An empirical study of pretrained representations for few-shot classification

Tiago Ramalho, Thierry Sousbie, Stefano Peluchetti
Cogent Labs
Tokyo, Japan
tramalho@cogent.co.jp

Abstract

Recent algorithms with state-of-the-art few-shot classification results start their procedure by computing data features output by a pretrained model. In this paper we systematically investigate which models provide the best representations for a few-shot image classification task when pretrained on the Imagenet dataset. We test their representations when used as the starting point for different few-shot classification algorithms. We observe that models trained on a supervised classification task have higher performance than models trained in an unsupervised manner even when transferred to out-of-distribution datasets. Models trained with adversarial robustness transfer better, while having slightly lower accuracy than supervised models.

1 Introduction

Deep learning systems achieve remarkable performance for several classification problems when given large enough datasets [15, 32]. In the case of small amounts of training data, these models can easily overfit as they contain a very large number of parameters [3]. There are two main techniques to address these issues: if we have a medium-sized dataset, we can fine-tune the weights of a model trained on a larger dataset [12]; for very small datasets we rely on methods which learn from the 'training' dataset without multiple steps of gradient descent over all parameters in the model, known as meta-learning or few-shot classification [25].

Just as in the fine-tuning regime, in the few-shot classification case we want to leverage large datasets for better final performance. This is almost always achieved by feeding the data representation produced by a deep neural network as input to the algorithm. For example, in the case of image classification practitioners will use a deep residual convolutional network pre-trained on a larger dataset to compute the features [30, 31].

If the representations produced by the pre-trained model are more discriminative of the different classes under consideration at test time, it will be easier for a few-shot classification method to produce better results [30]. Previous work has shown that deeper convolutional models have higher accuracy and their feature quality appears to be the limiting factor in performance of most few-shot classification algorithms [4].

In this paper we will perform a systematic exploration of whether deep convolutional networks pretrained on the ImageNet dataset without few-shot classification in mind can transfer well to this task. In the case of fine-tuning it was shown that models with better accuracy on the base task transfer better to other tasks [12]. Such an analysis has not been performed for few-shot classification methods, where systematic surveys have instead focused on the adaptation algorithm or architectural choices [4, 34].

After evaluating multiple models for 14 different datasets, we have come to the following conclusions:

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

- Models pre-trained on a supervised classification task on the ImageNet dataset transfer well to other natural image datasets, but poorly to other types of image (e.g. MNIST, SVHN). The more data models are trained on, the better their transfer performance.
- Models trained with adversarial robustness [37, 11] suffer a smaller performance decrease when transferred to out-of-distribution datasets. Their absolute performance is slightly lower than non-robust models.
- Models trained with an unsupervised learning loss [10, 2] do not reach the performance of models trained in a supervised manner, and do not seem to transfer better.
- Unlike what was reported in [31] for simpler architectures trained from scratch, similarity-based few-shot classification methods used with pretrained models perform best with cosine similarity.

2 Preliminaries

The few-shot classification problem consists of a set of datasets $\mathcal{D} = \{d_0, d_1, ...\}$ where d_n is an individual dataset with image and target label pairs $d_n = \{(x_0, y_0), (x_1, y_1), ...\}$. We consider each dataset d split into two parts: a support set $s = \{(x_s, y_s)\}$ and a query set $q = \{(x_q, y_q)\}$. The model can access both data and labels for all examples in s and is asked to predict labels for $x_q \in q$. The model is trained to minimize the cross-entropy loss:

$$\mathcal{L} = \mathbb{E}_{d\sim\mathcal{D}}\left[-\sum_{(x_q,y_q)\in q} y_q \log\left(f_{\theta,s}(x_q)\right)\right], \tag{1}$$

where $f_{\theta,s}(x)$ stands for a parametric model which outputs the vector of inferred class probabilities. Following established nomenclature, we define k-shot, N-way classification learning as the case where the support set s contains N classes, and for each class we have k observations.

Following [30, 28], we decompose the meta-classification model into two modules: the representation network (also called the convolutional backbone) ϕ, and the adaptation method a such that $f_{\theta,s}(x) = a_\theta(\phi(x_q), \{\phi(x_s), y_s\})$.

