

Enhancing Accuracy in Deep Learning Using Random Matrix Theory

Leonid Berlyand¹, Etienne Sandier², Yitzchak Shmalo¹, and Lei Zhang^{* 3}

¹Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA.

²LAMA-CNRS UMR 8050, Université Paris-Est Créteil, Créteil 94010, France.

³Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.

Abstract. We explore the applications of random matrix theory (RMT) in the training of deep neural networks (DNNs), focusing on layer pruning that reduces the number of DNN parameters (weights). Our numerical results show that this pruning leads to a drastic reduction of parameters while not reducing the accuracy of DNNs and convolutional neural network (CNNs). Moreover, pruning the fully connected DNNs actually increases the accuracy and decreases the variance for random initializations. Our numerics indicate that this enhancement in accuracy is due to the simplification of the loss landscape. We next provide rigorous mathematical underpinning of these numerical results by proving the RMT-based Pruning Theorem. Our results offer valuable insights into the practical application of RMT for the creation of more efficient and accurate deep-learning models.

Keywords:

Deep learning,
Marchenko-Pastur distribution,
Random matrix theory,
Increasing accuracy,
Pruning.

Article Info.:

Volume: 3
Number: 4
Pages: 347 - 412
Date: December/2024
doi.org/10.4208/jml.231220

Article History:

Received: 20/12/2023
Accepted: 29/08/2024

Communicated by:

Zhi-Qin John Xu

Contents

1	Introduction	349
2	Background on deep learning	351
3	Numerical algorithm and experiments	352
3.1	Numerical algorithm	353
3.1.1	An overview of the Marchenko-Pastur (MP) distribution and its applications in machine learning	353
3.1.2	Using MP for pruning DNN weights	354
3.1.3	MP and Tracy-Widom distribution for DNN training	356
3.2	Numerical experiments	356
3.2.1	Training of fully connected DNNs on MNIST: Simplifying the loss landscape	357
3.2.2	MP-based pruning of CNNs on MNIST and Fashion MNIST	365

*Corresponding author. lzhang2012@sjtu.edu.cn

3.2.3	Numerics for training DNNs on CIFAR-10: Reducing parameters via MP-based pruning	368
4	Mathematical underpinning of numerical results	370
4.1	The classification confidence	370
4.2	How pruning affects classification confidence (for deterministic weight layer matrices)	371
4.3	Assumptions on the random matrix R and the deterministic matrix S	372
4.4	Key technical lemma: Removing random weights for DNN with arbitrary many layers does not affect classification confidence	375
4.5	Pruning Theorem for DNN with arbitrary many layers: How pruning random weights using PM distribution affects the classification confidence .	377
4.6	Simple example of DNN with one hidden layer	379
4.7	Pruning Theorem for accuracy: How pruning affects accuracy	381
A	Some known results on perturbation of matrices	383
A.1	Asymptotics of singular values and singular vectors of deformation matrix	383
A.2	Gershgorin's circle theorem	384
B	An approximation lemma – pruned matrix W' approximates the deterministic matrix S	385
B.1	Numerics for Example 4.2	388
B.2	Details for Example 4.3	389
C	Proof for Pruning Theorem	389
C.1	Proof for key technical Lemma 4.2	389
C.2	Proof of Pruning Theorem for accuracy	394
D	Other algorithms required for implementing RMT-SVD based pruning of DNN	394
D.1	BEMA algorithm for finding λ_+	394
D.2	The role of singular value decomposition in deep learning	397
D.3	Eliminating singular values while preserving accuracy	397
D.4	MP fit criteria: Checking if the ESD of X fits a MP distribution	400
E	Some of the proofs and numerics	401
E.1	Proof of Lemma 4.1	401
E.2	Effectiveness of MP-based pruning for different initialization methods . . .	403
E.3	A regression problem: MP-based pruning in regression	404
E.4	Numerical example used to calculate δX	406
E.5	Hyperparameters for Section 3.2.1	407
E.6	Hyperparameters for Section 3.2.1	407
E.7	Hyperparameters for Section 3.2.2	408
E.8	CNN architecture description	408
E.8.1	Pooling and regularization details	409
E.9	The hyperparameters for Section 3.2.3	409