When Does Self-supervision Improve Few-shot Learning?

Jong-Chyi Su
UMass Amherst
jcsu@cs.umass.edu

Subhransu Maji
UMass Amherst
smaji@cs.umass.edu

Bharath Hariharan
Cornell University
bharathh@cs.cornell.edu

1. Introduction

Current machine learning algorithms require enormous amounts of training data to learn new tasks. This is an issue for many practical problems across domains such as biology and medicine where labeled data is hard to come by. In contrast, we humans can quickly learn new concepts from limited training data by relying on our past "visual experience". Recent work attempts to emulate this by training a feature representation to classify a training dataset of "base" classes with the hope that the resulting representation generalizes not just to unseen examples of the same classes but also to novel classes, which may have very few training examples (called few-shot learning). However, training for base class classification might discard semantic information that is irrelevant for base classes but critical for novel classes. This might be especially true when the base dataset is small or when the class distinctions are challenging.

One way to recover this useful semantic information is to leverage representation learning techniques that do not use class labels, namely, *unsupervised* or *self-supervised learning*. The key idea is to learn about statistical regularities within images, such as the spatial relationship between patches, or its orientation, that might be a cue to semantics. Despite recent advances, these techniques have only been applied to a few domains (*e.g.*, entry-level classes on internet imagery), and under the assumption that large amounts of unlabeled images are available. Their applicability to the general few-shot scenario is unclear. In particular, can these techniques help prevent overfitting to base classes and improve performance on novel classes in the few-shot setting?

We show that with *no additional training data*, adding a self-supervised task as an auxiliary task (Figure 1) improves the performance of existing few-shot techniques on benchmarks across a multitude of domains (Figure 2). One might surmise that as with traditional SSL, additional unlabeled images might improve performance further. We find that adding more unlabeled images improves performance *only* when the images used for self-supervision are within the *same domain* as the base classes; otherwise, they can even *negatively* impact the performance of the few-shot learner (Figure 3). Based on this analysis we present a simple approach that uses a domain classifier to pick similar-domain unlabeled images for self-supervision from a large, generic pool of images. The resulting method improves over the performance of a model trained with self-supervised learning from images within the dataset (Figure 4). Taken together, this results in a general and practical approach for improving few-shot learning on small datasets in novel domains.

2. Related Work

Few-shot learning. Few-shot learning aims to learn representations that generalize well to the novel classes where only a few images are available. To this end, several meta-learning approaches have been proposed that evaluate representations by sampling many few-shot tasks within the domain of a *base* dataset. These include optimization-based meta-learners, *e.g.* model-agnostic meta-learner (MAML) [4], and distance-based classifiers such as matching networks [13] and prototypical networks (ProtoNet) [11]. While the literature is rapidly growing, a recent study by [3] has shown that the differences between meta-learners are diminished when deeper networks are used. We build our experiments on top of this work and show that auxiliary self-supervised tasks provide additional benefits across a large array of few-shot benchmarks and meta-learners.

Self-supervised learning. Human labels are expensive to collect and hard to scale up. To this end, there has been increasing research interest to investigate learning representations from unlabeled data. Goyal *et al.* [7] and Kolesnikov *et al.* [8] compared various self-supervised learning tasks at scale and concluded that solving jigsaw puzzles [10] and predicting image rotations [6] are among the most effective, motivating our choice of self-supervised tasks.

The focus of most prior works on self-supervised learning is to supplant traditional supervised representation learning with unsupervised learning on large unlabeled datasets for downstream tasks. Crucially in almost all prior works, self-supervised representations consistently lag behind fully-supervised representations trained on the same dataset with the same architecture [7, 8]. *In contrast, our work focuses on*

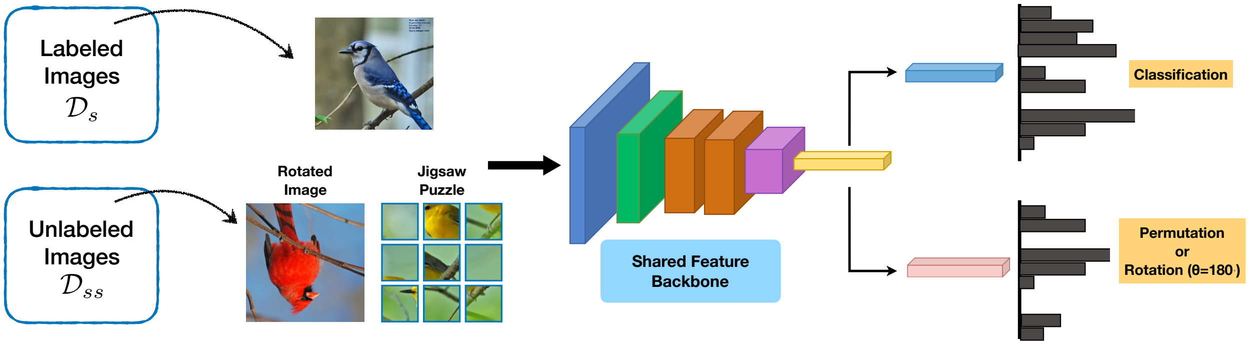
Labeled Images $\mathcal{D}_s$
Unlabeled Images $\mathcal{D}_{ss}$
Rotated Image
Jigsaw Puzzle
Shared Feature Backbone
Classification
Permutation or Rotation (θ=180◦)