We restrict our study to deep convolutional residual networks [9] as representation networks. For all our experiments we take the pre-logits activation vector as the representation to be fed to the adaptation method. We consider a number of recent architectures trained on the Imagenet ILSVRC2012 dataset [29] including: non-robust supervised learning (`resnet50` [9], `efficientB0`, `efficientB7` [33], and `wsl` [18] which is pretrained on a larger dataset), robust supervised learning (`denoise` [37], `robust50` [6]), and unsupervised learning (`amdim` [2]). We provide full details of all architectures considered in the Appendix.

There are three main approaches for adaptation methods in the literature: Distance based [35, 31, 8, 24]; recurrent networks [27, 20] and weight adaptation [7, 30, 21]. We focus our analysis on three simple adaptation methods: Matching networks [35] and Prototype networks [31] as distance based methods; and Logistic regression with SGD as an upper bound for the performance of weight adaptation methods. We avoid recurrent based methods as they are comparatively harder to train and have more parameters to fit.

3 Experiments

We evaluate the performance of all methods on 14 natural image datasets: ILSVRC2012 Imagenet validation set [29]; MNIST [17]; Omniglot [16]; VGG flowers [23]; FGVC-Aircraft [19]; Cars [13]; SVHN [22]; CIFAR-10 and CIFAR-100 [14]; DTD [5]; Fungi [1]; Caltech Birds [36]; ImageNet-v2 [26]; and a subset of Imagenet classes not contained in the ILSVRC2012 1000-class set (we provide a full list of synsets in the Appendix).

For each dataset, we generate a new episode by sampling N classes ($N \in \{5, 10, 20, 50, 100\}$), and create a k-shot support set ($k \in \{1, 2, 3, 5, 10, 20\}$) using the features generated by the network under

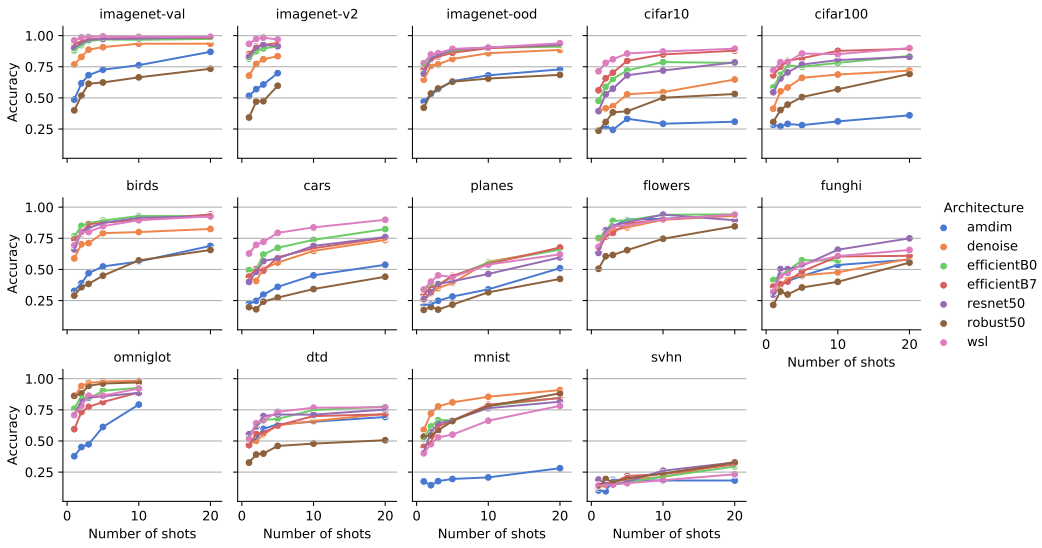

Figure 1: Accuracy comparison for different datasets as a function of number of shots (elements in the support set) for 10-way classification. The accuracy value reported is the average accuracy over 25 unique episodes (where each episode is a random sample of 10 classes, from which the examples are then randomly split into a support set with k examples and a query set with max $(z_c - k, 32)$ examples, with z_c the total number of examples available in class c). Top row: we show imagenet and cifar datasets, where the best performance is obtained by networks trained on the supervised classification task. Middle row: more specialized natural image datasets where the performance decreases but supervised architectures still do better. Note that better supervised architectures and the `wsl` network trained on more data perform better. Bottom row: out-of-distribution datasets. Here networks trained with adversarial robustness do better.

study. Unless otherwise noted, we always use the last layer's features. The accuracy of an episode is calculated over a batch of new query datapoints (batch size 32).

