

Снижение размерности пространства обучаемых параметров в задаче адаптации к домену

Ремизова Анна Вадимовна

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Научный руководитель к.ф.-м.н. А. В. Грабовой

Москва
2024 г

Цель: Исследовать методы снижения размерности пространства обучаемых параметров при помощи сингулярного разложения матриц, а также показать корректность применения изучаемых методов к задаче классификации текстов.

Методы

- 1 Низкоранговое разложение применяемое к матрицам весов (англ. Low Rank Adaptation) в больших языковых моделях.
- 2 Статистические методы оценки минимизации функции потерь, а также свойства матриц для доказательства применимости предложенного метода к задаче классификации.

Теоретическая значимость. В работе проведен теоретический анализ проблемы снижения размерности пространства обучаемых параметров. Доказана теорема об применимости исследуемого метода низкорангового разложения к задаче многоклассовой классификации.

Практическая значимость. Проведен вычислительный эксперимент, показывающий улучшение качества и экономию ресурсов при решении задачи классификации текстов.

Для задачи классификации текстов:

$$f_{\theta} : \hat{V} \rightarrow [N_c], \quad (1)$$

где $\hat{V} \subset V^*$; V — словарь токенов и V^* — его замыкание или множество всех последовательностей над V , $[N_c]$ — множество классов. Таким образом, модель f_{θ} отображает текст из \hat{V} в класс из $[N_c]$.

Тогда $(X_i, c_i) \in \hat{V} \times [N_c]$ является парой текст — класс, выбранной из $P(X, c)$ и наша цель — оценить $P(c|X)$.

При дообучении модель инициализируется предварительно обученными весами Φ_0 и обновляется до $\Phi_0 + \Delta\Phi$, где $\Delta\Phi$ — набор дообучаемых параметров. При использовании LoRA $\Delta\Phi = \Theta$, где $|\Theta| \ll |\Phi_0|$. Задача минимизации функции потерь:

$$\begin{aligned} \min_{\Theta} \left(- \sum_{X_i \in \hat{V} \subset V^*} \sum_{c_i \in [N_c]} \log(P_{\Phi_0 + \Theta}(c_i | X_i)) \right) = \\ = \max_{\Theta} \sum_{X_i \in \hat{V} \subset V^*} \sum_{c_i \in [N_c]} \log(P_{\Phi_0 + \Theta}(c_i | X_i)). \end{aligned} \quad (2)$$

Подробнее о предложенном методе

В данной работе LoRA применяется к задаче классификации. Структура обновления весов при использовании LoRA адаптера описана в таблице 1,

Таблица: Структура обновления весов при использовании LoRA адаптера

Fine tuning	LoRA fine tuning
$W_{\text{upd}} = W + \Delta W$ $\hat{y} = xW_{\text{upd}} = x(W + \Delta W)$ $\hat{y} = xW + x\Delta W$	$W_{\text{upd}} = W + AB$ $\hat{y} = xW_{\text{upd}} = x(W + AB)$ $\hat{y} = xW + xAB$

где $W \in \mathbb{R}^{d \times k}$ — предобученные веса, $\Delta W \in \mathbb{R}^{d \times k}$ — матрица обновленных весов. ΔW приближается с помощью метода LoRA произведением AB , где $A \in \mathbb{R}^{d \times r}$, $B \in \mathbb{R}^{r \times k}$ и r — ранг матрицы, являющийся гиперпараметром модели. Здесь $A \sim \mathcal{N}(0, \sigma^2)$ и $B = [0]_{r \times k}$.

Состоятельность модели была доказана ранее для задачи генерации. Здесь теорема 1 сформулирована для задачи классификации.

