The Impact of Rarely-firing Nodes in Neural Networks on Representational Geometry and Predictions of Human Similarity Judgments

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

Anna Bavaresco (anna.bavaresco@unitn.it)
Center for Mind/Brain Sciences, University of Trento
Rovereto, Trento, Italy

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

Abstract

Vision deep neural networks (DNNs) learn many sparse features that are activated rarely or not at all. It is however unknown how these features structure the representational geometry of a DNN's embedding space. Previous research used supervised iterative-magnitude pruning ("lottery-ticket") to remove less-important features, and concluded that even minor feature-pruning strongly alters layer representations. Here we investigate DNN's representational geometry, but using an unsupervised approach, where pruning is guided by how frequently a node is inactivated across samples. Using representational similarity analysis, we find that for CIFAR-10 and MNIST, 20% of the features can be removed without any impact on the representational space. However, these redundant features do contain distributed information: when used alone, they account for 10%-50% of the variance in the non-pruned embeddings. Additionally, we find that for some natural image-sets, the removal of sparse features improves the prediction of human similarity judgments. Finally, we show that for a given set of images belonging to an object category, never-activated features encode meaningful semantics that is irrelevant for representing the category. Overall, our findings contribute to the understanding of how sparse features shape objects' representations in DNNs and how they impact their effectiveness as a model of human behavior.

Keywords: deep neural networks, features, pruning, representational geometry, human similarity judgments

Introduction

DNNs trained for image classification learn many sparse features which seldom activate, with some never activating for any image ("ghost" features). For example, up to 70% of VGG-16 model's features trained on CIFAR-10/100 do not activate for any image in the training set (Mehta, Kim, & Theobalt, 2019). Removing infrequently-firing nodes by computing their Percentage of Zeros (PoZ) statistic over a training batch is an effective pruning strategy (Hu, Peng, Tai, & Tang, 2016).

The impact of ghost features and high-PoZ nodes on a trained model's representational geometry is currently unknown. Computing representational dissimilarity matrices (RDMs) is a common approach to characterize a neural network's geometry, capturing all pair-wise similarities between objects (Kriegeskorte, Mur, & Bandettini, 2008). While the removal of high-PoZ features minimally affects or even improves classification (Hu et al., 2016), retaining them in image embeddings can significantly impact the pairwise similarity values. Consider three images (A, B, C) with hypothetical embeddings below, where node #6 is always 0. The full embeddings produce pairwise Pearson similarity values of $0.62, 0.56, -0.18$ for (A,B; B,C; A,C). However, removing feature #6 yields values of $-0.06, 0.40, -0.94$. Thus, $Sim(A,B) > Sim(B,C)$ with the full embeddings, but $Sim(A,B) < Sim(B,C)$ without the 0 feature.

Image A :	0.91	0.76	0.3	0.7	0.9	0
Image B :	0.4	0.7	0.6	0.3	0.7	0
Image C :	0.02	0.4	0.9	0.2	0.2	0

This will affect any analysis in which RDMs computed from DNN features are used as models of human behavior, for example, when used as models of human similarity judgments. Psychologically, removing a feature that is non-activated for all images should not impact perceived similarity, however, it impacts RDM values. Pruning these features prior to constructing RDMs may improve the prediction of human similarity judgments, and compress the embedding space itself.

In our work, we investigate the following questions: 1) How does pruning high and low-PoZ features impact a DNN's representational geometry? 2) Are pruned DNNs better or worse models of human behavior? and 3) Can identifying high-PoZ features be used to explain what dimensions are *not* relevant for coding differences within a semantic category?

