Internship weekly report

Mathis Embit

MILES, LAMSADE

July 1, 2024

Introduction

This week: more code + new loss idea

New loss idea

Continuous optimization of x leads to $p(y|x) \approx 1$. If we want to leverage the divergence of output distributions between the continous optimized prompt and the discrete prompt to be optimized we need to artificially add information in the continous prompt output distribution. To do so we can regularize by the entropy:

$$\mathcal{L} = -\log p(y|x) - H(p(.|x))$$

Another idea

Inspired by [Chu+24], maybe we can also use hidden layers to match the discrete prompt behavior to the one of the continuous prompt.

References I

[Chu+24] Yung-Sung Chuang et al. *DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models*. 2024. arXiv: 2309.03883 [cs.CL]. URL: https://arxiv.org/abs/2309.03883.