Figure 1: **Combining supervised and self-supervised losses for few-shot learning.** Self-supervised tasks such as jigsaw puzzle or rotation prediction act as a data-dependent regularizer for the shared feature backbone. Our work investigates how the performance on the *target task domain* ($\mathcal{D}_s$) is impacted by the choice of the *domain used for self-supervision* ($\mathcal{D}_{ss}$).

an important counterexample: self-supervision can in fact augment standard supervised training for few-shot transfer learning in the low training data regime without relying on any external dataset.

Concurrent works. Several concurrent works appear to draw similar conclusions to ours. Asano *et al*. [1] find that self-supervised pre-training is useful even with one large image when combined with extreme data augmentation. Zhai *et al*. [14] show that self-supervision can be used to improve recognition in a semi-supervised setting and present results on a partially labeled version of the ImageNet dataset. Carlucci *et al*. [2] use self-supervision to improve domain generalization. The most related work to ours is that of Gidaris *et al*. [5] who also show that self-supervision improves few-shot learning. Although the initial results are similar, we further show these benefits on several datasets with harder recognition problems (fine-grained classification) and with deeper models. Moreover, we also present a novel analysis of the impact of the domain of unlabeled data, and a new, simple approach to automatically select similar-domain unlabeled data for self-supervision.

3. Method

Our framework combines meta-learning approaches for few-shot learning that rely on supervised labels with *self-supervised learning* on unlabeled images. Denote a training dataset as $\{(x_i, y_i)\}_{i=1}^n$ consisting of pairs of images $x_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$. A feed-forward convolutional network $f(x)$ maps the input to an embedding space, which is then mapped to the label space using a classifier g. The objective can be written as:

$$\mathcal{L}_s := \sum_i \ell\big(g \circ f(x_i), y_i\big) + \mathcal{R}(f, g),$$

where $\mathcal{R}$ is the regularization term. We consider self-supervised losses $\mathcal{L}_{ss}$ based on labeled data $x \rightarrow (\hat{x}, \hat{y})$ that can be derived from inputs x alone. Figure 1 shows two examples: the *jigsaw task* rearranges the input image and uses the index of the permutation as the target label, while the *rotation task* uses the angle of the rotated image as the target label. In our framework, a separate function h is used to predict these labels from the shared feature backbone f with a self-supervised loss:

$$\mathcal{L}_{ss} := \sum_i \ell\big(h \circ f(\hat{x}_i), \hat{y}_i\big).$$

Our final loss combines the two losses: $\mathcal{L} := \mathcal{L}_s + \mathcal{L}_{ss}$. The self-supervised losses act as a data-dependent regularizer for representation learning. Note that the domain of images used for supervised and self-supervised losses denoted by $\mathcal{D}_s$ and $\mathcal{D}_{ss}$ need not be identical.

4. Experiments

4.1. Results on Few-shot Learning

We experiment with datasets across domains: Caltech-UCSD birds, Stanford cars, FGVC aircrafts, Stanford dogs, Oxford flowers, and the widely-used mini-ImageNet and tiered-ImageNet benchmarks for few-shot learning. We follow the best practices and use the codebase for few-shot learning described in [3]. In particular, we use a prototypical network [11] with a ResNet18 as the feature backbone. We use 5-way (classes) and 5-shot (examples per-class) with 16 query images for training.

In Figure 2, we show that with *no additional training data*, the jigsaw puzzle task improves the ProtoNet baseline on all seven datasets. Predicting rotations also yields improvements on most of the datasets, and combining SSL tasks can be beneficial for some datasets. We also find that the benefits of self-supervision *increase* with the difficulty of the task, and the improvements also generalize to other meta-learners (please see the full version of this paper).