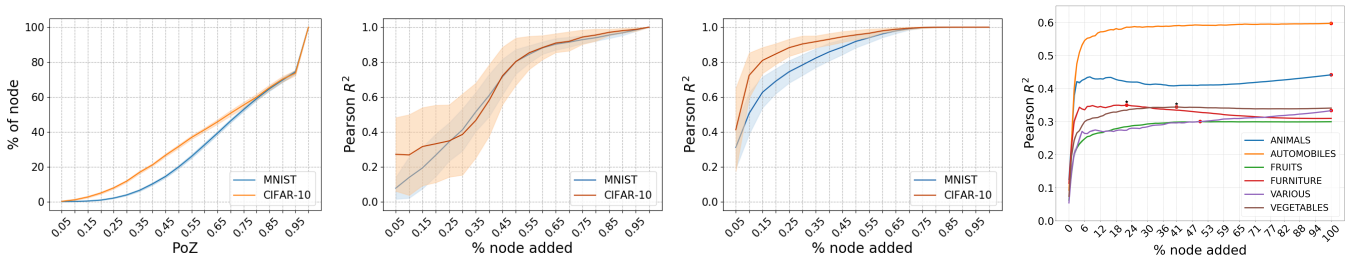
Prior work (Ansuini, Medvet, Pellegrino, & Zullich, 2020) used supervised pruning to address the first of our questions, concluding that: "pruning, even at small rates, produces layer representations which are different from the unpruned network ones". Similar findings were reported by Blakeney, Yan, and Zong (2020) and one of their findings was that as pruning progresses through training, deeper layers undergo substantial transformations in representation, which are then maintained throughout the rest of the training.

Experiments and Results

Study of the representational geometry

Fifty instances of LeNet5, a small DNN model, were trained on MNIST and CIFAR-10 with Adam optimizer until they converged on typical accuracy levels. The models were then used to generate embeddings for out-of-sample data using post-ReLU activations from the penultimate layer. Figure 1a shows a PoZ histogram, indicating that about 40% of nodes have a $PoZ > 80\%$, meaning that they are mostly zeros.

To quantify how a feature's PoZ impacts its contribution to representational geometry, we sorted the features from high to low PoZ. We then incrementally added features from high- to low-PoZ, each time recomputing the (partial) DNN RDM and computing the Pearson correlation with the full embeddings' RDM (figure 1b). Figure 1c shows a similar analysis, but where the features were inserted from low- to high-PoZ (that is, from most to least informative). The R^2 arrives at 1.0 after inserting 80% of the nodes, meaning the 20% highest-PoZ nodes are not required. However, these high-PoZ features in some cases produced an R^2 of over 0.4 for CIFAR-10 when used alone (figure 1b). While substantial, this is much lower than found when using the 30% of the features with low-PoZ values, which produced a R^2 of over 0.9 for CIFAR-10 (1c). In conclusion, high-PoZ features are less informative than low-PoZ ones with respect to capturing a DNN's overall representational geometry, and may not capture unique information.



(a) (b) (c) (d)

Figure 1: a) Cumulative histogram of PoZ statistic; b) and c) Pearson correlation R^2 between RDMs of pruned and full DNN embeddings; d) Pearson correlation R^2 between RDM from pruned DNN embeddings and RDM from human similarity judgments. The red dots indicate the maximum scores, while the black stars indicate the significant improvements.

Predicting human similarity judgments

We used data from Peterson, Abbott, & Griffiths, 2018, which provides pairwise-similarity judgments for six datasets, each with 120 images. DNN embeddings in each dataset were extracted from pretrained VGG-19's penultimate layer containing 4096 nodes (Simonyan & Zisserman, 2014). Figure 1d illustrates the relationship between RDMs computed from DNN embeddings and the *human* RDMs when features are inserted from lowest-PoZ to highest-PoZ. For most datasets, the first 20% of features were found to be sufficient for approximating the predictive capacity provided by the full embeddings. Notably, adding high-PoZ features produced small but statistically significant *reductions* in predictive accuracy for two datasets (Furniture and Vegetables). That is, they decreased accuracy as compared to a smaller set of lower-PoZ features.

Semantics of ghost features

For each of the six above-mentioned datasets we identified those features that fired 0 for all 120 images. We then passed ImageNet's test-set through VGG-19, retaining the embeddings of the these features alone, and performed PCA to generate 10 scores per image. We then identified those images that scored highly on each Principle Component (PC). Figure 2 shows that for Fruit-ghost-features, the images scoring most strongly on PC1 were associated with strong repetitive vertical pattern; here, dogs, people, or their combinations. The highest scoring images on PCs2-6 have foreground patterns or foreground/background combinations inconsistent with fruits. For Furniture-ghost-features, images scoring highly on PC1 have a single cohesive foreground over patterned background, while images scoring highly on PCs 2-6 have foreground patterns or color combinations not typical for furniture.

