Transformers Trained via Gradient Descent Can Provably Learn a Class of Teacher Models

Chenyang Zhang[1] Qingyue Zhao[2] Quanquan Gu[2] Yuan Cao[1]

[1]The University of Hong Kong
[2]University of California, Los Angeles

ICLR International Conference On Learning Representations

Introduction

- Transformers achieve remarkable success across NLP, vision, and reinforcement learning. Recent theoretical works focus on validating their capacities in solve certain tasks.
- However, prior works usually study isolated tasks. Moreover, a central observation is that several statistical tasks considered in prior works share a common **bilinear structure**, which makes a unified analysis possible.

Our goal: establishing a unified learning framework of transformers towards a broad class of teacher models

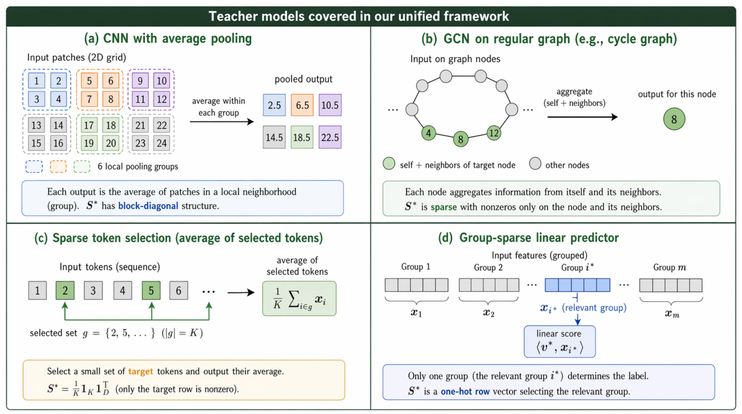
Problem setups

- **Unified bilinear teacher model**

$$f^*(\mathbf{X}) = \sigma(\mathbf{V}^*\mathbf{X}\mathbf{S}^*), \qquad \mathbf{X} \in \mathbb{R}^{d\times D},\ \mathbf{V}^* \in \mathbb{R}^{M\times d},\ \mathbf{S}^* \in \mathbb{R}^{D\times D}.$$

Each column of $\mathbf{S}^*$ has K nonzero entries, each equal to $1/K$.

- ▷ **Examples included in the theory**
  1. single convolutional layer with average pooling,
  2. single graph convolution layer on a regular graph,
  3. sparse token selection,
  4. group-sparse linear predictor.



- **Input with positional encodings**

$$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_D] = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_D \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_D \end{bmatrix} \in \mathbb{R}^{(d+D)\times D},$$

where $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_D] \in \mathbb{R}^{D\times D}$ is an orthogonal positional encoding matrix.

- **Simplified Position-only attention models**

$$\mathrm{TF}(\mathbf{Z}; \mathbf{W}_V, \mathbf{W}_{KQ}) = \sigma\left(\mathbf{W}_V\mathbf{X}\,\mathcal{S}\left(\frac{\mathbf{P}^\top\mathbf{W}_{KQ}\mathbf{P}}{\sqrt{D}}\right)\right) = \sigma(\mathbf{W}_V\mathbf{X}\mathbf{S}).$$

- ▷ This simplification keeps the actively trained blocks while removing empirically negligible ones, aligning with a line of theoretical works.

Main theoretical results

- **Training objective and algorithm**
  - ▷ Considering minimizing the population MSE, defined as

$$\mathcal{L}(\mathbf{W}_V; \mathbf{W}_{KQ}) = \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\|\mathbf{Y} - \mathrm{TF}(\mathbf{Z}; \mathbf{W}_V, \mathbf{W}_{KQ})\|_F^2\right],$$

where $\mathbf{Y} = f^*(\mathbf{X}) + \mathcal{E}$, and the entries of $\mathbf{X}, \mathcal{E}$ follows Gaussian distribution.

  - ▷ Optimize the objective by gradient descent from 0 initialization as

$$\mathbf{W}_V^{(t+1)} = \mathbf{W}_V^{(t)} - \eta\nabla_{\mathbf{W}_V}\mathcal{L}(\mathbf{W}_V^{(t)}, \mathbf{W}_{KQ}^{(t)}); \mathbf{W}_{KQ}^{(t+1)} = \mathbf{W}_{KQ}^{(t)} - \eta\nabla_{\mathbf{W}_{KQ}}\mathcal{L}(\mathbf{W}_V^{(t)}, \mathbf{W}_{KQ}^{(t)}).$$

- **Parameter recovery, training convergence, and O.O.D generalization**

Suppose $D \geq \Omega(\mathrm{poly}(M, K))$ and $\eta \leq O(M^{-1}D^{-5/2})$. Then for all $T \geq T^* = \Theta\left(\frac{KD^2}{\eta\|\mathbf{V}^*\|_F^2}\right)$, the following hold.

1. **Attention recovery**

$$\|\mathbf{S}^{(T)} - \mathbf{S}^*\|_F = \Theta\left(\frac{D^{5/2}\|\mathbf{V}^*\|_F}{\sqrt{\eta T}}\right).$$

2. **Value recovery**

$$\|\mathbf{W}_V^{(T)} - \mathbf{W}_V^*\|_F = \Theta\left(\frac{D^2\sqrt{K}}{\sqrt{\eta T}}\right).$$

3. **Tight optimization rate**

$$c\frac{KD^4}{\eta T} \leq \mathcal{L}(\mathbf{W}_V^{(T)}; \mathbf{W}_{KQ}^{(T)}) - \mathcal{L}_{\mathrm{opt}} \leq \bar{c}\frac{KD^4}{\eta T},$$

where $\mathcal{L}_{\mathrm{opt}}$ is the irreducible loss.

