Max-Margin Token Selection in Attention Mechanism

Davoud Ataee Tarzanagh[1] Yingcong Li[2] Xuechen Zhang[2]
Samet Oymak[3]

[1] University of Pennsylvania

[2] University of California, Riverside

[3] University of Michigan

Transformer (Vaswani et al. in 2017)

Transformer is a neural network architecture that includes:

1. **Tokenization**:
   Treating input as a sequence of tokens.
2. **Attention Mechanism**:
   Computing token similarities using dot-products.

Text input: "This is a sample sentence."

Tokens: ["This", "is", "a", "sample", "sentence"]

Visual input:

Tokens are patches:

Goal: Understanding transformer and attention through optimization theory.

Attention Model in Transformer

For the input token sequences $\boldsymbol{X} \in \mathbb{R}^{T \times d}$ and $\boldsymbol{Z} \in \mathbb{R}^{T \times d}$, attention model f maps an input sequence to an output sequence as follows:

$$f(\boldsymbol{X}; \boldsymbol{Z}) = \mathbb{S}(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{K}^\top \boldsymbol{Z}^\top)\boldsymbol{X}\boldsymbol{V}.$$

Here,

- $\boldsymbol{K}, \boldsymbol{Q} \in \mathbb{R}^{d \times m}$, $\boldsymbol{V} \in \mathbb{R}^{d \times v}$ are the trainable key, query, value matrices respectively;
- $\mathbb{S}(\cdot)$ is softmax function.

Our setting:

1. Replace $\boldsymbol{Z}$ with $\boldsymbol{p}$, where $\boldsymbol{p} \in \mathbb{R}^d$ is [CLS] token or tunable prompt;
2. Replace $\boldsymbol{K}\boldsymbol{Q}^\top$ with a combined $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{V}$ with $\boldsymbol{v} \in \mathbb{R}^d$.

Classification with Attention

- The training dataset $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ has binary labels $Y_i \in \{-1, 1\}$ and token sequences $\boldsymbol{X}_i \in \mathbb{R}^{T\times d}$.
- Let $\boldsymbol{K}_i = \boldsymbol{X}_i \boldsymbol{W}^\top$. For a decreasing loss ℓ, define

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{p}) := \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p})\right). \tag{ERM}$$

Token Score

Given $\boldsymbol{v} \in \mathbb{R}^d$, the score of token $\boldsymbol{x}_{it}$ of input $\boldsymbol{X}_i$ is defined as

$$\gamma_{it} := Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_{it}.$$

Optimal Tokens

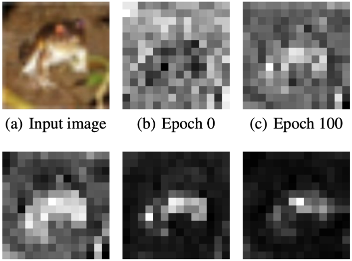
The optimal tokens for input $\boldsymbol{X}_i$ are those tokens with highest scores:

$$\alpha_i \in \arg\max_{t\in[T]} \gamma_{it}.$$

Intuition: Optimal tokens $\{\alpha_i\}_{i=1}^n$ minimize (ERM).

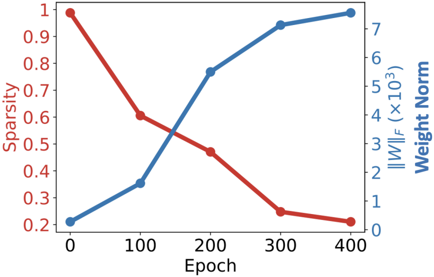
Empirical Insights

1. Attention mechanism selects a few tokens that are most relevant for prediction.
2. As we select fewer tokens, the norm of the weights $\boldsymbol{W}$ (or $\boldsymbol{p}$) grows.



(a) Input image (b) Epoch 0 (c) Epoch 100

(d) Epoch 200 (e) Epoch 300 (f) Epoch 400



Our optimization theory rigorizes these observations via "optimal tokens" & Attention-SVM equivalence.

