A Baseline for Few-Shot Image Classification

Guneet S. Dhillon[1], **Pratik Chaudhari**[2,*], **Avinash Ravichandran**[1], **Stefano Soatto**[1,3]
[1]Amazon Web Services, [2]University of Pennsylvania, [3]University of California, Los Angeles
{guneetsd, ravinash, soattos}@amazon.com, pratikac@seas.upenn.edu

Abstract

Fine-tuning a deep network trained with the standard cross-entropy loss is a strong baseline for few-shot learning. When fine-tuned transductively, this outperforms the current state-of-the-art on standard datasets such as Mini-Imagenet, Tiered-Imagenet, CIFAR-FS and FC-100 with the same hyper-parameters. The simplicity of this approach enables us to demonstrate the first few-shot learning results on the Imagenet-21k dataset. We find that using a large number of meta-training classes results in high few-shot accuracies even for a large number of few-shot classes. We do not advocate our approach as the solution for few-shot learning, but simply use the results to highlight limitations of current benchmarks and few-shot protocols.

1 Introduction

Cost of annotating data and the difficulty of procuring it for rare categories has fueled interest in few-shot learning. Fig. 1 shows a boxplot of the performance of state-of-the-art algorithms for 1-shot 5-way classification on Mini-ImageNet [1]. We estimated this plot using published numbers of the confidence internal of the mean accuracy and the number of few-shot episodes. However, the error in the estimate of the median (notches in the boxplot) does not completely reflect the standard deviation of the accuracy, the former goes to zero asymptotically even if the latter is arbitrarily large. The large standard deviation suggests that progress in few-shot learning may be illusory.

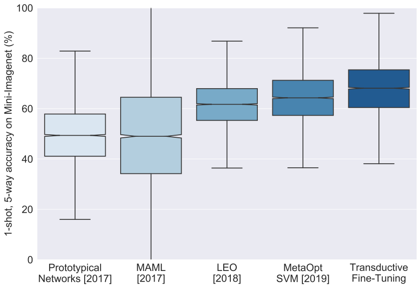


Figure 1: **Are we making progress?** This box plot shows that state-of-the-art few-shot learning methods have enjoyed steady, if limited, improvements in the mean accuracy.

Further, many algorithms report results using different models for different few-shot protocols[2] Our goal is to *develop a simple baseline for few-shot learning*, one that does not require specialized training or hyper-parameter tuning for changing few-shot protocols.

Our contribution is then to develop a *transductive fine-tuning* baseline for few-shot learning. Our baseline outperforms the state-of-the-art on a variety of benchmark datasets such as Mini-ImageNet [1], Tiered-ImageNet [4], CIFAR-FS [5] and FC-100 [3], all with the same hyper-parameters. We report the first few-shot learning results on the Imagenet-21k dataset [6] which contains 14.2 million images across 21,814 classes, with its rare classes forming a natural benchmark for few-shot learning. The strong performance of such a simple baseline indicates that we need to interpret existing results[3] with a grain of salt, and be wary of methods that tailor to the benchmark.

[*] Work done while at Amazon Web Services

[2] For instance, [2] tune specifically for different few-shot protocols, with parameters changing by up to six orders of magnitude; [3] uses a different query shot for different few-shot protocols.

[3] For instance, [1, 7] use different versions of Mini-ImageNet; [3] report results for a backbone pre-trained on the training set while [8] use both the training and validation sets; [9] use full-sized images from the parent Imagenet-1k dataset [6]; [10, 11, 12, 3, 2] all use different architectures for the backbone, of varying sizes, which makes it difficult to disentangle the effect of their algorithmic contributions.

3rd Workshop on Meta-Learning at NeurIPS 2019, Vancouver, Canada.

We give the problem definition and discuss related work in Appendix A.

