“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], and Zhenyu Liao+[1]

[1]EIC, Huazhong University of Science and Technology, Wuhan, China
[2]Hikvision Research Institute, Hangzhou, China

*Equal contribution ((gulingyu, yongqi_du)@hust.edu.cn)
+Corresponding author (zhenyu_liao@hust.edu.cn)

Introduction

Questions:

- modern deep neural networks (DNNs) are powerful
- however, require *massive* storage and computation
- DNN compression: to remove redundancy in the net
- *little* is known about DNNs, challenging to find redundancy
- understanding DNNs should be the first step! But how?

Neural tangent kernel helps!

Set up:

✓ input data $x_1, \dots, x_n \in \mathbb{R}^p$ drawn from a K-class Gaussian mixture model (GMM), $X \in \mathbb{R}^{p\times n}$, $p/n \to c \in (0,\infty)$.

✓ L-layer fully-connected network (width d_i for i-th layer) with weight matrices $W_1 \in \mathbb{R}^{d_1\times d_0}, \dots, W_L \in \mathbb{R}^{d_L \times d_{L-1}}$, output $f_\theta(x) \in \mathbb{R}$, and $\theta = (vec(W_1), \dots, vec(W_L), w)$

✓ activations $\sigma_1, \dots, \sigma_L$ at least four-times differentiable for the standard normal measure

Neural tangent kernel (NTK)

- NTK matrix $K_{NTK} = (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X)) \in \mathbb{R}^{n\times n}$
- only depends on input data, network structure, and (the distribution of) random initialization
- characterizes the convergence and generalization of networks (via its eigenspectrum) [2]
- builds a connection between network structure, input data, weights initialization, and network performance
- NTK can help us understand the DNNs!

Results

Theoretical Result: Asymptotic spectral equivalence for NTK matrices

With random matrix theory (RMT), for fully-connected network and high dimensional GMM data where the number of data n and their dimension p are both large ($n, p \to \infty$, $p/n \to c \in (0,\infty)$), we have, for NTK matrix $K_{NTK,\ell}$ of layer ℓ, that $\|K_{NTK,\ell} - \tilde{K}_{NTK,\ell}\| \to 0$, in which

$$\tilde{K}_{NTK,\ell} = \beta_{\ell,1} X^T X + V B_\ell V^T + (\kappa_\ell^2 - \tau_0^2 \beta_{\ell,1} - \tau_0^4 \beta_{\ell,3}) I_n$$

with $V \in \mathbb{R}^{n\times(K+1)}$, $B_\ell = \begin{bmatrix} \beta_{\ell,2} t t^T + \beta_{\ell,3} T & \beta_{\ell,2} t \\ \beta_{\ell,2} t^T & \beta_{\ell,2} \end{bmatrix} \in \mathbb{R}^{(K+1)\times(K+1)}$, and some statistics of input data τ_0, t, T.

As such, the NTK matrix

- depends on activations via *only* four parameters $\beta_{\ell,1}, \beta_{\ell,2}, \beta_{\ell,3}, \kappa_\ell$
- *independent* of the distribution of weights if they have zero mean and unit variance

✦ The precise form of the activation functions and the distribution of weights do not affect the spectrum of NTK!

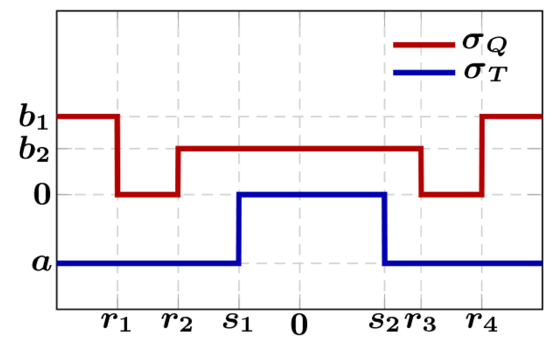
Compression Algorithm:

- Weights distribution

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

✓ both sparse and ternary
✓ of zero mean and unit variance
✓ freely choose sparsity level ε

- Activation function



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

✓ some coefficients to be determined so as to “match” *any* given DNN!

