Unsupervised Learning via Meta-Learning

Kyle Hsu[†]
University of Toronto
kyle.hsu@mail.utoronto.ca

Sergey Levine, Chelsea Finn
University of California, Berkeley
{svlevine,cbfinn}@eecs.berkeley.edu

Abstract

A central goal of unsupervised learning is to acquire representations from unlabeled data or experience that can be used for more effective learning of downstream tasks from modest amounts of labeled data. Many prior unsupervised learning works aim to do so by developing proxy objectives based on reconstruction, disentanglement, prediction, and other metrics. Instead, we develop an unsupervised meta-learning method that explicitly optimizes for the ability to learn a variety of tasks from small amounts of data. To do so, we construct tasks from unlabeled data in an automatic way and run meta-learning over the constructed tasks. Surprisingly, we find that, when integrated with meta-learning, relatively simple task construction mechanisms, such as clustering embeddings, lead to good performance on a variety of downstream, human-specified tasks. Our experiments across four image datasets indicate that our unsupervised meta-learning approach acquires a learning algorithm without any labeled data that is applicable to a wide range of downstream classification tasks, improving upon the embedding learned by four prior unsupervised learning methods.

1 Introduction

Unsupervised learning is a fundamental, unsolved problem (Hastie et al., 2009) and has seen promising results in domains such as image recognition (Le et al., 2013) and natural language understanding (Ramachandran et al., 2017). A central use case of unsupervised learning methods is enabling better or more efficient learning of downstream tasks by training on top of unsupervised representations (Reed et al., 2014; Cheung et al., 2015; Chen et al., 2016) or fine-tuning a learned model (Erhan et al., 2010). However, since the downstream objective requires access to supervision, the objectives used for unsupervised learning are only a rough proxy for downstream performance. If a central goal of unsupervised learning is to learn *useful* representations, can we derive an unsupervised learning objective that explicitly takes into account how the representation will be used?

The use of unsupervised representations for downstream tasks is closely related to the objective of meta-learning techniques: finding a learning procedure that is more efficient and effective than learning from scratch. However, unlike unsupervised learning methods, meta-learning methods require large, labeled datasets and hand-specified task distributions. These dependencies are major obstacles to widespread use of these methods for few-shot classification.

To begin addressing these problems, we propose an unsupervised meta-learning method: one which aims to learn a learning procedure, without supervision, that is useful for solving a wide range of new, human-specified tasks. With only raw, unlabeled observations, our model's goal is to learn a useful prior such that, after meta-training, when presented with a modestly-sized dataset for a human-specified task, the model can transfer its prior experience to efficiently learn to perform the new task. If we can build such an algorithm, we can enable few-shot learning of new tasks without needing any labeled data nor any pre-defined tasks.

[†]Work done as a visiting student researcher at the University of California, Berkeley.

2nd Workshop on Meta-Learning at NeurIPS 2018, Montréal, Canada.

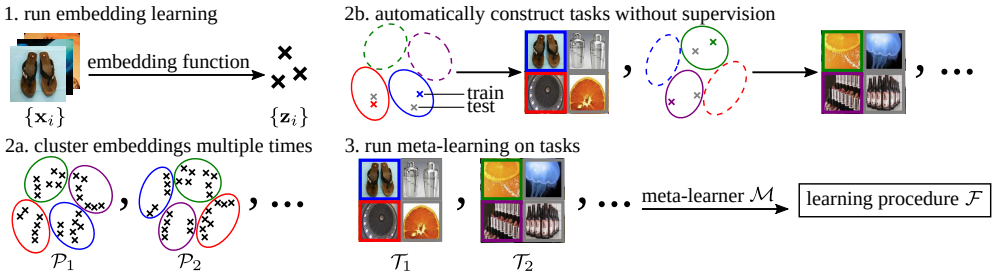

Figure 1: Illustration of the proposed unsupervised meta-learning procedure. An out-of-the-box unsupervised representation is clustered with k-means to construct partitions ($k = 4$ here), which give rise to classification tasks. Each task involves distinguishing between examples from $N = 2$ clusters, with $K_{\text{m-tr}} = 1$ example from each cluster being a training input. The aim is to produce a learning procedure that can solve these tasks.