2 Approach

Let (x, y) denote an image and its ground-truth label respectively. The training (support) and test (query) sets are $\mathcal{D}_\mathrm{s} = \{(x_i, y_i)\}_{i=1}^{N_\mathrm{s}}$ and $\mathcal{D}_\mathrm{q} = \{(x_i, y_i)\}_{i=1}^{N_\mathrm{q}}$ respectively, where $y_i \in C_\mathrm{t}$ for some set of classes C_t. The training and test sets together are called an episode. The number of *ways*, or classes, is $|C_\mathrm{t}|$. *Shots* refers to the number of labeled images per class. One typically also has a *meta-training* set, $\mathcal{D}_\mathrm{m} = \{(x_i, y_i)\}_{i=1}^{N_\mathrm{m}}$, where $y_i \in C_\mathrm{m}$, with classes C_m disjoint from C_t.

2.1 Meta-training

The goal of meta-training is to use $\mathcal{D}_\mathrm{m}$ to infer the parameters of the few-shot learning model: $\hat{\theta}(\mathcal{D}_\mathrm{s}; \mathcal{D}_\mathrm{m}) = \arg\min_\theta \frac{1}{N_\mathrm{m}} \sum_{(x,y)\in\mathcal{D}_\mathrm{m}} \ell(y, F_\theta(x; \mathcal{D}_\mathrm{s}))$, where ℓ is a meta-training loss that depends on the specific method. The simplest form of meta-training is pre-training with the cross-entropy loss: $\hat{\theta} = \arg\min_\theta \frac{1}{N_\mathrm{m}} \sum_{(x,y)\in\mathcal{D}_\mathrm{m}} - \log p_\theta(y|x)$, where $p_\theta(\cdot|x)$ is the probability distribution on the set of classes C_t. The model predicts logits $z(x;\theta) \in \mathbb{R}^{|C_\mathrm{m}|}$; distribution $p_\theta(\cdot|x)$ is computed using the softmax operator on these logits.

2.2 Support-based initialization

Given a pre-trained model (backbone) $p_{\hat{\theta}}$ trained on $\mathcal{D}_\mathrm{m}$, we append a new fully-connected "classifier" layer that takes the logits of the backbone as input (after ReLU non-linearity) and predicts the labels in C_t. For a support sample (x, y), denote the logits of the backbone by $z(x;\ \theta) \in \mathbb{R}^{|C_\mathrm{m}|}$. We denote the weights and biases of the classifier by $w \in \mathbb{R}^{|C_\mathrm{t}|\times|C_\mathrm{m}|}$ and $b \in \mathbb{R}^{|C_\mathrm{t}|}$ respectively. The ReLU non-linearity is denoted by $(\cdot)_+$.

With the classifier's logits given by $z' = wz(x;\theta)_+ + b$, the first term in the cross-entropy loss: $-\log p_\Theta(y|x) = -w_y z(x;\theta)_+ - b_y + \log \sum_k e^{w_k z(x;\theta)_+ + b_k}$, would be the cosine distance between w_y and $z(x;\theta)_+$ if both were normalized to unit ℓ_2 norm and bias $b_y = 0$. This suggests

$$w_y = \frac{z(x;\theta)_+}{\|z(x;\theta)_+\|} \quad \text{and} \quad b_y = 0 \tag{1}$$

as a good candidate for initialization to maximize the cosine similarity between w_y and $z(x;\theta)_+$. For multiple support samples per class, we take the Euclidean average of features $z(x;\theta)_+$ for each class in C_t, before normalization in (1). The logits of the classifier are thus given by $\mathbb{R}^{|C_\mathrm{t}|} \ni z(x;\Theta) = w\frac{z(x;\theta)_+}{\|z(x;\theta)_+\|} + b$, where $\Theta = \{\theta, w, b\}$. All parameters Θ are trainable in the fine-tuning phase.

