Отчет о выполнении исследовательской работы Logit Lens in Multimodal Models

Игнатова Мария

Цель работы:

Целью данной работы являлось изучение метода Logit Lens и исследования применения его к мультимодальным моделям типа Qwen2-VL. В частности, в данной работе стаим целью посмотреть и выявить особенности поведения модели (предсказывания токенов) на разных слоях для разных датасетов, выявить проблемы и паттерны поведения, если таковые пронаблюдаются.

I. Вычисление логитов:

В качестве мультимодельно модели была взята Qwen2-VL. По своей архитектуре модель в декодере имеет 29 скрытых слоев, с ними и была произведена работа: к каждому слою прменялась голова логитов lm head, которая делала преобразование каждого слоя так, словно это выходной слой (чтобы мы смогли посмотреть, что творится внутри). В качестве датасетов мной были выбраны `https://huggingface.co/datasets/ByteDance/MTVQA` **MTQVA от ByteDance** и `https://huggingface.co/datasets/linxy/LaTeX_OCR` **LaTeX OCR.** Где первый датасет состоит из изображений с текстом на разных языках и пар question answer к каждому изображению, а второй датасет - пара изображение формулы и ее вариант записи в LaTex. Поставила гипотезы: что в зависимости от языка логиты могут вести себя по-разному и генерировать правильные токены с разной скоростью. А также в зависимости от изображения логиты могут вести себя иначе.

II. Проверка гипотез:

С помощью специально подготовленного скрипта для каждого изображения из каждого датасета были проведены следующие действия:

1. Вычисление и запись в .csv файл декодированных токенов по слоям
2. Вычисление энтропии
$$H(P) = -\sum_i p_i \cdot \log(p_i),$$
по слоям для каждого токена и усреднение по токенам для каждого изображениие
3. Построение усредненной по токенам энтропии по слоям
4. Построение тепловой карты токенов по слоям

Первый датасет дал следующие тепловые карты и энторию:

По энтропии видно(рис.1), что при генерации токенов есть точка (слой), начиная с которого энтропия начинает заметно уменьшаться. Этот слой для данного изображения - 22 (layer 21). Это же подтверждает изменение цветовой палитры на тепловой карте - начиная с 22 слоя карта зеленеет, а с 25, считая все с нуля - синеет, что означает, что хаотические генерации сменяются более уверенными и близкими к истине. Красным цветом выделены

Рис. 1: Средняя энтропия для тестового изображения модели Qwen2-VL

ячейки, тоены которых на потокенной генерации и на генерации ответа моделью совпали.

Тепловые карты для **MTQVA** (рис.3 - рис.5), вне зависимости от языка (был выбран русский и арабский) повторяют этот тренд: только на последниех скрытыех слоях происходит осознанная генерация токенов. Вне зависимости от языка промпта (пара изображение и текст), по слоям можно видеть, что модель "думает"на разных языках, и уже под конец, механизм attention склоняет модель генерить токены на языке промпта.

Данные с тепловых карты для датасета **LaTeX OCR** с изображениями формул также повторяют эту закономерность. Тут будет интереснее посмотреть на изменение энтропии по слоям в зависимости от датасета:

III. Выводы:

Интересное наблюдение: энтропия вне зависимости от датасета (рис. 6)(изображениие + текстовый промпт) для модели Qwen2 Vl имеет пороговое значение, которое определяется слоем ‘layer 21‘ , начиная с которого она падает. Это падение энтропии означает "адекватность"предсказанных моделью токенов.

Скорее всего это можно считать особенностью того, как работает конкретная мультимодальная модель, однако статьях и публикаций, как

1. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`,
2. `https://medium.com/@adjileyeb/unlocking-visual-insights-applying-the-logit-lens-to-` (открывается с vpn),
3. `https://arxiv.org/pdf/2503.11667v1`,
4. `https://arxiv.org/pdf/2406.11193`

также проследивается характерное поведение логитов - наало предсказания имеющих смысл токенов на последних слоях модели.

