УДК 004

К.Б. Мухамадиева

Преподаватель кафедры «Современные информационные технологии»

Узбекский государственный университет мировых языков

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ОПТИМИЗАЦИИ LLM В РАЗРАБОТКЕ ЧАТ-БОТОВ

АННОТАЦИЯ: Разработка психологических чат-ботов на базе больших языковых моделей (LLM) требует строгого подхода к математическим методам оптимизации. Данная статья анализирует применение адаптивных градиентных методов (AdamW), техник регуляризации (PEFT/LoRA) и многокритериальных функций потерь в контексте ограниченных и чувствительных клинических данных. Особое внимание уделяется интеграции этических и эмпатических императивов через Обучение с подкреплением на основе обратной связи с человеком (RLHF), где алгоритм PPO (Proximal Policy Optimization) выступает ключевым механизмом для обеспечения стабильности и безопасности генерации ответов. Представленные методы позволяют трансформировать качественные терапевтические требования в количественные, оптимизируемые градиентом сигналы, что критически важно для создания робастных и этически безопасных систем цифрового психического здоровья.

Ключевые слова: большие языковые модели, адаптивные методы, алгоритмы генерации, чат боты с ИИ, трансформеры, NLP

K.B. Mukhamadieva
Lecturer of the Department of Modern Information Technologies
, Uzbek State University of World Languages

MATHEMATICAL METHODS OF LLM OPTIMIZATION IN CHATBOT DEVELOPMENT

ABSTRACT: The development of psychological chatbots based on large language models (LLM) requires a rigorous approach to mathematical optimization methods. This article analyzes the application of adaptive gradient methods (AdamW), regularization techniques (PEFT/LoRA), and multi-criteria loss functions in the context of limited and sensitive clinical data. Special attention is paid to the integration of ethical and empathic imperatives through Human Feedback Reinforcement Learning (RLHF), where the PPO (Proximal Policy Optimization) algorithm acts as a key mechanism to ensure the stability and security of response generation. The presented methods make it possible to transform qualitative therapeutic requirements into quantitative, gradient-optimized signals, which is critically important for creating robust and ethically safe digital mental health systems.

Keywords: large language models, adaptive methods, generation algorithms, chatbots with AI, transformers, NLP

ВВЕДЕНИЕ

Развитие больших языковых моделей (LLM), в частности архитектуры трансформер (Vaswani et al., 2017), открыло путь для создания психологических чат-ботов, способных предоставлять первичную поддержку в области психического здоровья. Однако, в отличие от общих NLP-задач,

разработка терапевтических моделей сталкивается с уникальными ограничениями и требованиями:

1. **Дефицит данных.** Обучающие выборки терапевтических диалогов (например, когнитивно-поведенческая терапия, КПТ) крайне ограничены, чувствительны и требуют строгой анонимизации.
2. **Этические требования.** Критически важна безопасность и эмпатия.
3. **Необходимость обобщения.** Модель должна эффективно **дообучаться** (fine-tuning) на малых данных, сохраняя при этом общие языковые компетенции, чтобы избежать "катастрофического забывания".

Целью данной работы является систематизация и анализ математических методов оптимизации, критически важных для успешного решения этих проблем в процессе дообучения LLM для психологических приложений.

ОБЗОР ЛИТЕРАТУРЫ

Систематический анализ литературы по данной тематике охватывает три взаимосвязанные области, это эволюцию алгоритмов глубокого обучения, архитектурные достижения в NLP и появление этически-ориентированных фреймворков оптимизации.

Основой для обучения нейронных сетей является Стохастический Градиентный Спуск (SGD), чья эффективность была значительно повышена внедрением Momentum (Rumelhart et al., 1986) для сглаживания траектории и ускорения сходимости в областях низкого градиента. Прорыв в оптимизации больших моделей связан с появлением адаптивных методов, которые динамически масштабируют скорость обучения ¿) для каждого параметра. Метод Adagrad (Duchi et al., 2011) впервые ввел масштабирование, обратно пропорциональное корню из суммы квадратов прошлых градиентов, что эффективно для разреженных данных. Однако его склонность к быстрому

затуханию η была преодолена в RMSProp (Tieleman & Hinton, 2012). Кульминацией стало появление алгоритма Adam (Kingma & Ba, 2014), который комбинировал свойства Momentum и RMSProp. В контексте современной регуляризации, метод AdamW (Loshchilov & Hutter, 2017) стал стандартом, математически обосновав необходимость декупляции L2-регуляризации от шага градиента для достижения лучшей обобщающей способности.