The core idea in this paper is that we can leverage unsupervised embeddings to propose tasks for a meta-learning algorithm, leading to an unsupervised meta-learning algorithm that is particularly effective as pre-training for human-specified downstream tasks. We instantiate our method with two meta-learning algorithms and compare to prior state-of-the-art unsupervised learning methods. Across four image datasets, we find that our method consistently leads to effective downstream learning of a variety of human-specified tasks without requiring any labels or hand-designed tasks during meta-learning. We show that, even though our unsupervised meta-learning algorithm trains for one-shot generalization, one instantiation of our approach performs well not only on few-shot learning, but also when learning downstream tasks with up to 50 training examples per class. In fact, some of our results begin to approach the performance of fully-supervised meta-learning techniques trained with fully-specified task distributions.

2 Unsupervised Meta-Learning

We assume access to an unlabeled dataset $\mathcal{D} = \{\mathbf{x}_i\}$ during meta-training. After learning from the unlabeled data, which we will refer to as unsupervised meta-training, we want to apply what was learned towards learning a variety of downstream, human-specified tasks from a modest amount of labeled data, potentially as few as a single example per class. These downstream tasks may, in general, have different underlying classes or attributes (in contrast to typical semi-supervised problem assumptions), but are assumed to have inputs from the same distribution as the one from which datapoints in $\mathcal{D}$ are drawn. Concretely, we assume that downstream tasks are M-way classification tasks, and that the goal is to learn an accurate classifier using K labeled datapoints $(\mathbf{x}_k, \mathbf{y}_k)$ from each of the M classes, where K is relatively small (i.e. between 1 and 50).

The unsupervised meta-training phase aligns with the unsupervised learning problem in that it involves no access to information about the downstream tasks, other than the fact that they are M-way classification tasks, for variable M upper-bounded by N. The upper bound N is assumed to be known during unsupervised meta-training, but otherwise, the values of M and K are not known *a priori*. As a result, the unsupervised meta-training phase needs to acquire a sufficiently general prior for applicability to a range of classification tasks with variable quantities of data and classes. This problem definition is our prototype for a practical use-case in which a user would like to train an application-specific image classifier, but does not have an abundance of labeled data.

We aim to construct classification tasks from the unlabeled data and then learn how to efficiently learn these tasks. If such unsupervised tasks are adequately structured and diverse, then meta-learning these tasks should enable fast learning of new, human-provided tasks. In the unsupervised learning literature, common distance functions operating in learned embedding spaces have been shown to qualitatively correspond to semantic meaning (e.g., see Cheung et al. (2015); Bojanowski & Joulin (2017); Donahue et al. (2017)). We consider using such an embedding space to construct tasks with internal structure. We note that, while a given representation may not be directly suitable for highly-efficient learning of new tasks (which would require the representation to be precisely aligned or adaptable to the classes of those tasks), we can still leverage it for the construction of structured and diverse tasks, a process for which requirements are less strict.

We call our method clustering to automatically construct tasks for unsupervised meta-learning (CACTUs). We detail the task construction algorithm in Algorithm 1, and provide an illustration of the complete unsupervised meta-learning approach for classification in Figure 1.

Algorithm 1 CACTUs for classification

procedure CACTUs($\mathcal{E}, \mathcal{D}, P, k, T, N, K_{\text{m-tr}}, Q$)
  Run embedding learning algorithm $\mathcal{E}$ on $\mathcal{D}$ and produce embeddings $\{\mathbf{z}_i\}$ from observations $\{\mathbf{x}_i\}$.
  Run k-means on $\{\mathbf{z}_i\}$ P times (with random scaling) to generate a set of partitions $\{\mathcal{P}_p = \{\mathcal{C}_c\}_p\}$.
  for t from 1 to the number of desired tasks T **do**
    Sample a partition $\mathcal{P}$ uniformly at random from the set of partitions $\{\mathcal{P}_p\}$.
    Sample a cluster $\mathcal{C}_n$ uniformly without replacement from $\mathcal{P}$ for each of the N classes desired for a task.
    Sample an embedding $\mathbf{z}_r$ without replacement from $\mathcal{C}_n$ for each of the $R = K_{\text{m-tr}} + Q$ training and query examples desired for each class, and record the corresponding datapoint $\mathbf{x}_{n,r}$.
    Sample a permutation $(\boldsymbol{\ell}_n)$ of N one-hot labels.
    Construct $\mathcal{T}_t = \{(\mathbf{x}_{n,r}, \boldsymbol{\ell}_n)\}$.
  return $\{\mathcal{T}_t\}$