Attention SVMs

$\boldsymbol{p}$-SVM

$$\boldsymbol{p}(\alpha) = \arg\min_{\boldsymbol{p}} \|\boldsymbol{p}\|$$
$$\text{s.t. } \boldsymbol{p}^\top(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it}) \geq 1, \ \forall i, t \neq \alpha_i.$$

$\boldsymbol{v}$-SVM

$$\boldsymbol{v}(\alpha) = \arg\min_{\boldsymbol{v}} \|\boldsymbol{v}\|$$
$$\text{s.t. } \ Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_{i\alpha_i} \geq 1, \ \forall i \in [n].$$

Joint $\boldsymbol{v}$ and $\boldsymbol{p}$-SVMs

$$\boldsymbol{v}(\alpha) = \arg\min_{\boldsymbol{v}} \|\boldsymbol{v}\| \ \text{ s.t. } \ Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_{i\alpha_i} \geq 1, \ \forall i \in [n].$$
$$\boldsymbol{p}(\alpha) = \arg\min_{\boldsymbol{p}} \|\boldsymbol{p}\| \ \text{ s.t. } \ \boldsymbol{p}^\top(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it}) \geq \begin{cases} 1 & t \neq \alpha_i, \ i \in \mathcal{S} \\ 0 & t \neq \alpha_i, \ i \in \bar{\mathcal{S}}. \end{cases}$$

- $\mathcal{S} \subset [n]$ is the set of indices that $\boldsymbol{x}_{i\alpha_i}$ is a support vector when solving $\boldsymbol{v}$–SVM; and $\bar{\mathcal{S}} = [n] - \mathcal{S}$.

Convergence of Gradient Descent on $\boldsymbol{p}$

Assumption $\mathcal{A}$

Over any bounded interval: i) $\ell : \mathbb{R} \to \mathbb{R}$ is strictly decreasing. ii) ℓ' is M_0-Lipschitz continuous and $|\ell'(u)| \leq M_1$.

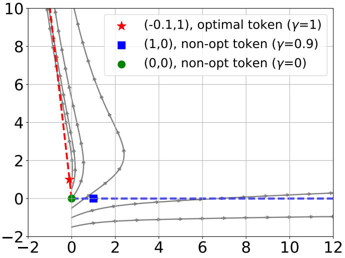
Theorem I: Convergence of GD

Suppose Assumption -$\mathcal{A}$ holds. Then, the gradient descent (GD) iterates

$$\boldsymbol{p}(k+1) = \boldsymbol{p}(k) - \eta \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{v}(k), \boldsymbol{p}(k)),$$

with proper η and $p(0)$ satisfies

- $\lim_{k\to\infty} \|\boldsymbol{p}(k)\| = \infty$; and
- $\lim_{k\to\infty} \frac{\boldsymbol{p}(k)}{\|\boldsymbol{p}(k)\|} = \frac{\boldsymbol{p}(\boldsymbol{\alpha})}{\|\boldsymbol{p}(\boldsymbol{\alpha})\|}$



- (-$>$-): GD trajectories from different $p(0)$.
  - Global (- - -) and Local (- - -) $\boldsymbol{p}(\boldsymbol{\alpha})$.

Joint Convergence of Regularized Path on $\boldsymbol{v}$ and $\boldsymbol{p}$

Assumption $\mathcal{B}$

Let $\Gamma = 1/\|\boldsymbol{v}(\alpha)\|$ *and define* $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})$. *For all* $\boldsymbol{p}$, *solving* $\boldsymbol{v}$*-SVM with* $\boldsymbol{x}_i^{\boldsymbol{p}} := \boldsymbol{X}_i^\top \boldsymbol{s}_i$ *results in a label margin of at most*

$$\Gamma - \nu \cdot \max_{i\in[n]}(1 - \boldsymbol{s}_{i\alpha_i}), \quad \text{for some} \quad \nu > 0.$$

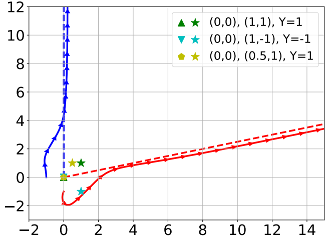
Theorem II: Convergence of RP

Consider the regularized path solutions $(\boldsymbol{v}_r, \boldsymbol{p}_R)$ of (ERM) defined as

$$(\boldsymbol{v}_r, \boldsymbol{p}_R) = \underset{\|\boldsymbol{v}\|\le r, \|\boldsymbol{p}\|\le R}{\arg\min} \mathcal{L}(\boldsymbol{v}, \boldsymbol{p}).$$

Under Assumptions -$\mathcal{A}$ and -$\mathcal{B}$, we have

- $\lim_{r\to\infty} \frac{\boldsymbol{v}_r}{r} = \frac{\boldsymbol{v}(\alpha)}{\|\boldsymbol{v}(\alpha)\|}$; and
- $\lim_{R\to\infty} \frac{\boldsymbol{p}_R}{R} = \frac{\boldsymbol{p}(\alpha)}{\|\boldsymbol{p}(\alpha)\|}$.



- $\boldsymbol{p}$ (—>—) and $\boldsymbol{v}$ (—>—) trajectories.
- $\boldsymbol{p}(\alpha)$(- - -) and $\boldsymbol{v}(\alpha)$ (- - -) directions.

Thank You!