One-layer transformer for NTP task

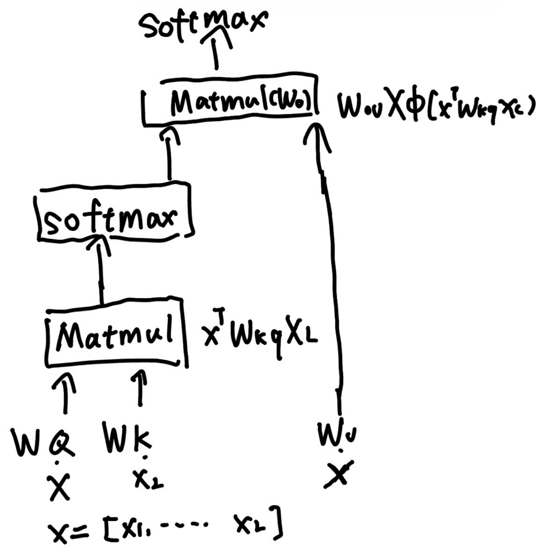
Li Yunai

September 3, 2024

Task Set Up

- **Vocabulary set:**$\mathcal{V} \subset \mathbb{R}^d$
  Sentence:$X = [x_1, \ldots, x_L] \in \mathcal{V}^L \subset \mathbb{R}^{d \times L}$,length $L < L_{max}$
- **Ground truth model:** $p_L^* : X \in \mathcal{V}^L \rightarrow x_{L+1} \in \mathcal{V}$
- **training data set:**
  $\mathcal{D}_0 = \{(X, x_{L+1}) | L < L_{\max}, X \in \mathcal{V}^L, x_{L+1} \in \mathcal{V}\}$

Model Architecture



Output:

$$\mathrm{T}_\theta(X) := \phi(W_{\mathrm{ov}} X \phi(X^T W_{\mathrm{kq}} x_L)) \in [0,1]^{|\nu|}$$

X_q partial order

- Motivation:training loss $\to 0$
- **"Collocation":** Consider the one-token in which only the feed forward layer works (with parameter W_t at time t).Then if the loss fuction $\to$ 0,i.e $\lim_{t\to\infty} -\sum_{x\in\mathcal{D}_0} \log e_{\iota}(x)^\top \phi(W_t x) = 0$ then ι is injective, here $\iota(x)$ is the index of the next token, indicating the unique next token n(x) for one-token input. Call $\{x, \mathsf{n}(x)\}_{x\in\mathcal{V}}$ as **collocation**. $(p_1^* = n)$
- **"order"**: $\mathbf{n}^{-1}(x_{L+1}) \in \{x_\ell\}_{\ell\le L}$. Denote $\varphi_\ell \propto \exp(x_\ell^\top W_k q x_L)$ (Weight from the attention layer). Then $\varphi_\ell > \varphi_{\ell'}$ if $n(x_\ell) = x_{L+1} \neq n(x_{\ell'})$

query-dependent partial orders

Definition

Let $\mathcal{D}_0^{x^q}$ be the set of all sentences in the training dataset that has the final token x^q. Then, for any pair of tokens $x, x' \in \mathcal{V}$, $x >_{x^q} x'$ if there exists a sentence $X = [x_1, \ldots, x_L] \in \mathcal{D}_0^{x^q}$ and x, x'*are tokens in* X such that $n(x) = x_{L+1} \neq n(x')$,where x_{L+1} is the next token of X.

token classification:
optimal/non-optimal/non-comparable/confused token (non-exist by assumption)
Realizable dataset: 1) collocation 2) well-defined $>_{x^q}$ partial order without confused token

Algorithm Design

Training W_{OV} (by single token prediction)

$$\mathcal{L}_0(W_{ov}) = -\sum_{x \in \mathcal{V}} \log \frac{\exp(e_{In(x)}^\top W_{ov} x)}{\sum_{v \le |\mathcal{V}|} \exp(e_v^\top W_{ov} \; x)}$$

$$W_{\text{ov}}^{(t+1)} = W_{\text{ov}}^{(t)} - \eta_0 \frac{\nabla_{W_{\text{ov}}} \mathcal{L}_0(W_{\text{ov}}^{(t)})}{\|\nabla_{W_{\text{ov}}} \mathcal{L}_0(W_{\text{ov}}^{(t)})\|}$$

Algorithm Design

Training W_{kq} Considering cross entropy loss:

$$\mathcal{L}(\theta) = -\sum_n \pi^{(n)} \left(\log \left(\sum_v e_{\mathrm{In}(X^{(n)})}^\top \bar{\mathcal{T}}_\theta(X^{(n)}) \right) - e_{\mathrm{In}(X^{(n)})}^\top \bar{\mathcal{T}}_\theta(X^{(n)}) \right)$$

Where $X^{(n)}$ denotes the n-th sentence, $X_{-1}^{(n)}$ is the last token, $\pi^{(n)} = \frac{\sum_{(X, x_{L+1}) \in \mathcal{D}_0} \mathbf{1}\{X = X^{(n)}\}}{|\mathcal{D}_0|}$, $\bar{\mathrm{T}}_\theta(X) = W_{\mathsf{ov}} X \phi(X^\top W_{\mathsf{kq}} X_{-1}^{(n)})$.

Then the parameter is trained by:

$$W_{\mathrm{kq}}^{(t+1)} = W_{\mathrm{kq}}^{(t)} - \eta \frac{\nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)})\|}, \text{where } \theta^{(t)} = (W_{\mathrm{ov}}^{(T)}, W_{\mathrm{kq}}^{(t)})$$

From the perspective of hard margin problem

W_{ov}

$$W^*_{\text{ov}} = \arg\min \|W\|$$

$$\text{s.t.} \quad (e_{v^*} - e_v)Wx \geq 1, \quad \forall v^* = \mathsf{In}(x), v \neq \mathsf{In}(x)$$

W_{kq}

$$W^*_{\mathsf{kq}} = \arg\min \|W\|$$

$$\text{s.t.} \quad (x^{(n)}_{\ell_*} - x^{(n)}_{\ell})WX^{(n)}_{-1} \geq 1, \quad \forall \ell_* \in l^{(n)}, \ell \notin l^{(n)}, \forall n$$

l(n) is the set of indices of the optimal tokens

Generalization ability

If the trained transformer takes input X with query x^q that consists of a non-comparable and non-optimal tokens, then the prediction made by $\mathrm{T}_{\theta^{(t)}}(X)$ is $n(x_0)$with high probability:

Theorem

With certain assumptions in effect and $t = \Omega(\log(1/\epsilon))$. *Then there exists a constant* C_0 such that

$$(x_* - x_0)^\top W_{\mathrm{kq}}^{(t)} x^q \geq C_0 t, \quad (x_0 - x)^\top W_{\mathrm{kq}}^{(t)} x^q \geq C_0 t, .$$

$\forall x_* \in \mathcal{O}_{x^q}, x_0 \in \mathcal{M}_{x^q}, x \in \mathcal{N}_{x^q}.$

Subject,Verb, Object& Punctuation mark

V=SVOP, VOP, OPP, PSV Collocation:(S, V),(V, O),(O, P),(P, S).
Partial order under query S. $S >_S P$.
Partial order under query V. $V >_V S$.
Partial order under query O. $O >_O S$ $O >_O V$.
Partial order under query P. $O >_P P$.

Generalization to unseen data

Example 1 Input:SP under$>_P$, S non comparable, p non-optimal, then Φ_s is larger prediction: n(S)=V

Example 2 Input:OSP $O>_P P$ prediction:n(O)=P