Custom Expressivity without the Degeneracy
MAT1510H1F: Deep Learning: Theory & Data Science

Robert Wu rupert@cs.toronto.edu
Department of Computer Science
University of Toronto

1. Introduction

Initialization is an important consideration in deep learning (DL) to avoid vanishing/exploding gradients and local optima. Despite the long history of this topic, most techniques still incorporate *randomness*. Indeed, such delicate schemes therefore introduce variability and difficulty in reproducing experiments. Conversely, *deterministic* (i.e. non-random) schemes can directly address the latter issues but are often suboptimal or less generalizable. Finding the "sweet spot" remains an area of interest.

2. ZerO Initialization: Initializing Neural Networks with only Zeros and Ones

Zhao et al. (2022) introduced `ZerO`, a deterministic scheme that initializes layers of a neural network with a product of identity (depending on shape changes) and Hadamard transforms.

Definition 1 (Partial Identity) *The partial identity matrix* $\mathbf{I}^* \in \mathbb{R}^{l\times r}$ *is defined:*

$$\mathbf{I}^* := \begin{cases} (\mathbf{I}, \mathbf{0}), \text{where } \mathbf{I} \in \mathbb{R}^{l\times l}, \mathbf{0} \in \mathbb{R}^{l\times r-l} & \text{(contraction) if } l < r \\ \mathbf{I}, \text{where } \mathbf{I} \in \mathbb{R}^{l\times l} & \text{(passthrough) if } l = r \\ (\mathbf{I}, \mathbf{0})^T, \text{where } \mathbf{I} \in \mathbb{R}^{r\times r}, \mathbf{0} \in \mathbb{R}^{r\times l-r} & \text{(expansion) if } l > r \end{cases} \tag{1}$$

Definition 2 (Hadamard) *With* $\mathbf{H}_0 = 1$*, any Hadamard* $\mathbf{H}_m \in \mathbb{R}^{2^m\times 2^m}$ *is expressed recursively:*

$$\mathbf{H}_m := \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \mathbf{H}_{m-1} \in \mathbb{R}^{2^m\times 2^m} \tag{2}$$

The authors remark: **1)** adding random noise to **I** near-guarantees the layer is full-rank; **2)** simply applying $\mathbf{I}*$ upper-bounds every layer's expressivity to the rank of the input; this is sometimes referred to as feature symmetry and results in training degeneracy; and **3)** introducing the Hadamard transform $\mathbf{H}_m$ at expansion layers $W_{\text{expand}} \in \mathbb{R}^{l\times r}$ breaks said degeneracy.

$$W_{\text{expand}} \leftarrow c\mathbf{I}^*\mathbf{H}_m\mathbf{I}^*, \text{where } m := \lceil \log_2(l) \rceil \text{and } c := 2^{\frac{1-m}{2}} \tag{3}$$

While they used a multi-layer perceptron (MLP) on MNIST (LeCun et al., 1998) to validate their theorem, Zhao et al. (2022) also experimented on ResNet-18 and ResNet-50 (He et al., 2015) for CIFAR-10 (Krizhevsky, 2009) and ImageNet Deng et al. (2009), respectively. `ZerO` consistently outperformed Kaiming and Xavier initialization. On CIFAR-10 `ZerO` was on-par with or exceeded ReZero (Bachlechner et al., 2020), Fixup (Zhang et al., 2019), SkipInit (De and Smith, 2020), and ConstNet* (Blumenfeld et al., 2020). The authors also claim superior results for the Transformer on WikiText-2 (Vaswani et al., 2017).

In addition to improving reproducibility and allowing training ultra-deep networks without batch normalization (Zhao et al., 2022), `ZerO` exhibits a *monotonic low-rank learning trajectory*. The authors posit that it's a pioneering case of greedy low-rank learning (GLRL), whereby low rank parameters are maintained and slowly relaxed as needed during training to restrict model complexity.

3. Custom Expressivity via Pruning `ZerO`

Low rank is (weakly) correlated with sparsity; the latter can potentially reduce redundancy and be optimized at a systems-level. But they are roughly at odds with expressivity, a quality lacking in models trained with degenerate regimes. A lower-rank and/or sparser model might not learn as much from new data. This has implications for robustness (to outliers or out-of-distribution data) and transfer learning.

