Forward projection of the embeddings

Mathis Embit

June 19, 2024

Abstract

Here is a short explanation on another possible way to project continuous embeddings onto the embedding matrix.

From a general point of view we want to find a sequence of tokens x that maximize the probability of generating a target y. If we write E the sequences of embeddings of tokens of the vocabulary we want to find

$$\arg\max_{x \in E} p(y|x) \tag{1}$$

Let's suppose we have n tokens in the prompt and m tokens in the target:

$$\arg\max_{x_{1:n} \in E} p(y_{1:m}|x_{1:n}) \tag{2}$$

x can be discrete tokens (AutoPrompt and Universal attacks) or continuous embeddings (prompt tuning, FluentPrompt).

Let's suppose $x_{1:n}$ are continuous.

Let $\mathcal{L}$ be the loss we need to minimize. We have two possible ways to optimize $x_1, \dots, x_n$:

- Updating the n tokens at each step: compute $\mathcal{L}(x_{1:n})$ and use this information to update $x_{1:n}$.
- Updating one token at each step: for some i compute $\mathcal{L}(x_{1:n})$ and use this information to update x_i so this change will be taken into account when doing the same for another i.

Additionally we can add some randomness that will help explore the candidate space.

With a loss such as $\mathcal{L}(x) = -\log p(y|x) = -\sum_{i=1}^{m} p(y_i|x_{1:n}, y_{<i})$ the classic discrete optimization step is

$$x_i = \arg\min_{e \in \mathcal{E}} \langle \nabla_{x_i} \mathcal{L}, e \rangle$$

Additionally we can add randomness (like in GCG) and reevaluation of $\mathcal{L}$ on a tok-k of $\langle \nabla_{x_i} \mathcal{L}, e \rangle$ to pick the best one.

We can think of another way to project. Since we want to project while minimizing the effect the projection will have on the projection and that the LLM is not smooth at all we may think of this:

Rather than projecting like

$$\arg\min_{e \in W_E} d(x, e)$$

we may do

$$\arg\min_{e \in W_E} D(f(x), f(e))$$

where f is the output of the LLM, for example the probabilities of generating the target, and D is a distance on the output space.

And considering the LLM case:

However doing so suppose we have access the response m optimal distributions (contrary to $\mathcal{L}$). Hence we need to assume that a first continuous optimization process has given us the optimal $x_1, \dots, x_n$. From there we can compute $\mathcal{D}$.

$$x_i = \arg\min_{e \in \mathcal{E}} \langle \nabla_{x_i} \mathcal{D}, e \rangle$$

where

$$\nabla_{x_i}\mathcal{D} = \nabla_{x_i}\mathcal{D}(p\|q(x_i))$$

where p is the fixed $p(y \mid x) = \begin{bmatrix} p(y_1 \mid x) \\ p(y_2 \mid x, y_1) \\ \vdots \\ p(y_m \mid x, y_{<m}) \end{bmatrix}$

and $q(x_i)$ is the variable $p(y \mid x_1, \ldots, x_i, \ldots, x_n) = \begin{bmatrix} p(y_1 \mid x_1, \ldots, x_i, \ldots, x_n) \\ p(y_2 \mid x_1, \ldots, x_i, \ldots, x_n, y_1) \\ \vdots \\ p(y_m \mid x_1, \ldots, x_i, \ldots, x_n, y_{<m}) \end{bmatrix}$

and $\mathcal{D}(p\|q(x_i)) = \frac{1}{m}\sum_{j=1}^{m}\mathcal{D}(p_j\|q(x_i)_j)$

Practically, $\nabla\mathcal{D} = \begin{bmatrix} \nabla_{x_1}\mathcal{D} \\ \vdots \\ \nabla_{x_n}\mathcal{D} \end{bmatrix}$ is computed in a single forward/backward pass.

A new algorithm could be:

1. Continuous optimization gives us $x_1^*, \ldots, x_n^*$ the continuous solution to $\underset{x_{1:n}\in E}{\arg\max}\, p(y_{1:m}|x_{1:n})$.
2. Discrete optimization of $x_1, \ldots, x_n$ with the $\mathcal{D}$ loss.

However the first step migth convergence to a solution such that $p(y_{1:m}|x_{1:n}^*) \approx 1$. Then, writting $p = p(.|x_1, \ldots, x_n) \in [0,1]^{|V|}$ and $q_i(e) = p(.|x_1, \ldots, e, \ldots, x_n) \in [0,1]^{|V|}$, we have

$$\begin{aligned} \max_{e\in E}\mathcal{D}(p\|q(x_i)) &\iff \max_{e\in E}\sum_{w\in V} p(w)\log\frac{p(w)}{q_i(e)(w)} \\ &\iff \max_{e\in E} -\sum_{w\in V} p(w)\log q_i(e)(w) \end{aligned}$$

If $p(y) = 1, \iff \max_{e\in E} -\log q_i(e)(y) = \max_{e\in E} -\log p(y|x_1, \ldots, e, \ldots, x_n)$.

Therefore it might not be useful for our loss $\mathcal{L}$ but if we want to add to it some terms such as the likelihood of x or the fact that x_i are attracted to words of the vocab, it might be interesting.

References