Experiments

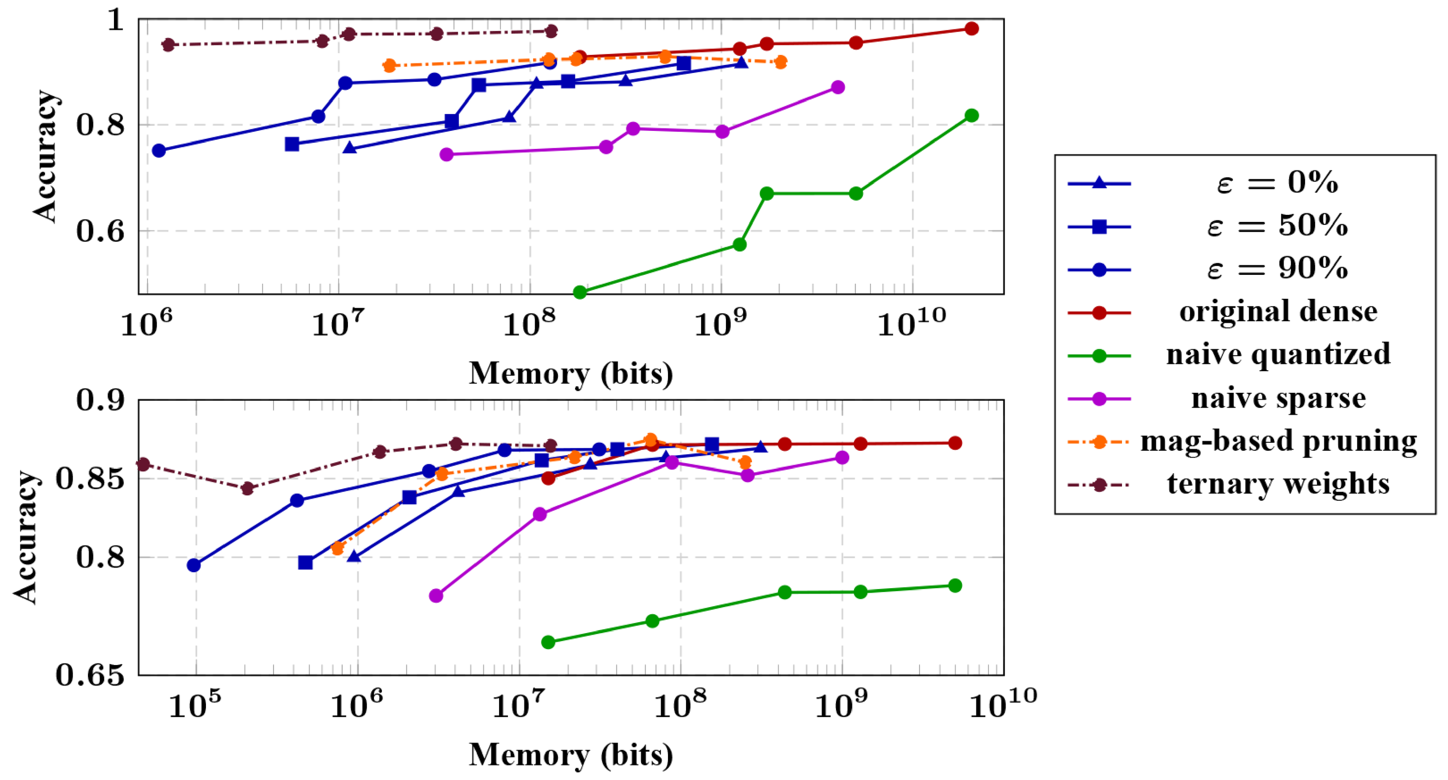


Figure: Classification accuracies of different compressed fully-connected nets on MNIST (top) and CIFAR10 (bottom) datasets

Compared to original or heuristically compressed nets (with, e.g., popular *magnitude-based* approach), the proposed “lossless” compression scheme (blue and brown)

- achieve comparable performance compared to some other compression methods.
- occupy (up to) a factor of 10^3 less memory
- produce significantly sparser networks (up to 90% of weights set to zero) with minimal performance loss

Outlook & Reference

Outlook:

- extend to more involved settings, e.g., convolutional nets
- apply asymptotic characterizations for NTK to analyze learning dynamics of ultra-wide fully-connected DNNs

Reference:

[1] Jacot Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." *Advances in neural information processing systems* 31 (2018).
[2] Fan Zhou, and Zhichao Wang. "Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks." *Advances in neural information processing systems* 33 (2020): 7710-7721.
[3] Lingyu Gu, Yongqi Du, Yuan Zhang, Di Xie, Shiliang Pu, Robert C. Qiu and Zhenyu Liao. "“Lossless” Compression of Deep Neural Networks: A High-dimensional Neural Tangent Kernel Approach." (accepted in) *Advances in neural information processing systems* 35 (2022)