Pruning sparse features for cognitive modeling

Nhut Truong, Uri Hasson
Center for Mind/Brain Sciences (CIMeC), University of Trento
Rovereto, Trento, 38068, Italy
leminhnhut.truong, uri.hasson@unitn.it

Abstract

While deep neural networks are increasingly adopted in cognitive sciences, they are often computationally expensive and contain irrelevant information for downstream tasks. In contrast to pruning approaches that aim to maintain classification accuracy, we present a pruning method to compress entire models while preserving their representation geometry. The target representational space can derived from a neural network or from human similarity space. Our method involves eliminating sparse, rarely activated components throughout the entire network architecture, employing both top-down and bottom-up directions. We show that a deep model's representational space can be preserved or minimally altered when sparse features are removed, producing a compact model for network distillation and predicting human similarity judgments. Furthermore, since our method is structured pruning, it can identify modular structures within pre-trained models.

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

Introduction

Pruning is an important topic in deep learning, where the typical objective is to compress large models while preserving their performance. In this work, our first aim is to prune deep neural networks (DNNs) but using a different objective, which is maintaining the model's representational geometry. This geometry is described by the pairwise distances between objects in the model's feature space. Success for this aim is defined as producing a structurally pruned model where the original object-to-object distances are maintained. Additionally, our second aim is to extend this method, but when using the model to approximate an external object-to-object similarity matrix obtained from human similarity judgments (HSJs). The success of this aim is defined as finding a structurally pruned version of the original model which approximates a set of HSJs as well, or better than, the original model.

There are a few benefits of preserving the geometry of the model's representations in pruning, from both machine learning and cognitive science perspective. For instance, it helps transferring the representations in teacher-student distillation, in which a larger network exports its knowledge to a smaller one (Tung & Mori, 2019; Chen et al., 2020). Moreover, aligning the model's representations with those of humans can help in constructing ecological cognitive models that explain human similarity space, which is a fundamental cognitive function that allows people to make sense of the world - from categorizing objects to forming memories and making decisions (Battleday et al., 2020; Roads & Love, 2023). More specifically, achieving these aims by learning sub-parts of the original model allows representing human knowledge of different concepts or domains as masks over an initial model, enabling flexible model reuse.

In the literature on modeling HSJs, pruning is an effective method to select a sub-part of the network that not only improves the task performance, but also enhances model interpretability (Tarigopula et al., 2023; Manrique et al., 2023; Bao & Hasson, 2024; Truong et al., 2024). This approach assumes that pre-trained DNNs have learnt modular structures where information about different categories is encoded in different sub-spaces in the model, which can be detected via pruning. Pruning differs from the common practice of utilizing all features to construct a representational space or adjusting activations through transformation or re-weighting (Peterson et al., 2018; Kaniuth & Hebart, 2022; Jha et al., 2023).

Logic of approach

Extensive sparsity has been observed across various deep architectures (Blalock et al., 2020; Neill, 2020). Hu et al. (2016) show it is possible to use the Percent of Zeros (PoZ) metric, which indicates the tendency of nodes or feature maps to remain inactive (zero) for a given set of objects, to guide pruning while maintaining or improving classification accuracy. Truong et al. (2023) further showed that PoZ can guide removal of nodes in the fully connected layer in several models with minimally impact the network's representational geometry, but do not offer a general approach for pruning the entire network including feature maps.

In this study, we present a pruning procedure that considers the entire network, where pruning is employed using two search directions: top-down (deep to shallow layers) and bottom-up (shallow to deep). This approach allows us to not only prune the penultimate layer but also earlier layers, such as convolutional ones, which may have even greater sparsity (Hu et al., 2016). Beyond producing compact models, an advantage of this approach is that, it can identify the more important feature maps in a given layer, which then allows studying their representations using activation maximization techniques (Erhan et al., 2009; Zeiler & Fergus, 2014).

