

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РЕШЕНИЙ ПРОБЛЕМЫ ЗАТУХАЮЩИХ ГРАДИЕНТОВ В ТРАНСФОРМЕРНЫХ АРХИТЕКТУРАХ: LAYER NORM И NEURAL ODE ПОДХОДЫ

И.А. Власов¹, С.Н. Исаев¹

¹Самарский университет, г. Самара
e-mail: igorvlasov732@gmail.com

Аннотация

В работе проведен сравнительный анализ между способами борьбы с затухающими градиентами в нейросетевых моделях на примере задачи NLP. Как правило, для решения задачи затухающих градиентов используются skip connections в связке с нормировкой слоев, но, с недавним развитием Neural ODE, идеи рассмотрения задач как дифференциальных уравнений и их решения находят свое применение в обработке естественного языка. Были проведены эксперименты, в ходе которых выяснилось, что ODE подход показывает лучшие метрики обучения по сравнению со стандартным residual connections, используемым в построение трансформерных архитектур.

Ключевые слова: residual connections, Neural ODE, трансформеры, vanishing gradients, exploding gradients.

Введение

Проблема затухающих градиентов является актуальной в области глубоких нейронных сетей. Для ее решения применяют различные техники, одной из которых является skip connections. В частности, в трансформерных архитектурах используются residual connections вместе с различными стратегиями нормировки слоев [1]. Мы предлагаем изменить подход к решению проблемы взрывающихся градиентов, путем применения метода, вдохновленного идеей построения Neural ODE.

В работе выполняются эксперименты с использованием трансформеров, сравнивается качество непрерывного и дискретного подхода, делаются выводы по их применимости. В качестве наборов данных используются wikitext-2-raw-v1 и первые 100000 строк из wikitext-103-raw-v1, токенизатор взят из модели GPT-2.

1. Использование residual connections в трансформерных архитектурах

Применение residual connections заключается в преобразовании блока нейронной сети по следующей формуле:

$$h_{t+1} = h_t + F_t(h_t) \quad [2]. \quad (1)$$

В случае трансформерной нейронной сети преобразованиями F являются attention и MLP блоки. То есть, мы добавляем изначальное скрытое состояние к скрытому состоянию после прохождения блока внимания и добавляем получившиеся эмбединги к изначальному скрытому состоянию после прохождения через MLP. В общем случае подход со skip connection в виде residual connection помогает решить проблему затухания

градиентов [2, 3]. Однако, такое решение может вызвать обратную проблему взрыва градиентов. Для того, чтобы это не допустить применяется нормировка слоев. Существуют два основных метода тренировки: Post-LN и Pre-LN. Проблема первого метода заключается в затухании градиентов нижних блоков, вызванное многократной нормализацией, а второго в провале представлений, следствием которого является уменьшение емкости модели [1]. Существуют различные модификации этих методов, разработанные для решения этих проблем [1]. Таким образом, residual connections является мощным инструментом решения проблемы затухания градиентов, но его реализация, основанная на применении нормировки слоев кажется неэффективной.

2. Использование дифференциальных уравнений для решения проблемы затухающих градиентов в трансформерных архитектурах

Рассмотрим еще раз формулу 1 и будем добавлять слои, уменьшая шаг. В пределе получаем дифференциальное уравнение: $\frac{dh(t)}{dt} = F(h(t), t)$ [4]. То есть выполняется переход от дискретного метода к непрерывному.

Нейронные сети, использующие такой способ вычисления слоев называются Neural ODE и имеют ряд преимуществ [4]. Попробуем применить идею Neural ODE для решения проблемы взрывающихся градиентов.

