Unraveling Attention via Convex Duality: Analysis and Interpretations of Vision Transformers

Arda Sahiner[1], Tolga Ergen[1], Batu Ozturkler[1], Burak Bartan[1], John Pauly[1], Morteza Mardani[2], Mert Pilanci[1]

[1]Department of Electrical Engineering, Stanford University, Stanford, CA
[2]NVIDIA Corporation, Santa Clara, CA

Introduction

- Vision transformer architectures very successful at image tasks
- A variety of attention mechanisms have been proposed: self-attention, MLP-Mixer, Fourier Neural Operator (FNO), and more!
- However, these architectures are not theoretically understood

Contributions

- Prove that self-attention, MLP-Mixer, and FNO with linear and ReLU activation can be solved to their global optima by demonstrating their equivalence to convex optimization problems.
- Provide interpretability to the optimization objectives of these attention modules.
- Validate the (convex) vision transformers perform better than baseline convex methods in a transfer learning task (see paper).

Preliminaries: Background

- Training data $\{\mathrm{X}_i \in \mathbb{R}^{s \times d}\}_{i=1}^{n}$, corresponding labels of arbitrary size $\{\mathrm{Y}_i \in \mathbb{R}^{r \times c}\}_{i=1}^{n}$
- Solve the optimization problem

$$p^* := \min_{\theta} \sum_{i=1}^{n} \mathcal{L}\left(f_\theta(\mathrm{X}_i), \mathrm{Y}_i\right) + \mathcal{R}(\theta) \qquad (1)$$

- Can include classification, where $r = 1$, or for regression, where $r = s$, one can directly use squared loss or other convex loss functions.
- One may also use this formulation to apply to both supervised and self-supervised learning.

Single Block of Multi-Head Self-Attention

- jth self-attention head given by

$$f_j(\mathsf{X}_i) := \sigma\left(\frac{\mathsf{X}_i\mathsf{Q}_j\mathsf{K}_j^\top\mathsf{X}_i^\top}{\sqrt{d}}\right)\mathsf{X}_i\mathsf{V}_j, \tag{2}$$

- Multi-head self-attention

$$\begin{aligned} f_{MHSA}(\mathsf{X}_i) &:= \begin{bmatrix} f_1(\mathsf{X}_i) & \cdots & f_m(\mathsf{X}_i) \end{bmatrix}\mathsf{W} \\ &= \sum_{j=1}^{m}\sigma\left(\frac{\mathsf{X}_i\mathsf{Q}_j\mathsf{K}_j^\top\mathsf{X}_i^\top}{\sqrt{d}}\right)\mathsf{X}_i\mathsf{V}_j\mathsf{W}_j \end{aligned} \tag{3}$$

- Can simplify as

$$f_{MHSA}(\mathsf{X}_i) := \sum_{j=1}^{m}\sigma\left(\frac{\mathsf{X}_i\mathsf{W}_{1j}\mathsf{X}_i^\top}{\sqrt{d}}\right)\mathsf{X}_i\mathsf{W}_{2j}. \tag{4}$$

Convexity of Linear Multi-Head Self-Attention

Theorem

Pose the non-convex weight-decay linear-activation multi-head self-attention training problem

$$p^*_{SA} := \min_{\mathsf{W}_{1j},\mathsf{W}_{2j}} \sum_{i=1}^{n} \mathcal{L}\Big(\sum_{j=1}^{m} \mathsf{X}_i \mathsf{W}_{1j} \mathsf{X}_i^\top \mathsf{X}_i \mathsf{W}_{2j}, \mathsf{Y}_i\Big) + \frac{\beta}{2} \sum_{j=1}^{m} \|\mathsf{W}_{1j}\|_F^2 + \|\mathsf{W}_{2j}\|_F^2. \quad (5)$$

Then, for $\beta > 0$ *and* $m \geq m^*$ *where* $m^* \leq \min\{d^2, dc\}$, *this is equivalent to a convex optimization problem*

$$p^*_{SA} = \min_{\mathsf{Z} \in \mathbb{R}^{d^2 \times dc}} \sum_{i=1}^{n} \mathcal{L}\left(\sum_{k=1}^{d} \sum_{\ell=1}^{d} \mathsf{G}_i[k,\ell] \mathsf{X}_i \mathsf{Z}^{(k,\ell)}, \mathsf{Y}_i\right) + \beta \|\mathsf{Z}\|_* \quad (6)$$

where $\mathsf{G}_i := \mathsf{X}_i^\top \mathsf{X}_i$ *and* $\mathsf{Z}^{(k,\ell)} \in \mathbb{R}^{d \times c}$.

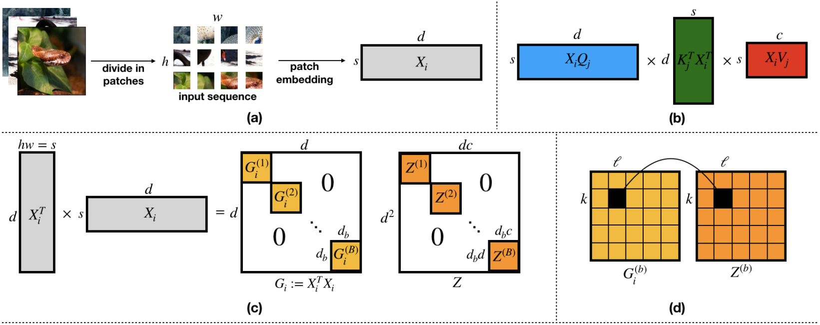

Figure 1: (a) Input image is first divided into $hw = s$ patches, where each patch is represented by a latent vector of dimension d. (b) The (non-convex) scaled dot-product self-attention applies learnable weights Q_j, K_j, V_j to the patch embeddings X_i. (c) In the equivalent convex optimization problem for the self-attention training objective, the Gram matrix G_i is formed that groups latent features in B different blocks, (d) and accordingly the nuclear norm regularization is imposed on the dual variables Z based on the similarity scores $G_i[k, l]$.

Conclusion

A similar procedure can be used to analyze self-attention with ReLU, as well as

- MLP-Mixer
- FNO
- A modification of FNO called block-FNO (BFNO)

with both ReLU and linear activations (see paper). In summary,

- Studied the vision transformer problem by finding convex equivalents to single attention blocks.
- These dual forms provide new interpretations, and provide better convex solvers than previously formulated