Internship weekly report

Mathis Embit

MILES, LAMSADE

June 21, 2024

Introduction

This week: baselines code + "forward projection"

AutoPrompt/GCG observation

In my experiments GCG ([Zou+23]) works whereas AutoPrompt ([Shi+20]) does not.

GCG was tested on Llama-2-7B-Chat (and others) while AutoPrompt was tested on $BERT_{BASE}$ (110M parameters) and $RoBERTa_{LARGE}$ (355M parameters).

Hence it is not so surprising but it highlights the fact that the randomness in the GCG algorithm makes a real difference.

Forward projection

Formalization of the $\arg\min_{e \in W_E} D(f(x), f(e))$ idea resulted in the following observation:
Let $p(.) = p(.|x_1, \dots, x_n) \in [0,1]^{|V|}$ and
$q_i(e)(.) = p(.|x_1, \dots, e, \dots, x_n) \in [0,1]^{|V|}$, we then have

$$\begin{aligned}
\max_{e \in E} \mathcal{D}(p \| q_i(e)) &\iff \max_{e \in E} \sum_{w \in V} p(w) \log \frac{p(w)}{q_i(e)(w)} \\
&\iff \max_{e \in E} - \sum_{w \in V} p(w) \log q_i(e)(w) \\
&\iff \max_{e \in E} - \log q_i(e)(y) \text{ if } p(y) = 1 \\
&= \max_{e \in E} - \log p(y|x_1, \dots, e, \dots, x_n)
\end{aligned}$$

Forward projection

Since a continuous optimization gives us $x^*_{1:n}$ such that $p(y_{1:m}|x^*_{1:n}) \approx 1$, doing the forward projection might not be useful. However it might be useful with another loss, such as

$$\mathcal{L}(x_i) = -\log p(y|x_{1:n}) + H(\text{softmax}(W_E x_i)) - p(x_{1:n})$$

References I

[Shi+20] Taylor Shin et al. *AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts*. 2020. arXiv: 2010.15980.

[Zou+23] Andy Zou et al. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. 2023. arXiv: 2307.15043.