Training-Inference Mismatch In LLM KD

Alephia 25/6/25

BACKGROUND

模型q_θ依据前缀w^{t-1}生成文本的时候，loss可以表示为

$$l(q_\theta, w^{t-1}; o) = \mathop{\mathbb{E}}_{w_t \sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}$$

由目标分布o采样下一个token w_t，再进行KLD对齐。

可以展开得到

$$L(q_\theta; o) \approx \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1} \sim d_o^{t-1},\, w_t \sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}$$

TRAIN-INFERENCE MISMATCH

由于模型能力不足，生成token的分布与目标分布存在差距，进而模型训练和推理时面对的前缀是不同的

- Distribution Mismatch (Exposure Bias)
- Error Accumulation

Distribution Mismatch会导致Error Accumulation，且Accumulation行为无法被已有的loss捕捉

Arora, K., Asri, L.E., Bahuleyan, H., & Cheung, J.C. Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation. In ACL, 22

TRAIN-INFERENCE MISMATCH

模型与目标的总偏差表示为

$$
\begin{aligned}
L(q_\theta; o) &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1}\sim d_o^{t-1},\, w_t\sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})} \\
&= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1}\sim d_o^{t-1}} D_{KL}(o(\cdot|w^{t-1})||q_\theta(\cdot|w^{t-1}))
\end{aligned}
$$

记生成第t个token的期望误差为

$$
\epsilon_t = \mathop{\mathbb{E}}_{w_0^{t-1}\sim d_o^t,\, w_t\sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}
$$

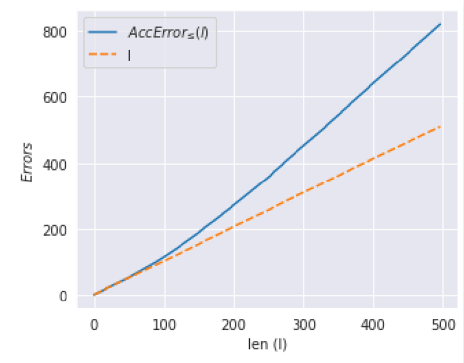
Arora, K., Asri, L.E., Bahuleyan, H., & Cheung, J.C. Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation. In ACL, 22

TRAIN-INFERENCE MISMATCH

$$l\epsilon_{\leq l} \leq L_{\leq l}(q_\theta) \leq l^2\epsilon_{\leq l}, \;\; \epsilon_{\leq l} = \frac{1}{l}\sum_{t=1}^{l}\epsilon_t$$

$$AccErr_{\leq}(l) = \frac{L_{\leq l}(q_\theta)}{\epsilon_{\leq l}}$$

如果Distribution Mismatch确实会导致误差累积的话，应当观察到AccErr值是随着序列长度增加而超线性增长的。

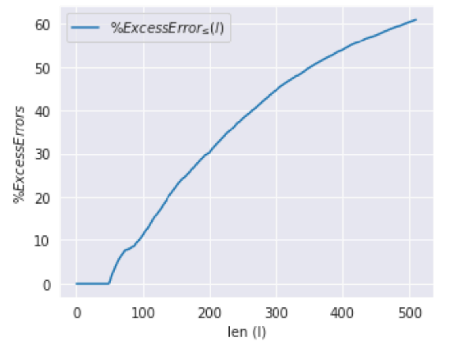


(a) $AccErr_{\leq}(l)$ vs l

TRAIN-INFERENCE MISMATCH

$$ExAccErr_{\leq}(l) = \frac{L_{\leq l}(q_\theta) - l\epsilon_{\leq l}}{l\epsilon_{\leq l}} \cdot 100$$

如果一个模型能够做到每一个的损失不会累积的话，那么这个值应当一直在0左右，否则，就会呈不断上升的趋势。



(b) %ExError≤(l) vs l.

TRAIN-INFERENCE MISMATCH

$$
\begin{aligned}
\epsilon &= \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{w_0^{t-1}\sim d_o^t \; w_t\sim o(\cdot|w_0^{t-1})} \log \frac{o(w_t|w_0^{t-1})}{q_\theta(w_t|w_0^{t-1})} \\
&\approx -\frac{1}{|D|}\sum_{(w_0^{i-1},w_i)\in D} \log q_\theta(w_i|w_0^{i-1}) + c \\
&= H(q_\theta; D) + c'
\end{aligned}
$$