3 Experiments

In this section, we present an abridged set of experiments. For the full set, including results on two more datasets and unsupervised learning methods, as well as ablations on the task construction and the problem setting, see Hsu et al. (2018). We instantiate CACTUs with embedding-learning methods bidirectional GAN (BiGAN) (Donahue et al., 2017) and DeepCluster (Caron et al., 2018), and meta-learning methods model-agnostic meta-learning (MAML) (Finn et al., 2017) and prototypical networks (ProtoNets) (Snell et al., 2017). We train on unlabeled training splits of the miniImageNet and CelebA datasets, and evaluate on tasks derived from the labeled testing splits. We compare to i) four non-meta-learning algorithms that leverage the embeddings directly, ii) fitting a model from scratch using the MAML architecture for each downstream task, and iii) a meta-learning oracle which trains on tasks derived from labeled versions of the training splits.

As discussed by Oliver et al. (2018), keeping proper experimental protocol is particularly important when evaluating unsupervised and semi-supervised learning algorithms. Our foremost concern is to avoid falsely embellishing the capabilities of our approach by overfitting to the specific datasets and task types that we consider. To this end, we adhere to two key principles. We do not perform any architecture engineering: we use architectures from prior work as-is, or lightly adapt them to our needs if necessary. We also keep hyperparameters related to the unsupervised meta-learning stage as constant as possible across all experiments, including the MAML and ProtoNets model architectures as well as the number of clusters used for CACTUs. We assume knowledge of an upper bound on the number of classes N present in each downstream meta-testing task for each dataset. However, regardless of the number of shots K, we do not assume knowledge of K during unsupervised meta-learning. We use N-way 1-shot tasks during meta-training, but test on larger values of K during meta-testing. When we experiment with the embedding-plus-supervised-learning methods used as fair comparisons to unsupervised meta-learning, we err on the side of providing more supervision and data than technically allowed. Specifically, we separately tune the supervised learning hyperparameters for each dataset and each task difficulty on the labeled version of the meta-validation split.

Results are summarized in Table 1. CACTUs-MAML consistently yields a learning procedure that results in more successful downstream task performance than all other unsupervised methods, including those that learn on top of the embedding that generated meta-training tasks. However, as noted by Snell et al. (2017), ProtoNets perform best when meta-training shot and meta-testing shot are matched; this characteristic prevents ProtoNets from improving upon DeepCluster for 50-shot miniImageNet. We attribute the success of CACTUs-based meta-learning over the embedding-based methods to two factors: its practice in distinguishing between many distinct sets of clusters from modest amounts of signal, and the underlying classes of the testing split data being out-of-distribution. In principle, the latter factor is solely responsible for the success over embedding cluster matching. This algorithm uses the downstream task's training examples to label training split clusters, and can be viewed as a meta-learner on embeddings that trivially obtains perfect accuracy (via memorization) on the meta-training tasks. The same factor also helps explain why training from standard network initialization is, in general, competitive with directly using the task-generating embedding as a representation.

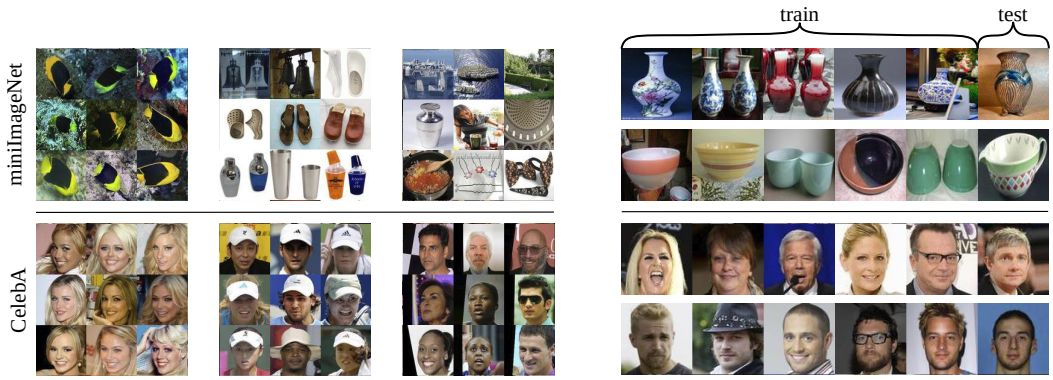

Figure 2: Examples of three DeepCluster-embedding cluster-based classes (a) and a 2-way 5-shot evaluation task (b) for two datasets. (a) Some of the clusters correspond well to unseen labels (top left, bottom left). Others exhibit semantic meaning despite members not being grouped as such in the labeled version of the dataset (top middle: pair of objects, bottom middle: white hat). Still others are uninterpretable (top right) or are based on image artifacts (bottom right). (b) We evaluate unsupervised learning based on the ability to learn downstream test tasks with held-out images and underlying classes.