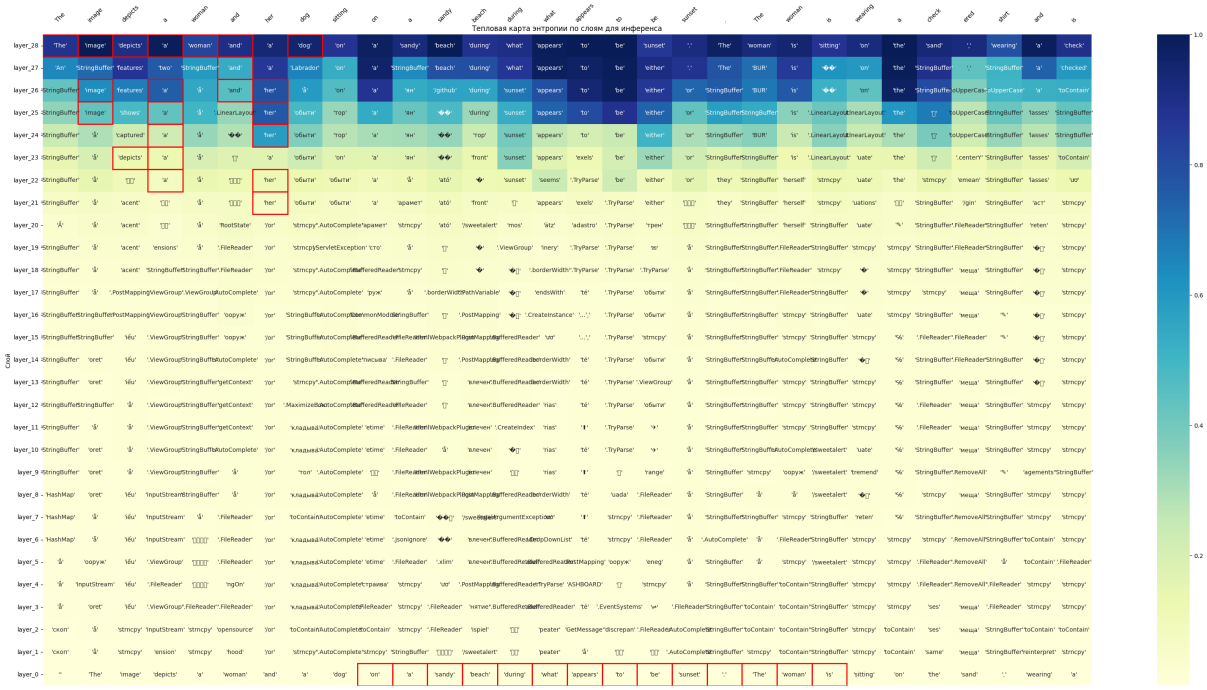

Рис. 2: Тепловая карта для тестового изображения модели Qwen2-VL

В статье 4 было проведено исследование не только логитов, но и исследование того, каике нейроны включаются при определенном датасете. Было замечено, что определенные нейроны отключаются, при работе с определенными предметами (датасетами). но это не влияет на логиты!

Открытые проблемы интерпретации нейросетей описаны Less Wrong `https://www.lesswrong.com/posts/LbrPTJ4fmABEdEnLf/200-concrete-open-problems-in-mechanistic-interpr` но интересно подчеркнуть, что при работе с логит линзами мы так и не понимаем "несогласованность"ранних слоёв (что было получено в ходе данного исследования).
Возможные предложения по решению :

- Искать интерпретируемые базисы. SVD/sparse coding выделит ортогональные компоненты (например, отдельно «цвет», отдельно «форма»), которые легче интерпретировать.

- Использовать проекции на семантические пространства (например, через word2vec). Можно сравнить, как logits слоя N соотносятся с семантикой word2vec (например, если слой 5 выдаёт высокие logits для «фруктов», а его активации близки к «яблоку», это подтверждает интерпретацию).

