{\boldsymbol{B}^{(t)}} \mathcal{L}.$$

Делаем

$$\boldsymbol{A}_{i*}^{(t+1)} = \mathcal{T}_A(\tilde{\boldsymbol{A}}_{i*}^{(t)}, \bar{I}^{(t)}(\boldsymbol{A}_{i*}^{(t)})),$$
$$\boldsymbol{B}_{*i}^{(t+1)} = \mathcal{T}_B(\tilde{\boldsymbol{B}}_{*i}^{(t)}, \bar{I}^{(t)}(\boldsymbol{B}_{*i}^{(t)})),$$

$$\left(\mathcal{T}_A\left(\tilde{\boldsymbol{A}}_{i*}^{(t)}, \bar{I}^{(t)}\left(\boldsymbol{A}_{i*}^{(t)}\right)\right)\right)_j = \begin{cases} \tilde{\boldsymbol{A}}_{ij}^{(t)}, & \bar{I}^{(t)}\left(\boldsymbol{A}_{ij}^{(t)}\right) \text{ is top- } \tau^{(t)} \text{ in } \bar{I}^{(t)}\left(\boldsymbol{A}_{i*}^{(t)}\right), \\ 0, & \text{otherwise}, \end{cases}$$

$$\left(\mathcal{T}_B\left(\tilde{\boldsymbol{B}}_{*i}^{(t)}, \bar{I}^{(t)}\left(\boldsymbol{B}_{*i}^{(t)}\right)\right)\right)_j = \begin{cases} \tilde{\boldsymbol{B}}_{ji}^{(t)}, & \bar{I}^{(t)}\left(\boldsymbol{B}_{ji}^{(t)}\right) \text{ is top- } \tau^{(t)} \text{ in } \bar{I}^{(t)}\left(\boldsymbol{B}_{*i}^{(t)}\right), \\ 0, & \text{otherwise}, \end{cases}$$

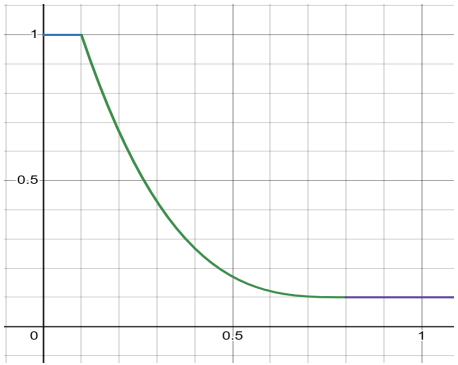
Обновление весов: бюджет

$$\tau^{(t)} = \begin{cases} 1, & 1 \le t \le t_i, \\ \tau + (1-\tau)\left(1 - \dfrac{t - t_i}{t_f - t_i}\right)^3, & t_i \le t \le t_f, \\ \tau, & t_f \le t \le T, \end{cases}$$

Прожиг

Разреживание

Дообучение



Зависимость доли ненулевых элементов от числа итераций

Бенчмарки

Для сравнения подходов было выбрано несколько моделей и датасетов из разных областей NLP

