Simplex LoRA: Expand important adapters

Давыденко Григорий, **Шалыгин Игорь**

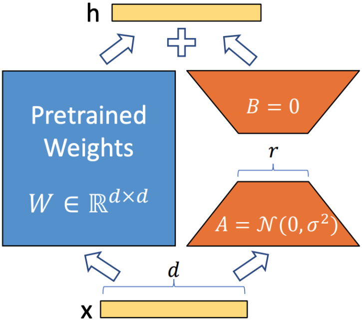
Научный руководитель: Безносиков А. Н.
Научный консультант: Веприков А. С.

Московский физико-технический институт

18 марта 2025 г.

Содержание доклада

Что такое LoRA

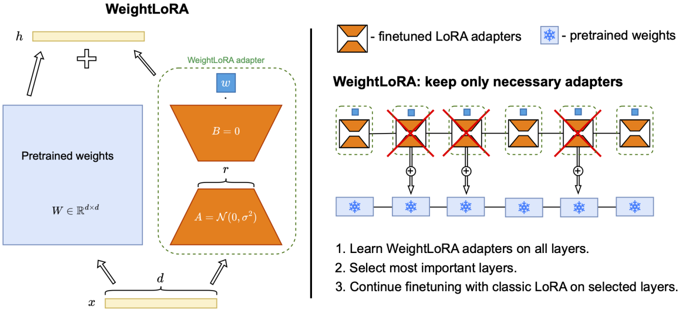


- Тюнинг линейных слоев больших моделей через низкоранговые матрицы $A \in \mathbb{R}^{d\times r}$ и $B \in \mathbb{R}^{r\times d}$
- $r \ll d$ (например, $r = 8$ для моделей с $d = 1024$)

Идея проекта

Для разных слоев требуются LoRA разных рангов.
Такие методы уже есть, например, AdaLoRA [arxiv], однако у этого метода проблема в скорости, для получения рангов совершается подсчет полного SVD разложения.

WeightLoRA* и WeightLoRA+



- Основная идея WeightLoRA: избавиться с помощью параметров важности ω от лишних LoRA
- Основная идея WeightLoRA+: увеличить ранги оставшихся LoRA

* Статья с этими алгоритмами подали на конференцию ACL

SimplexLoRA

Теперь ω - это веса на симплексе $(\sum_{i=1}^{n} \omega_i = 1)$.

За время обучения мы совершаем раз в несколько шагов изменения рангов LoRA в соответствии с их обучаемыми весами:

1. Сначала все ранги равны гиперпараметру $r_i^{(0)} = r^0$, веса - $\omega_i^{(0)} = \frac{1}{n}$
2. После шага в оптимизаторе $\omega^{(k)} = proj_{simplex}(\omega^{(k)})$
3. Через фиксированное количество шагов x меняем ранги LoRA согласно новым весам:

$$r_i^{(u)} = ceil(r^0 \cdot n \cdot \omega_i^{(u \cdot x)})$$

4. Пункты 2 и 3 повторяем несколько раз.
5. Запускаем стандартное обучение LoRA с подобранными рангами.

Преимущества такого подхода

- Физически обоснованный смысл у новых параметров ω_i.
  Мы ищем центр масс между всеми LoRA.
- Количество параметров не увеличивается (уменьш. при ceil).
  Можем задавать при инициализации:

$$\sum_{i=1}^{n} r_i^{(u)} \approx \sum_{i=1}^{n} r^0 \cdot n \cdot \omega_i^{(u \cdot x)} = r^0 \cdot n$$

- Продолжаем обучение с места остановки - локальный минимум, к которому идем, остается тем же.

Расширение рангов

Пусть есть LoRA адаптер с $A_{old} \in \mathbb{C}^{n \times r_{old}}$ и $B_{old} \in \mathbb{C}^{r_{old} \times m}$, который мы хотим расширить до ранга $r > r_{old}$

- **Наивный подход**
  Дополнение обученных матриц A и B нормальной и нулевой

$$A = [A_{old} \quad N_{n \times (r - r_{old})}] \qquad B = [B_{old} \quad 0_{m \times (r - r_o ld)}]$$

- **Через QR-разложение**
  QR-разложения: $A_{old} = Q_A R_A$

$$A = [Q_A \quad (I - Q_A \times Q_A^*) \times N_{n \times (r - r_{old})}] = [Q_A \quad N^{new}_{n \times (r - r_{old})}]$$

$$B = [B_{old} \times R_A^* \quad 0_{m \times r_{old}}]$$

В обоих случаях у нас соблюдается равенство $A_{old} \times B_{old} = A \times B$

Сжатие рангов

Пусть есть LoRA адаптер с $A_{old} \in \mathbb{C}^{n \times r_{old}}$ и $B_{old} \in \mathbb{C}^{r_{old} \times m}$, который мы хотим уменьшить до положительного ранга $r < r_{old}$

Ставим оптимизационную задачу

$$||A_{old} \times B^*_{old} - A \times B^*||_F \longrightarrow min_{A \in \mathbb{C}^{n \times r}, B \in \mathbb{C}^{r \times m}}$$

Решение: эффективно обрезать SVD разложение $A_{old} \times B^*_{old}$.

$$A_{old} = Q_A \times R_A \quad B_{old} = Q_B \times R_B \quad U \times \Sigma \times V^* = SVD(R_A \times R^*_B)$$

$$\begin{cases} U \in \mathbb{C}^{r_{old} \times r_{old}} \\ V \in \mathbb{C}^{r_{old} \times r_{old}} \\ \Sigma \in \mathbb{C}^{r_{old} \times r_{old}} \end{cases} \longrightarrow \begin{cases} U_r \in \mathbb{C}^{r_{old} \times r} \\ V_r \in \mathbb{C}^{r_{old} \times r} \\ \Sigma_r \in \mathbb{C}^{r \times r} \end{cases}$$

Итоговые матрицы A и B

$$A = Q_A \times U_r \qquad B = \Sigma_r \times V^*_r \times Q^*_B$$