ZERO-SHOT GENERALIZATION ACROSS ARCHITECTURES FOR VISUAL CLASSIFICATION

Evan Gerritz*, **Luciano Dyballa***, **Steven W. Zucker**
Dept. of Computer Science, Yale University, New Haven, CT, USA * Equal contribution

ABSTRACT

Generalization to unseen data is a key desideratum for deep networks, but its relation to classification accuracy is unclear. Using a minimalist vision dataset and a measure of generalizability, we show that popular networks, from deep convolutional networks (CNNs) to transformers, vary in their power to extrapolate to unseen classes both across layers and across architectures. Accuracy is not a good predictor of generalizability, and generalization varies non-monotonically with layer depth. Our code is available at github.com/dyballa/generalization.

1 INTRODUCTION

In deep learning for classification, generalization is typically considered in the context of training versus test sets (Zhang et al., 2021), where both contain examples from the same set of classes, and is measured via test set accuracy. Our goal is to investigate the capacity of a network to generalize its classification power to similar classes absent in the training set (unseen). We work within the domain of Chinese calligraphy. It consists exclusively of ink on paper, making it a particularly challenging instance of the artist recognition problem and also related to the classic problem of separating style from content (Tenenbaum & Freeman, 2000). It is easy to show that many networks can classify characters by artist, as expected, from the training set (Lucas, 2021), but generalization remains largely unexamined. We adopt an embedding-based zero-shot paradigm, commonly applied to text-to-image models (Xian et al., 2018), but here applied to pure vision models, i.e., with no external semantic information. In the end, (*i*) our approach leads to a method to quantify generalization in this minimalist domain; (*ii*) accuracy of classification is not a good predictor of classification generalizability, i.e., the ability to generalize to out-of-sample (unseen) artists; and (*iii*) different architectures yield surprisingly different latent representations.

2 METHODS & RESULTS

We fine-tuned an ensemble of pretrained networks[1] to classify artists from the calligraphy dataset of Wang (2020). Using the intuition that learned classes should appear more clustered in the space of intermediate representations, we develop a proxy for generalization based on the degree of separability that unseen classes achieve in the latent embeddings. To quantify zero-shot classification, we define a *generalization index*, g, for a given network, based on the normalized mutual information (Danon et al., 2005) between a k-means cluster assignment computed on the embedding of unseen examples and the ground truth: $g = \max_i \left\{\mathrm{NMI}\left(\mathcal{C}^i_{\mathrm{unseen}}, \mathcal{C}^\star\right)\right\}$ where i indexes the intermediate layer outputs, and $\mathcal{C}_{\mathrm{unseen}}$ and $\mathcal{C}^\star$ denote cluster assignments for unseen examples and ground truth, respectively. To establish a baseline for generalization, we also computed g for the classes the network was trained on (g_{seen}) (Table 1). To validate our choice of metric, we tested an additional metric based on k-nearest neighbors (see Appendix A.1); finally, to ensure our findings were consistent with a standard dataset, we repeated our tests using CIFAR-100 (Krizhevsky, 2009) (see Appendix A.2).

The calligraphy dataset contained over 100,000 64×64 images from 20 calligraphers (classes). All networks were fine-tuned using Pytorch from pretrained weights (ImageNet) until achieving high

[1] ResNet50 (He et al., 2016), ViT-base (Dosovitskiy et al., 2020), Swin Transformer (Liu et al., 2021), Pyramid ViT (PViT) (Wang et al., 2021), CvT-21 (Wu et al., 2021), PoolFormer-S12 (Yu et al., 2022), ConvNeXt V2 (Woo et al., 2023)

Table 1: Generalization g of classification networks for unseen and seen classes after fine-tuning.