3.1 GLRL/Expressivity Tuning with Pruning

Because data is typically not homogeneous, it might be possible to prune `ZerO`-initialized matrices without losing too much expressivity. As an exploration, I conducted experiments by modifying the initialization schemes used by Zhao et al. (2022) in their MLP experiments on MNIST (LeCun et al., 1998). I came up with some pruning techniques for every weight matrix in the MLP (Figure 1). Most of the MNIST digits were roughly centred, which inspired me to target the outer edges of initialized matrices.

$$\begin{bmatrix} a & & & \\ & b & & \\ & & c & \\ & & & d \end{bmatrix} \rightarrow \begin{bmatrix} a & & & \\ & 0 & & \\ & & c & \\ & & & 0 \end{bmatrix}, \begin{bmatrix} 0 & & & \\ & b & & \\ & & 0 & \\ & & & d \end{bmatrix}, \begin{bmatrix} 0 & & & \\ & 0 & & \\ & & c & \\ & & & d \end{bmatrix}, \begin{bmatrix} a & & & \\ & b & & \\ & & 0 & \\ & & & 0 \end{bmatrix}, \begin{bmatrix} 0 & & & \\ & b & & \\ & & c & \\ & & & 0 \end{bmatrix}$$

Figure 1: 4 × 4 matrix → {even/odd-indexed, top-left/bottom-right corner, trim-around} pruning.

Zhao et al. (2022) used a four-layer MLP with input dimension $n_x = 784$ (MNIST images are 28×28) and hidden dimension $n_h = 2048$. The weights were of dimensions $W_1 \in \mathbb{R}^{n_x \times n_x}$, $W_2 \in \mathbb{R}^{n_x \times n_h}$, $W_3 \in \mathbb{R}^{n_h \times n_h}$, $W_4 \in \mathbb{R}^{n_h \times n_c}$ where $n_c = 10$ corresponds to classes of MNIST. All models were `ZerO`-initialized, pruned as below, and trained to a stable $96\% - 97\%$ test accuracy.

Regular Interval Pruning I attempted to sparsify the matrices with a regular interval (roughly uniform distribution) along the main diagonal. An interval of 2 means pruning even/odd-indexed entries (Figure 1). The results show the matrix ranks immediately increase, supposedly to a point of redundancy; I suspect this is perhaps due to the indiscriminate loss of information (Figure 2).

Corner Pruning Motivated by the many corners of some example images, I pruned entries from either corner on the main diagonal. Matrix rank appears to finely and monotonically correlate with number of entries pruned (Figure 3), implying that the the model can still learn expressively while remaining sparse.

Trim-Around Pruning Equivalent to corner pruning both the top-left and bottom-right corners. Similarly, rank roughly increases in relation to trim-around pruning (Figure 4). In fact, trim-around pruning might be more consistent compared to the assymetric results of corner pruning (Figure 3).

Due to MNIST's simple and centred structure, it appears pruning from the edges provides the best "tuning knob" to finely control expressivity/rank. Based on these experiments, custom-pruning weight matrices initialized with `ZerO` (or elsehow) can: **1)** take advantage of the deterministic nature (and benefits) of `ZerO`; **2)** increase sparsity where appropriate; **3)** provide fine-grained control over expressivity; **4)** be tailored and optimized to each dataset's structure; **5)** all while avoiding decreasing accuracy or exploding matrix rank. The result is a potentially powerful tuning tool for model efficiency/expressivity. However, generalization becomes a problem: the pruning technique must be engineered for each dataset and potentially each model. Furthermore, it's unlikely that the GLRL trajectory and tuning process are as trivial in training on more complex data. Work remains to be done to investigate this idea further.

References

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W. Cottrell, and Julian J. McAuley. Rezero is all you need: Fast convergence at large depth. *CoRR*, abs/2003.04887, 2020. URL https://arxiv.org/abs/2003.04887.

Yaniv Blumenfeld, Dar Gilboa, and Daniel Soudry. Beyond signal propagation: Is feature diversity necessary in deep neural network initialization? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 960–969. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/blumenfeld20a.html.

Soham De and Sam Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19964–19975. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e6b738eca0e6792ba8a9cbcba6c1881d-Paper.pdf.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Y. LeCun, C. Cortes, and C.J. Burges. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/. 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Residual learning without normalization via better initialization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gsz30cKX.

