Топологический анализ математических доказательств, генерируемых большими языковыми моделями на формальном языке Lean4

Студент: Линич А.А.
Научный руководитель: Баранников С.А.

Московский физико-технический институт

Москва,
2024 г.

Цель исследования

Разработка метода для оценки корректности и качества математических доказательств, генерируемых большими языковыми моделями на формальном языке Lean4, с использованием топологического анализа матриц внимания.

Задачи исследования

- Разработать способ применения топологического анализа данных для предсказания качества и корректности доказательств в Lean4.
- Выявить головы с разным поведением на правильных и неправильных доказательствах в Lean4.

Введение: Lean4

Lean4 основан на лямбда-исчисления. Это типовая теория, позволяющая выражать математические понятия и строить доказательства в формализованной форме. В последние годы происходит активная формализация математики с помощью Lean4 и других языков доказательств.

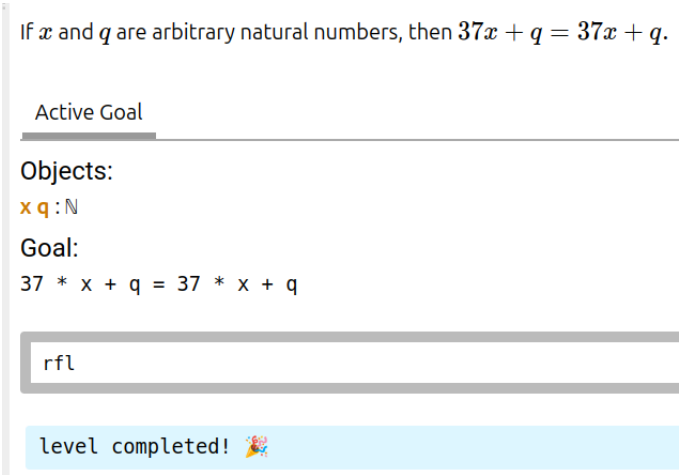

Рис.: Пример доказательства простого утверждения на Lean4

Введение: MTD(Manifold Topology Divergence)

MTD является методологией, основанной на Cross-Barcode, для сравнения облаков точек в высокоразмерном пространстве. Для вычисления MTD между облаками точек P, Q расстояния между всеми точками множества Q необходимо приравнять к нулю[1]:

$$Cross\text{-}Barcode_i(P, Q) = Barcode_i((P \cup Q)/Q).$$

Сэмплируя подмножества P_i, Q_i из P, Q, можно вычислить MTD по формуле:

$$MTD = \frac{1}{n} \sum_{i=1}^{n} mtd_i,$$

где mtd_i вычисляется как сумма длин Cross-Barcodes для i-ого подмножества.

Введение: MTD(Manifold Topology Divergence)

Пример подсчета MTD на синем и красном облаках точек.