Network	ResNet	ViT	Swin	PViT	CvT	PoolFormer	ConvNeXtV2
g_{unseen}	0.62	0.70	0.62	0.77	0.67	**0.79**	0.63
g_{seen}	0.88	**0.95**	0.80	0.93	0.94	0.91	0.92
accuracy	0.95	0.98	0.98	0.98	0.99	0.99	0.99

accuracy (>95%) on a subset of 15 classes (in-sample). Hyperparameters: learning rate 2e−4, batch size 72; AdamW optimizer. To compute g, we took a subset of 500 images chosen at random from each of the 5 classes not used during training or validation. As the intermediate representation $\Phi^i(x)$ at each layer i, we used the classification token for the transformers and the vector of the mean activations of each feature map for the CNNs. In Table 1, we show that *generalization varies considerably across different network architectures*. In Figure 1, we show that *generalization varies unpredictably across layers and epochs*.

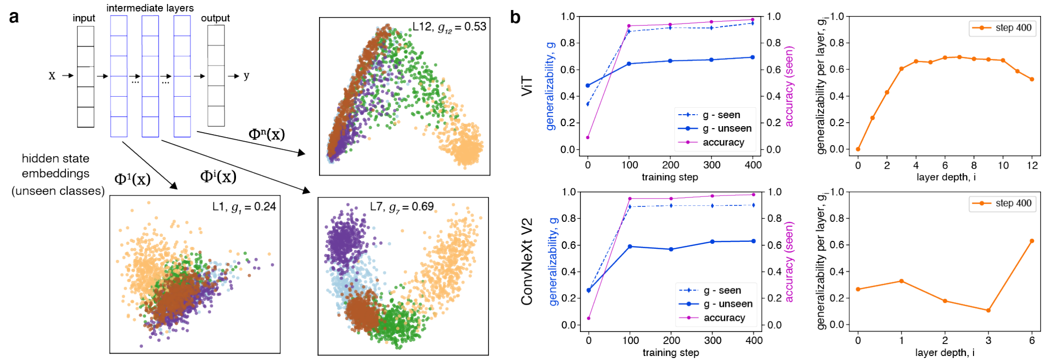

Figure 1: **Generalizability and accuracy are loosely coupled.** (**a**) An overview of our method of assessing generalizability through out-of-sample embeddings using intermediate layers, visualized using PCA. Embeddings from different hidden states of the Vision Transformer (ViT) produce widely varying results. Color labels indicate ground truth: clustered unseen classes indicate better generalization (g). (**b**) Results for ViT (top) and ConvNeXtV2 (bottom). Across epochs, test-set accuracy monotonically increases while generalizability may plateau or even decrease (left). Across layers, there are no predictable trends (right). For additional results, see Appendix A.

3 Conclusion

The relationship between classification accuracy and generalization remains complex. Our experiments demonstrated the pivotal but unpredictable role architecture plays: different models maximize generalization at different layer depths; in other words, we found no consistent encoding strategy emerging after fine-tuning. Furthermore, we found that higher accuracy does not necessarily imply higher generalizability: although all models reached high classification accuracy after fine-tuning (at least 95%), they demonstrated differing capacities to generalize.

In our case study of calligraphy classification, we posit that the generalization index can be used as a measure of how much the network learns a general "language" of stroke patterns. Such a representation can then be applied to adequately represent characters from out of sample artists, as opposed to memorizing characters from in-sample artists.

More generally, we have developed a framework for quantifying the robustness of a network's representation within a domain. Our results leave us with the following question: What capabilities might be revealed if deep learning classification performance were based on generalization rather than typical classification loss functions?

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

Acknowledgments

Supported by NIH Grant 1R01EY031059, NSF Grant 1822598.

References

Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and experiment*, 2005(09):P09008, 2005.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Kauvin Lucas. Calligraphy style classification, 2021. URL `https://www.kaggle.com/code/kauvinlucas/calligraphy-style-classification`.

Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

Yuanhao Wang. Chinese calligraphy styles by calligraphers. `https://www.kaggle.com/datasets/yuanhaowang486/chinese-calligraphy-styles-by-calligraphers`, 2020.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31, 2021.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2018.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10819–10829, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A Appendix

A.1 Alternative generalization metric

Because NMI compares cluster assignments from k-means (section 2), our g metric implicitly assumes that the unseen classes should form Gaussian-like distributions, thus penalizing latent embeddings in which the classes may still be reasonably separable, but more uniformly distributed. Therefore we tested an additional metric without this assumption, based on the traditional notion that nearest neighbors should belong to the same class. For each data point, we compute its k-nearest neighbors (kNN) in an embedding, with k set to the number of examples in each class. We then take the mean, over all data points, of the fraction of a data point x_i's k nearest neighbors belonging to the same class as x_i. The results are shown in Figure 5.

A.2 CIFAR-100 dataset

In addition to the calligraphy dataset, we also tested a standard natural scenes dataset, CIFAR-100. Because ImageNet-1k (the dataset used for pre-training all networks) also consists of natural scenes, we needed to ensure that there were five classes not seen during pretraining or fine-tuning. We observed that CIFAR-100 contains five classes corresponding to flower species that are not present in ImageNet-1k, namely: sunflower, tulip, orchid, poppy, and rose. Moreover, in order to have results comparable to those from the calligraphy dataset, we limited the number of CIFAR classes seen during fine-tuning to 15. Interestingly, the resulting trends were largely similar between the two datasets for all networks; see Figure 5.

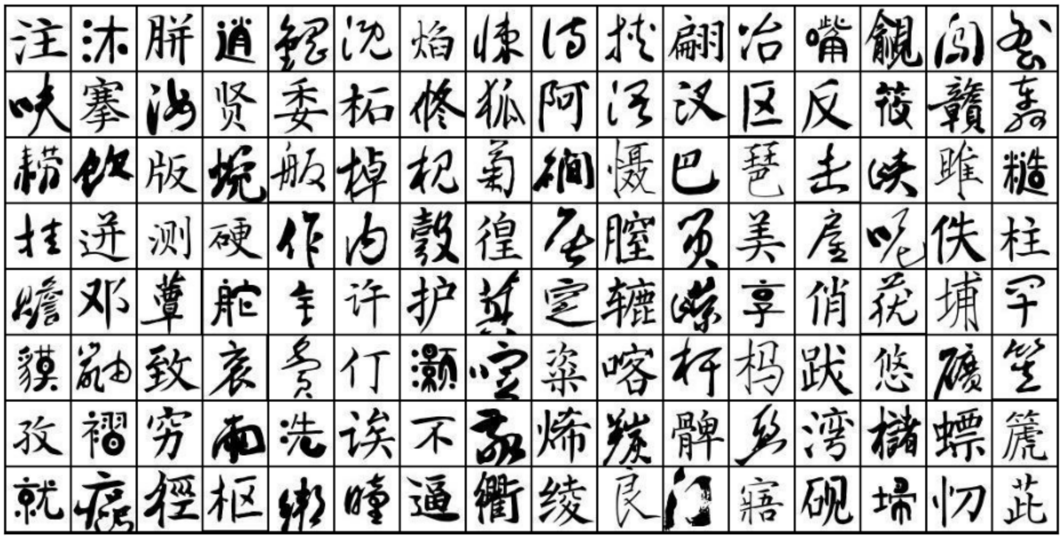

Figure 2: Subset of dataset consisting of 128 individual characters drawn by 20 different calligraphers. We invite the reader to identify characters that appear to be have been written by the same person, and to noticing the nuances of this task. Fine-tuned networks can easily perform this task with high accuracy for artists on which they have been trained (Table 1).