IV. Приложение:

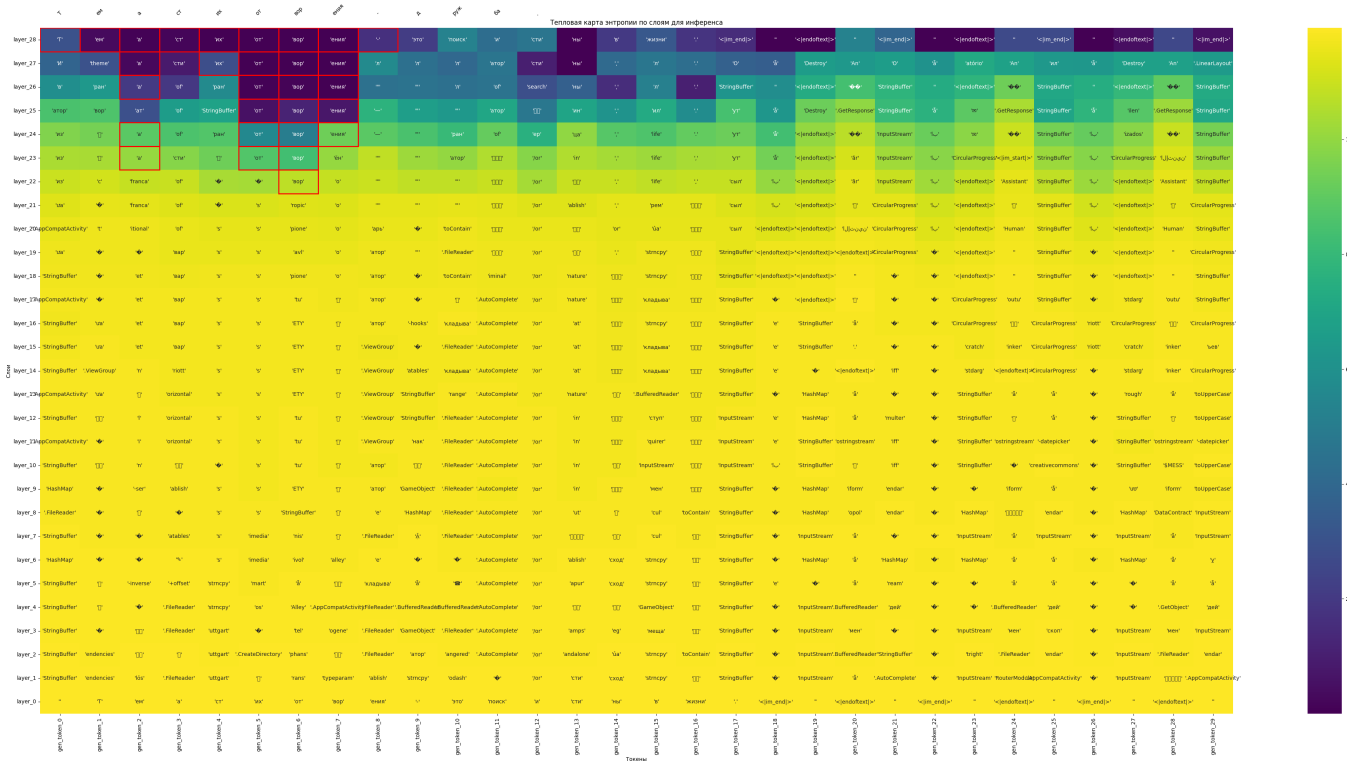

Рис. 3: Тепловая карта для изображения 1 датасета MTQVA (русский язык)