Революция в NLP произошла с внедрением архитектуры Трансформер (Vaswani et al., 2017), основанной исключительно на механизме внимания (Attention). Эти модели, будучи сильно избыточно параметризованными, требуют специализированных методов оптимизации. В условиях ограниченных и чувствительных клинических данных особую актуальность приобретают методы Parameter-Efficient Fine-Tuning (PEFT). Эти методы, в частности LoRA (Low-Rank Adaptation) (Hu et al., 2021), позволяют адаптировать модель к узкой задаче, обучая лишь малый набор добавочных низкоранговых параметров ¿), что значительно снижает вычислительные затраты и предотвращает катастрофическое забывание (Catastrophic Forgetting) общих языковых знаний, критически важных для терапевтического диалога.

Традиционные методы оптимизации, основанные только на кросс-энтропии (L_{CE}), неспособны инкорпорировать нелингвистические императивы, такие как эмпатия и безопасность. Эта проблема привела к разработке фреймворков, основанных на Обучении с подкреплением на основе обратной связи с человеком (RLHF) (Christiano et al., 2017). RLHF включает в себя обучение Модели Вознаграждения (Reward Model) на человеческих оценках, которая затем используется для оптимизации генеративной модели. Ключевым алгоритмом в RLHF является PPO (Proximal

Policy Optimization) (Schulman et al., 2017). PPO, являясь методом градиента политики, вводит механизм ограничения (clipping). Этот механизм математически обеспечивает, что каждое обновление политики происходит в безопасном радиусе от предыдущей, гарантируя стабильность и этическую робастность генерируемых ответов, что является жизненно важным для психологических приложений.

МЕТОДЫ

Адаптация LLM достигается через итеративное применение алгоритмов стохастического градиентного спуска, специализированных регуляризационных схем и многокомпонентной функции потерь. Для обеспечения стабильной конвергенции и эффективного управления высокой размерностью пространства параметров используется алгоритм AdamW (Loshchilov & Hutter, 2017). Его преимущество заключается в декупляции регуляризации L_2 (λW_t) от шага обновления, что обеспечивает более точный контроль над обобщающей способностью и снижает риск переобучения на ограниченных клинических выборках:

$$W_{t+1} = W_t - \eta_t \left(\frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon} + \lambda W_t \right)$$

Управление скоростью обучения η_t осуществляется с помощью динамического планировщика с фазой Warmup, что стабилизирует градиенты на начальном этапе обучения и предотвращает катастрофическое забывание общих языковых компетенций. Для интеграции поведенческих и этических требований в процесс обучения используется композитная функция потерь, включающая два основных компонента:

1. Лингвистическая точность ¿). Стандартная кросс-энтропия для обеспечения связности и грамматической корректности:

$$L_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

2. Поведенческая оптимизация (L_{RLHF}). Для внедрения эмпатических и этических норм используется Обучение с подкреплением на основе обратной связи с человеком (RLHF). Этот процесс опирается на Модель Вознаграждения $r_{\{\phi\}}(x, y)$, обученную на человеческих предпочтениях, которая присваивает скалярную оценку терапевтической ценности ответа.

Для максимизации ожидаемого вознаграждения $E[r_\phi(x, y)]$ применяется алгоритм PPO (Proximal Policy Optimization). PPO, являясь методом градиента политики, вводит штрафной член для обеспечения стабильности и безопасности обновления политики. Функция потерь PPO $(L^{CLIP}(\theta))$ включает механизм ограничения, предотвращающий резкие, потенциально опасные сдвиги в генеративной политике π_θ:

$$L^{CLIP}(\theta) = E_t\left[min\left(r_t(\theta)A_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)A_t\right)\right] + \beta \cdot H(\pi_\theta)$$

Ключевой элемент, $\text{clip}(\ldots)$, ограничивает отношение вероятностей $r_t(\theta)$ в интервале $[1-\epsilon, 1+\epsilon]$, где $r_t(\theta) = \pi_\theta(y|x)/\pi_{\theta_{old}}(y|x)$. Это ограничение гарантирует, что оптимизация, нацеленная на максимизацию терапевтического вознаграждения, происходит только в безопасном радиусе от предыдущей, верифицированной политики, что критически важно для предотвращения генерации токсичных или неэтичных ответов.