		miniImageNet				CelebA
Algorithm	(way, shot)	(5, 1)	(5, 5)	(5, 20)	(5, 50)	(2, 5)
Training from scratch		27.59%	38.48%	51.53%	59.63%	63.19%
BiGAN k_{nn}-nearest neighbors		25.56%	31.10%	37.31%	43.60%	56.15%
BiGAN linear classifier		27.08%	33.91%	44.00%	50.41%	58.44%
BiGAN MLP with dropout		22.91%	29.06%	40.06%	48.36%	56.26%
BiGAN cluster matching		24.63%	29.49%	33.89%	36.13%	56.20%
BiGAN CACTUs-MAML (ours)		36.24%	51.28%	61.33%	66.91%	**74.98%**
BiGAN CACTUs-ProtoNets (ours)		36.62%	50.16%	59.56%	63.27%	65.58%
DeepCluster k_{nn}-nearest neighbors		28.90%	42.25%	56.44%	63.90%	61.47%
DeepCluster linear classifier		29.44%	39.79%	56.19%	65.28%	59.57%
DeepCluster MLP with dropout		29.03%	39.67%	52.71%	60.95%	60.65%
DeepCluster cluster matching		22.20%	23.50%	24.97%	26.87%	51.51%
DeepCluster CACTUs-MAML (ours)		**39.90%**	**53.97%**	**63.84%**	**69.64%**	**73.79%**
DeepCluster CACTUs-ProtoNets (ours)		**39.18%**	**53.36%**	61.54%	63.55%	**74.15%**
Oracle-MAML (control)		46.81%	62.13%	71.03%	75.54%	87.10%
Oracle-ProtoNets (control)		46.56%	62.29%	70.05%	72.04%	85.13%

Table 1: Unsupervised learning on miniImageNet and CelebA, averaged over 1000 downstream classification tasks. CACTUs experiments use $k = 500$ for each of $P = 50$ partitions. Cluster matching uses the same k.

4 Discussion

We demonstrate that meta-learning on tasks produced using a simple mechanism based on unsupervised representations improves upon the utility of these representations in learning downstream tasks. We empirically show that this holds across instances of datasets, task difficulties, and unsupervised representations, while fixing key hyperparameters across all experiments. Since MAML and ProtoNets produce nothing more than a learned representation, our method can be viewed as deriving, from a previous unsupervised representation, a new representation particularly suited for learning downstream tasks. While we have demonstrated that k-means is a broadly useful mechanism for constructing tasks from embeddings, it is unlikely that combinations of k-means clusters in learned embedding spaces are universal approximations of arbitrary class definitions. An important direction for future work is to find examples of datasets and human-designed tasks for which CACTUs-based meta-learning results in ineffective downstream learning. Beyond visual classification tasks, the notion of using unsupervised pre-training is generally applicable to a wide range of domains, including regression, speech (Oord et al., 2018), language (Howard & Ruder, 2018), and reinforcement learning (Shelhamer et al., 2017). Hence, our unsupervised meta-learning approach has the potential to improve unsupervised representations for a variety of such domains, an exciting avenue for future work.

References

Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning (ICML)*, 2017.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Neural Information Processing Systems (NIPS)*, 2016.

Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. In *International Conference on Learning Representations (ICLR)*, 2015.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations (ICLR)*, 2017.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research (JMLR)*, 2010.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The Elements of Statistical Learning*. Springer, 2009.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Association for Computational Linguistics (ACL)*, 2018.

Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. *arXiv preprint arXiv:1810.02334*, 2018.

Quoc V Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Gregory S. Corrado, Kai Chen, Jeffrey Dean, and Andrew Y Ng. Building high-level features using large scale unsupervised learning. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *International Conference on Learning Representations (ICLR)*, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Prajit Ramachandran, Peter J Liu, and Quoc V Le. Unsupervised pretraining for sequence to sequence learning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning (ICML)*, 2014.

Evan Shelhamer, Parsa Mahmoudieh, Max Argus, and Trevor Darrell. Loss is its own reward: Self-supervision for reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2017.

Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems (NIPS)*, 2017.