Jiawei Zhao, Florian Tobias Schaefer, and Anima Anandkumar. Zero initialization: Initializing neural networks with only zeros and ones. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=1AxQpKmiTc.

Appendix A. Results of Pruning

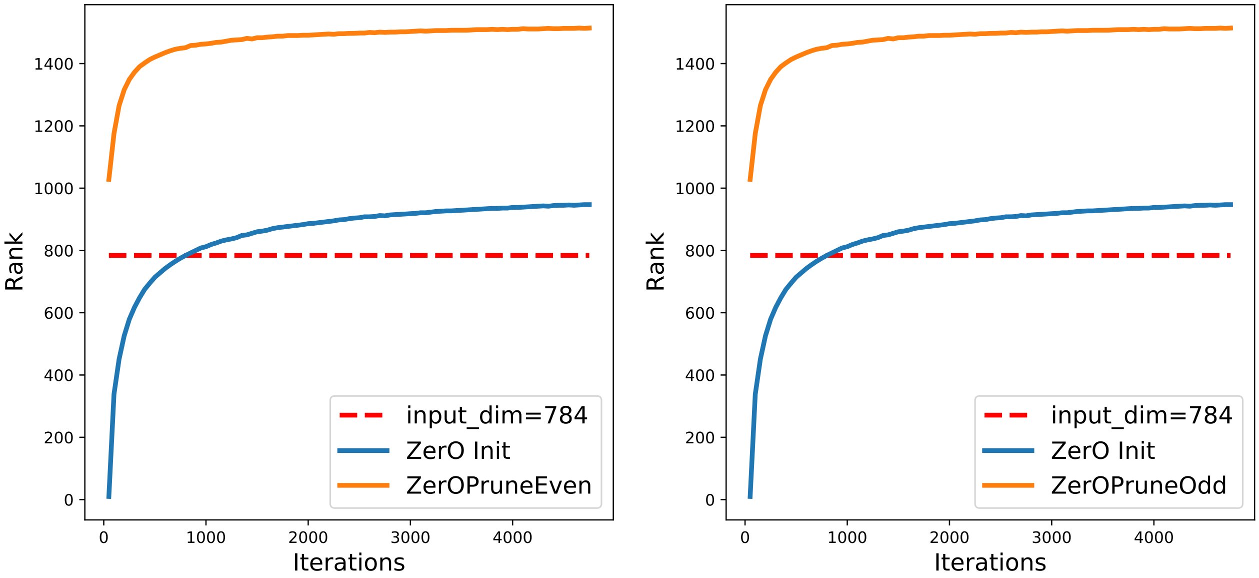

Figure 2: Results from regular interval (even/odd-indexed) pruning.