Aim 1: Pruning to maintain the representational geometry of the full networks

In this experiment, we used a relatively shallow DNN, LeNet5 (LeCun et al., 1998), trained on two small datasets (MNIST and CIFAR-10), which is repeated using 50 different weight initializations of the models. MNIST (LeCun, 1998) contains small grayscale images of 0-9 digits in handwritten form. CIFAR-10 (Krizhevsky et al., 2009) contains small color images of 10 object categories. Each model was trained until their accuracy's on the test sets converges. Then, we computed Pearson-correlation representational similarity matrix (RSM) from the post-ReLU penultimate activations of the first 5000 images in the two training sets, resulting in two baseline matrices, RSM_0.

Next, we visit each layer in one network and progressively remove feature maps (in convolutional layers) or nodes (in fully connected layers) based on their activation frequency ranking. We adopted Percentage of Zeros (PoZ), which is the percentage of zero activations of a feature across a dataset (Hu et al.,

	conv1	conv2	fc1	fc2	CR
Num. of fmaps/nodes	6	16	120	84	
Top-down	5.7 ± 0.5	11 ± 2	67 ± 8	19 ± 2	2.92 ± 0.60
Bottom-up	5 ± 1	10 ± 2	58 ± 9	53 ± 7	3.28 ± 0.76
Jaccard	90 ± 11	91 ± 7	83 ± 10	38 ± 5	

Table 1: Results on CIFAR-10 dataset. The Top-down and Bottom-up row show the number of retained feature maps or nodes. CR: compression rate.

2016). A network component (feature map or node) with high PoZ activates less frequently and may contribute less to the representational geometry (Truong et al., 2023).

Specifically, the pruning procedure is describe as follows. We traverse through the network one layer at a time. In each layer, feature maps or node are rank ordered according to PoZ. Then, we removed components from highest to lowest PoZ in an accumulated manner. At each step, we compute the Pearson correlation R^2 between RSM_0 and the RSM from the pruned matrix, resulting in the fit between the original and pruned geometries. This procedure is commonly known as Representational Similarity Analysis. We prune a component as long as its removal does not result in a divergence from the original RSM_0, i.e. R^2 remains above the pre-defined threshold R^2_{target}. Otherwise, we stop pruning that layer and move to the subsequent one. Note that the RSMs are always computed from activations propagated to the penultimate layer. This entire procedure is repeated for every layer, starting from the first convolutional layer (bottom-up) or the fully connected penultimate layer (top-down).

The average results of 50 model instances for CIFAR-10 and MNIST are presented in Table 1 and Table 2, respectively. Overall, the architecture used for CIFAR-10 was compressed by a factor of 3 and that used for MNIST by a factor of 74, while maintaining strong similarity with the original representational space all the time (our target value was $R^2_{target} = 0.8$). There is no consistent pattern regarding the number of retained or removed components across the layers. The overlap of pruned components between the two directions, quantified by the Jaccard index, is high (i.e. more overlap) for three out of four layers, except for the penultimate layer (fc2). This difference may stem from the fact that the fc2 layer in the top-down approach are removed first, whereas in the bottom-up approach, they are removed last while the earlier layer's activation profile underwent significant changes in the course of pruning. In conclusion, we find that PoZ can effectively select a small subset of the model while preserving the representational geometry. Furthermore, fine-tuning the pruned models recovered the classification accuracies.