这里$H(q_\theta; D)$代表log Perplexity

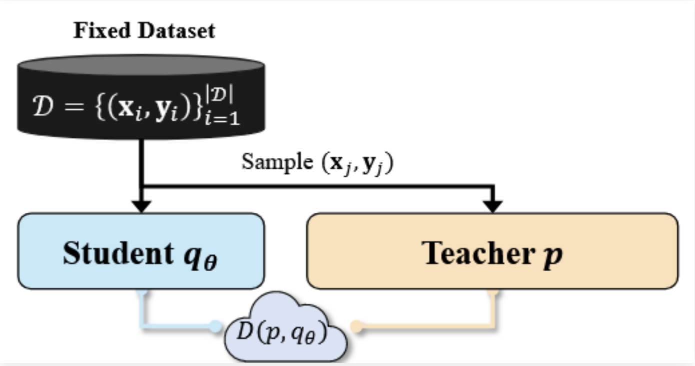
不管是CE Loss，还是Perplexity，都无法监督error的累加过程

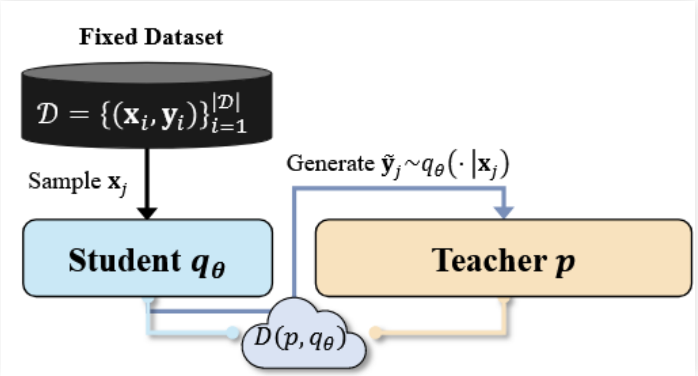
THE UTILIZATION OF SGO

引入模型自己推理生成的内容用于训练(**S**tudent **G**enerated **O**utput)

$$L_{SGO}(q_\theta; o) = \sum_{t=1}^{T} \underset{w^{t-1} \sim d_{q_\theta}^{t-1},\, w_t \sim o(\cdot|w^{t-1})}{\mathbb{E}} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}$$

每次有λ的概率使用SGO，$1-\lambda$的概率使用训练集样本

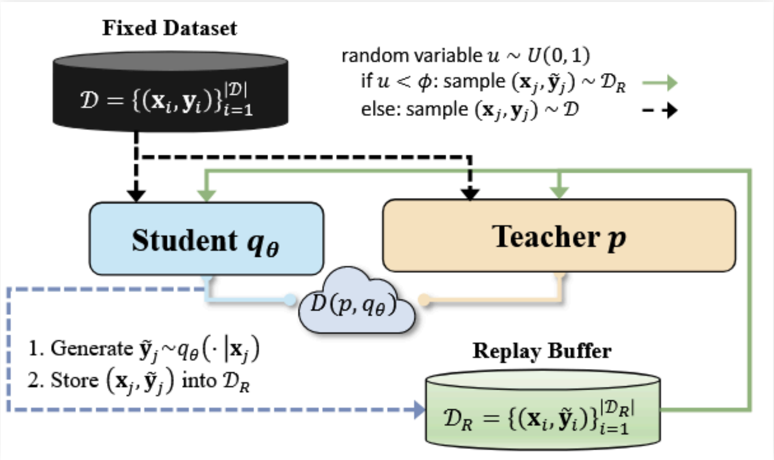




Agarwal, R., Vieillard, N., Zhou, Y., Stańczyk, P., Ramos, S., Geist, M., & Bachem, O. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In ICLR, 24

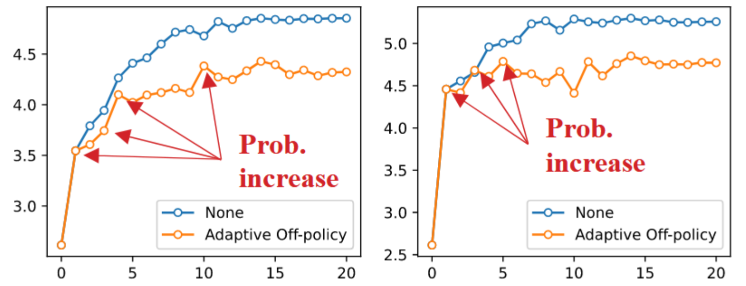
THE UTILIZATION OF SGO

- 使用SGO时，Teacher也面临Train-Inference Mismatch，会带来噪声
    - 💡 更加保守地使用SGO
- 每次都要学生重新生成SGO，利用率低，计算开销大
    - 💡 on-policy -> off-policy



Ko, J., Kim, S., Chen, T., & Yun, S. DistiLLM: Towards Streamlined Distillation for Large Language Models. In ICML, 24

THE UTILIZATION OF SGO



(a) SKL (b) SRKL

Figure 10. Plot of validation loss values (y-axis) across each validation iteration (x-axis). Although validation loss tends to increase as training progresses, employing SGO effectively prevents this increase. This is the core philosophy of our adaptive SGO scheduler (orange line).

