Internship weekly report

Mathis Embit

MILES, LAMSADE

July 8, 2024

Introduction

This week: mainly universal attack code adaptation.

Unembedding idea

Use the pseudo-inverse of the embedding matrix to go from the embedding space to the token space.

Loss

Final loss before we test:

$$\mathcal{L}(x_i) = -\log p(y|x) + H(\text{softmax}(Ex_i)) - \log p(x) - H(p(.|x))$$

where $H(p(.|x)) = \sum_{i=1}^{m} H(p(.|x, y_{<i}))$, if $y = y_{1:m}$.