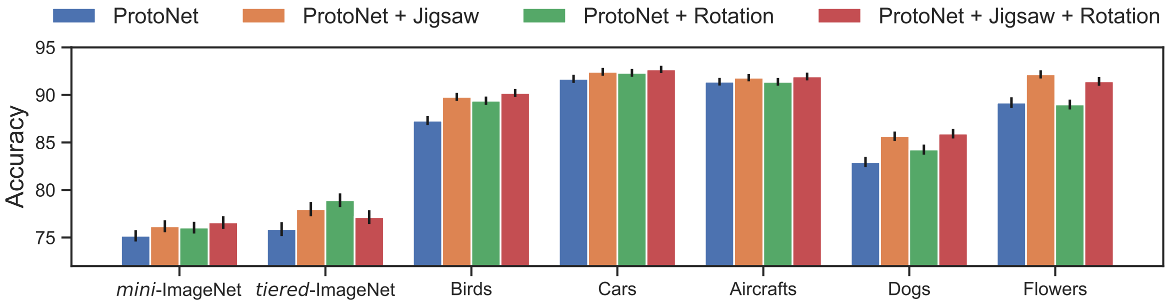

Figure 2: **Benefits of SSL on 5-way 5-shot few-shot learning tasks.** The jigsaw task results in an improvement of the 5-way 5-shot classification accuracy across datasets. Combining SSL tasks can be beneficial for some datasets.

4.2. Analyzing the Effect of Domain Shift for SSL

A promising direction of research in SSL is by scaling it to massive unlabeled datasets that are readily available for some domains. *However, do more unlabeled data always help for a task in hand?* This question hasn't been sufficiently addressed yet as most prior works study the effectiveness of SSL on a curated set of images, such as ImageNet, and their transferability to a handful of tasks. Here we conduct a series of experiments to characterize the effect of size and distribution $\mathcal{D}_{ss}$ of images used for SSL in the context of few-shot learning on domain $\mathcal{D}_s$.

First, we investigate if SSL on unlabeled data from the same domain improves the meta-learner. We use 20% of the images in the base categories for meta-learning. The labels of the remaining 80% data are withheld and only the images are used for SSL. We systematically vary the number of images used by SSL from 20% to 100%. The results are presented on the top of Figure 3. The accuracy improves with the size of the unlabeled set with diminishing returns. Note that 0% corresponds to no SSL and 20% corresponds to using only the labeled images for SSL ($\mathcal{D}_s = \mathcal{D}_{ss}$).

The bottom of Figure 3 shows an experiment where a fraction of the unlabeled images are replaced with images from other four datasets. For example, 20% along the x-axis for birds indicate that 20% of the images in the base set are replaced by images drawn uniformly at random from other datasets. Since the number of images used for SSL is identical, the x-axis from left to right represents increasing amounts of domain shifts between $\mathcal{D}_s$ and $\mathcal{D}_{ss}$. The effectiveness of SSL decreases as the fraction of out-of-domain images increases.

4.3. Selecting Images for Self-supervision

Based on the above analysis we propose a simple method to select images for SSL from a large, generic pool of unlabeled images in a dataset dependent manner. We use a "domain weighted" model to select the top images based on a domain classifier, in our case a binary logistic regression model trained with images from the source domain $\mathcal{D}_s$ as the positive class and images from the pool $\mathcal{D}_p$ as the negative class based on ResNet-101 image features. The top images are selected according to the ratio $p(x \in \mathcal{D}_s)/p(x \in \mathcal{D}_p)$. These *importance weights* account for the domain shift. The top of Figure 4 shows an overview of the selection process.

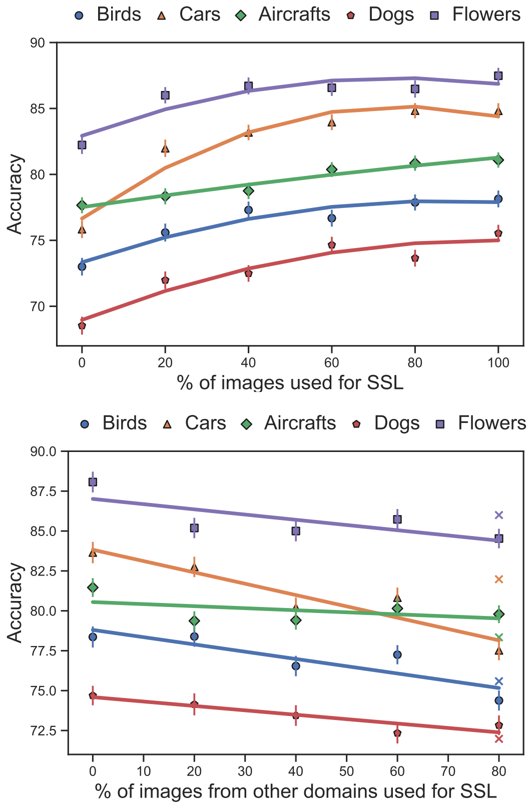

Figure 3: **Effect of size and domain of SSL on 5-way 5-shot classification accuracy. Top:** Effect of number of images on SSL. **Bottom:** Effect of domain shift on SSL.