Loss↑，ROUGE_L↑

"Our observations indicate that training on a diverse range of SGOs, rather than solely on a fixed dataset, mitigates training-inference mismatch and consequently lowers validation loss"

过拟合？train-inference mismatch？ 指标与loss的不匹配？

INTRODUCE SAMPLE-WISE WEIGHT

P为真实分布，Q为合成数据分布，q_θ为模型预测分布

$$E_Q[-\log q_\theta(y|x;\theta)]$$

$$E_Q\left[-\frac{P(y|x)}{Q(y|x)}\log q_\theta(y|x;\theta)\right] = E_P[-\log q_\theta(y|x;\theta)]$$

$P(y|x)$大，数据点与真实分布高度相关且明确，有参考意义。
$Q(y|x)$越小，数据点在分布Q中所包含的信息越多。

Kuo, H., Liao, Y., Chao, Y., Ma, W., & Cheng, P. Not All LLM-Generated Data Are Equal: Rethinking Data Weighting in Text Classification. In ICLR, 25

INTRODUCE SAMPLE-WISE WEIGHT

加权策略更新为

$$E_Q\left[-\frac{q_\theta(y|x;\theta,D_{P'})}{q_\theta(y|x;\theta)}\log q_\theta(y|x;\theta)\right]$$

$q_\theta(y|x)$越小，模型在这个数据点上学的越差，越应注重

Dataset	Method	Financial Acc	Financial F1	Tweet Irony Acc	Tweet Irony F1	MRPC Acc	MRPC F1
	GPT-3.5 few-shot	79.46	81.6	63.39	69.39	69.28	71.75
Small real world	CE-Loss (quality checker)	78.05	75.26	62.5	62.38	73.16	68.69
	Focal-Loss	78.47	76.2	67.73	62.32	73.10	66.64
	DIMP-Loss (Ours)	**79.87**	**77.05**	**69.01**	**67.05**	**74.84**	**66.80**
GPT-3.5 generated	CE-Loss	77.39	74.01	76.91	76.8	72	65.47
	Focal-Loss	79.29	75.32	74.87	74.82	72.17	62.77
	Hu et al.'s	71.7	61.93	71.42	70.18	67.13	50.08
	SunGen	80.45	76.87	78.96	75.06	71.65	66.08
	IMP-Loss (Ours)	82.09	**79.40**	**81.89**	**81.71**	**75.83**	**70.52**
	DIMP-Loss (Ours)	**82.67**	**79.53**	78.44	78.14	**75.83**	**70.04**
	- w/o diversity checker	81.35	77.94	77.68	77.62	74.72	69.34
Large real world	CE-Loss	84.74	82.69	68.75	68.41	80.92	77.73
	Focal-Loss	**84.98**	81.98	67.6	67.19	80.35	76.28
	Hu et al.'s	80.19	76.58	60.33	37.63	71.36	67.78
	SunGen	84.65	82.51	63.9	62.66	80.81	78.78
	IMP-Loss (Ours)	**85.3**	**83.27**	**70.15**	**70.08**	81.33	78.3
	DIMP-Loss (Ours)	**85.4**	**82.79**	69	68.78	**82.84**	**80.49**

THINKINGS

🤔 Add weight to SGO ?

$$L_{WSGO}(q_\theta; o) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1} \sim d_o^{t-1}} \lambda(t-1, w^{t-1}) D_{KL}(o(\cdot|w^{t-1})||q_\theta(\cdot|w^{t-1}))$$

🤔 When using SGO, add adaptive weight for L_{Base} w.r.t L_{KD} ?

$$L = L_{KD} + \phi_{epoch} L_{Base}$$

🤔 Deeper research in teacher's response to SGO ?

🤔 How to solve error accumulation?

LLM KD WITH DIFFERENT VOCABULARIES

教师$(m \times D)$向学生$(n \times d)$对齐：

$$Q = P^q([e^s_{1:n}; e^s_{2:n+1}]; \theta^q_P) \in R^{n \times 2D}$$

$$K = [e^t_{1:m}; e^t_{2:m+1}] \in R^{m \times 2D}$$

$$V = P^v(e^t_{2:m+1} + h^t_{1:m}; \theta^v_P) \in R^{m \times d}$$

Zhang, S., Zhang, X., Sun, Z., Chen, Y., & Xu, J. Dual-Space Knowledge Distillation for Large Language Models. In EMNLP, 24

LLM KD WITH DIFFERENT VOCABULARIES

教师变换后的emd可以表示为

$$h_{1:n}^{t\to s} = softmax(\frac{QK^T}{\sqrt{2D}}V) \in R^{n\times d}$$

最后过学生的映射头得到概率分布

$$p^t = softmax(h_{1:n}^{t\to s}\color{red}{W_S})$$

THANKS!