"Lossless" Compression of Deep Neural Networks: A High-dimensional Neural Tangent Kernel Approach

Lingyu Gu*[1] Yongqi Du*[1] Yuan Zhang[2] Di Xie[2]
Shiliang Pu[2] Robert C. Qiu[1] Zhenyu Liao†[1]

[1]EIC, Huazhong University of Science and Technology, Wuhan, China

[2]Hikvision Research Institute, Hangzhou, China

NeurIPS, 2022

*Equal contribution, †Corresponding author.

Motivation: Model Compression and Its Difficulties

- Deep Neural Network (DNN)
  - powerful framework
  - massive storage and computing consumption
  - over-parameterized
- Model Compression
  - compress DNN, maintain performance
  - pruning, quantization, knowledge distillation . . .

Difficulties

- limited theoretical understanding of DNN
- unclear the trade-off between performance and complexity

Motivation: Model Compression and Its Difficulties

- Deep Neural Network (DNN)
  - powerful framework
  - massive storage and computing consumption
  - over-parameterized
- Model Compression
  - compress DNN, maintain performance
  - pruning, quantization, knowledge distillation . . .

Difficulties

- limited theoretical understanding of DNN
- unclear the trade-off between performance and complexity

Understanding DNN model first!

Neural Tangent Kernel

- Neural Tangent Kernel (NTK) [JGH18]
  - the NTK matrix $\mathbf{K}_{NTK} = \mathbf{J}^\top \mathbf{J} = (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X))$
  - only depends on input data, network structure, and (law of) random initialization
  - characterizes convergence and generalization properties of network (via its **eigenspectrum**)

How Can NTK Help Compression?

- For high-dimensional Gaussian mixture data (number n and dimension p) and fully-connected multi-layer neural nets

Asymptotic spectral equivalence between $\mathbf{K}_{\rm NTK}$ and $\tilde{\mathbf{K}}_{\rm NTK}$

- For the NTK matrix $\mathbf{K}_{{\rm NTK},\ell}$ of layer ℓ, as $n, p \to \infty$, one has that

$$\left\| \mathbf{K}_{{\rm NTK},\ell} - \tilde{\mathbf{K}}_{{\rm NTK},\ell} \right\| \to 0,$$

- $\tilde{\mathbf{K}}_{{\rm NTK},\ell}$ with explicit expression.
- Proof via an induction on the layer $\ell = 0, 1, \ldots, L$.

How Can NTK Help Compression?

Explicit expression of $\tilde{\mathbf{K}}_{\mathrm{NTK},\ell}$

$$\tilde{\mathbf{K}}_{\mathrm{NTK},\ell} \equiv \beta_{\ell,1}\mathbf{X}^\top\mathbf{X} + \mathbf{V}\mathbf{B}_\ell\mathbf{V}^\top + \left(\kappa_\ell^2 - \tau_0^2\beta_{\ell,1} - \tau_0^4\beta_{\ell,3}\right)\mathbf{I}_n$$

with $\mathbf{V} \in \mathbb{R}^{n\times(K+1)}, \mathbf{t} \in \mathbb{R}^K, \mathbf{T} \in \mathbb{R}^{K\times K}, \tau_0$ some statistics for input data, and

$$\mathbf{B}_\ell \equiv \begin{bmatrix} \beta_{\ell,2}\mathbf{t}\mathbf{t}^\top + \beta_{\ell,3}\mathbf{T} & \beta_{\ell,2}\mathbf{t} \\ \beta_{\ell,2}\mathbf{t}^\top & \beta_{\ell,2} \end{bmatrix} \in \mathbb{R}^{(K+1)\times(K+1)},$$

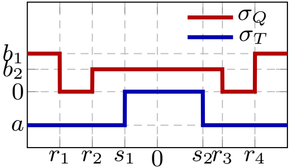
- depends on activations with only four parameters $\beta_{\ell,1}$, $\beta_{\ell,2}$, $\beta_{\ell,3}$, κ_ℓ
- independent of the distribution of weights (satisfying zero mean and unit variance)

How to Compress Weights and Activation Functions?

- Sparsity and Ternary Weights W with sparsity rate $\varepsilon \in [0, 1)$

$$[W]_{ij} = \begin{cases} 0 & p = \varepsilon \\ (1-\varepsilon)^{-1/2} & p = 1/2 - \varepsilon/2 \\ -(1-\varepsilon)^{-1/2} & p = 1/2 - \varepsilon/2 \end{cases}$$

- Quantized Activations



$$\sigma_T(t) = a \cdot (1_{t<s_1} + 1_{t>s_2}),$$
$$\sigma_Q(t) = b_1 \cdot (1_{t<r_1} + 1_{t>r_4}) + b_2 \cdot 1_{r_2 \le t \le r_3}.$$

Figure: Visual representations of activations σ_T and σ_Q.

Numerical Experiments

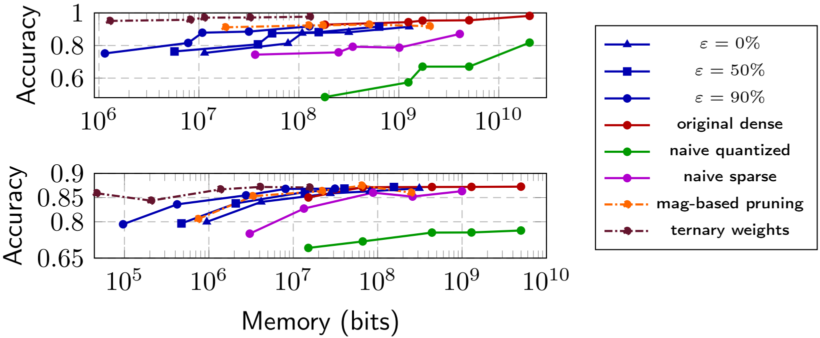


Figure: Classification accuracies of different compressed fully-connected nets on MNIST (**top**) and CIFAR10 (**bottom**) datasets. Blue curves represent the proposed compression approach with different levels of sparsity $\varepsilon \in \{0\%, 50\%, 90\%\}$, purple curves represent the heuristic sparsification approach by uniformly zeroing out 80% of the weights, green curves represent the heuristic quantization approach using the binary activation $\sigma(t) = 1_{t<-1} + 1_{t>1}$, red curves represent the original network, brown curves represent the proposed compression approach *without* activation quantization, with $\varepsilon = 90\%$ for MNIST (**top**) and $\varepsilon = 95\%$ for CIFAR10 (**bottom**), and orange curves represent magnitude-based pruning with the same sparsity level ε as brown. Memory varies due to the **change of layer width** of the network.

Conclusion and Outlook

- Conclusion
  - **Theoretical Result**: precise characterizations of the eigenspectra of NTK matrix
  - **Compression Algorithm**: sparsify and quantize fully-connected deep nets
- Outlook
  - apply asymptotic characterizations for NTK for some analysis for dynamics of fully-connected DNN models
  - extend to more involved settings, like convolutional nets

Reference I

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31 (2018).

Thank You!

And welcome to come to talk with us at (virtual) poster session for more details!