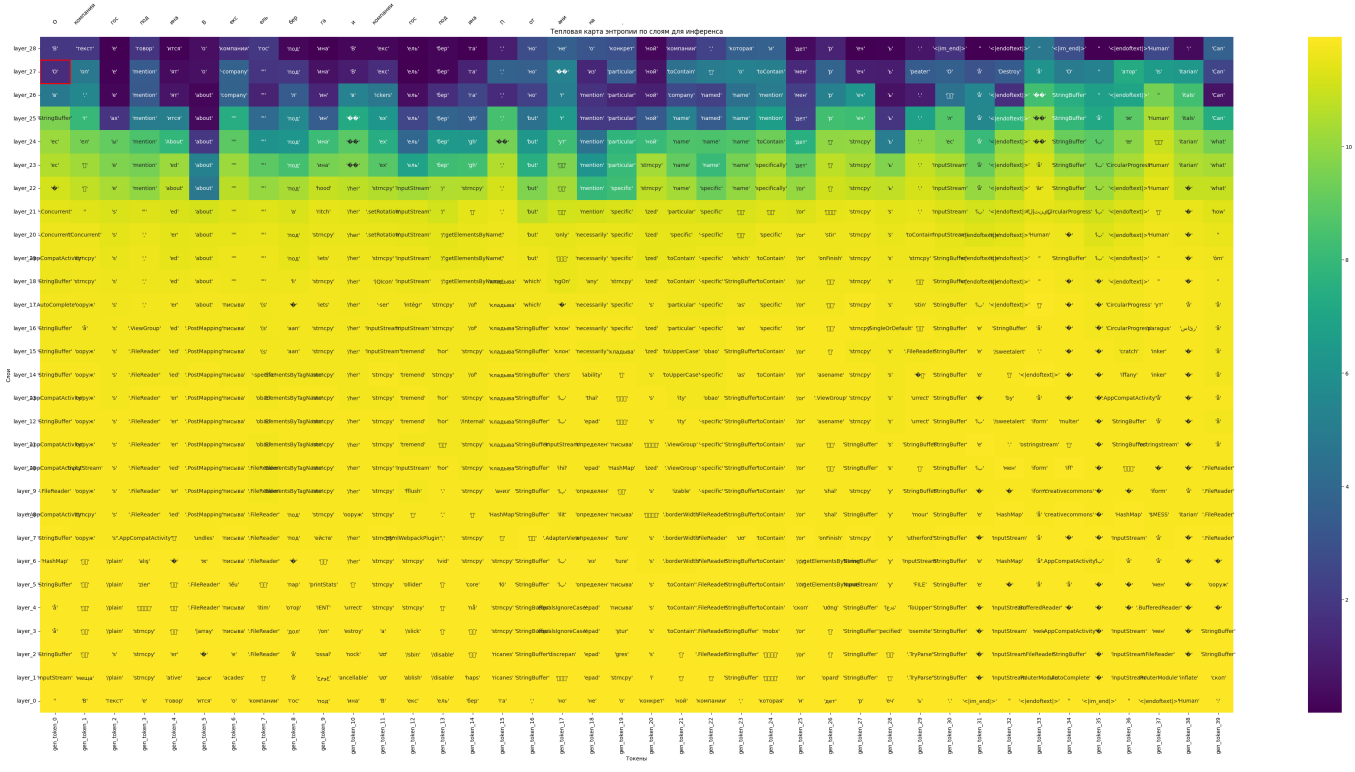

Рис. 4: Тепловая карта для изображения 2 датасета MTQVA (русский язык)

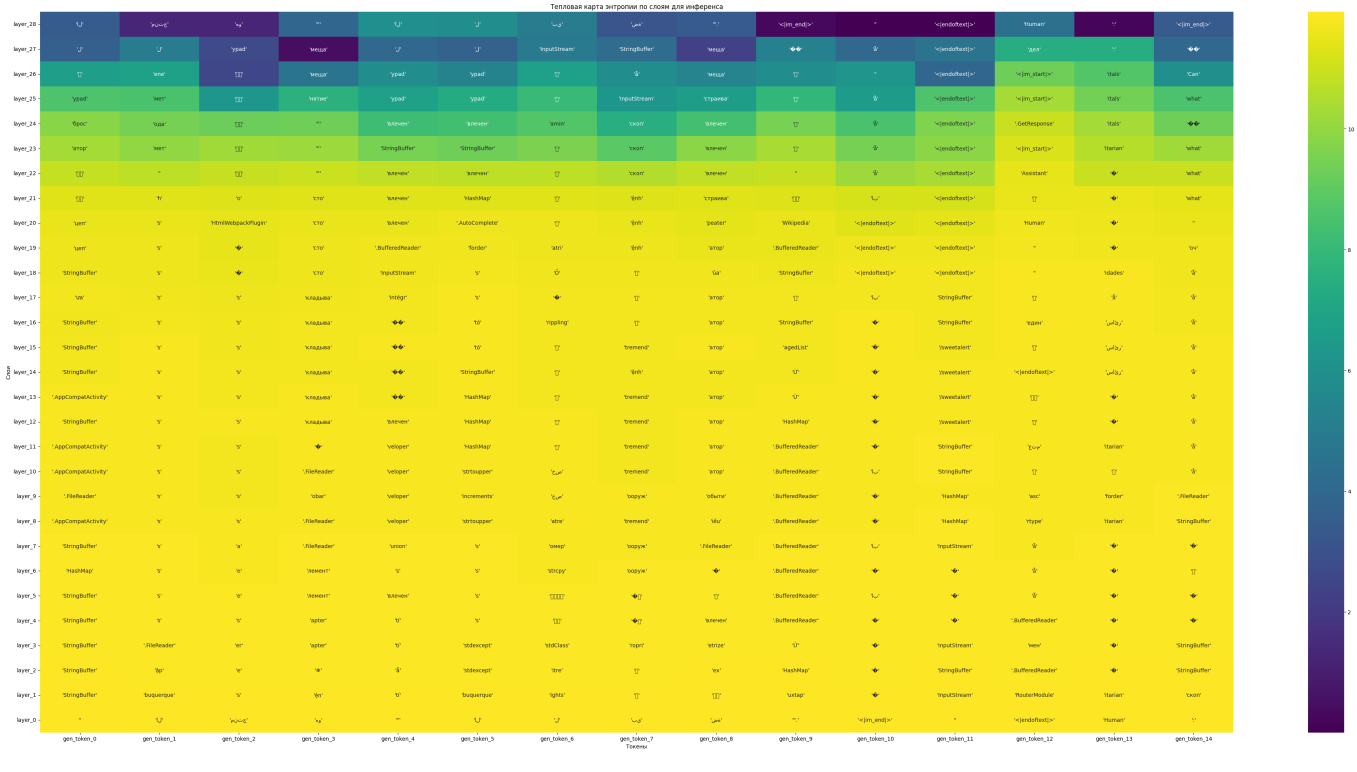

Рис. 5: Тепловая карта для изображения 3 датасета MTQVA (арабский язык)

Рис. 6: Зависимость усредненной по токенам энтропии от слоев ($layer_i$) для каждого изображения