# Attention with Trained Embeddings Provably Selects Important Tokens

Diyuan Wu*[1] Aleksandr Shevchenko*[2] Samet Oymak[3] Marco Mondelli[1]

[1]ISTA, [2]ETHZ, [3]University of Michigan

November 4, 2025

# Motivation

Consider sentences classification problem, some tokens are more important than others.

### Example from IMDB movie reviews

It's **terrific** when a funny movie doesn't make smile you. What a **pity**!! This film is very **boring** and so long. It's simply **painful**. The story is **staggering** without goal and **no fun**.You feel better when it's finished.

### Questions

When learn with attention based models on such task, can the model learn to select important tokens from others?

# Model

## Training set

We have $n$ data point $\mathcal{X}_n = \{(X^k, y^k)\}_{k=1}^n$ with each $X \in \mathcal{X}_n$ is $[x_1, \ldots, x_T]$, and $y = \pm 1$. Denote the vocab set is $\mathcal{S}$.

## Single-layer attention

Denote $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{S}| \times d}$ be the embedding layer. For each sequence $X$, the embedding is $\boldsymbol{E}_X = [E_{x_1}, \ldots, E_{x_T}]^\top \in \mathbb{R}^{T \times d}$

$$f(X; p, \boldsymbol{E}) = \mathtt{Softmax}(p^\top \boldsymbol{E}_X^\top)\boldsymbol{E}_X v = \frac{\sum_{i=1}^T \exp(p^\top E_{x_i}) E_{x_i}^\top v}{\sum_{j=1}^T \exp(p^\top E_{x_j})}$$

Fix $v$ be a unique vector, and we train $p, \boldsymbol{E}$.

## Logistic Loss

$\mathcal{L}(p, \boldsymbol{E}) = \widehat{\mathbb{E}}[\log(1 + \exp(-yf(X; p, \boldsymbol{E})))]$

# What tokens are important?

**Questions**

Given a training set without any prior information, how do we define what tokens are important?

**Empirical Importance**

Given a token $s$, we define:

$$\alpha_s = \text{total occurrance with label } '1' - \text{total occurrance with label } '-1'$$

1. Tokens with larger $|\alpha_s|$ is more important.
2. $s$ is positive/negative/irrelevant: $\alpha_s > 0/ < 0/ = 0$.
3. $s$ is completely positive/negative if $s$ only occur with label $1/-1$.

# One step of embedding training learns token importance

**Lemma**

*Under standard initialization where $E_s \overset{i.i.d}{\sim} \mathcal{N}(0, 1/d \cdot I)$, $p \sim \mathcal{N}(0, 1/d \cdot I)$, with large enough $d$, after one-setp of GD with step size $\eta_0$,:*

$$E_s^1 \approx E_s^0 + \frac{\eta_0}{2} \alpha_s v, \forall s \in \mathcal{S}$$

Recall the model

$$f(X; p, \boldsymbol{E}) = \frac{\sum_{i=1}^{T} \exp(p^\top E_{x_i}) \color{red}{E_{x_i}^\top v}}{\sum_{j=1}^{T} \exp(p^\top E_{x_j})}$$

After one step: $E_{x_i}^\top v \approx \eta_0 \alpha_s / 2$

# Implicit bias

**Question**

Suppose we fix $\boldsymbol{E}$ after one-step and only train $p$ with GF, is there any implicit bias over $p$ ?

**Simple data model**

Each sequence in $\mathcal{X}_n$ contains either a *single completely positive* token or a *single completely negative* token, and all remaining tokens are *irrelevant*.

# What do we expect?

We want to make $\mathcal{L}(p, \boldsymbol{E}) = \widehat{\mathbb{E}}[\log(1 + \exp(-yf(X; p, \boldsymbol{E}))]$ small.

$\Rightarrow$ For each $(X, y)$, we want $yf(X; p, \boldsymbol{E}) = \frac{\sum_{i=1}^{T} \exp(p^\top E_{x_i}) y E_{x_i}^\top v}{\sum_{j=1}^{T} \exp(p^\top E_{x_j})}$ large.

