Internship weekly report

Mathis Embit

MILES, LAMSADE

August 2, 2024

Introduction

This week: code update, wandb logs and report beginning.

Code update

In the discrete optimization we should not forget to reevaluate each batch with the KL as $\mathcal{L}$: $x_{1:n} := \tilde{x}_{1:n}^{(b^\star)}, \quad$ where $b^\star = \arg\min_b \mathcal{L}(\tilde{x}_{1:n}^{(b)})$

We also train on a whole dataset rather than on an example.

Wandb logs

At the beginning:

- model_name
- user_prompt
- adv_string_init
- target
- epochs
- lr
- w_ce
- w_at
- w_nl
- w_en

Wandb logs

But also:

- matrix to matrix cosine similarity mean
- matrix to matrix dot product mean
- matrix to matrix L2 distance mean

And:

- generation before opt, with target
- generation before opt, without target

Wandb logs

For the continuous optimization:

For each iteration:

- loss
- cross entropy loss
- attraction loss
- negative log likelihood
- entropy
- attack success

And:

- iterate norms
- iterate metric with closest embedding **for metric in metrics**
- iterate metric closest embedding norms **for metric in metrics**

And at the end:

- generation after cont opt, with target
- generation after cont opt, without target

Wandb logs

For the discrete optimization:

- disc_num_steps
- batch_size
- topk

At each iteration:

- current_loss

And at the end:

- generation after discrete opt, without target

Report

Structure of the report:

1. Introduction and Problem Statement
    1. Context in LLMs
    2. In-Context Learning
    3. Parameter-Efficient Fine-Tuning
    4. Prompt Optimization
2. State-of-the-Art Study
    1. Attributing Output to Input Features
    2. Discrete Prompting
    3. Continuous Prompting
3. Internship Contribution
    1. Prompt Optimization Methods Comparison
    2. Analysis of the Embedding Space
    3. From Continuous Embeddings to Discrete Tokens
    4. Toward a Probabilistic Characterization of LLMs
4. Conclusion and Perspectives