Using deep neural networks for modeling representational spaces: the prevalence and impact of rarely-firing nodes

Nhut Truong (leminhnhut.truong@unitn.it)
Center for Mind/Brain Sciences (CIMeC), University of Trento

Uri Hasson (uri.hasson@unitn.it)
Center for Mind/Brain Sciences (CIMeC), University of Trento

Keywords: human similarity judgment, deep neural networks, representational similarity analysis, pruning

Background

Deep neural networks (DNNs) are increasingly being used as computational models of human vision and higher-level cognition. Many studies have shown that after training these networks to categorize objects, the latent representations they form, quantified via image-similarity in multidimensional space, moderately approximate those produced by human similarity judgments. For example, Peterson, Abbott, and Griffiths (2018) showed that it is possible to improve the prediction of human similarity spaces from DNN embeddings by learning a reweighting of the saliency of each feature (or node). This suggests that DNNs learn relevant features for modeling human knowledge, but those features have the wrong level of saliency, which can be adjusted via reweighting.

Recently, Tarigopula, Fairhall, and Hasson (2021) have shown that it is possible to improve prediction of human similarity judgments not via reweighting, but via supervised pruning of DNN models. Pruning outperforms reweighting in learning human similarity spaces. Because pruning does not alter the original activations of retained features, its success suggests that DNNs may learn a relevant basis function at adequate levels of salience, but that only a subset of features is relevant when modeling human representational space.

Current study

While the work of Tarigopula et al. (2021) used *supervised* pruning, in this work we examine to what extent we can achieve the same goal with an *unsupervised* method. Our work was inspired by a study by Hu, Peng, Tai, and Tang (2016). In their work, they show that in a trained DNN a substantial subset of nodes do not activate for the majority of stimuli, with some nodes not firing for over 90% of all images; moreover, removing such nodes has little impact on the network classification accuracy. In our work, we investigated how the removal of infrequently-activated nodes impacts the representational space of DNNs, and how useful they are for modeling human similarity spaces. For each node we computed the percentage of images in the dataset for which the node's activation was 0. We call this node-wise measure the Percentage of Zeros (*PoZ*) as in Hu et al. (2016).

In **Experiment 1** we trained LeNet5, a small DNN, to classify the CIFAR-10 dataset. The dataset consists of 60000 small images drawn from 10 object categories. We then extracted the representations for 10000 test images from the penultimate layer (containing 84 nodes) to obtain a matrix size of 10000×84. From each matrix we computed a Baseline Representational Similarity Matrix (baseline RSM) from the average representations of each category, and the *PoZ* of each feature sorted from highest to lowest. We then iteratively removed 10%, 20%, ..., 90% of features according to their *PoZ* ranking, each time 1) recomputing an RSM from the pruned network, and computing the match between the pruned RSM and baseline RSM (quantified by Pearson correlation R^2 fit between the two RSMs, a.k.a representational similarity analysis); and 2) storing the maximum *PoZ* value in the remaining features. We repeated the entire process 50 times to start from different initialization positions to obtain means and standard deviations for the two measurements.

As Figure 1 (blue line) shows, keeping the bottom 80% of *PoZ*-rank features had almost no impact on R^2, with values remaining very close to 1. A sharp drop only occurs once 30% of features and less are retained. It can also be seen (yellow line) that some features have *PoZ* values nearing 100%, and that, e.g., when ranked by *PoZ*, the top-ranked 20% features all had $PoZ > 60\%$. The findings show that even for a relatively heterogeneous dataset, there is a substantial subset of features with mostly-zero firing, which contributes minimally to model the similarity space.

In **Experiment 2** we applied *PoZ*-based pruning to a more realistic dataset, but here we examined whether non-supervised *PoZ*-based pruning can improve the match between RSMs produced from a DNN and RSMs produced from human similarity judgments. The dataset included images from six different categories (Animals, Automobiles, Fruits, Furniture, Vegetables and Various), each consisting of 120 images. Human similarity matrices were obtained for all image-pairs within each category and provided to us by Peterson et al. (2018). For each set of 120 images we obtained DNN embeddings from the penultimate layer of the Pytorch ImageNet-pretrained VGG-16, computed and ranked the *PoZ* of each node, and then iteratively removed nodes based on *PoZ* ranking. After each removal we quantified the fit between the human RSM and the RSM for the DNN

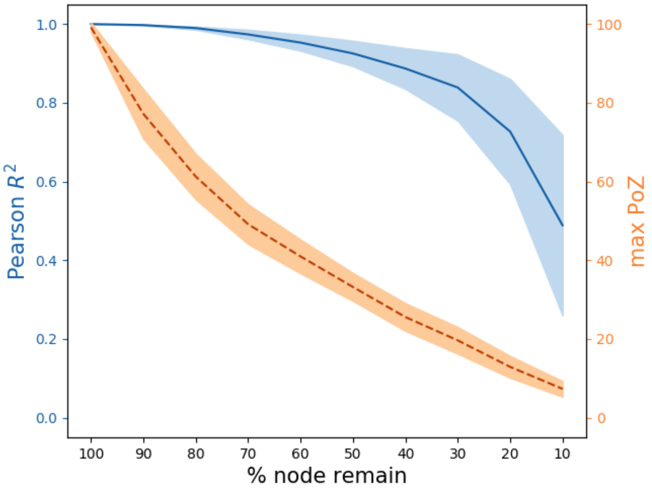

Figure 1: **Experiment 1**. Impact of PoZ-based pruning on representational space. Blue line: match between baseline RSM and RSM of each pruning level. Yellow line: maximal PoZ value remaining in the set at each level of pruning.

pruned at that level.