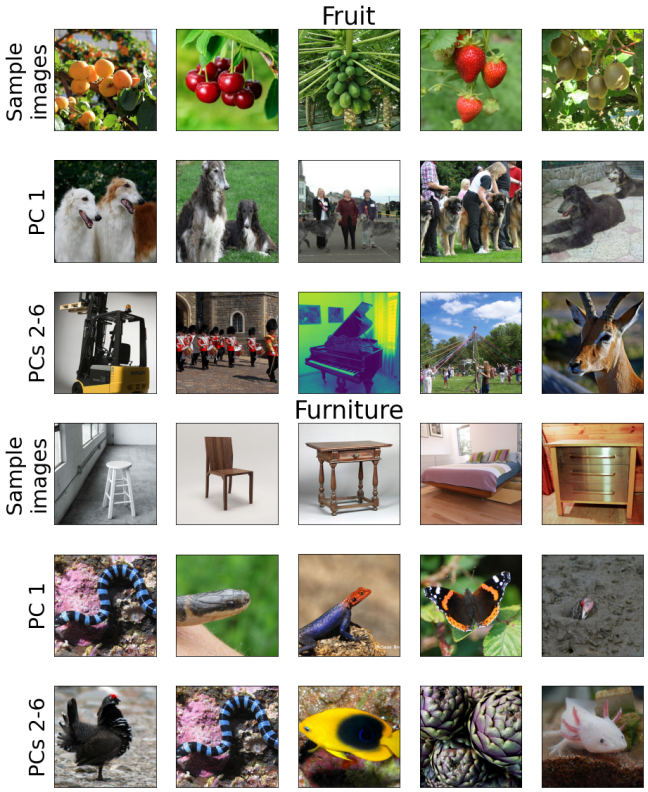


Figure 2: **Content coded by ghost features.** Sample images of representative Fruits and Vegetables (rows 1, 4) and of images that scored highly on information coded by ghost-features for these two categories (rows 2-3; 5-6).

Discussion

We extended previous research that studied how supervised pruning affects DNN's representational geometry (Ansuini et al., 2020; Blakeney et al., 2020), but here focused on unsupervised pruning that is guided by a feature's PoZ. High-PoZ features captured a significant proportion of variance in a DNN's original geometry, but less effectively than low-PoZ features. Interestingly, in some cases, removing high-PoZ features from the full embeddings improved the network's ability to predict human similarity judgments. We show that inspecting ghost features is an explainable-AI tool for describing visual dimensions that do not differentiate between objects within a given category, but distinguish them from other categories. Retrieving images based on ghost features could be a strategy for creating a negative query (i.e. not fruits, not furniture) in image search engine applications.

References

Ansuini, A., Medvet, E., Pellegrino, F., & Zullich, M. (2020). On the similarity between hidden layers of pruned and unpruned convolutional neural networks:. In *Proceedings of the 9th international conference on pattern recognition applications and methods* (p. 52–59). Valletta, Malta: SCITEPRESS - Science and Technology Publications. doi: 10.5220/0008960300520059

Blakeney, C., Yan, Y., & Zong, Z. (2020, Mar). Is pruning compression?: Investigating pruning via network layer similarity. In *2020 ieee winter conference on applications of computer vision (wacv)* (p. 903–911). Snowmass Village, CO, USA: IEEE. Retrieved from https://ieeexplore.ieee.org/document/9093318/ doi: 10.1109/WACV45572.2020.9093318

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]*. Retrieved from http://arxiv.org/abs/1607.03250 (arXiv: 1607.03250)

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Mehta, D., Kim, K. I., & Theobalt, C. (2019, Apr). On implicit filter level sparsity in convolutional neural networks. *arXiv:1811.12495 [cs, eess, stat]*. Retrieved from http://arxiv.org/abs/1811.12495 (arXiv: 1811.12495)

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.