Q&A

Dashygo

September 11, 2024

Pre-training

Q1

- 训练时有一个参数max_length，它是做什么的
- 在真正开始训练时有一个warm up，它是用来做什么的

A1

max_length: In pre-training stage, max_length denotes the maximum length of input token sequences.If length of seqs are less than max_length,then they will be padded into max_length.If length of seqs are more than max_length,then they will be cut off into max_length.(Well actually these can be chosen sometimes as options)

warm up: In the paper of ResNet,there is a warmup epoch with $lr = 0.01$ on the training of 110 layers ResNet on training set cifar10,different from the following epochs with $lr = 0.1$.This increases the stability of neural networks,especially those deep neural networks,for there will be risks of bringing severe oscillations to models if we just do epochs with large lr on a newborn network whose weights are randomly generated.So as the LLMs.

Post-training

Q2.1

- 什么是instruction tuning？为什么需要instruction tuning?
- Llama3 的instruction tuning的格式是怎么样的？

A2.1

instruction tuning: Instruction tuning is as what its name denotes,model is given a descriptive instruction to find out what it should do on most of times, multi-tasks datasets for tuning.These instructions might be, for example, "Referring to the passage above,answer the question:xxx", "Give some examples on instruction tuning" or "You know what I'm saying, print it out."

Take prompt tuning for comparison,instruction tuning provides a better method to instruct the model to perform better on zero-shot of multi-tasks or something other tasks that the model has even not known before.So instruction tuning is a tuning that teaches the model to learn about "learning" itself to some degrees.

Llama3: format of instruction tuning: We can easily find out how these special tokens work in *tokenizer.py* and *test_tokenizer.py* on the official Meta Llama3 GitHub site.

Algorithm 1: Dialog Example

```
1 <|begin_of_text|><|start_header_id|>system<|end_header_id|>
2 This is a test sentence. <|eot_id|> <|start_header_id|> user
    <|end_header_id|>
3 This is a response. <|eot_id|> <|start_header_id|> assistant
    <|end_header_id|>
```

Q2.2 —— SFT

- 为什么要将llm与人类偏好对齐？不这么做会出现什么问题？

A2.2

SFT(Supervised Fine-Tuning): Unsupervised LLMs learn from a large scale of unsupervised dataset which makes it hard to control how well it can present suitable information with high-quality Q&A,provide expert and beautiful codings,or prevent it from giving some dangerous statements with abuse and discrimination.So these are exactly what SFT is aimed at.

Q2.3/Q2.4 —— RLHF/PPO/DPO

- rlhf的偏好数据集是如何构造的？
- reward model是做什么的？它是如何被训练的？
- DPO和PPO相比优势在哪里？请你详细阐述一下

1 RL: A2.3/A2.4

I am somewhat interested in RL so we take a quick glance at it.

1.1 Policy Gradient:

RL(Reinforcement Learning) is a paradigm for learning strategies through feedback.We aim to describe it in mathematics.In *Agent* $\longleftrightarrow$ *Environment* case , given state of timestep t,denoting s_t,agent takes action a_t and get reward r_t form the environment, before getting into state s_{t+1}.While in this process , agent accumulates experience to fit itself to get the most of reward by adjusting strategies.Denote $\pi_\theta(a_t|s_t;\theta)$ or simply $\pi_\theta(a_t|s_t)$ the agent actions distribution on timestep t.So our target function is :

$$maximum\ \ J(\theta) = \widehat{\mathbb{E}}_{\tau\sim p_\theta(\tau)}\left[\sum_i r(s_i, a_i)\right]$$

in which τ is called *trajectory* denotes stochastic process of

$$\tau : s_1 \to \langle a_1, r_1\rangle \to s_2 \to ... \to s_t \to\ \langle a_t, r_t\rangle$$

while $p_\theta(\tau) = p_\theta\left(s_1, a_1, \ldots, s_T, a_T\right) = p_\theta\left(s_1\right)\prod_{t=1}^{T}\pi_\theta\left(a_t \mid s_t\right)p\left(s_{t+1} \mid s_t, a_t\right)$

We want to apply GD or SGD to solve this optimization problem,so we introduce *likelihood ratio* as score function to present $\bigtriangledown_\theta J(\theta)$.

In policy-based cases, we take

$$J^{PG}(\theta) = \widehat{\mathbb{E}}\left[log\, \pi_\theta(a_t|s_t)\widehat{A}_t\right]$$

where $\widehat{A}_t$ is *is an estimator of the advantage function at timestep t* as PPO paper notes.For example, in one-step MDP, $\widehat{A}_t$ is exactly the reward of a single step.While in multiple steps, this can also represent the Monte-Calro process reward.But as PPO papers notes, that*...doing so is not well-justified, and empirically it often leads to destructively large policy updates.*So in most times we just take a single step.

1.2 RLHF(Reinforcement Learning from Human Feedback)

Well , there are too many details to tell along RL,so let's get our problems first.I may write it later if i have time....

A2.3

Datasets: In DPO paper, we see $(y_1, y_2) \sim \pi^{SFT}(y|x)$ in preference sampling phase, with $y_w|x \succ y_l|x$ in which x is the prompt and y_w,y_l denote preferred and dispreferred completion respectively.And for model preference, there is a popular model named BT(Bradley-Terry) for stipulating the preference distribution p^* as the form of softmax oprator:

$$p^*(y_w \succ y_l|x) = \frac{exp(r^*(y_w|x))}{exp(r^*(y_w|x) + exp(r^*(y_l|x)}$$

where $r^*(y|x)$ is fenerated by reward of model.Therefore we have dataset$\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^n$ sampled from p^*

RM(Reward Model): Alike DPO paper,denote the params of reward model as θ, so we have $r_\theta(y|x)$ for our reward model. We use likelihood method to estimate the params θ from the loss function(negative log-likelihood) to be optimized.That is:

$$\mathcal{L}_R(r_\theta, \mathcal{D}) = -\widehat{\mathbb{E}}_{(x,y_w,y_l)\sim\mathcal{D}}\left[log\ \ \sigma(r_\theta(x, y_w) - r_\theta(x, y_l))\right]$$

where σ is sigmoid.

This formula means minimizing the loss to distinguish between y_w and y_l,and this is what reward model is designed for.We apply some basic optimizers like GD,SGD,Adam etc to solve this problem.

A2.4

DPO/PPO: In the paper, researchers give the evaluations on IMDB set and TL;DR Summarization Win Rate vs Reference, from which we can see that,in the first case, DPO stands out on the approach to the highest reward and performs well in low KL values(KL value stands for the consistence between trained and referred distribution), while in the second case,DPO show its high robustness on summerization with temperature changed(temperature stands for the significance of output,take the softmax for example,we often add a coefficient inside the softmax, if the temperature is high then the differnce between two numbers are significantly greater.)

What'more DPO(Direct Preference Optimization) means this algorithm is more lightweight while in the same time simpler to train.