We evaluate this approach using a pool of images $\mathcal{D}_p$ consisting of (1) the training images of the "bounding box" subset of Open Images V5 [9] which has 1,743,042 images

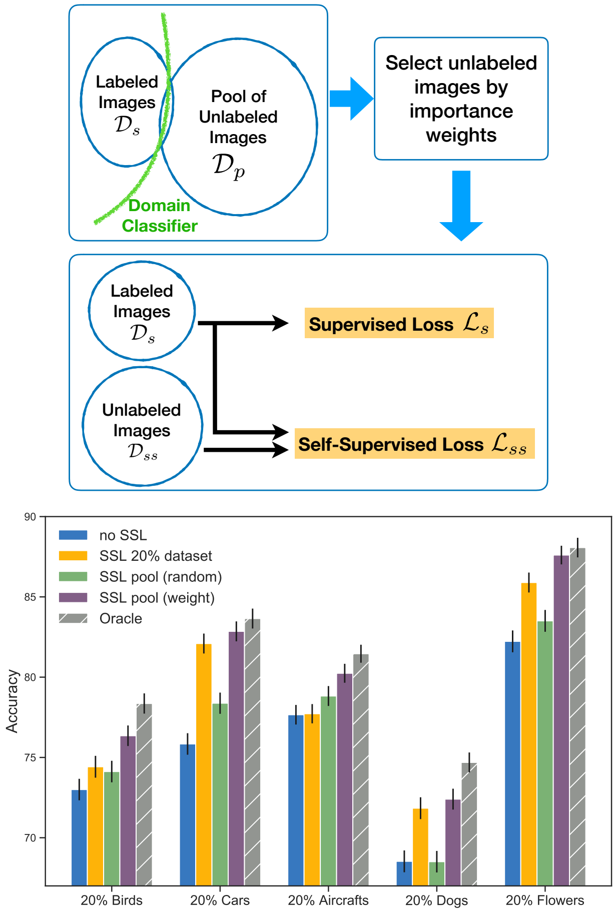

Figure 4: **Selecting images from a pool of unlabeled images for SSL. Top:** Overview of our image selection process. **Bottom:** With random selection, the extra unlabeled data often hurts the performance, while those sampled using the *importance weights* improve performance on all 5 datasets.

from 600 classes, and (2) iNaturalist 2018 dataset [12] which has 461,939 images from 8162 species. For each dataset, we use 20% of the labeled images as $\mathcal{D}_s$. The rest 80% of the data are only used as the "oracle" where the unlabeled data are drawn from the exact same distribution as $\mathcal{D}_s$.

The bottom of Figure 4 shows the results of ProtoNet trained on 20% labeled examples with jigsaw puzzle as self-supervision. To have a fair comparison, for methods of selecting images from the pool, we select the same number (80% of the original labeled dataset size) of images as $\mathcal{D}_{ss}$. We report the mean accuracy of five runs. "SSL with 20% dataset" denotes a baseline of only using $\mathcal{D}_s$ for self-supervision ($\mathcal{D}_s = \mathcal{D}_{ss}$), which is our reference "lower bound". SSL pool "(random)" and "(weight)" denote two approaches of selecting images for self-supervision. The former selects images uniformly at random, which is detrimental for cars, dogs, and flowers. The pool selected according to the *importance weights* provides significant improvements over "no SSL", "SSL with 20% dataset", and "random selection" baselines on all five datasets. The oracle is trained with the remaining 80% of the original dataset as $\mathcal{D}_{ss}$, which is a reference "upper bound".

5. Conclusion

We have shown that self-supervision improves the transferability of representations on few-shot learning tasks across a range of different domains. Surprisingly, we found that self-supervision is more beneficial for more challenging problems, especially when the number of images used for self-supervision is small, orders of magnitude smaller than previously reported results. This has a practical benefit that the images within small datasets can be used for self-supervision without relying on a large-scale external dataset. We have also shown that additional unlabeled images can improve performance only if they are from the *same or similar* domains. Finally, for domains where unlabeled data is limited, we present a novel, simple approach to automatically identify such similar-domain images from a larger pool.

References

[1] Asano, Y.M., Rupprecht, C., Vedaldi, A.: A critical analysis of self-supervision, or what we can learn from a single image. In: ICLR (2020) 2

[2] Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR (2019) 2

[3] Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C., Huang, J.B.: A closer look at few-shot classification. In: ICLR (2019) 1, 2

[4] Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017) 1

[5] Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: ICCV (2019) 2

[6] Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018) 1

[7] Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: ICCV (2019) 1

[8] Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: CVPR (2019) 1

[9] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982 (2018) 3

[10] Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016) 1

[11] Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017) 1, 2

[12] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The iNaturalist species classification and detection dataset. In: CVPR (2018) 4

[13] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NeurIPS (2016) 1

[14] Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4L: Self-supervised semi-supervised learning. In: ICCV (2019) 2