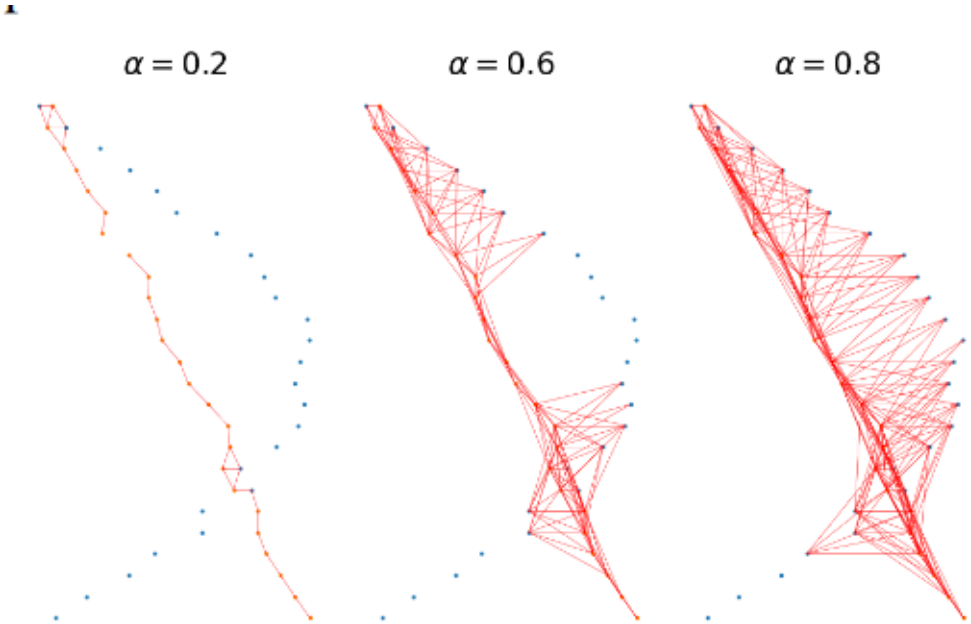

Рис.: Ребра (красные), соединяющие точки P (красные) с точками Q (синие), а также точки P между собой, добавлены для трех пороговых значений: $\alpha = 0.2, 0.4, 0.6$.

Способы примения MTD к задаче

1. Применение к графу, построенному по матрице внимания модели на слое L, голове Q.
2. Применение к векторизованным тактикам и состояниям доказательства.

1. Адаптация MTD к графу

Для применения MTD к графу, построенному на матрицах внимания, будем считать токены промпта точками множества Q, токены генерации – точками множества P в некотором пространстве с расстояниями между точками, заданными через матрицу внимания. Тогда значения матрицы внимания зануляются для токенов промпта. Кроме того, упростим подсчет MDT:

- будем подсчитывать сумму длин Cross-Barcodes только для нулевой гомологии,
- откажемся от сэмплирования подмножеств Q_i, P_i.

1. Шаги эксперимента, первый вариант

1. Подадим на вход модели Qwen-coder2.5 правильные и неправильные доказательства в Lean4.
2. Посчитаем упрощенный MTD для каждой головы на каждом слое.
3. Для каждой головы получим распределение MDT на верных и неверных доказательствах.
4. Оценим расстояние между получившимися распределениями для каждой головы.
5. Определим предсказательную силу MTD голов с наибольшим расстоянием из предыдущего пункта на тестовом датасете.

1. Шаги эксперимента, второй вариант

1. Подадим на вход модели Qwen-coder2.5 правильные и неправильные доказательства в Lean4 с просьбой определить, верно ли доказательство.
2. Посчитаем упрощенный MTD для каждой головы на каждом слое.
3. Для каждой головы получим распределение MDT на доказательствах с правильно и неправильно предсказанным типом.
4. Оценим расстояние между получившимися распределениями для каждой головы.
5. Определим предсказательную силу MTD голов с наибольшим расстоянием из предыдущего пункта на тестовом датасете.

1. Предварительные результаты.

Я использовала датасет из 50 правильных и неправильных доказательств, модель Qwen-coder2.5-1.5B. Расстояния оценивала как абсолютную разность выборочных средних.