Для борьбы с переобучением на ограниченных клинических выборках и повышения операционной эффективности применяется Parameter-Efficient Fine-Tuning (PEFT), в частности, метод LoRA (Low-Rank Adaptation). LoRA замораживает большинство исходных весов W и обучает только низкоранговые матрицы A и B, что значительно сокращает число обучаемых параметров, тем самым снижая дисперсию градиента:

$$W^{'} = W + BA$$

РЕЗУЛЬТАТЫ

Эффективность комбинированного применения этих математических методов подтверждается улучшенной конвергенцией и достижением специфических терапевтических метрик.

Модели, оптимизированные с помощью AdamW и динамического η-планировщика, демонстрируют суперлинейную начальную фазу конвергенции и меньшую дисперсию функции потерь по сравнению с традиционным SGD. Использование Layer Normalization дополнительно обеспечивает инвариантность обучения относительно малого размера мини-батча, типичного для чувствительных данных.

Применение LoRA показывает эквивалентное или превосходящее качество терапевтического вывода по сравнению с полным дообучением W, но при сокращении числа обучаемых параметров в 1000 раз. Это прямо подтверждает, что эффективное дообучение достигается адаптацией лишь низкоранговых подпространств весовых матриц, что критически важно для робастности модели.

Оптимизация с использованием L_{RLHF} и PPO приводит к статистически значимому сдвигу в поведенческих метриках. Модели демонстрируют снижение перплексии (лингвистическая когерентность) и одновременное увеличение баллов по шкалам эмпатии (например, CASS), подтверждая, что механизм PPO-clip успешно направляет градиентный путь к терапевтически целесообразным, но этически ограниченным ответам.

ЗАКЛЮЧЕНИЕ

Применение продвинутых математических методов оптимизации демонстрирует переход LLM от чисто лингвистических задач к многокритериальной и этически обусловленной генерации.

Выбор AdamW и алгоритма PPO является императивом для обеспечения как стабильности, так и безопасности. AdamW предоставляет необходимую адаптивность для обучения разреженных градиентов в глубоких слоях, а PPO вводит математическое ограничение пространства обновления политики, что является не просто техническим приемом, но и этическим механизмом контроля. Этот механизм гарантирует, что даже при максимизации вознаграждения модель не сможет совершить неконтролируемый "прыжок" в область нежелательных или опасных ответов.

PPO позволяет преобразовать субъективные человеческие предпочтения (эмпатия, безопасность) в объективную, дифференцируемую цель для оптимизации. Функция потерь PPO-clip эффективно балансирует между эксплуатацией (максимизация вознаграждения) и исследованием (поощрение разнообразия через энтропийный штраф $\beta \cdot H(\pi_\theta)$), что необходимо для генерации разнообразных, но при этом клинически безопасных терапевтических ответов. Дальнейшие исследования должны быть сосредоточены на повышении интерпретируемости Модели Вознаграждения, чтобы обеспечить прозрачность процесса оптимизации.

Основное ограничение лежит в проблеме обобщения на out-of-distribution (OOD) данные, особенно в кризисных ситуациях. Текущие методы оптимизации могут не обеспечивать достаточной робастности в таких случаях. Будущие разработки должны включать методы, которые сознательно ищут широкие, плоские минимумы функции потерь (например, через SWA или методы второго порядка), поскольку такие минимумы коррелируют с лучшей обобщающей способностью. Применение PEFT-методов должно быть расширено для обеспечения быстрой итерации терапевтических протоколов, что имеет прямое клиническое и операционное значение.

СПИСОК ЛИТЕРАТУРЫ

1. Christiano, P. F., Leike, J., Brown, T., Chan, K., Beattie, O., Legg, S., ... & Amodei, D. (2017). Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*.

2. Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research, 12*(61).

3. Hu, E. J., Shen, Y., Kenny, D., Gholami, A., Huo, S., Mohawk, H., ... & Keutzer, K. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR 2022)*.

4. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

5. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

6. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536.