3.1 Pretrained features comparison

Firstly we quantify how the features calculated by the various pretrained backbones under consideration affect the final performance of the few-shot classification task across a range of number of shots and number of classes under consideration. Are certain backbones universally better than others? Do unsupervised models and models trained with adversarial robustness generalize better?

Our findings are summarized in Figure 1. We observe that supervised classification architectures all perform best on the three datasets derived from ImageNet, with the `wsl` model coming on top. Performance on the cifar datasets is also high, in spite of a very different image resolution (we upscale all images where the resolution does not match the original model's input resolution).

Performance decreases in natural image based datasets such as birds, cars, flowers, as the features might lack information to discriminate between different subsets of the same class (e.g. `planes` and `funghi`). Out-of-distribution datasets such as MNIST or SVHN are a harder challenge for all networks. In this case we do observe that networks trained with adversarial robustness generally suffer less of a performance drop than their non-robust counterparts.

3.2 Adaptation method comparison

Next we quantify the effect of the adaptation method on the final performance of few-shot classification when used with pre-trained models. Since pretrained models were not trained with few-shot classification in mind, conclusions reached by previous studies [4, 34, 31] may not hold true. Furthermore, it is not clear which similarity measure is most discriminative in this fixed feature space.

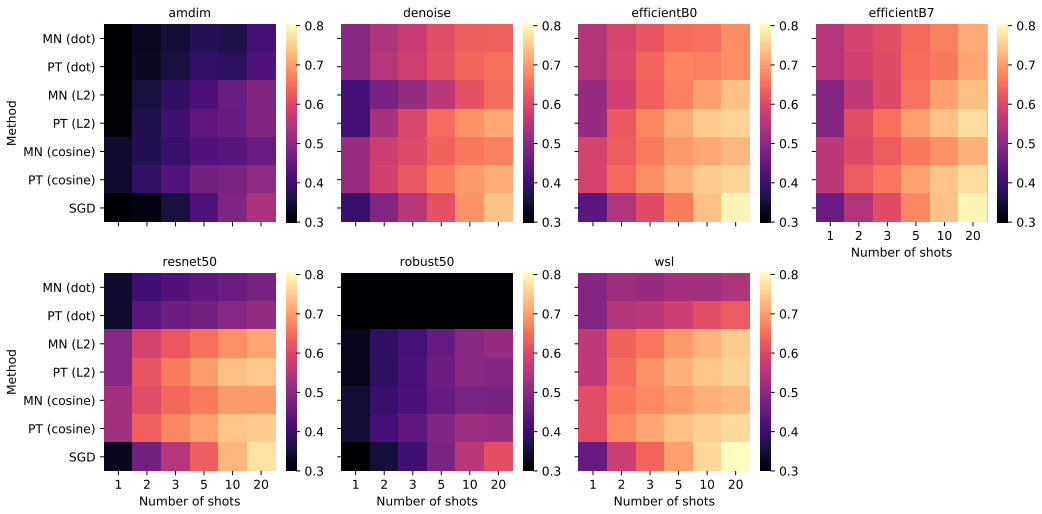

Figure 2: Accuracy comparison for different datasets as a function of adaptation method and similarity function for 10-way classification. The abbreviations MN, PT and SGD are respectively used for Matching Networks, Prototype Networks, and Logistic Regression with Stochastic Gradient Descent. The accuracy value reported is the average accuracy over all datasets.

In Figure 2 we compare the accuracy averaged over all datasets as a function of adaptation method and similarity function for each of the pretrained models. We find that Prototype Networks seems to consistently beat Matching Networks, and cosine similarity is superior to L_2 distance. This is in contrast to the results originally reported in Prototype Networks, where L_2 worked better for a convolutional backbone pretrained from scratch. The unnormalized dot product performs worse than other similarity measures.

Logistic regression with SGD is the best performer for 10 shots and above, while being significantly worse than other adaptation methods for a small number of shots. We present a further breakdown of these results by dataset in the Appendix.

