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 коректность 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 модели BERT [1] с адаптером LoRA к задаче многоклассовой классификации.

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

Постановка задачи классификации текстов

Для задачи классификации текстов:

$$f_\theta : \hat{V} \to [N_c], \tag{1}$$

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 $(X_i, c_i) \in \hat{V} \times [N_c]$ для $i \in [N_{data}]$ является парой текст —класс, выбранной из $P(X, c)$, где X_i — входной текст, а c_i — его класс. Таким образом, наша цель — оценить $P(c|X)$.

Согласно [2] при дообучении модель инициализируется предварительно обученными весами Φ_0 и обновляется до $\Phi_0 + \Delta\Phi$, где $\Delta\Phi$ — набор дообучаемых параметров такой, что $|\ \Delta\Phi\ | = |\ \Phi_0\ |$. Тогда задача минимизации функции потерь имеет вид:

$$\begin{aligned} \min_{\Phi} \left(- \sum_{X_i \in \hat{V} \subset V^*} \sum_{c_i \in [N_c]} \log\left(P_\Phi\left(c_i \mid X_i\right)\right)\right) = \\ = \max_{\Phi} \sum_{X_i \in \hat{V} \subset V^*} \sum_{c_i \in [N_c]} \log\left(P_\Phi\left(c_i \mid X_i\right)\right), \end{aligned} \tag{2}$$

В то время как при использовании LoRA $\Delta\Phi$ задается набором параметров Θ намного меньшего размера: $\Delta\Phi = \Theta$, где $|\ \Theta\ | \ll |\ \Phi_0\ |$ и задача минимизации функции потерь имеет вид:

$$\begin{aligned} \min_{\Theta} \left(- \sum_{X_i \in \hat{V} \subset V^*} \sum_{c_i \in [N_c]} \log\left(P_{\Phi_0+\Theta}\left(c_i \mid X_i\right)\right)\right) = \\ = \max_{\Theta} \sum_{X_i \in \hat{V} \subset V^*} \sum_{c_i \in [N_c]} \log\left(P_{\Phi_0+\Theta}\left(c_i \mid X_i\right)\right). \end{aligned} \tag{3}$$

Предложенный метод

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

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

Таблица: Структура обновления весов при использовании LoRA адаптера

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 трансформер была доказана в работе [3]. Здесь теорема 1 сформулированна 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:*

$$\exists \Theta^* : \Theta^* = \arg\min_{\Theta} D_{KL}(P_{true} \mid\mid P_{model}(\cdot, \Theta)), \tag{4}$$

2. *При увеличении размера выборки $\hat{V}$ эмпирическое распределение данных $P_{model}(\cdot, \Theta^*)$ приближается к истинному распределению, генерирующему данные.*

3. *Функция ошибки $\mathscr{L}(\Theta)$ — непрерывная, дифференцируемая. Где*

$$\mathscr{L}(\Theta) = -\frac{1}{\mid \hat{V} \mid} \sum_{X_i \in \hat{V}} \sum_{c_i \in [N_c]} \log \left(P_{\Phi_0 + \Theta} \left(c_i \mid X_i \right) \right). \tag{5}$$

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

$$\lim_{|\hat{V}| \to \infty} \arg\min_{\Theta} \mathscr{L}(\Theta) = \Theta^*. \tag{6}$$

Theorem (Ремизова Анна, 2024)

В рамках задачи классификации, при заданных условиях:

1. *Модель семейства BERT с дополнительным слоем*

$$\hat{\mathbf{y}} = \text{softmax}\left(W_{upd}^T \mathbf{x}\right) = \frac{\exp\left(W_{upd}^T \mathbf{x}\right)}{\sum_{i=1}^{k} \exp\left(W_{upd}^T \mathbf{x}\right)_i}, \tag{7}$$

   где

$$W_{upd} = \underset{(d\times k)}{W} + \underset{(d\times k)}{\Delta W}, \tag{8}$$

   и x —это выходной результат BERT, W — матрица весов, ΔW — матрица обновленных весов.
2. *Данная модель BERT без дополнительного слоя также корректно работает с аппроксимацией*

$$\underset{(d\times k)}{\Delta W} = \underset{(d\times r)}{A} \times \underset{(r\times k)}{B}, \tag{9}$$

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

Тогда можно утверждать, что при (9) заданная модель BERT с дополнительным слоем гарантирует корректную выходную матрицу.

Результаты эксперимента

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.

В эксперименте на данном датасете обучались модели DistilRoBERTa без использования LoRA адаптера, DistilRoBERTa & LoRA, мультиклассовая классификация и параллельно три DistilRoBERTa & LoRA для бинарной классификации с последующим усреднением результатов.

После обучения для оценки использовались метрики точности, полноты, f1 меры, а также для визуализации ошибки использовалась матрица ошибок (англ. Confusion matrix) как наиболее точно отображающие качетво моделей мультиклассовой классификации [4]. В матрице ошибок по вертикали указаны истинные метки классов, а по горизонтали — предсказанные.

Результаты эксперимента

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 качетва DistilRoBERTa, мультиклассовая классификация

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: Метрики качетва DistilRoBERTa & LoRA, бинарные классификаторы

Результаты эксперимента

	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

	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: Матрица ошибок,
DistilRoBERTa & LoRA

Результаты эксперимента

Итого, в эксперименте, использующем DistilRoBERTa & LoRA для бинарной классификации и последующего усреднения, качество классификации выросло, не потеряв во времени обучения, по сравнению с предобученной моделью DistilRoBERTa.
А также по результатам эксперимента с использованием DistilRoBERTa & LoRA для мультиклассовой классификации, модель сильно выиграла по времени обучения у модели DRoBERTa для мультиклассовой классификации, но проиграв ей в качестве.

Выносится на защиту

1. Сформулирована и доказана теорема о применимости LoRA адаптера к задаче классификации текстов
2. Переформулирована теорема о состоятельности модели для модели с использованием LoRA адаптера
3. Проведен эксперимент и показана эффективность предложенного метода

Список литературы


Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
Advances in neural information processing systems, 30, 2017.


Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.
Lora: Low-rank adaptation of large language models.
arXiv preprint arXiv:2106.09685, 2021.


Minhyeok Lee.
A mathematical investigation of hallucination and creativity in gpt models.
Mathematics, 11(10):2320, 2023.


Margherita Grandini, Enrico Bagli, and Giorgio Visani.
Metrics for multi-class classification: an overview.
arXiv preprint arXiv:2008.05756, 2020.