# Слой за слоем: раскрытие скрытых представлений в языковых моделях

Дата: 2025-02-04 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.02013

Рейтинг: 70
Адаптивность: 85

## Ключевые выводы:

Исследование направлено на анализ скрытых представлений в промежуточных слоях больших языковых моделей (LLM). Основной вывод: промежуточные слои часто содержат более богатые представления и превосходят финальные слои по эффективности на различных задачах, вопреки традиционному мнению о том, что финальные слои наиболее полезны.

## Объяснение метода:

Исследование показывает, что промежуточные слои LLM часто превосходят финальные по качеству эмбеддингов. Практическая ценность высока для понимания работы моделей и потенциального улучшения результатов, но ограничена доступностью промежуточных слоев в стандартных API. Концептуальная ценность значительна для формирования более эффективных запросов и понимания ограничений моделей.

## Ключевые аспекты исследования: 1. **Эффективность промежуточных слоев**: Исследование обнаружило, что промежуточные слои языковых моделей часто превосходят финальные слои по качеству представлений и производительности на различных задачах. Это противоречит общепринятому мнению о том, что финальные слои всегда дают лучшие представления.

**Единая система метрик оценки**: Авторы предлагают унифицированную структуру метрик для оценки качества представлений, основанную на теории информации, геометрии и инвариантности к возмущениям входных данных. Ключевые метрики включают энтропию, кривизну, инвариантность к аугментациям.

**Сжатие информации в промежуточных слоях**: В авторегрессивных моделях наблюдается характерный "провал сжатия" в средних слоях, где происходит оптимальное балансирование между сохранением важной информации и отбрасыванием шума.

**Архитектурные различия**: Исследование сравнивает различные архитектуры

(трансформеры, модели пространства состояний) и обнаруживает, что эффект превосходства промежуточных слоев проявляется во всех архитектурах, но с разной интенсивностью.

**Влияние обучения "цепочкой рассуждений"**: Модели, дообученные с использованием chain-of-thought, сохраняют более высокую энтропию в промежуточных слоях, что позволяет им лучше удерживать контекст для многошагового рассуждения.

## Дополнение: Для непосредственного применения методов данного исследования действительно требуется доступ к промежуточным слоям модели, что обычно недоступно через стандартные API. Однако многие концепции и подходы можно адаптировать для использования в стандартном чате без необходимости доступа к внутренним представлениям модели.

Вот ключевые концепции и подходы, которые можно адаптировать для стандартного чата:

**Chain-of-Thought (CoT)**: Исследование показывает, что модели, обученные с использованием CoT, сохраняют более высокую энтропию в промежуточных слоях, что позволяет им лучше удерживать контекст. Пользователи могут применять принцип "цепочки рассуждений" в своих запросах, побуждая модель постепенно разворачивать логику рассуждений шаг за шагом, что помогает ей использовать промежуточные представления более эффективно.

**Структурирование запросов**: Зная, что авторегрессивные модели обрабатывают информацию последовательно и создают "бутылочное горлышко" в средних слоях, пользователи могут структурировать запросы таким образом, чтобы важная информация была распределена более равномерно, а не сконцентрирована в начале или конце.

**Инвариантность к возмущениям**: Исследование показывает важность робастности представлений к небольшим изменениям во входных данных. Пользователи могут формулировать запросы с некоторой избыточностью, чтобы ключевые концепции были выражены несколькими способами, что повышает вероятность их правильного восприятия моделью.

**Работа с длинными запросами**: Исследование показывает, что с увеличением длины промпта необработанная энтропия растет, но нормализованная энтропия растет медленнее. Это означает, что каждый дополнительный токен вносит всё меньше уникальной информации. Пользователи могут оптимизировать длину запросов, избегая излишней вербальности и повторений, но обеспечивая достаточно контекста.

**Учет особенностей различных моделей**: Исследование выявило различия в обработке информации в разных архитектурах. Понимание того, что, например, авторегрессивные модели (типа GPT) имеют более выраженное "бутылочное горлышко" в средних слоях по сравнению с двунаправленными моделями (типа

BERT), может помочь пользователям адаптировать свои запросы под конкретную модель.