$\Rightarrow$ For each $(X, y)$, if $yE_{x_i}^\top v > yE_{x_j}^\top v$, we want $p^\top E_{x_i} \geq p^\top E_{x_j}$.

## Token selection

Given $p$, for each $X$, we defined the tokens in $X$ selected by $p$ as:

$$\mathcal{S}_X(p) = \{x_i : i \in \arg\max_i p^\top E_{x_i}\}$$

# Limiting token selection

## Lemma

*With high probability over the initialization,*

$$\|p_t\|_2 \to \infty$$

## Only selected tokens matter

After training long enough,

$$f(X; p_t, \boldsymbol{E}) \approx \frac{1}{|\mathcal{S}_X(p_t)|} \sum_{i \in \mathcal{S}_X(p_t)} E_{x_i}^\top v$$

## Lemma

*When $\eta_0$ is large enough, for each $X$, all the important tokens must be selected by $p_t$ for large enough $t$. (But can also select irrelevant ones)*

# Max-margin token selection

**Question**

There could be many $p$ that do the same selection for all $X$, is there any implicit bias over the $\lim_{t \to t} \frac{p_t}{\|p_t\|_2}$?

**Lemma (Max-margin token selection)**

*Suppose* $\frac{p_\infty}{\|p_\infty\|_2} = \lim_{t \to t} \frac{p_t}{\|p_t\|_2}$ *exists, then*

$$p_\infty = \arg\min_p \|p\|_2$$

$$s.t. \quad p^\top (E_s - E_{s'}) \geq 1, \forall s \in \mathcal{S}_X(p_\infty), \forall s' \in X \setminus \mathcal{S}_X(p_\infty), \forall X.$$

Synthetic data:

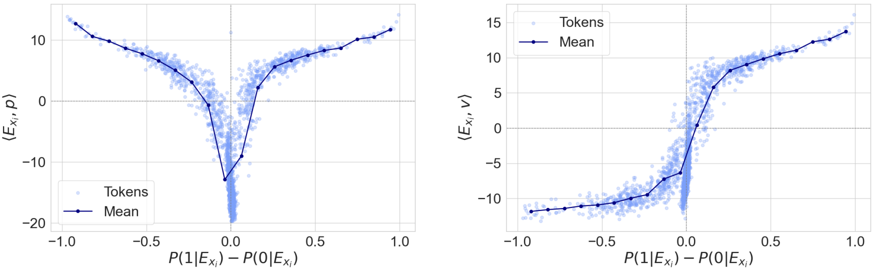


Figure: Dot-product of embedding tokens with $\langle \mathrm{cls} \rangle$ token $p$ (left) and regression coefficients $v$ (right), as a function of the token-wise difference in posterior probabilities for synthetic data. The concentrated cloud of points around zero corresponds to the tokens in the irrelevant set.

# Numerical experiments

IMDB and Yelp dataset:

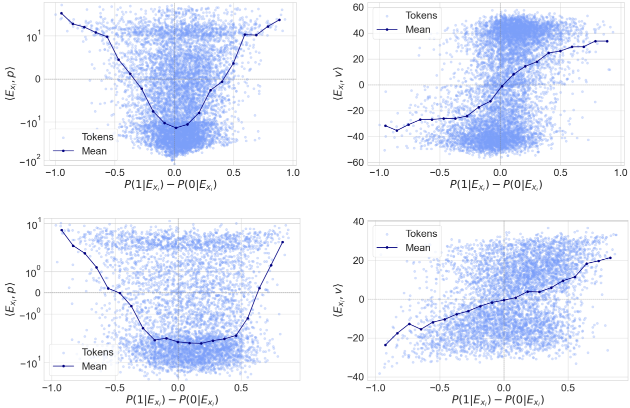


Figure: Dot-product of embedding tokens with CLS token $p$ and regression coefficients $v$ versus token-wise difference in posterior for IMDB dataset (top row) and Yelp dataset (bottom row).