Figure 2 presents the *PoZ* distribution per category. It shows that for all categories, more than 50% of features had $PoZ > 80\%$. Consistent with this observation, Figure 3 shows that for all categories, the large majority of nodes could be removed with very little impact on the fit between the human and DNN RSMs. A substantial drop only occurred when less than 12% of the features were retained. We also found that for three categories, at least one pruned RSM provided a better fit to human judgements than an RSM computed from the non-pruned network. Significance testing showed that this pattern departed from chance for the Furniture and Vegetables category, where the fit between the DNN and Human RSM improved linearly till 24% and 41% respectively of the features remained.

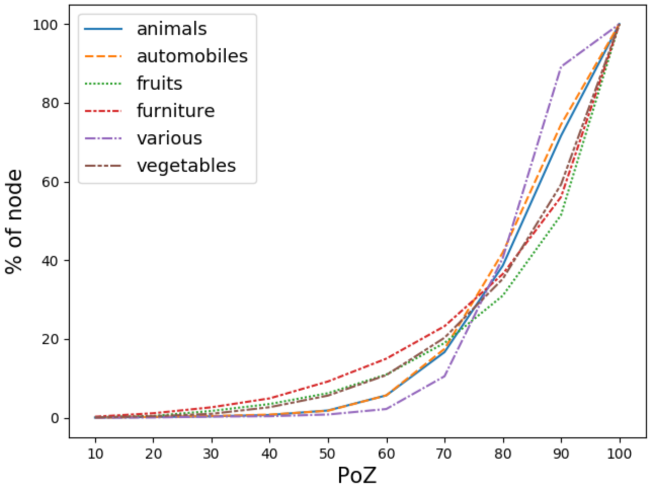

Figure 2: **Experiment 2**. Cumulative histograms of PoZ values computed for embeddings of each image category.

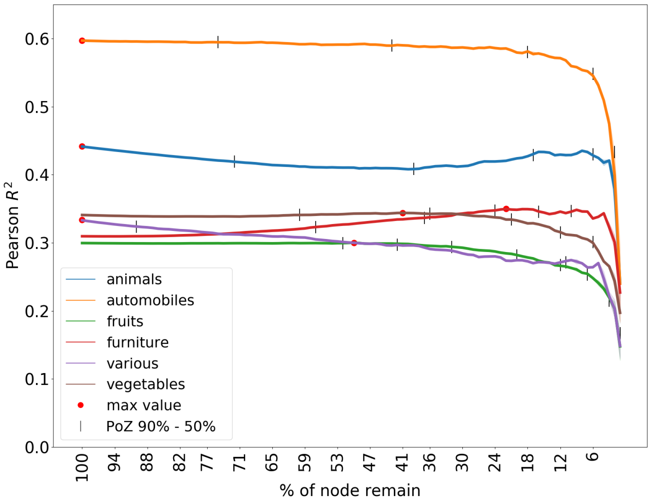

Figure 3: **Experiment 2**. Fit between RSMs from DNN embeddings and human RSM, for different levels of pruning. Red dots mark point of maximal fit per category. Small vertical lines mark quantiles of PoZ values per category.

Discussion: Experiment 1 showed it was possible to remove all nodes with $PoZ > 50\%$ with minimal impact on representational space. This was unexpected: removal of nodes with very high PoZ values should obviously not impact representational space, but the reason for why removal of lower PoZ nodes had a similarly-weak impact requires further study. Experiment 2 generalized the results to a more extensive, realistic dataset and suggested that PoZ-based pruning of DNN embeddings can in some cases improve the fit with human similarity judgments. Overall, our findings suggest that high-PoZ nodes are weakly-informative, and prevalent in image sets of natural categories. We suggest these nodes should be considered as a separate class when constructing encoding or decoding models of human cognition.

Reproducibility

The code to produce results and figures is available at github.com/tlmnhut/DNN_model_sim_space

References

Hu, H., Peng, R., Tai, Y.-W., & Tang, C.-K. (2016, Jul). Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv:1607.03250 [cs]*.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018, Nov). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669. doi: 10.1111/cogs.12670

Tarigopula, H. P., Fairhall, S. L., & Hasson, U. (2021). Improved prediction of behavioral and neural similarity spaces using pruned dnns. *bioRxiv*.