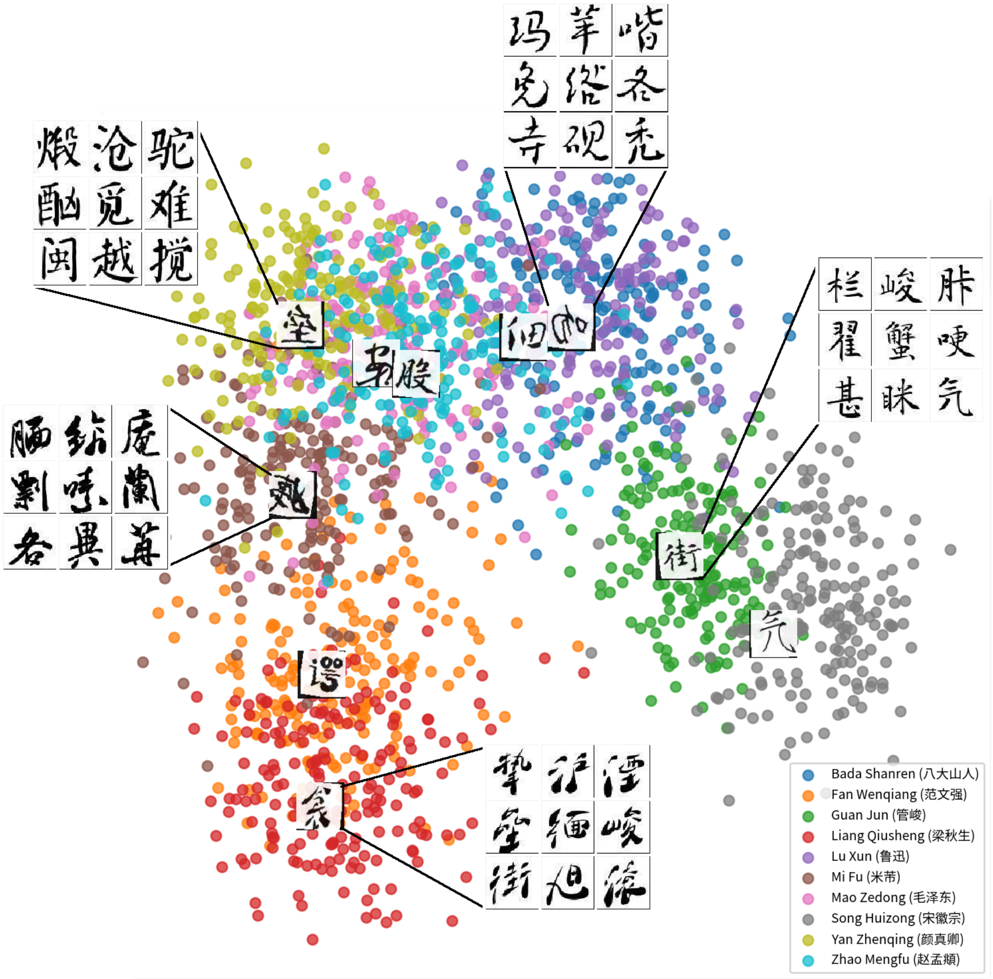

Figure 3: Embedding of 10 calligraphers (including in-sample and out-of-sample) obtained via ResNet and visualized using PCA, with representative character images for each of the ground-truth clusters shown. The images were chosen by computing the centroid of each cluster and selecting those corresponding to the 10 nearest points.

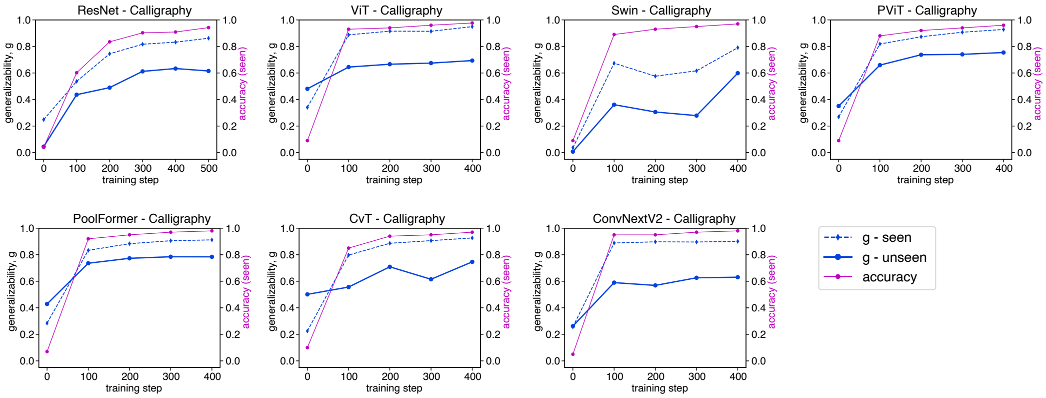

Figure 4: Generalization across fine-tuning epochs for all studied networks. We plot g_{seen} and test-set accuracy (using seen classes) for reference. Notice how g_{seen} is higher for all networks, as expected. This demonstrates that our g metric is able to capture the ability of the networks to generalize to unseen classes.

