Internship weekly report

Mathis Embit

MILES, LAMSADE

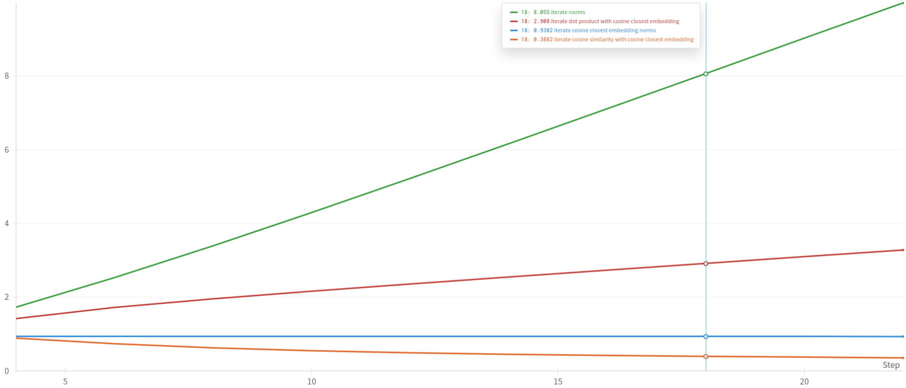
July 29, 2024

Introduction

This week: experiments

Experiments

When minimizing $H(\text{softmax}(Ex_i))$ we just maxiximizing the dot product with the closest embedding. Hence it just increases the norm of the vector but not its direction.



Cosine decomposition

Experiments

To avoid norm explosion we can do it with cosine or maybe simply use a distance or similarity to the closest embedding as a loss for attraction toward E?

Unembedding idea

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

Notations: Let V be the size of the vocabulary and d be the hidden dimension. Let $E \in \mathbb{R}^{V \times d}$ be the embedding matrix (llama2 convention). Let's suppose the weights are shared with the unembedding matrix $U = E \in \mathbb{R}^{V \times d}$. To embed a one hot encoded token $t \in \{0,1\}^{V \times 1}$ we compute $x = E^T t \in \mathbb{R}^{d \times 1}$.

Unembedding idea

We are interested in the unembedding operation: $x \in \mathbb{R}^d \mapsto w \in E$. If we consider the L^2 distance the problem is:

$$\arg\min_{w \in E} \|x - w\|_2^2$$

It's a linear regression: $\arg\min_{t \in \{0,1\}^V} \|x - Et\|_2^2$.
Without the $\in E$ constraint:

$$\arg\min_{y \in \mathbb{R}^d} \|x - Ey\|_2^2$$

the solution is $y = (E^T E)^{-1} E^T x$.
But with the $\in E$ constraint we still need to take the closest $w \in E$ so pseudo-inverse doesn't bring something new.