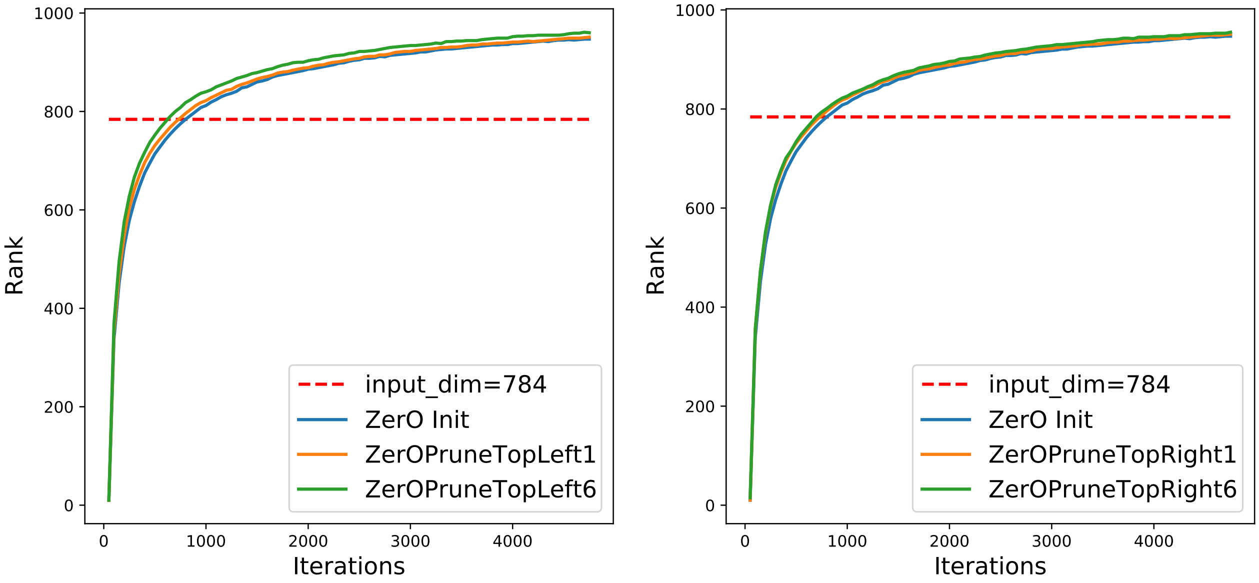

Figure 3: Results from corner (top-left/bottom-right) pruning.

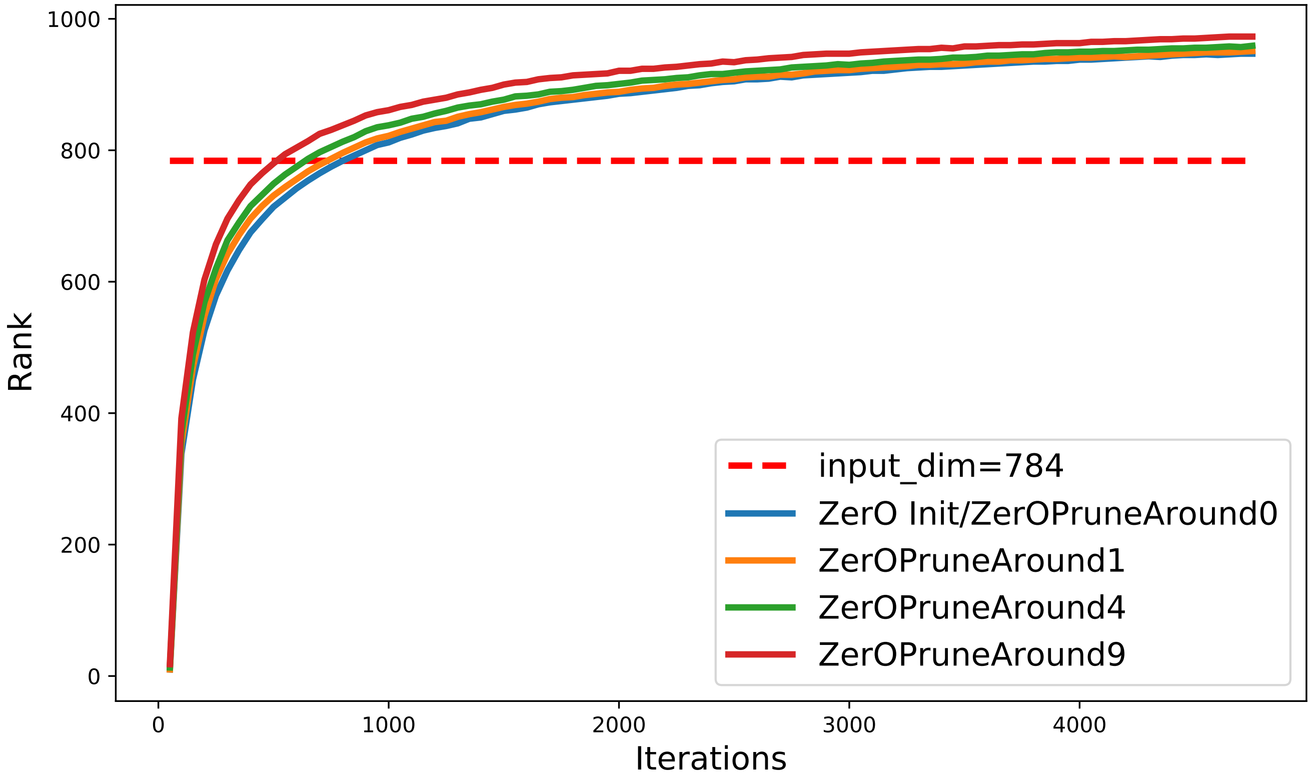

Figure 4: Results from trim-around pruning.