Алайнмент моделей: REINFORCE и Reward Model

Семен Безъязычный

Июнь 2025

1 Level 1: REINFORCE с reward model

Техническая часть

- SFT-модель: `HuggingFaceTB/SmolLM2-135M-Instruct` (на основе LLaMA, 135M параметров)
- Токенизатор: тот же, добавлен pad-токен `<eos>`, установлены `padding_side = left`, `truncation_side = left`
- Reward Model: SFT + `SequenceClassification`
- Алгоритм: REINFORCE с EMA-baseline

Данные Используется датасет `juyoungml/HelpSteer2-binarized` с полями `prompt`, `chosen`, `rejected`. Train size: 7224; validation size: 373.

Reward model Обучение проводится с LoRA ($r = 16$, $\alpha = 32$, dropout=0.05), максимальная длина 512, одна эпоха, learning rate $5 \cdot 10^{-5}$, оптимизатор AdamW.

Функция потерь:

$$\mathcal{L} = -\log \sigma(s_{\text{chosen}} - s_{\text{rejected}})$$

После одной эпохи mean loss составил 0.682.

REINFORCE. Параметры: 300 итераций, макс. длина ответа — 128 токенов, learning rate 10^{-5}, EMA-бейзлайн с $\alpha = 0.9$, clip_grad_norm = 1.0, сэмплирование с `top_p = 0.9`, `do_sample = True`.

Используется скалярная награда от RM. Advantage вычисляется как $A = r - \hat{b}$, где $\hat{b}$ — скользящее среднее.

Результаты. Для примера взятого из тренировочной подвыборки результаты:

Reward for chosen: 0.15

Reward for rejected: -0.13

То есть модель в positive-negative расставила оценки согласно тому как было задумано: выбранному тексту (chosen большая оценка)

Для промпта p = "What is RLHF?"reward модель выводит:

chosen = p + "Reinforcement Learning from Human Feedback is a method of training LLMs..."

rejected = p + "I don't know."
Reward for chosen: -0.6348
Reward for rejected: -2.4824

То есть осмысленному ответу chosen модель присвоила большую оценку, в отличие от ответа 'не знаю', что согласуется с логикой построения модели. То есть модель работает корректно.

После **REINFORCE**:
для 512 промтов
Avg reward SFT: -0.0537
Avg reward RL: -0.0366
delta improvement: +0.0171
то есть небольшое улучшение reinforcement model в среднем даёт.

Пример ответа базовой модели и модели RL:

test_prompt = "Question: Why is the sky blue? :"

SFT reply: The sky is blue because of a combination of the blue light that is emitted by the Earth's oceans and atmosphere, and the scattering of light by the air. As light passes through the atmosphere, it is scattered in all directions in every direction. However, there are certain parts of the sky that reflect more blue light than others, so the blue light from the oceans and atmosphere reaches the surface of the Earth in a wider range of directions than the blue light from the air. As a result, the blue light is scattered, and the sky appears blue. Answer:

RL reply: The sky appears blue because it contains oxygen, which is a gas that gives the sky its blue color. Oxygen is a byproduct of the combustion of fuels like fossil fuels, and the sun emits ultraviolet radiation that is not absorbed by the atmosphere. This ultraviolet radiation causes the color of the sky to shift, causing it to appear blue.

Модель RL ответила абсолютно некорректно, модель SFT содержит указание на рассеяние света. SFT модель дала более точный ответ.

Почему не RLOOTrainer. По рекомендации статьи *Back to Basics* используется EMA-baseline, поскольку:

- лишние вычисления и расход памяти: RLOO требует двух проходов модели по каждому примеру (основной + baseline), тогда как EMA-baseline — один.
- даёт результат, сопоставимый с RLOO.

Вывод. EMA-baseline требует меньше вычислений в отличие от RLOO (квадратичная сложность), даёт сопоставимое качество.

2 Level 2: Вероятностная Reward Model и REINFORCE

Цель. Модифицировать RM так, чтобы она выдавала распределение вероятностей по целым оценкам от 1 до 10, а не скаляр.

Формализация. Пусть p_c, p_r — вероятностные оценки для `chosen` и `rejected`. Вектор классов: $(1, 2, \ldots, 10)$.

Ожидаемая оценка:

$$E[p] = \sum_{i=1}^{10} i \cdot p_i$$

Функция потерь:

$$\mathcal{L} = -\log \sigma \left(E[p_c] - E[p_r] \right)$$

Параметры обучения. Модель: SmolLM2-135M-Instruct, 10 выходов, learning rate $5 \cdot 10^{-5}$, 3 эпохи, обучение в `float16` с AMP. Финальный loss ~ 0.52.

Интеграция в REINFORCE.

- $r = E[p]$ — непрерывная награда, быстрая сходимость.
- $r \sim p$ — дискретная награда, ближе к реальной, но более шумная.

Результаты. Примеры двух промптов: **Prompt:** Why is the sky blue?:

SFT: 'the sky blue in the movie Skyfall?' | Reward: 1.3632

RL : '1. **Blue light:** This is due to the blue range of wavelengths in the visible spectrum, which is more accessible and easier to perceive for humans than the red and orange wavelengths. Blue light is often emitted by the sun, stars, and other celestial objects.2. **Spectacular:** Sky blue is often associated with the vastness of the universe and the sky itself, creating a sense of awe and wonder. Sky blue also evokes a sense of harmony and balance, as the colors of the sky and the surrounding landscape seem to be in balance.that the sky blue is subjective and can depend on personal' | Reward: 3.0625

delta improvement : $+1.6993$

RL модель дала осмысленный ответ и получила значительно большую награду, что согласуется с логикой

Вывод. Вероятностная Reward-модель предоставляет больше информации, даёт гладкий градиент, а использование мат. ожидания $E[p]$ повышает стабильность.