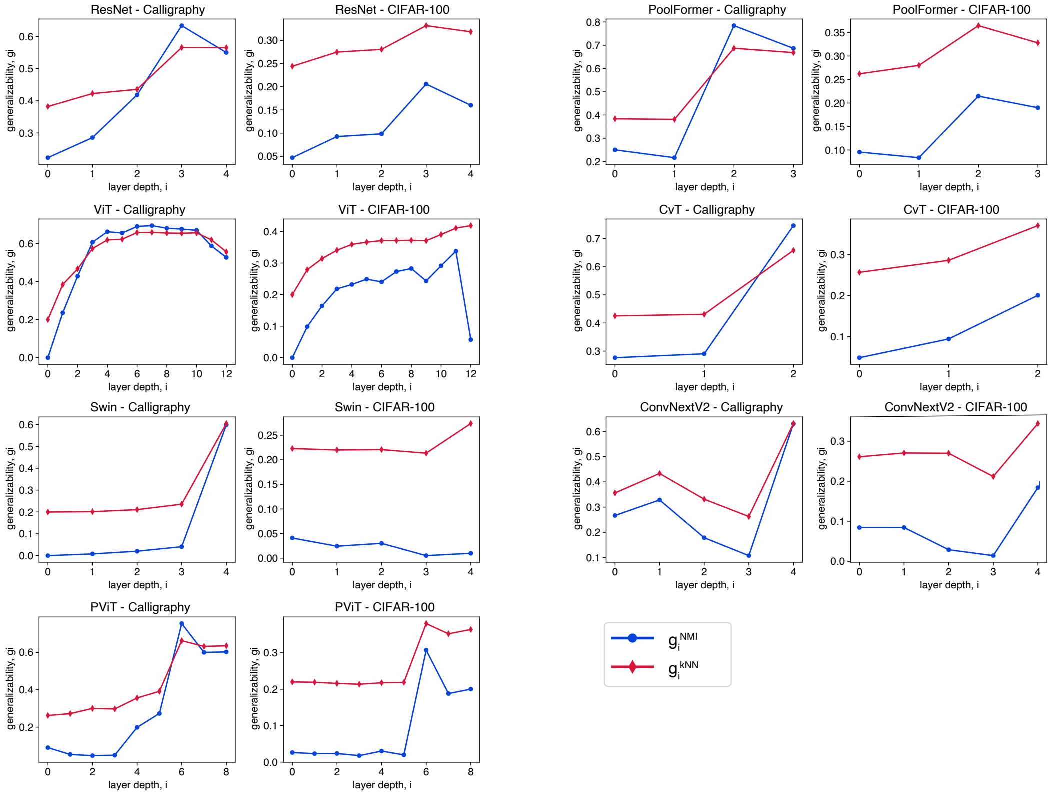

Figure 5: The generalizability trend across layers may vary considerably from model to model. Although deeper layers tend to better separate the examples from unseen classes, g_i does not increase monotonically with depth, and for most models (ResNet, ViT, PViT, and PoolFormer) the best generalization is found at intermediary layers. This hold true for both datasets used for fine-tuning (calligraphy or CIFAR), showing that this phenomenon does not seem dataset-specific. Moreover, similar trends were found regardless of the metric used (NMI or k-nearest neighbors) as demonstrated by qualitatively similar curve shapes, despite being based on completely different methods. In almost all cases, the two metrics identified the same layer as the one that best generalized to the unseen classes. Interestingly, the g_i curves are also qualitatively similar across datasets, indicating that the patterns observed are due to the network's architecture, and not the dataset.