4. **O.O.D. generalization**

For any O.O.D. pair $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ with bounded second moments, any small ε, and $T > T_\varepsilon$,

$$\mathcal{L}_{\mathrm{OOD}}(\mathbf{W}_V^{(T)}; \mathbf{W}_{KQ}^{(T)}) \leq \frac{1}{2}\mathbb{E}\left[\|\widetilde{\mathbf{Y}} - f^*(\widetilde{\mathbf{X}})\|_F^2\right] + \varepsilon.$$

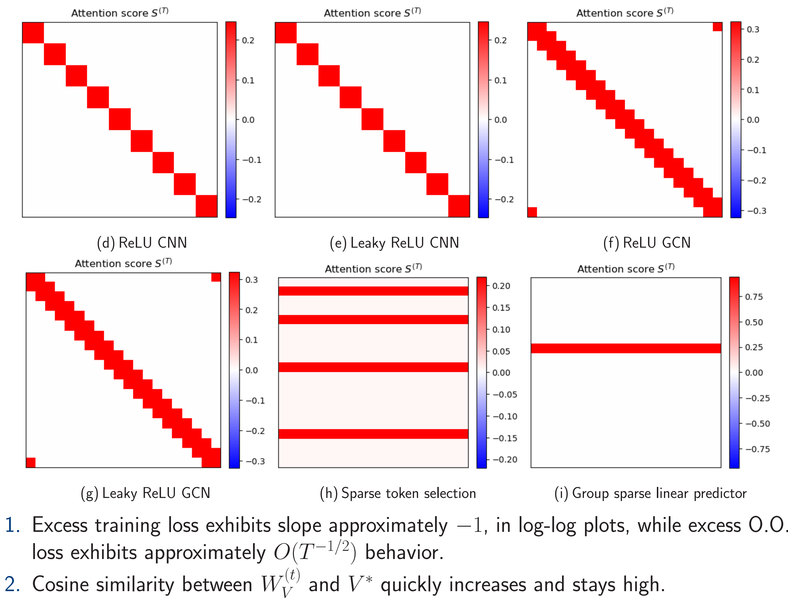
Synthetic experiments

- **Four kinds of teachers:** (i) CNN with average pooling, (ii) regular-graph GCN, (iii) sparse token selection, and (iv) group-sparse linear predictor.
  - ▷ CNN teacher: $D = 36$, $K = 4$, diagonal block softmax pattern.
  - ▷ GCN teacher: cycle graph with $D = 20$, $K = 3$, cyclic tridiagonal softmax pattern.
  - ▷ Sparse token selection / group-sparse predictor teacher: $D = 20$, with $K = 4$ and $K = 1$ respectively, and row-wise softmax pattern.
- **Training vs. O.O.D. data:** online gradient descent with fresh Gaussian batches for training and exponential-distribution inputs for O.O.D. test-time evaluation.
- **Excess training loss, excess O.O.D. test loss , and cosine similarity**

(a) Excess training loss (log-log) (b) Excess OOD test loss (log-log) (c) Cosine similarity

Synthetic experiments (continued)

- **Heatmap of attention score matrix $\mathbf{S}^{(T)}$**



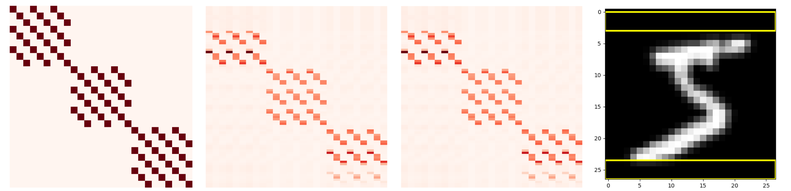
(d) ReLU CNN (e) Leaky ReLU CNN (f) ReLU GCN

(g) Leaky ReLU GCN (h) Sparse token selection (i) Group sparse linear predictor

1. Excess training loss exhibits slope approximately -1, in log-log plots, while excess O.O.D. loss exhibits approximately $O(T^{-1/2})$ behavior.
2. Cosine similarity between $W_V^{(t)}$ and V^* quickly increases and stays high.
3. Learned attention heatmaps match the target softmax structures for all synthetic tasks.

Real-data experiment: teacher CNN trained on MNIST

- **MNIST setup highlights**
  - ▷ Images are normalized and resized to 27×27, then partitioned into $D = 81$ patches.
  - ▷ Two-layer CNNs use $M = 16$ kernels. Kernel and average pooling size $d, K = 3 \times 3$.
  - ▷ After pre-training, the first convolution layer with pooling layer serves as the teacher.
- **Heatmaps of the average-pooling of teacher CNN, and learned attention scores**



(j) teacher CNN (k) ReLU result (l) Leaky ReLU result (m) Example image

1. Training loss rapidly converges to a small value for both ReLU and Leaky ReLU.
2. The cosine similarity between transformer value matrix and teacher convolution kernel matrix rises above 0.9.
3. Attention heatmaps recover the average-pooling pattern except near boundary patches that are mostly background in MNIST.