4 Conclusions

In this paper we systematically investigate the performance of pretrained models as backbones to calculate feature representations for few-shot image classification tasks. Models pre-trained on a supervised classification task on the ImageNet dataset transfer well to natural image datasets, but poorly to strongly out-of-distribution datasets or tasks with very fine discriminative requirements. Our results suggest that the more data a supervised model is trained on, the better their transfer performance. At the same time, architectural choices seem to matter less: all supervised networks trained on the standard ILSVRC2012 dataset have comparable performance.

Models with adversarial robustness or trained with an unsupervised loss do not seem to outperform non-robust models trained on the ImageNet classification task. Given that these research fields are relatively new, we wish to revisited these models' performance when the field matures, as we expect their features to be more transferable.

We also found that some best practices on few-shot classification do not transfer to the use of pre-trained models (e.g. we find that the cosine similarity provides better performance than L_2 distance).

We hope our empirical investigation will spur the use of pretrained models in applied few-shot classification and online learning tasks, as they can provide excellent performance with very little training cost. All convolutional backbones we experimented with have publicly available weights and code implementations and we thank all the respective authors for releasing their code for experimentation.

References

[1] Schroeder B. and Y. Cui. Fgvcx fungi classification challenge 2018. 2018.

[2] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning Representations by Maximizing Mutual Information Across Views. *arXiv:1906.00910 [cs, stat]*, June 2019. arXiv: 1906.00910.

[3] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A Closer Look at Few-shot Classification. *arXiv:1904.04232 [cs]*, April 2019. arXiv: 1904.04232.

[5] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[6] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Learning Perceptually-Aligned Representations via Adversarial Robustness. *arXiv:1906.00945 [cs, stat]*, June 2019. arXiv: 1906.00945.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv:1703.03400 [cs]*, March 2017. arXiv: 1703.03400.

[8] Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional Neural Processes. *arXiv:1807.01613 [cs, stat]*, July 2018. arXiv: 1807.01613.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. arXiv: 1512.03385.

[10] Olivier J. Hénaff, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. *arXiv:1905.09272 [cs]*, May 2019. arXiv: 1905.09272.

[11] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. *arXiv:1905.02175 [cs, stat]*, May 2019. arXiv: 1905.02175.

[12] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? *arXiv:1805.08974 [cs, stat]*, May 2018. arXiv: 1805.08974.

[13] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[14] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6, 2009.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, December 2015.

[17] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

[18] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the Limits of Weakly Supervised Pretraining. *arXiv:1805.00932 [cs]*, May 2018. arXiv: 1805.00932.

[19] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[20] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A Simple Neural Attentive Meta-Learner. July 2017.

[21] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to Adapt in Dynamic, Real-World Environments Through Meta-Reinforcement Learning. *arXiv:1803.11347 [cs, stat]*, March 2018. arXiv: 1803.11347.

[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[23] M-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1447–1454, 2006.

[24] Tiago Ramalho and Marta Garnelo. Adaptive Posterior Learning: few-shot learning with a surprise-based memory module. *arXiv preprint arXiv:1902.02527*, 2019.

[25] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[26] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

[27] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-Learning for Semi-Supervised Few-Shot Classification. *arXiv:1803.00676 [cs, stat]*, March 2018. arXiv: 1803.00676.

[28] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E. Turner. Fast and Flexible Multi-Task Classification Using Conditional Neural Adaptive Processes. *arXiv:1906.07697 [cs, stat]*, June 2019. arXiv: 1906.07697.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[30] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-Learning with Latent Embedding Optimization. *arXiv:1807.05960 [cs, stat]*, July 2018. arXiv: 1807.05960.

[31] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. *arXiv:1703.05175 [cs, stat]*, March 2017. arXiv: 1703.05175.

[32] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *arXiv:1707.02968 [cs]*, July 2017. arXiv: 1707.02968.

[33] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:1905.11946 [cs, stat]*, May 2019. arXiv: 1905.11946.

[34] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. *arXiv:1903.03096 [cs, stat]*, March 2019. arXiv: 1903.03096.

[35] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. *arXiv:1606.04080 [cs, stat]*, June 2016. arXiv: 1606.04080.

[36] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[37] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature Denoising for Improving Adversarial Robustness. *arXiv:1812.03411 [cs]*, December 2018. arXiv: 1812.03411.