به نام خدا

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

مدل‌های زبانی بزرگ

پاییز ۱۴۰۲

استاد: دکتر سلیمانی، دکتر رهبان، دکتر عسگری
گردآورندگان: فریدون مهری، محمد علی صدرایی جواهری، علی رازقندی، مهدی زکی‌زاده
بررسی و بازبینی: محمد علی صدرایی جواهری

تمرین چهارم — RLHF-DPO-LLM Decoding and Exploration — مهلت ارسال: ۱۵ دی

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است. برای انجام تمرین زمان کافی اختصاص داده شده است. انجام آن را به هیچ‌وجه به روزهای پایانی موکول نکنید.
- سوالات خود را فقط از طریق کوئرا درس و در نوشته‌ی مربوط به اطلاع‌رسانی این تمرین بپرسید.
- حتما در نام‌گذاری فایل‌های آپلودی خود از قالب $\{STD\_Number\}\_\{Name\}$ تبعیت کنید.
- در طول ترم ۵ روز تاخیرمجاز برای ارسال تکالیف دارید. پیشنهاد می‌شود تاخیرهای خود را برای مواقع ضروری نگه دارید.
- پاسخ‌های ارسالی باید منحصرا حاصل تلاش فردی شما باشد. در صورت استفاده از منابع خارجی یا همفکری، حتما این موارد را ذکر کنید. همچنین توصیه می‌شود آداب نامه‌ی انجام تمرین‌های درسی را مطالعه کنید. برای اطلاع از قوانین خاص این درس به فایل قوانین درس بر روی کوئرا مراجعه کنید.

توضیحات (۱۰۰ نمره)

سوال اول

آ) توضیح دهید که در روند آموزش مدل امتیازدهی [۱] چگونه تابع هزینه [۲] زیر می‌تواند از مشکل اختلاف امتیاز بسیار زیاد بین پاسخ ها جلوگیری کند و چرا این امر مفید است.

$$loss(\theta) = -\frac{1}{\binom{K}{2}}\mathbf{E}_{(x,y_w,y_l)\sim D}[log(\sigma(r_\theta(x,y_w) - r_\theta(x,y_l)))]$$

ب) می‌دانیم در نهایت تابع هدف [۳] یادگیری مدل به شکل زیر است.

$$objective(\phi) = \mathbf{E}_{(x,y)\sim D_{\pi_\phi^{RL}}}[r_\theta(x,y) - \beta log(\pi_\phi^{RL}(y|x)/\pi^{SFT}(y|x))] + \gamma \mathbf{E}_{x\sim D_{pretrain}}[log(\pi_\phi^{RL}(x))]$$

توضیح دهید هر یک از جملات در این تابع چه هدفی را دنبال می‌کنند و چرا در تابع ظاهر شده‌اند.

ج) در تابع هدف قسمت قبل مشتق گیری بر حسب ϕ است اما با این حال جمله $r_\theta(x,y)$ در تابع وجود دارد توضیح دهید چرا با وجود اینکه r تابعی از θ است اما با این حال در تابع هدف قرار دارد و مشتق آن صفر نیست.

د) حال روشی را معرفی کنید که بتوان مشتق تابع مورد نظر را به راحتی پیاده سازی کنیم. (راهنمایی: برای سادگی فرض کنید تابع مورد نظر به شکل $\mathcal{L}_\theta = \mathbb{E}_{\pi_\theta}[G_t]$ و نشان دهید مشتق این تابع برابر $\mathbb{E}_{\pi_\theta,\tau}[\nabla_\theta log\pi_\theta(\tau)G_\tau]$)

[۱] reward model
[۲] loss function
[۳] objective function

ه) جمله دیگری که در تابع هدف وجود دارد برابر $log(\pi_\phi^{RL}(y|x)/\pi^{SFT}(y|x))]$ است که KLD بین دو توزیع پالیسی اولیه و پالیسی در حال یادگیری است. در ابتدا توضیح دهید چرا نمی‌توان این جمله را به صورت مستقیم محاسبه کرد و سپس نشان دهید تخمینگر

$$\mathbb{D}_{KL}(q \parallel p) \approx \frac{۱}{N}\sum_{i=۱}^{N}\frac{۱}{۲}[log(q(x_i)) - log(p(x_i))]^{۲}, x_i \sim q(x)$$

تخمینگر مناسبی با واریانس کمتر نسبت به تخمینگر ساده زیر است.

$$\mathbb{D}_{KL}(q \parallel p) \approx \frac{۱}{N}\sum_{i=۱}^{N}[log(q(x_i)/p(x_i))], x_i \sim q(x)$$

و) با اینکه تخمینگر معرفی شده در قسمت قبل دارای واریانس کمتری نسبت به حالت ساده است اما این تخمینگر دارای بایاس است. حال با تغییر در این تخمینگر بایاس آن را نیز کم کنید.

سوال دوم

در این سوال سعی در رسیدن به تابع هدف الگوریتم DPO داریم.

آ) نشان دهید جواب تابع زیر

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}[r_\phi(x,y)] - \beta\mathbb{D}_{KL}[\pi_\theta(y|x) \parallel \pi_{ref}(y|x)]$$

برابر با

$$\pi_r(y|x) = \frac{۱}{Z(x)}\pi_{ref}(y|x)\exp(\frac{۱}{\beta}r(x,y))$$

است.

ب) حال با استفاده از نتیجه قسمت قبل و جایگذاری آن در تابع هزینه مدل امتیازدهی به تابع هزینه DPO برسید.

ج) حال که تابع هزینه DPO را بدست آوردید گرادیان آن را محاسبه کنید و توضیح دهید هر جمله در گرادیان به شکل شهودی چه هدفی را دنبال می‌کنند.

Practical Exercises

Included with this exercise are four notebooks. Please complete the tasks as directed in each notebook and attach the corresponding files upon completion.

Grading Guidelines

The total grading points are from $60 + 40 = 100$, where there are 40 bonus points. The breakdown is as follows:

Theory: 35 points

Practical: 65 points

Part 1 Analyzing Sycophancy in Large Language Models (LLMs)

- 20 points

Part 2 Loss Visualization

- 10 points

Part 3 Generative Model Simple Decoding - Translation - Visualization

- 20 points

Part 4 Advanced Generation Decoding

- 15 points