Aim. 2: Pruning to approximate human representation space

In this experiment, rather than maintaining the representational geometry of the unpruned network, we assess whether it is possible to use sparsity to align the representation of a pruned model with HSJs. The data are kindly provided by Peterson et al. (2018), in which participants were asked to rate the similarity of paired images on a scale from completely dissimilar (0) to identical (10). The algorithm follows the same steps in Aim 1, with the exception that the threshold is set as the R^2 computed from the RSMs of the unpruned model and HSJs. In other words, we aim to maintain the representational space between the full model and human data. To demonstrate the method, we focused on one dataset from Peterson's collection - Animals. We used a VGG-16 model pre-trained on natural images (Simonyan & Zisserman, 2014).

	conv1	conv2	fc1	fc2	CR
Num. of fmaps/nodes	64	64	256	256	
Top-down	44 ± 4	3 ± 2	81 ± 21	30 ± 5	74 ± 43
Bottom-up	33 ± 6	4 ± 2	74 ± 21	89 ± 25	74 ± 41
Jaccard	76 ± 12	89 ± 18	83 ± 10	35 ± 8	

Table 2: Results on MNIST dataset.

Examining the results, the top-down pruning approach compressed 7/15 layers, retaining 4-91% of the components, while the bottom-up approach compressed 11/15 layers, retaining 62-98% of the components. Interestingly, the first three convolutional layers, which encode low-level visual features, remained intact. The threshold R^2_{target} was set at 0.54, corresponding to the prediction of HSJs from the full model. Overall, both top-down and bottom-up approaches achieved a compression rate of 1.6.

We applied activation maximization technique (Lucent package at github.com/greentfrapp/lucent) to generate images activating the most for 4 selected components (Figure 1). The pruning method can retain components resembling animal features (image 1 and 3) while discarding those that do not (image 2 and 4).

Figure 1: Two left: generated images for the feature maps in the last convolutional layer with lowest and highest PoZ. Two right: the same for the nodes in the last fully connected layer.

Conclusion

In this study, we present a method for pruning sparse features in DNNs to create a sub-network while preserving the representational geometry, which is a goal inspired by cognitive science. The approach is potential for building computational models of category-selective areas in the brain, such as those dedicated to faces, words, and so on. Given that our search method is greedy, future work could involve developing effective heuristic guidelines to avoid exhaustive searches.

References

Bao, W., & Hasson, U. (2024). Identifying and interpreting non-aligned human conceptual representations using language modeling. *arXiv preprint arXiv:2403.06204*.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, *11*(1), 5418.

Blalock, D., Gonzalez Ortiz, J. J., Frankle, J., & Guttag, J. (2020). What is the state of neural network pruning? *Proceedings of machine learning and systems*, *2*, 129–146.

Chen, H., Wang, Y., Xu, C., Xu, C., & Tao, D. (2020). Learning student networks via feature embedding. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 25–35.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1.

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

Jha, A., Peterson, J. C., & Griffiths, T. L. (2023). Extracting low-dimensional psychological representations from convolutional neural networks. *Cognitive science*, *47*(1), e13226.

Kaniuth, P., & Hebart, M. N. (2022). Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage*, *257*, 119294.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* .

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Manrique, N. F., Bao, W., Herbelot, A., & Hasson, U. (2023). Enhancing interpretability using human similarity judgements to prune word embeddings. *arXiv preprint arXiv:2310.10262*.

Neill, J. O. (2020). An overview of neural network compression. *arXiv preprint arXiv:2006.03669*.

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.

Roads, B. D., & Love, B. C. (2023). Modeling similarity and psychological space. *Annual Review of Psychology*, *75*.

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

Tarigopula, P., Fairhall, S. L., Bavaresco, A., Truong, N., & Hasson, U. (2023). Improved prediction of behavioral and neural similarity spaces using pruned dnns. *Neural Networks*, *168*, 89–104.

Truong, N., Bavaresco, A., & Hasson, U. (2023). Unsupervised feature selection methods for modeling human similarity judgments with deep neural networks. *Journal of Vision*, *23*(9), 4975–4975.

Truong, N., Pesenti, D., & Hasson, U. (2024). Explaining human comparisons using alignment-importance heatmaps. In *Iclr 2024 workshop on representational alignment.*

Tung, F., & Mori, G. (2019). Similarity-preserving knowledge distillation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1365–1374).

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part i 13* (pp. 818–833).