Проанализируем почему именно происходит взрыв градиентов при использовании residual connections: $\frac{\partial L}{\partial x_n} = \frac{\partial L}{\partial x_{n+1}} \frac{\partial x_{n+1}}{\partial x_n}$, $\frac{\partial x_{n+1}}{\partial x_n} = \frac{\partial(x_n + F(x_n))}{\partial x_n} = 1 + \frac{\partial F(x_n)}{\partial x_n}$. Поскольку 1 является малой константой, проблемой может стать $\frac{\partial F(x_n)}{\partial x_n}$. Применим Neural ODE подход к $F_t(t)$, который в нашей задаче является механизмом внимания и следующий за ним MLP. Таким образом, требуется решить дифференциальное уравнение $\frac{dx(t)}{dt} = F(x(t))$.

3. Сравнение residual connections со слоями нормировки и ODE подходы

Во всех экспериментах будут строиться модели с одинаковым количеством слоев, голов внимания, размерами эмбеддингов. В качестве метрики оценки нашей модели будем использовать perplexity. В первой серии экспериментов посмотрим на модели, применяющие Post-LN и Pre-LN.

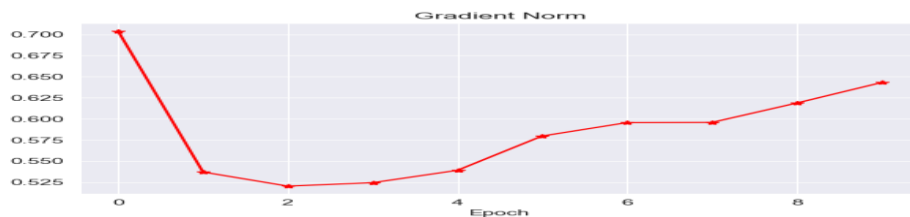


Рисунок 1 – L2 норма градиентов для Post-LN

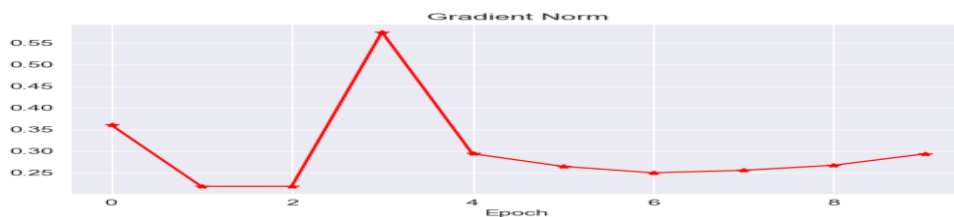


Рисунок 2 – L2 норма градиентов для Pre-LN

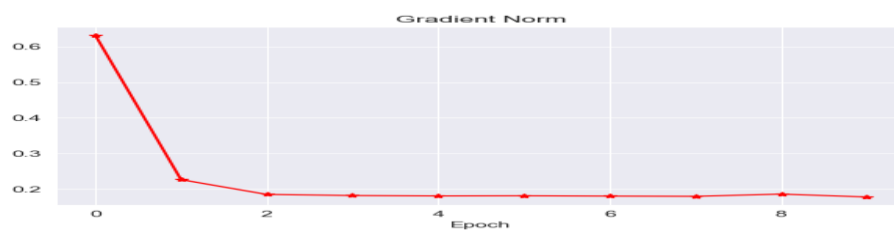


Рисунок 3 – L2 норма градиентов для модели, использующую нормировку из [1]

Проблемы нестабильности градиентов для Post-LN и Pre-LN видны на рисунках 1 и 2. Модель со способом нормировки описанным в [1] показал наилучший результат, продемонстрированный на рисунке 3.

Во второй серии экспериментов используем ODE подход. На рисунках 4 и 5 видно, что удалось решить проблему взрыва градиентов и сделать обучение более стабильным. В таблице 1 приведены метрики моделей.