2.3 Transductive fine-tuning

Since there are very few labeled examples available in few-shot learning, it is important to extract as much extra information from both the labeled support and the unlabled query samples. Inspired from the semi-supervised learning literature [13], we seek solutions that yield models with low Shannon Entropy $\mathbb{H}$ on the query samples. This phase solves for

$$\Theta^* = \arg\min_\Theta \frac{1}{N_\mathrm{s}} \sum_{(x,y)\in\mathcal{D}_\mathrm{s}} - \log p_\Theta\ (y \mid x) + \frac{1}{N_\mathrm{q}} \sum_{(x,y)\in\mathcal{D}_\mathrm{q}} \mathbb{H}(p_\Theta(\cdot \mid x)). \tag{2}$$

The first term uses the labeled support samples whereas the second term uses the unlabeled query samples. We can incorporate a coefficient to tune the entropic term above. However, in line with our goal of developing a simple baseline, we set it equal to 1 for all experiments on benchmark datasets.

3 Experimental results

This section shows results of transductive fine-tuning on benchmark datasets in few-shot learning, namely Mini-ImageNet [1], Tiered-ImageNet [4], CIFAR-FS [5] and FC-100 [3]. We also show large-scale experiments on the Imagenet-21k dataset [6] in Section 3.2. We include analysis of our advocated baseline in Section 3.3. Further details of the experimental setup are sketched in Appendix B.

Table 1: **Few-shot accuracies on benchmark datasets for 5-way few-shot episodes.** Best results in each column are shown in bold. Results where the support-based initialization is better than or comparable to existing algorithms are denoted by †. The notation (train + val) indicates that the backbone was pre-trained on both training and validation classes of the datasets; the backbone is pre-trained only on the training classes when not indicated. The authors in [14] use a 1.25× wider ResNet-12 which we denote as ResNet-12 *.

		Mini-ImageNet		**Tiered-ImageNet**		**CIFAR-FS**		**FC-100**	
Algorithm	Architecture	1-shot (%)	5-shot (%)	1-shot (%)	5-shot (%)	1-shot (%)	5-shot (%)	1-shot (%)	5-shot (%)
Prototypical Networks [11]	conv $(64)_{\times 4}$	49.42 ± 0.78	68.20 ± 0.66						
MAML [10]	conv $(32)_{\times 4}$	48.70 ± 1.84	63.11 ± 0.92						
TADAM [3]	ResNet-12	58.5 ± 0.3	76.7 ± 0.3					40.1 ± 0.4	56.1 ± 0.4
Transductive Propagation [15]	conv $(64)_{\times 4}$	55.51 ± 0.86	69.86 ± 0.65	59.91 ± 0.94	73.30 ± 0.75				
MetaOpt SVM [14]	ResNet-12 *	62.64 ± 0.61	**78.63 ± 0.46**	65.99 ± 0.72	81.56 ± 0.53	72.0 ± 0.7	84.2 ± 0.5	41.1 ± 0.6	55.5 ± 0.6
Support-based initialization (train)	WRN-28-10	56.17 ± 0.64	73.31 ± 0.53	67.45 ± 0.70†	82.88 ± 0.53†	70.26 ± 0.70	83.82 ± 0.49†	36.82 ± 0.51	49.72 ± 0.55
Fine-tuning (train)	WRN-28-10	57.73 ± 0.62	**78.17 ± 0.49**	66.58 ± 0.70	**85.55 ± 0.48**	68.72 ± 0.67	**86.11 ± 0.47**	38.25 ± 0.52	**57.19 ± 0.57**
Transductive fine-tuning (train)	WRN-28-10	**65.73 ± 0.68**	**78.40 ± 0.52**	**73.34 ± 0.71**	**85.50 ± 0.50**	**76.58 ± 0.68**	**85.79 ± 0.50**	**43.16 ± 0.59**	**57.57 ± 0.55**
LEO (train + val) [2]	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09				
MetaOpt SVM (train + val) [14]	ResNet-12 *	64.09 ± 0.62	**80.00 ± 0.45**	65.81 ± 0.74	81.75 ± 0.53	72.8 ± 0.7	85.0 ± 0.5	47.2 ± 0.6	62.5 ± 0.6
Support-based initialization (train + val)	WRN-28-10	58.47 ± 0.66	75.56 ± 0.52	67.34 ± 0.69†	83.32 ± 0.51†	72.14 ± 0.69†	85.21 ± 0.49†	45.08 ± 0.61	60.05 ± 0.60
Fine-tuning (train + val)	WRN-28-10	59.62 ± 0.66	**79.93 ± 0.47**	66.23 ± 0.68	**86.08 ± 0.47**	70.07 ± 0.67	**87.26 ± 0.45**	43.80 ± 0.58	64.40 ± 0.58
Transductive fine-tuning (train + val)	WRN-28-10	**68.11 ± 0.69**	**80.36 ± 0.50**	**72.87 ± 0.71**	**86.15 ± 0.50**	**78.36 ± 0.70**	**87.54 ± 0.49**	**50.44 ± 0.68**	**65.74 ± 0.60**