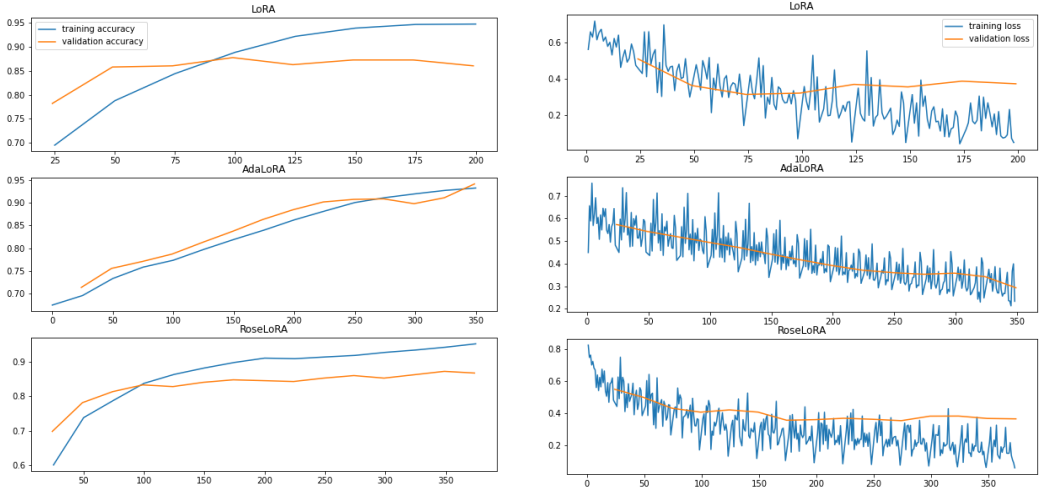
Датасет	Задача	Модель	PEFT методы
MRPC (GLUE)	Text classification	BERT (bert-base-uncased)	LoRA, AdaLoRA, RoseLoRA
AqUA	Multiple choice	BERT (bert-base-uncased)	LoRA, RoseLoRA
ZRSE	Knowledge editing	GPT-2	LoRA, RoseLoRA

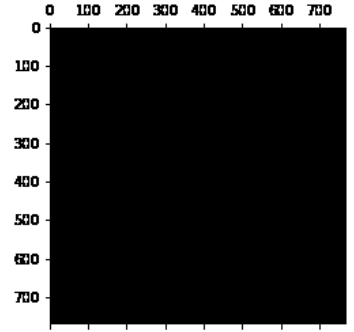
Бенчмарки: GLUE

Задача: Эквивалентны ли 2 утверждения?

Пример: Q1: “The DVD-CCA then appealed to the state Supreme Court”

Q2: “The DVD CCA appealed that decision to the U.S. Supreme Court” - эквивалентны





Процесс разреживания матрицы весов

Бенчмарки: GLUE

Задача: Эквивалентны ли 2 утверждения?

Пример: Q1: “The DVD-CCA then appealed to the state Supreme Court”

Q2: “The DVD CCA appealed that decision to the U.S. Supreme Court” - эквивалентны

Метрика: Binary Accuracy

Точность бинарной классификации на тестовой выборке			
Pre-trained сеть	LoRA	AdaLoRA	RoseLoRA
0.664	0.821	0.823	0.816

