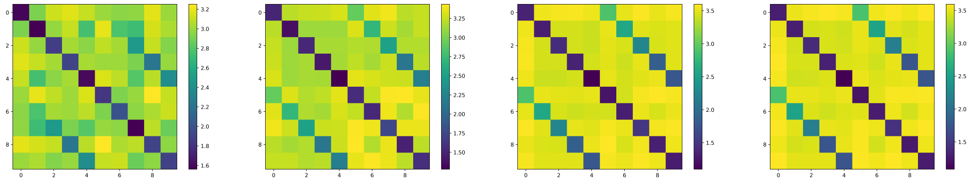
Rebuttal for "Are Neurons Actually Collapsed? On the Fine-Grained Structure in Neural Representations"

Anonymous Authors[1]

I. Experiments on SVHN and FashionMNIST

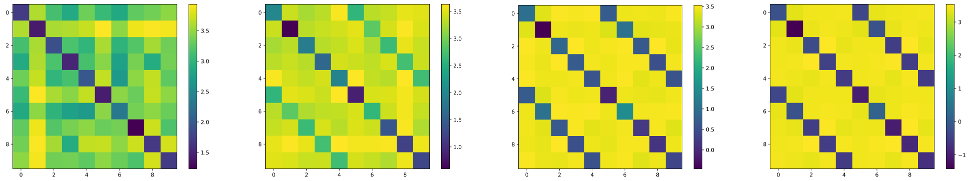
In this section we provide extended experiment results on SVHN and FashionMNIST, both of which are used in the original Neural Collapse paper (Papyan et al., 2020). Given limited time of rebuttal, here we reproduce our experiments in Section 4.1 and 4.2 of the main paper with ResNet-18 under one set of hyper-parameters. For FashionMNIST we use the same hyper-parameter setting with the main paper (learning rate = 0.1, weight decay = 5×10^4) while for SVHN we lower the weight decay to 5×10^{-5} since under the original weight decay the training doesn't reach 100% accuracy. We will include the full result in the revised version.

The results of distance matrices (introduced in Section 4.1) of the two new dataset are plotted in Figures 1 and 2. The t-SNE visualization results (introduced in Section 4.2) are plotted in Figures 3 and 4. It is clear that the added experiments supports our findings in the main paper.



(a) epoch = 20 (b) epoch = 120 (c) epoch = 240 (d) epoch = 350

Figure 1. The heatmap of class distance matrices on SVHN.



(a) epoch = 20 (b) epoch = 120 (c) epoch = 240 (d) epoch = 350

Figure 2. The heatmap of class distance matrices on FashionMNIST.

J. Per Class Linear-Probe Accuracy

In this section, we perform linear probe with original labels which was introduced in Section 4.3 and report per-class test accuracy v.s. the per-class mean square distance ratio (MSDR). MSDR was introduced in Section B of the appendix of the

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

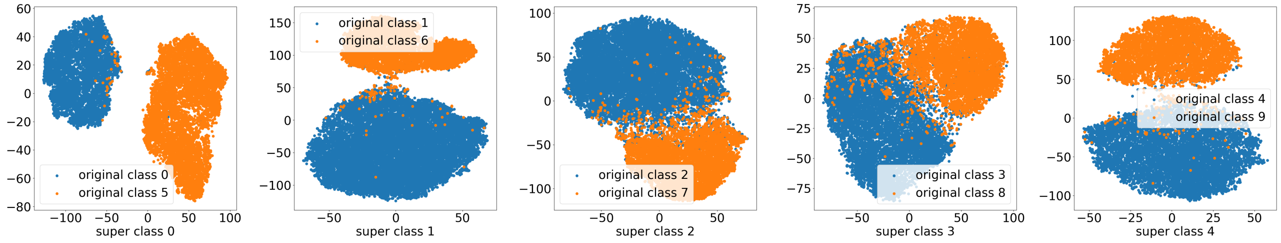


Figure 3. The t-SNE visualization result of representation learned on SVHN.

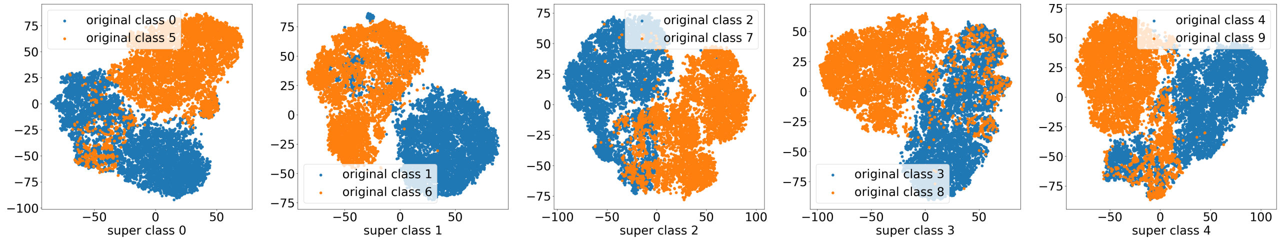


Figure 4. The t-SNE visualization result of representation learned on FashionMNIST.

original paper. Here we calculate MSDR within each super-class. A lower the per-class MSDR indicates a more collapsed representation while a higher per-class MSDR indicates a clearer fine-grained structure within the super-class. The results is displayed in Figure 5. We also perform a linear regression on the data points (shown as the dahsed line in the figure). We observe a positive correlation between linear probe accuracy and MSDR.

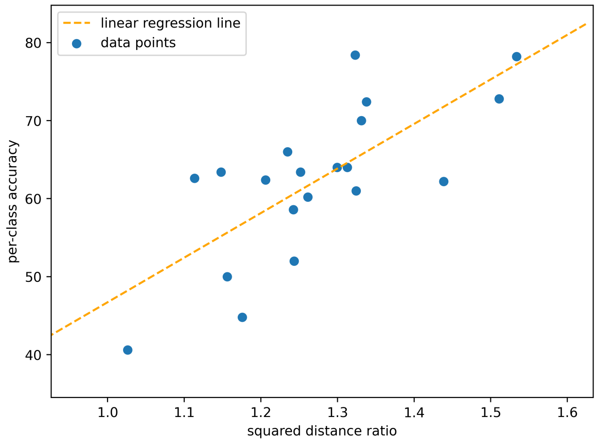


Figure 5. The per-class Linear Probe accuracy v.s. per-class mean square distance ratio.

References

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.