3.1 Results on benchmark datasets

Table 1 shows the results of transductive fine-tuning on benchmark datasets using standard few-shot protocols. We see that this simple baseline is uniformly better than state-of-the-art algorithms.

The **support-based initialization is sometimes better than or comparable to state-of-the-art** algorithms, these entries are marked using †. For large backbones even standard cross-entropy pre-training and support-based initialization work well. A similar observation was made by [12].

For the 1-shot 5-way setting, fine-tuning using only the support samples leads to minor improvement over the initialization, and sometimes marginal degradation. However, **for the 5-shot 5-way setting non-transductive fine-tuning is better than the state-of-the-art**.

In both (train) and (train + val) settings, **transductive fine-tuning leads to 2-7% improvement for the 1-shot 5-way setting** over the state-of-the-art for all datasets. It results in an **increase of 1.5-4% for the 5-shot 5-way setting** except for the Mini-ImageNet dataset, where the performance is matched. This suggests that the **use of the unlabeled query samples is vital for low-shot settings**.

For the Mini-ImageNet, CIFAR-FS and FC-100 datasets using additional data from the validation set to pre-train the backbone results in 2-8% improvements on the few-shot episodes; the improvement is smaller for Tiered-ImageNet. This suggests that **having more training classes leads to improved few-shot performance** as a consequence of a better embedding.

3.2 Large-scale few-shot learning

The Imagenet-21k dataset [6] with 14.2M images across 21,814 classes is an ideal large-scale few-shot learning benchmark due to the high class imbalance. We use 7,491 classes for meta-training and 13,007 classes for constructing few-shot datasets. Appendix C provides more details of the setup.

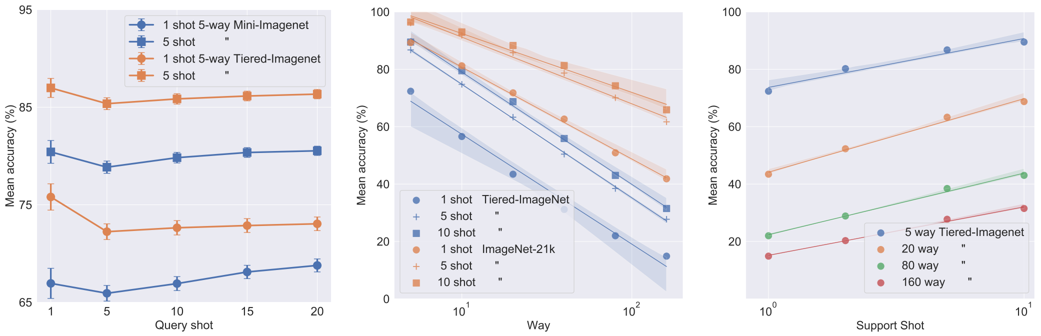
Table 2: **Accuracy (%) on the few-shot data of Imagenet-21k**. The confidence intervals are large because we compute statistics only over 80 few-shot episodes.

