RLHF и Direct Preference Optimization

Мурадян Денис Степанович

Предмет: «Вероятностные алгоритмы»
Санкт-Петербургский государственный университет
Математико-механический факультет
Искусственный интеллект и наука о данных
Бакалавриат, 3 курс

2025

Введение: от NTP к SFT и задаче выравнивания

- **Предобучение (NTP):** модель обучается предсказывать следующий токен

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t})$$

- **Ограничение:** правдоподобный текст не всегда соответствует ожиданиям человека
- **SFT:** дообучение по парам «запрос — эталонный ответ»

$$\max_\theta \; \mathbb{E}_{(x,y^*)} \big[\log \pi_\theta(y^* \mid x) \big]$$

- **Интуитивная цель:** сдвинуть распределение ответов модели к желаемому поведению

Актуальность: от эталонных ответов к предпочтениям

- После **SFT** остаётся множество допустимых, но неравноценных ответов
- Качество ответа часто задаётся **относительно**, а не абсолютно
- Цель: изменить распределение $\pi_\theta(y \mid x)$ так, чтобы предпочтительные ответы становились более вероятными
- **RLHF (Reinforcement Learning from Human Feedback)**

Общая идея RLHF: место в пайплайне обучения

1. **Pretraining (NTP):** обучение на больших текстовых корпусах
2. **SFT:** формирование reference-политики π_{ref}
3. **Preference data:** тройки (x, y_w, y_l)
   - **RLHF (PPO):** генерации (x, y), оценки reward model $r(x, y)$
   - **DPO:** фиксированные пары предпочтений $(x, y_w in, y_l ose)$
4. **Alignment:** RLHF (reward model + PPO) **или** DPO

Здесь: x — запрос (prompt), y — ответ модели.

Вероятностная модель LLM как политика π_θ

Определение. $\pi_\theta(y \mid x)$ — политика (policy): распределение вероятностей ответов y при условии запроса x.

Автогрессивная факторизация:

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t})$$

Лог-вероятность ответа:

$$\log \pi_\theta(y \mid x) = \sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t})$$

RLHF: оптимизация ожидаемой награды с контролем отклонения

Цель RLHF:

$$\max_{\theta} \; \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \big[r(x, y) \big] \; - \; \beta \, \mathrm{KL}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x))$$

- $r(x, y)$ — функция вознаграждения (reward model)
- π_{ref} — reference-политика после SFT
- KL — мера отклонения распределений
- $\beta > 0$ — коэффициент регуляризации

Почему не всегда PPO: мотивация Direct Preference Optimization

- **RLHF** через **PPO** решает задачу выравнивания, но требует сложного RL-цикла
- Необходимы:
  - **reward model** для оценки качества ответов
  - **critic** для оценки ожидаемой награды
  - **on-policy генерация** и стабилизация обновлений (clipping)
- Инженерная и вычислительная сложность обучения
- **Идея DPO**: решить ту же задачу alignment без reward model и RL-цикла

Direct Preference Optimization: обучение через предпочтения

- Обучение проводится на фиксированном датасете предпочтений
- Формат данных: (x, y_w, y_l)
  - y_w — предпочтительный (winner) ответ
  - y_l — менее предпочтительный (loser) ответ
- Цель: сделать y_w более вероятным, чем y_l, при том же x
- Reward model и RL-цикл не требуются

Пример данных:

```
{ "prompt": "Как приготовить кофе?
"chosen": "Возьмите кофе, налейте горячую воду, перемешайте.
"rejected": "Кофе - это вкусный напиток, он бывает разных видов."}
```

Direct Preference Optimization: вероятностная формулировка

Логит предпочтения:

$$s_\theta = \beta \left(\log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right)$$

Функция потерь:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma(s_\theta)$$

- π_{ref} — reference-политика после SFT
- $\beta > 0$ — коэффициент силы предпочтений

Примеры использования RLHF и DPO

Alignment чат-моделей

- повышение полезности и связности ответов
- контроль стиля, тона и безопасности

Доменные ассистенты

- поддержка пользователей и FAQ-системы
- образовательные и аналитические ассистенты

Практический пайплайн

- SFT $\rightarrow$ сбор предпочтений $\rightarrow$ DPO
- обучение без reward model и RL-цикла

Индустриальный пример

- **Яндекс Нейро**: alignment языковой модели с использованием DPO

Выводы

- LLM задаёт распределение ответов $\pi_\theta(y \mid x)$
- Alignment — это целенаправленный сдвиг этого распределения
- RLHF формулируется как оптимизация ожидаемой награды с KL-контролем
- PPO реализует RLHF, но требует сложного RL-цикла
- DPO напрямую оптимизирует предпочтения и упрощает обучение

RLHF и Direct Preference Optimization

Мурадян Денис Степанович

Предмет: «Вероятностные алгоритмы»
Санкт-Петербургский государственный университет
Математико-механический факультет
Искусственный интеллект и наука о данных
Бакалавриат, 3 курс

2025