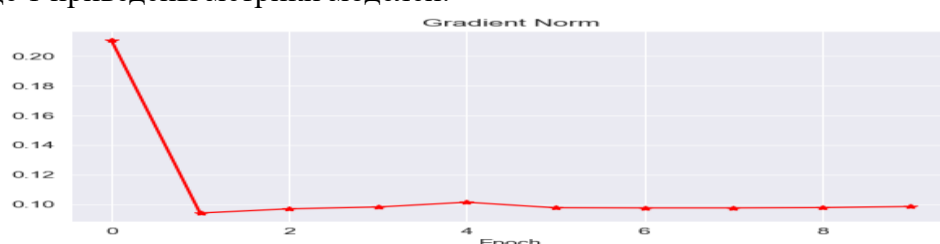


Рисунок 4 – L2 норма градиентов для модели с ODE слоями, решающимися методом Эйлера

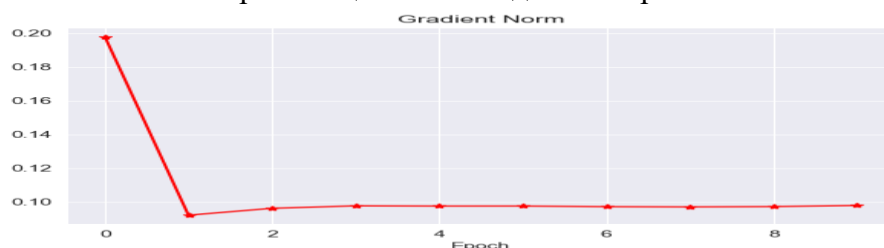


Рисунок 5 – L2 норма градиентов для модели с ODE слоями, решающимися методом Рунге-Кутты 2 порядка

Таблица 1 – Метрики моделей

Вид модели	PPL(wikitext-2-raw-v1)	Время обучения(минуты)	PPL(wikitext-103-raw-v1)	Время обучения(минуты)
Post-LN Transformer	33,76	14,08	32,62	50,19
Pre-LN Transformer	10,31	21,51	8,01	59,54
ResiDual Transformer	10,10	18,47	8,08	50,59
ODE Euler Transformer	9,39	16,76	7,68	47,33
ODE RK2 Transformer	9,37	28,69	7,67	78,88
Pre-LN Transformer	10,31	21,51	8,01	59,54

Заключение

Эксперименты показывают, что стандартные методы борьбы со взрывом градиентов в виде нормировки слоев являются неэффективными: в первом случае, несмотря на нормировку происходит взрыв, во втором также происходит неустойчивости, однако не такие сильные как в первом случае. Различные приемы, могут решить проблему неустойчивости при нормировке [1]. Однако при использовании таких приемов итоговые качества моделей хуже, чем при использовании neural ODE. При чем такая тенденция сохранится при увеличении количества данных что видно в таблице 1. В подходе с ODE важно правильно оценить сложность задачи, поскольку при применении более сложных решателей ODE время обучения увеличивается значительно, а итоговое качество почти не растет, что можно наблюдать в таблице 1.

Список использованных источников

1. Residual: Transformer with dual residual connections / S. Xie, H. Zhang, J. Guo [et al.] // arXiv. – 2023. – arXiv:2304.14802.
2. Residual connections encourage iterative inference / S. Jastrzębski, D. Arpit, N. Ballas [et al.] // arXiv. – 2017. – arXiv:1710.04773.
3. Orhan, A. Pitkow X. Skip connections eliminate singularities / A. Orhan, X. Pitkow // arXiv – 2017. – arXiv:1701.09175.
4. Neural Ordinary Differential Equations / R. Chan, Y. Rubanova, J. Bettencourt [et al.] // Advances in neural information processing systems. – 2018. – V. 31.

COMPARATIVE ANALYSIS OF SOLUTIONS TO THE PROBLEM OF DECAYING GRADIENTS IN TRANSFORMER ARCHITECTURES: LAYER NORM AND NEURAL ODE APPROACHES

I.A. Vlasov¹, S.N. Isaev¹

¹*Samara University, Samara
email: igorvlasov732@gmail.com*

Abstract

The paper compares methods for dealing with vanishing gradients in NLP. Traditional skip-connections and layer normalization are compared with the Neural ODE approach, where the part of a network is trained as a differential equation. Experiments show that the ODE approach provides better training metrics than standard residual connections in transformers.

Keywords: residual connections, Neural ODE, transformers, vanishing gradients, exploding gradients.