			Way					
Algorithm	Model	Shot	5	10	20	40	80	160
Support-based initialization	WRN-28-10	1	87.20 ± 1.72	78.71 ± 1.63	69.48 ± 1.30	60.55 ± 1.03	49.15 ± 0.68	40.57 ± 0.42
Transductive fine-tuning	WRN-28-10	1	89.00 ± 1.86	79.88 ± 1.70	69.66 ± 1.30	60.72 ± 1.04	48.88 ± 0.66	40.46 ± 0.44
Support-based initialization	WRN-28-10	5	95.73 ± 0.84	91.00 ± 1.09	84.77 ± 1.04	78.10 ± 0.79	70.09 ± 0.71	61.93 ± 0.45
Transductive fine-tuning	WRN-28-10	5	95.20 ± 0.94	90.61 ± 1.03	84.21 ± 1.09	77.13 ± 0.82	68.94 ± 0.75	60.11 ± 0.48

Table 2 shows the mean accuracy of transductive fine-tuning evaluated over 80 few-shot episodes on Imagenet-21k. It shows that the accuracy is extremely high as compared to corresponding results in Table 1 even for large ways. E.g., the 1-shot 5-way accuracy on Tiered-ImageNet is 72.87 ± 0.71% while it is 89 ± 1.86% here. This indicates that pre-training with a large number of classes may be an effective strategy to build large-scale few-shot learning systems.

3.3 Analysis

This section presents an analysis of transductive fine-tuning on the Mini-ImageNet, Tiered-ImageNet and Imagenet-21k datasets. All experiments use the (train + val) setting, pre-training the backbone on both the training and validation data of the corresponding datasets. More ablation experiments and further details are discussed in Appendix D.



(a) (b) (c)

Figure 2: **Mean accuracy of transductive fine-tuning for different query shot, way and support shot.** Fig. 2a shows the mean accuracy (with 95% confidence interval) and suggests that a larger query shot helps if the support shot is low; this effect is minor for Tiered-ImageNet. The accuracy for query shot of 1 is high because transductive fine-tuning can specialize the network specifically for the single query shot, possible if there are few query samples. Fig. 2b shows that the mean accuracy degrades logarithmically with way with fixed support shot and a query shot of 15; both Tiered-ImageNet and Imagenet-21k follow this trend with a similar slope. Fig. 2c suggests that the mean accuracy improves logarithmically with the support shot (1, 2, 5, 10) for fixed way and a query shot of 15. The trends in Figs. 2b and 2c suggest thumb rules for building few-shot systems.

Robustness of transductive fine-tuning to query shot: Fig. 2a shows the effect of changing the query shot on the mean accuracy. We observe that the mean accuracy increases with query shot, except for when query shot is 1 and the network specializes to this *one* query shot.

Performance for different way and support shot: Figs. 2b and 2c, show the performance of transductive fine-tuning with changing way and support shot. The mean accuracy changes logarithmically with the way and support shot which provides thumb rules for building few-shot systems.

Computational complexity: Transductive fine-tuning is slower than non-transductive approaches at inference time. Specifically, for 1-shot 5-way, transductive fine-tuning is about 300× slower (20.8 vs. 0.07 seconds) for 15 query shots, and about 60× slower (4 vs. 0.07 seconds) for 1 query shot, as compared to a prototypical network [11] with the same backbone. The 1 query shot numbers are more reasonable, considering the performance of the two algorithms compare 66.2% vs. 58% in our implementation. These factors reduce with higher support shot. A number of recent approaches such as [15, 2] also perform test-time processing and are expected to be slow.

4 Discussion

Our aim is to provide grounding to the practice of few-shot learning. The current literature is in the spirit of increasingly sophisticated approaches for modest improvements in mean accuracy using inadequate evaluation methodology. This is why we set out to establish a baseline, namely transductive fine-tuning. We would like to emphasize that this baseline is not novel and yet performs better than existing algorithms on all standard benchmarks. This is indeed surprising and indicates that we need to take a step back and re-evaluate the status quo in few-shot learning. We hope to use the results in this paper as guidelines for the development of new algorithms.

References

[1] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[2] 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*, 2018.

[3] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 719–729, 2018.

[4] 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*, 2018.

[5] Luca Bertinetto, João F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv:1805.08136*, 2018.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[8] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.

[9] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. 2018.

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[11] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[12] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[14] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. *arXiv:1904.03758*, 2019.

[15] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, and Yi Yang. Transductive propagation network for few-shot learning. *arXiv:1805.10002*, 2018.