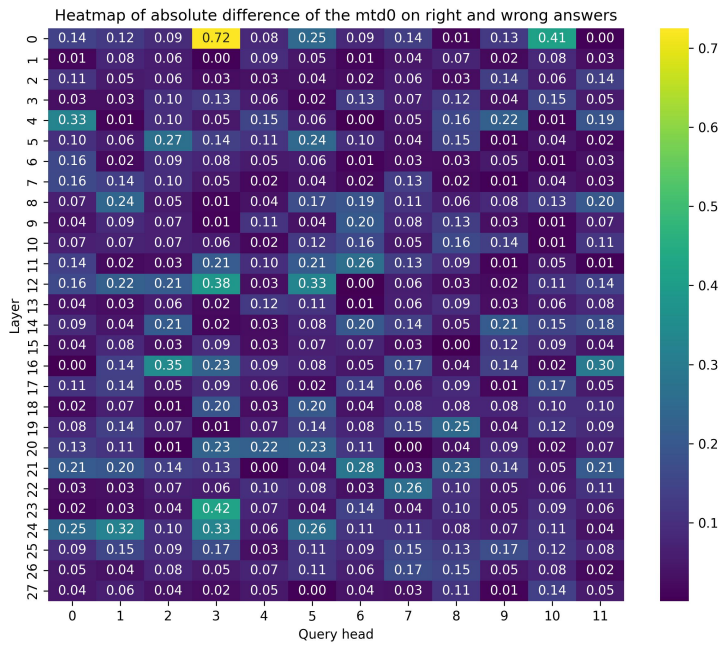

Рис.: Результаты эксперимена в первом варианте.

1. Предварительные результаты.

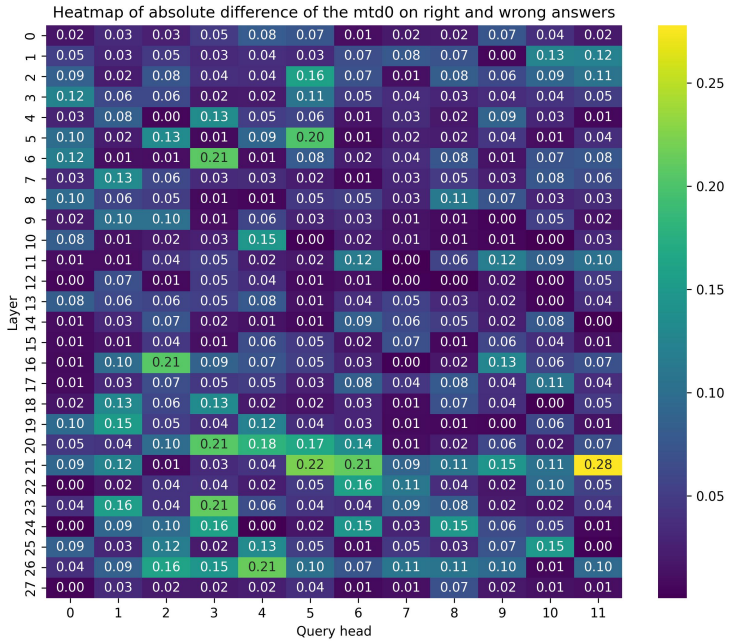

Рис.: Результаты эксперимена во втором варианте.

Предварительные выводы

Есть головы, для которых значение MTD различно на верных и неверных доказательствах Lean4, а также на доказательствах с правильно и неправильно определенным классом.

На 21ом слое чаще встречаются головы с различиями в MTD на доказательствах с правильно и неправильно определенным классом. Возможно, некоторые из этих голов "умнее"других.

Дальнейшие планы

1. Повторить проведенные эксперименты на большем датасете
2. Аккуратнее оценить расстояние между распределениями
3. Провести эксперимент с предсказанием правильности доказательств в Lean4 на тестовом датасете на основании MTD "лучших"голов
4. Применить метод MTD к облакам точек векторизованных состояний и тактик

Список литературы

1. Barannikov, Serguei, et al. "Manifold Topology Divergence: a Framework for Comparing Data Manifolds."Advances in neural information processing systems 34 (2021): 7294-7305.
2. Zelikman, Eric, et al. "Parsel: Algorithmic Reasoning with Language Models by Composing Decompositions."Advances in Neural Information Processing Systems 36 (2023): 31466-31523.
3. Taiwo, Funmilola Mary, Umar Islambekov, and Cuneyt Gurcan Akcora. "Explaining the Power of Topological Data Analysis in Graph Machine Learning."arXiv preprint arXiv:2401.04250 (2024).
4. Poesia, Gabriel, and Noah D. Goodman. "Peano: learning formal mathematical reasoning."Philosophical Transactions of the Royal Society A 381.2251 (2023): 20220044.
5. Lightman, Hunter, et al. "Let's verify step by step."arXiv preprint arXiv:2305.20050 (2023).