Огляд статті "Clinical Reading Comprehension with Encoder-Decoder Models Enhanced by Direct Preference Optimization"

Мазур Маркіян КІ-41мп

Тема та актуальність дослідження

Стаття присвячена задачі автоматичного знаходження відповідей на запитання у клінічних текстах, що є різновидом машинного читання з розумінням (Machine Reading Comprehension, MRC). Така проблема особливо актуальна в медицині, адже лікарі мають справу з великими обсягами текстових звітів і хотіли б швидко отримувати конкретні відповіді зі своїх нотаток замість переглядати десятки сторінок. Для оцінки моделей використовується датасет RadQA (Radiology Question Answering) – корпус із ~6148 пар запитань-відповідей на основі радіологічних звітів пацієнтів із MIMIC-III. Раніше для цієї задачі застосовувалися переважно моделі типу BERT (encoder-only), але автори досліджують переваги сучасніших великих encoder-decoder моделей (наприклад, сімейства T5) в поєднанні з методами оптимізації під уподобання користувача.

Основна ідея дослідження

Автори пропонують покращити точність клінічного QA, поєднавши потужну encoder-decoder модель T5 з методом Direct Preference Optimization (DPO). DPO – це підхід, запозичений з RLHF, що дозволяє безпосередньо оптимізувати модель на основі «бажаних» проти «небажаних» відповідей (первинно застосовувався для узгодження відповіді LLM з людськими вподобаннями). У роботі вперше продемонстровано застосування DPO до завдання читання з розумінням: замість ручного фідбеку згенеровано дані переваг (пари прикладів «правильна vs. неправильна відповідь») автоматично за допомогою евристик. Така стратегія дозволила суттєво підвищити ефективність моделі: загальний виграш станови.

Ключові методи

1. Supervised Fine-Tuning (SFT): Початкове донавчання базового трансформера T5 на тренувальному наборі RadQA для навчання моделі екстрактивного QA. Це створює базову модель, здатну знаходити відповідь-спан у документі за питанням.

2. Автоматична генерація даних переваг: Для методу DPO підготовано корпус пар запитань із прикладами *preferred* (правильна) та *rejected* (неправильна) відповідей без ручної розмітки. Використано два підходи: модельний, де напіводнана модель T5 сама згенерувала помилкові відповіді (і таким чином

“виявила” свої слабкі місця), що фіксуються як невдалі приклади; та правиловий, де застосовано набір евристичних правил для вибору неправильної відповіді з контексту (наприклад, випадковий фрагмент, що не містить тексту правильної відповіді, часткова відповідь або надто довга відповідь, відповідь на інше питання в тому ж документі тощо). У кожній парі запитання контексту gold-відповідь з RadQA слугує як *preferred* (еталон), а згенерований неправильний варіант – як *rejected*, утворюючи елемент даних переваг для навчання DPO.

3. Навчання з DPO: Після формування таких даних модель T5 додатково донавчається методом Direct Preference Optimization, який збільшує ймовірність правильної відповіді порівняно з відхиленою для кожного прикладу. На відміну від класичного RLHF, DPO не вимагає навчання окремої *reward model* (моделі винагороди для оцінки відповіді) – оптимізація відбувається напряму за допомогою стандартної функції втрат (binary cross-entropy) на парах «preferred vs. rejected». Таким чином досягається вирівнювання моделі під задані переваги без складних схем підкріплення.

Результати

1. Покращення точності vs. базові моделі: Усі варіанти T5 перевершили попередні BERT-моделі (BERT-MIMIC) на RadQA: показник F1 зріс приблизно на 12–15 відсоткових пунктів порівняно з найкращим попереднім результатом. Наприклад, модель Flan-T5-3B після звичайного SFT досягла ~76.4% F1 на тестовому наборі, що на ≈13 пунктів вище за найкращу BERT-базовану модель з попередніх досліджень.

2. Виграш від DPO: Донастройка моделей через DPO дала додатковий приріст ~1–3 пунктів F1 поверх supervised fine-tuning для більш потужних моделей (T5-3B, Flan-T5-3B). Найвищий результат отримано зв’язкою Flan-T5-3B + DPO – близько 77.5% F1 на тесті, що є новим рекордом для задачі RadQA. Для меншої моделі T5-large виграш від DPO був мінімальним, тоді як більші 3-мільярдні моделі змогли суттєво покращитися за рахунок оптимізації на складних прикладах.

Переваги та обмеження DPO

Метод DPO має кілька важливих переваг. По-перше, він спрощує налаштування моделі під потрібні відповіді: на відміну від RLHF, не потрібно навчати окрему модель винагороди для оцінювання виходу, що зменшує складність і обчислювальні витрати. По-друге, DPO явно використовує інформацію про негативні приклади: модель отримує сигнал не тільки *що генерувати*, але й *чого слід уникати*, навчаючись на відхилених варіантах відповіді. Це допомагає уникнути типових помилок і підвищує узгодженість вихідних відповідей із вимогами. Обмеження методу полягає в тому, що його

ефективність сильно залежить від якості набору *даних переваг*: для успішного навчання DPO потрібен репрезентативний корпус пар правильних/неправильних відповідей. У відсутності ручного аннотованого фідбеку це вимагає продуманої автоматичної генерації таких прикладів; до того ж, для кожної нової моделі бажано формувати свій набір «складних» негативних випадків під її слабкі місця (недолік модельно-орієнтованого підходу).

Висновок

Дослідження продемонструвало, що поєднання великих encoder-decoder моделей з оптимізацією переваг (DPO) дозволяє встановити новий рівень якості для задачі читання з розумінням у клінічній сфері. Зокрема, показано ~10% абсолютний виграш від переходу з BERT-подібних архітектур на T5, а також додаткові до +3% від застосування DPO поверх уже донавченої моделі, що в сумі дало ~12–15 пунктів F1 приросту на RadQA і оновило SoTA. Запропоновані методи автоматичного створення даних переваг та DPO-тренування є достатньо універсальними і можуть бути пристосовані до інших завдань інформаційного вилучення (наприклад, видобування сутностей чи відношень). Отримані результати відкривають можливості для створення більш ефективних систем question-answering на основі електронних медичних записів та впровадження підходу DPO в інших доменних застосуваннях.