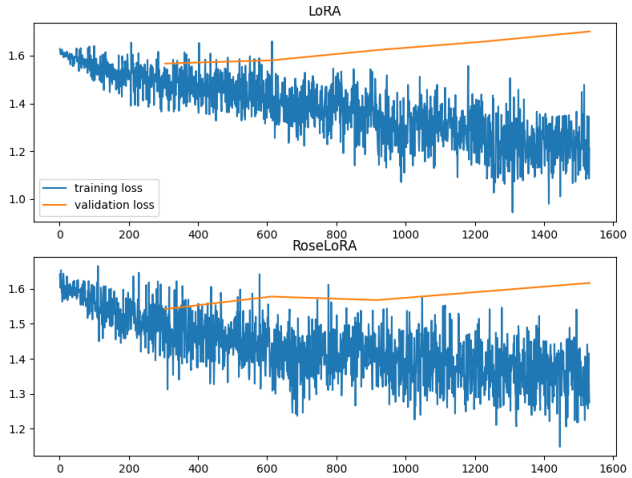
Бенчмарки: AqUA

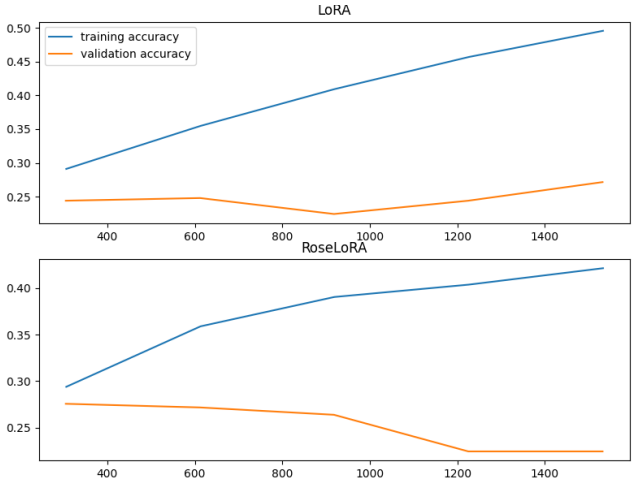
Задача: Выбор правильного ответа

Пример: Q: “If a/b=3/4 and 8a+5b=22,then find the value of a.”

A: “["A)1/2", "B)3/2", "C)5/2", "D)4/2", "E)7/2"]”

Метрика: Multiclass Accuracy





Бенчмарки: AqUA

Задача: Выбор правильного ответа

Пример: Q: “If a/b=3/4 and 8a+5b=22,then find the value of a.”

A: “["A)1/2", "B)3/2", "C)5/2", "D)4/2", "E)7/2"]”

Метрика: Multiclass Accuracy

Точность многоклассовой классификации на тестовой выборке		
Pre-trained сеть	LoRA	RoseLoRA
0.215	0.291	0.279

Бенчмарки: ZsRE Knowledge editing

Задача: оценка восприимчивости LLM-моделей к редактированию знаний.
Устаревание данных, устранение неточностей, актуализация.

Пример: Q: “Где родился Альберт Эйнштейн?”

A: “Альберт Эйнштейн родился в городе *Ульм, Германия.*”

New A: “Альберт Эйнштейн родился в городе ***Мюнхен, Германия.***”

Locality: “В каком году Альберт Эйнштейн получил Нобелевскую премию?”

Portability: “В каком городе находились родители Альберта, когда он родился?”

Бенчмарки: ZsRE Knowledge editing

Метрики

Edit Success	Насколько успешно изменился целевой факт	Альберт Эйнштейн родился в городе Ульм, Германия
Locality	Насколько изменение ограничено целевым фактом	Альберт Эйнштейн получил Нобелевскую премию по физике в 1921 году.
Portability	Насколько изменение распространяется на различные формулировки	Родители Альберта находились в Мюнхене на момент его рождения.
Fluency	Насколько ответ модели грамматически и стилистически корректен	Качество и естественность формулировок в ответах модели.

Бенчмарки: ZsRE Knowledge editing

Результаты

Метрика	LoRA	RoseLoRA
Edit success	96.84	100
Portability	31.42	41.74
Locality	15.88	35.30
Fluency	218.45	230.64

Результаты

- Реализован алгоритм RoseLoRA, реализация интегрирована с библиотеками от HuggingFace и с фреймворком EasyEdit для задач knowledge editing.
- Разреживание матриц дает заметное преимущество в задачах knowledge editing, однако не улучшает результаты в остальных бенчмарках - метод не универсален!

Возможные улучшения:

- Изучение sparse реализации
- Другие бенчмарки
- Изучение эффекта на сети с большим числом параметров

Ссылки

1. https://arxiv.org/abs/2406.10777 - авторская статья по RoseLoRA
2. https://github.com/emil2001/RoseLoRA - репозиторий с реализацией и бенчмарками
3. https://github.com/zjunlp/EasyEdit - An Easy-to-use Knowledge Editing Framework for Large Language Models
4. https://huggingface.co/datasets/zjunlp/KnowEdit - Knowledge Editing датасеты на huggingface

Распределение работы

- Шематович Павел – основа алгоритма, внедрение в EasyEdit, бенчмарк ZsRE
- Абасов Эмиль – интеграция в пайплайны huggingface, бенчмарки AQUA и GLUE
- Очиров Очир – интеграция с EasyEdit, правки алгоритма