Theorem

Будем считать, что:

- 1 Задана модель с набором параметров Θ^* , генерирующая эмпирическое распределение данных $P_{model}(\cdot, \Theta^*)$, которое аппроксимирует истинное распределение данных P_{true} с минимальным расхождением по KL-дивергенции:

$$\exists \Theta^* : \Theta^* = \arg \min_{\Theta} D_{KL}(P_{true} \parallel P_{model}(\cdot, \Theta)), \quad (3)$$

- 2 При увеличении размера выборки \hat{V} эмпирическое распределение данных $P_{model}(\cdot, \Theta^*)$ приближается к истинному распределению, генерирующему данные.
- 3 Функция ошибки $\mathcal{L}(\Theta)$ — непрерывная, дифференцируемая. Где

$$\mathcal{L}(\Theta) = -\frac{1}{|\hat{V}|} \sum_{X_i \in \hat{V}} \sum_{c_i \in [N_c]} \log(P_{\Phi_0 + \Theta}(c_i | X_i)). \quad (4)$$

Тогда минимизация функции потерь $\mathcal{L}(\Theta)$ приводит к состоятельной оценке истинного распределения, порождающего данные. Т.е.:

$$\lim_{|\hat{V}| \rightarrow \infty} \arg \min_{\Theta} \mathcal{L}(\Theta) = \Theta^*. \quad (5)$$

Theorem (Ремизова Анна, 2024)

В рамках задачи классификации, при заданных условиях:

- 1 *Задана модель с дополнительным слоем*

$$\hat{\mathbf{y}} = \text{softmax} \left(W_{upd}^T \mathbf{x} \right) = \frac{\exp \left(W_{upd}^T \mathbf{x} \right)}{\sum_{i=1}^k \exp \left(W_{upd}^T \mathbf{x} \right)_i}, \quad (6)$$

где

$$W_{upd} = \underset{(d \times k)}{W} + \underset{(d \times k)}{\Delta W}, \quad (7)$$

и \mathbf{x} — это выходной результат модели, W — матрица весов, ΔW — матрица обновленных весов.

- 2 *Данная модель без дополнительного слоя также корректно работает с аппроксимацией*

$$\underset{(d \times k)}{\Delta W} = \underset{(d \times r)}{A} \times \underset{(r \times k)}{B}, \quad (8)$$

- 3 *Выполняются условия теоремы 1. (можно считать данную модель состоятельной).*

Тогда можно утверждать, что при (8) заданная модель с дополнительным слоем гарантирует корректную выходную матрицу.

Открытый исходный датасет для мультиклассовой классификации текстов, написанных человеком и различными языковыми моделями. В выборке рассматривается 4 класса: ChatGPT, Davinci, Cohere, Humans. Всего в датасете 47327 текстов с разметкой по классам.

Таблица: Вес каждого класса

имя класса	вес, б/р
chatGPT	0.986
cohere	1.043
davinci	0.986
human	0.986

Таблица: Метрики качества DistilRoBERTa, мультиклассовая классификация

имя класса	точность	полнота	f1 мера
chatGPT	1.000	0.993	0.997
cohere	0.963	0.999	0.981
davinci	0.986	0.996	0.991
human	0.991	0.952	0.971

Таблица: Метрики качества DistilRoBERTa & LoRA, бинарные классификаторы

имя класса	точность	полнота	f1 мера
chatGPT	1.000	0.891	0.942
cohere	0.999	0.837	0.911
davinci	0.996	0.851	0.918
human	0.875	0.999	0.933

Таблица: Матрица ошибок, DistilRoBERTa

	chatGPT	Cohere	Davinci	Human
chatGPT	0.993	0.002	0.0	0.005
Cohere	0.0	0.999	0.0	0.001
Davinci	0.0	0.001	0.996	0.003
Human	0.0	0.035	0.013	0.952

Таблица: Матрица ошибок,
DistilRoBERTa & LoRA

	chatGPT	Cohere	Davinci	Human
chatGPT	0.79	0.01	0.08	0.12
Cohere	0.0	0.94	0.06	0.003
Davinci	0.001	0.03	0.98	0.0
Human	0.002	0.43	0.25	0.32

Таблица: chatGPT vs Human

	chatGPT	Human
chatGPT	0.892	0.108
Human	0.0	1.00

Таблица: Cohere vs Human

	Cohere	Human
Cohere	0.837	0.163
Human	0.001	0.999

Таблица: Davinci vs Human

	Davinci	Human
Davinci	0.852	0.148
Human	0.003	0.997

- ❶ Сформулирована и доказана теорема о применимости LoRA адаптера к задаче классификации текстов
- ❷ Переформулирована теорема о состоятельности модели для модели с использованием LoRA адаптера
- ❸ Проведен эксперимент и показана эффективность предложенного метода