AL2: Progressive Activation Loss for Learning General Representations in Classification Neural Networks

Majed El Helou, Frederike Dümbgen and Sabine Süsstrunk

IEEE International Conference on Acoustics, Speech, and Signal Processing

ICASSP 2020

Introduction

- Deep neural networks achieve increasingly-better results on a wide range of tasks.
- As the capacity of deeper networks increases, so does their potential to memorize[1].
- In turn, increased memorization is detrimental to network performance and generalization.

[1] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al., “A closer look at memorization in deep networks,” in ICML, 2017, pp. 233–242.

Motivation

- Larger, varied training sets can improve generalization, but increase training time and are expensive and time-consuming to collect.
- Neural network regularization is a valuable alternative that remains an open problem[1,2].
- In this work, we address neural network regularization.

[1] M. Blot, T. Robert, N. Thome, and M. Cord, “Shade: Information-based regularization for deep learning,” in ICIP, 2018, pp. 813–817.
[2] X. Li, S. Chen, X. Hu, and J. Yang, “Understanding the disharmony between dropout and batch normalization by variance shift,” in CVPR, 2019, pp. 2682–2690.

Background: Regularization

- Regularization methods are commonly used to reduce network overfitting:

1. **Batch Normalization** (BN): attempts to stabilize the output of one layer to aid the learning of the following one
2. **Dropout** (DO): attempts to increase robustness by forcing random signal ablations during training
3. **Weight Decay** (WD): reduces network complexity by penalizing the norm of some or all optimization weights

- It is recently shown that BN and DO actually have opposite effects on feature variance between training and inference[1].

[1] X. Li, S. Chen, X. Hu, and J. Yang, "Understanding the disharmony between dropout and batch normalization by variance shift," in CVPR, 2019, pp. 2682–2690.

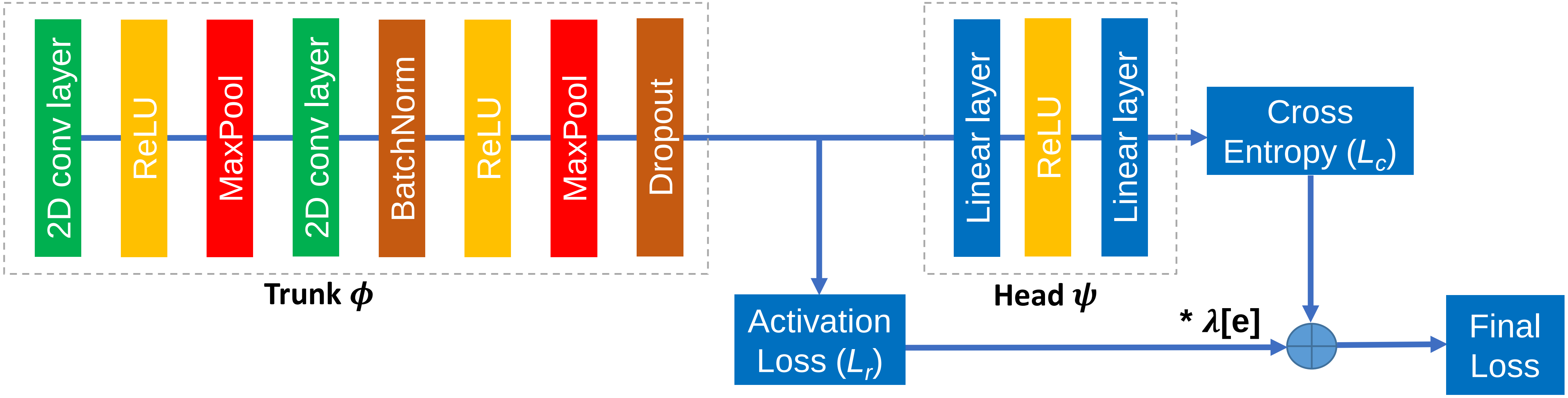
Background: Generalization

- Generalization in neural networks remains an open question[1].
- To assess the quality of feature representations learned by a network, we evaluate memorization.
- This is achieved by training with a portion of randomized class labels, which can only be predicted/learned by memorization [2].

[1] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, "Towards understanding the role of over-parametrization in generalization of neural networks," in ICLR, 2019.
[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in ICLR, 2017.