Применение этих концепций в стандартном чате может привести к следующим результатам: - Улучшение качества ответов на сложные запросы, требующие многошагового рассуждения - Повышение стабильности и последовательности ответов - Более эффективное использование контекстного окна модели - Лучшее понимание и учет ограничений конкретных моделей

Таким образом, хотя непосредственное использование промежуточных слоев недоступно в стандартном чате, понимание принципов их работы может значительно улучшить взаимодействие с LLM.

## Анализ практической применимости: 1. **Эффективность промежуточных слоев** - Прямая применимость: Высокая. Пользователи могут настроить модели для извлечения эмбеддингов из промежуточных слоев вместо финальных для повышения производительности на различных задачах без необходимости переобучения моделей. - Концептуальная ценность: Высокая. Понимание того, что промежуточные слои часто содержат более богатые представления, меняет подход к использованию LLM для извлечения признаков. - Потенциал для адаптации: Высокий. Этот подход может быть применен к любой модели трансформерного типа.

**Единая система метрик оценки** Прямая применимость: Средняя. Обычные пользователи не смогут напрямую применить эти метрики, но они могут использовать результаты для выбора оптимальных слоев. Концептуальная ценность: Высокая. Понимание того, как измерять качество представлений, помогает лучше понять внутреннюю работу моделей. Потенциал для адаптации: Средний. Метрики могут быть адаптированы разработчиками для создания более эффективных моделей и инструментов.

**Сжатие информации в промежуточных слоях**

Прямая применимость: Низкая. Это теоретическое наблюдение, которое сложно напрямую использовать без специальных инструментов. Концептуальная ценность: Высокая. Понимание механизмов сжатия информации помогает лучше интерпретировать работу моделей. Потенциал для адаптации: Средний. Знание о "бутылочном горлышке" в средних слоях может быть использовано для разработки более эффективных промптов и настройки моделей.

**Архитектурные различия**

Прямая применимость: Низкая. Обычные пользователи редко выбирают между различными архитектурами. Концептуальная ценность: Средняя. Понимание различий между архитектурами может помочь при выборе модели для конкретной задачи. Потенциал для адаптации: Средний. Знание о том, как различные архитектуры обрабатывают информацию, может помочь в разработке специализированных решений.

**Влияние обучения "цепочкой рассуждений"**

Прямая применимость: Высокая. Пользователи могут предпочесть модели, обученные с помощью CoT, для задач, требующих сложных рассуждений. Концептуальная ценность: Высокая. Понимание того, как CoT влияет на внутренние представления, помогает лучше использовать такие модели. Потенциал для адаптации: Высокий. Принципы CoT могут быть адаптированы для различных типов запросов.

## Prompt:

Использование знаний из исследования "Слой за слоем" в промптах для GPT ## Ключевые инсайты из исследования

• Промежуточные слои LLM содержат более богатые представления, чем финальные

• Лучший слой часто находится примерно в середине сети (40-60% глубины)

• Авторегрессивные модели демонстрируют выраженное сжатие информации в средних слоях

• Метрики качества представлений коррелируют с производительностью на задачах

## Пример промпта, использующего эти знания

[======] Я хочу, чтобы ты решил следующую задачу, используя многошаговое рассуждение. Важно сохранять высокую энтропию информации на каждом шаге, как это происходит в промежуточных слоях языковых моделей.

Задача: [описание сложной задачи]

Пожалуйста: 1. Сначала запиши все ключевые факты и переменные 2. В промежуточных шагах сохраняй больше контекста, не отбрасывая информацию преждевременно 3. На каждом шаге рассуждай как о возможных направлениях решения, так и об ограничениях 4. Только в финальном шаге сделай сжатие информации до конкретного ответа

Это позволит использовать преимущество богатых представлений, которые формируются в промежуточных этапах обработки информации. [======]

## Почему это работает

Этот промпт использует понимание того, как работают внутренние слои языковых моделей. Прося модель сохранять высокую энтропию информации в промежуточных шагах, мы имитируем работу промежуточных слоев нейросети, которые, согласно

исследованию, содержат более богатые представления.

Мы также структурируем рассуждение так, чтобы финальное сжатие информации происходило только в конце, что соответствует естественной архитектуре авторегрессивных моделей, где сжатие информации происходит ближе к выходным слоям.

Такой подход особенно полезен для сложных задач, требующих многошагового рассуждения или обработки длинных последовательностей информации.