Proposed Method: AL2

- The network architecture is separated into a trunk ϕ for feature learning followed by a head ψ for classification:



$$\mathcal{L}_e(x, y; \Theta) = \sum_{x \in \mathcal{B}} \mathcal{L}_c(\psi(\phi(x)), y; \Theta_c) + \lambda_e \mathcal{L}_r(\phi(x); \Theta_r)$$

Proposed Method: AL2 Cont'd

- The final mini-batch loss is given by: $\mathcal{L}_e(x, y; \Theta) = \sum_{x \in \mathcal{B}} \mathcal{L}_c(\psi(\phi(x)), y; \Theta_c) + \underbrace{\lambda_e \mathcal{L}_r(\phi(x); \Theta_r)}_{\text{AL2}}$

- And the activation loss is given a progressively increasing weight, based on recent findings[1,2] in network learning:

$$\lambda_e = \lambda_{e-1} * (1.1 * u[5 - \lambda_{e-1}] + 1.01 * u[\lambda_{e-1} - 5])$$

- We begin with a value of 0.01 for the weight, and the sequence λ is the same for all our experiments.

[1] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in NeurIPS, 2018, pp.8527–8537.
[2] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in CVPR, 2018, pp. 9446–9454.

Experimental Results

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	84.20/95.25	45.30/94.92	25.25/93.07	23.83/88.76	26.07/79.64	26.45/75.88	25.84/**68.46**
	$\mathcal{L}_c$	2.15/2.22	1.78/2.19	0.89/2.15	0.19/2.11	0.04/2.08	0.01/2.07	0.00/2.08
	$\mathcal{L}_r$	3.20/0.24	10.93/0.10	26.12/0.06	54.42/0.03	74.49/0.02	103.26/0.01	119.10/0.00
BN [7]	TA	74.72/95.47	36.65/94.48	26.72/90.20	25.97/85.34	25.88/83.02	25.60/81.53	25.55/**81.16**
	$\mathcal{L}_c$	2.07/2.22	1.48/2.19	0.30/2.15	0.04/2.12	0.01/2.11	0.01/2.12	0.01/2.14
	$\mathcal{L}_r$	0.84/0.24	2.35/0.10	6.46/0.06	9.25/0.03	10.40/0.01	11.06/0.01	11.51/0.00
DO [8]	TA	96.13/94.43	96.47/95.03	95.93/95.03	92.74/94.79	81.96/92.15	68.12/92.69	55.39/**91.70**
	$\mathcal{L}_c$	2.22/2.23	2.20/2.22	2.17/2.20	2.13/2.20	2.05/2.20	1.94/2.21	1.79/2.23
	$\mathcal{L}_r$	0.26/0.24	0.30/0.09	0.41/0.04	0.61/0.02	1.00/0.01	1.50/0.00	1.92/0.00
WD [9]	TA	88.91/95.21	50.87/95.47	27.98/95.17	27.66/94.03	25.14/91.42	28.05/89.81	25.57/**86.98**
	$\mathcal{L}_c$	2.16/2.22	1.87/2.20	1.06/2.18	0.32/2.16	0.07/2.16	0.04/2.17	0.02/2.19
	$\mathcal{L}_r$	2.94/0.23	10.52/0.09	26.04/0.05	53.65/0.02	81.53/0.01	84.64/0.00	107.80/0.00

- Test accuracy (TA), training cross-entropy loss (L_c), and our regularization loss (L_r) (shown for AL2 multiplied by 100 for readability), on the MNIST dataset with 75% corrupt labels.
- We note the counter-intuitive effect of WD on activation values.

Feature Representation Analysis

- We analyze the evolution of feature representations with canonical correlation, which is based on the coefficients:

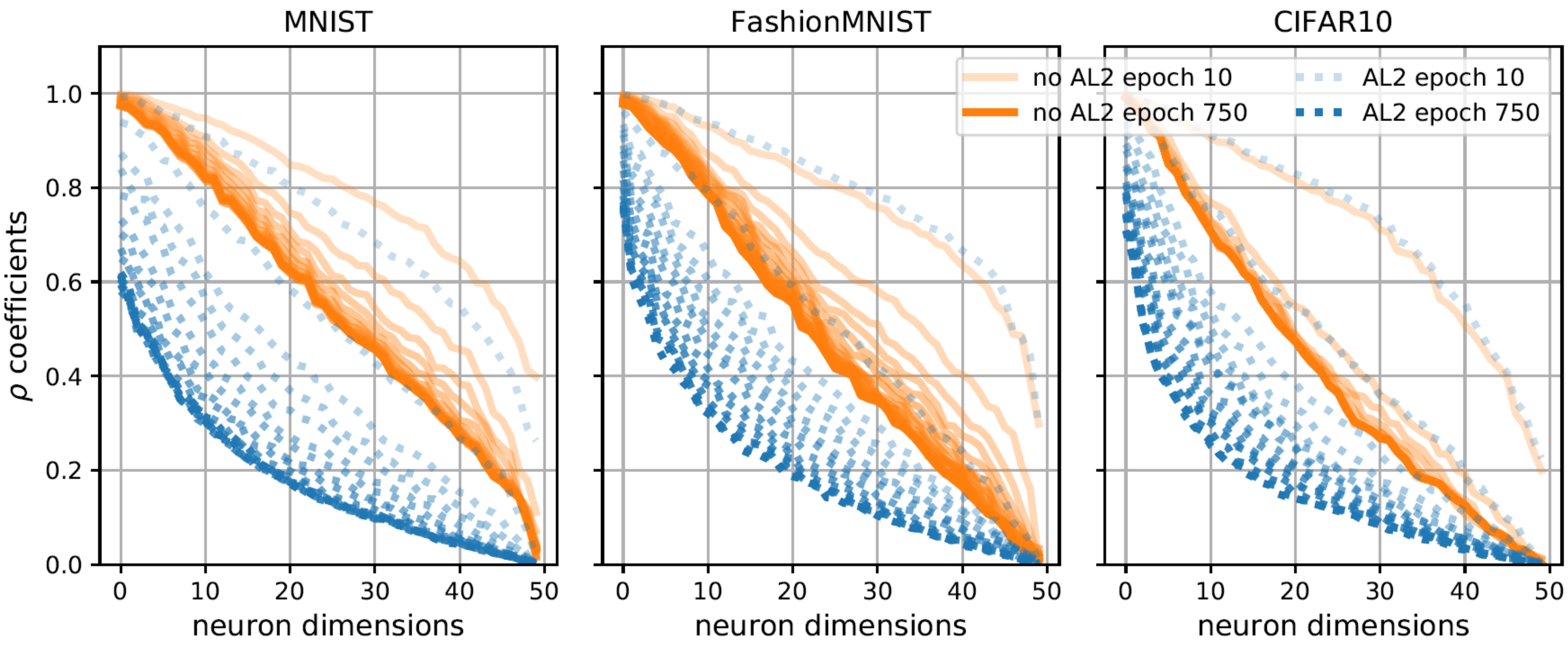
$$\rho = \max_{(\omega_1, \omega_2) \in (\mathbb{R}^a, \mathbb{R}^b)} \left(\frac{\langle \omega_1^T R_1, \omega_2^T R_2 \rangle}{||\omega_1^T R_1|| \cdot ||\omega_2^T R_2||} \right)$$

(where R_1 and R_2 hold feature activations per neuron and data sample)

- SVCCA[1] computes a weighted average of these coefficients, to assess the difference between two feature representations.
- To not lose any information, we visualize the entire sequences.

[1] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in NeurIPS, 2017, pp. 6076–6085.

Feature Representation Analysis Cont'd



- The baseline here is the network with DO.
- By analyzing the evolution in feature space across epochs, we see that AL2 significantly modifies the learning and the final learned representations.

Cumulative Ablations Analysis

- We further test the different networks by evaluating their average performance as we ablate increasing percentages of their feature activations during inference[1].

Area under cumulative ablation curve (/100) evaluated across training epochs (without/with AL2)

Baseline	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	35.44/77.81	19.17/72.67	15.52/69.44	14.73/64.11	15.36/55.08	15.36/51.73	15.19/**47.65**
BN [7]	35.08/77.01	19.17/71.23	15.80/63.48	15.79/57.42	15.64/55.67	15.69/54.97	15.60/**54.96**
DO [8]	81.66/78.52	79.90/78.74	76.23/78.80	70.38/78.31	60.17/73.57	49.86/73.30	41.39/**71.61**
WD [9]	39.50/78.18	20.74/74.83	15.94/74.39	16.09/72.97	15.40/67.62	16.12/64.85	15.35/**62.63**

[1] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick, "On the importance of single directions for generalization," in ICLR, 2018.

Conclusion

- We propose a novel progressive activation loss (AL2) to regularize neural networks.
- We use canonical correlation analysis to show the significant effect of AL2 on the learned feature representation.
- All results show that better performance can be obtained by combining standard regularization methods.

Thank you

https://github.com/majedelhelou/AL2