Flexible Few-Shot Learning of Contextual Similarity

Mengye Ren[*,1,2] **Eleni Triantafillou**[*,1,2] **Kuan-Chieh Wang**[*,1,2] **James Lucas**[*,1,2]

Jake Snell[1,2] **Xaq Pitkow**[3,4] **Andreas S. Tolias**[3,4] **Richard S. Zemel**[1,2]

[1]University of Toronto
[2]Vector Institute
[3]Baylor College of Medicine
[4]Rice University
{mren, eleni, wangkua1, jlucas, jsnell}@cs.toronto.edu
xaq@rice.edu, astolias@bcm.edu, zemel@cs.toronto.edu

Abstract

Existing approaches to few-shot learning deal with tasks that have persistent, rigid notions of classes. Typically, the learner observes data only from a fixed number of classes at training time and is asked to generalize to a new set of classes at test time. Two examples from the same class would always be assigned the same labels in any episode. In this work, we consider a realistic setting where the similarities between examples can change from episode to episode depending on the task context, which is not given to the learner. We define two new benchmark datasets for this flexible few-shot scenario, where the tasks are based on images of faces (Celeb-A) and shoes (Zappos50K). While classification baselines and episodic approaches learn representations that work well for standard few-shot learning, they suffer in our flexible tasks as novel similarity definitions arise during testing. We propose to build upon recent contrastive unsupervised learning techniques and use a combination of instance and class invariance learning, aiming to obtain general and flexible features. We find that our approach performs strongly on our new flexible few-shot learning benchmarks.

1 Introduction

Following the success of machine learning applied to fully-supervised settings, there has been a surge of interest in machine learning within more realistic, natural learning scenarios. Among these, few-shot learning [3] (FSL) has emerged as an exciting alternative paradigm. In the few-shot learning setting, the learner is presented with episodes of new learning tasks, where in each episode the learner must identify patterns in a small labeled support set and apply them to make predictions for an unlabeled query set. Since its inception, there has been significant progress on FSL benchmarks. However, standard supervised baselines are often shown to perform as well as carefully designed solutions [1, 8]. In this work, we argue that this observation is due in part to the rigidity in which FSL episodes are designed.

In this work, we define a new flexible few-shot learning (FFSL) paradigm. Instead of building episodes from classes, each episode is a binary classification problem that is constructed with some context that is hidden from the learner. In this way, the same data point may be given different labels across multiple episodes. For example, elephants and tables may belong to the same class if the

[*]Equal contribution

4th Workshop on Meta-Learning at NeurIPS 2020, Vancouver, Canada.

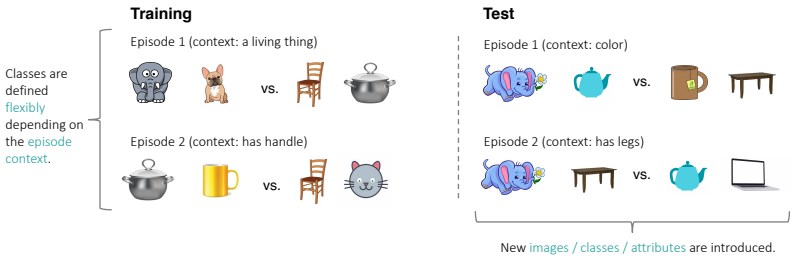

Figure 1: Illustration of the flexible few-shot learning tasks. Instead of having a fixed semantic class, each example may belong to different classes flexibly depending on the context of each episode. A context defines a positive class, based on a set of attribute values; all examples that do not match those values belong to the negative class. At test time contexts are defined based on different attributes than during training.

context is "has legs", but not when the context is "has ears". Importantly, the learner is not given direct access to the context and must infer it from the examples present in the episode.

We contribute two new benchmark datasets for this flexible few-shot scenario. The tasks are based on images of faces (Celeb-A) [4] and shoes (Zappos50K) [10]. We provide a thorough empirical evaluation of existing methods on these tasks. We find that successful approaches in the standard FSL setting fall short on the flexible few-shot tasks. Further, while supervised classification baselines can learn good representation in the standard FSL setting, they suffer in FFSL. Finally, we proposed to use a combination of a class invariance objective with supervised fine-tuning, which is able to provide improved performance on the flexible few-shot tasks.

2 FFSL: Flexible Few-Shot Learning

In this section, we define our FFSL paradigm, and then introduce our two new benchmark datasets on this new FFSL paradigm. Figure 1 shows some examples of different episodes in our FFSL setting. Each episode contains an image of a pot, but the class identity of the pot varies according to the hidden context. In Episode 1, the pot and the chair are given the same labels whereas in Episode 2 they belong to different classes. Moreover, at test time brand new concepts (e.g. tables) or criteria (e.g. color) may be introduced.

Celeb-A: The Celeb-A dataset [4] contains around 200K images of facial images of celebrities. We split half to training, and a quarter to validation and testing each. Each image is annotated with 40 binary attributes, detailing hair colour, facial expressions, and other descriptors. We picked 27 salient attributes and split 14 for training and 13 for both val and test. There is no overlap between training or test attributes but they may sometimes belong to a common category, e.g. blonde hair is in training and brown hair is in test. Split details are included in the supplementary materials.

Zappos-50K: The UT Zappos-50K dataset [10] contains just under 50K images of shoes annotated with attribute values, out of which we kept a total of 76 that we considered salient. We construct an image-level split that assigns 80% of the images to the training set, 10% to the validation and 10% to the test set. We additionally split the set of attribute values into two disjoint sets that are used to form the training and held-out FFSL tasks, respectively.

FFSL episode construction: For each episode, we randomly select one or two attributes (two for Zappos-50K) and look for positive example belonging to these attributes simultaneously. And we also sample an equal number of negative examples that don't belong to one or neither of the selected attributes. Sample episodes from each dataset are shown in Figure 2.

3 Learning Novel Contextual Similarity

Traditionally, the problem of few-shot classification was addressed via episodic models, so we adopt a number of representative models from this category as baselines: Prototypical Networks [7], Matching Networks [9], a MAML [2] variant that only optimizes the top-most layer in the inner loop [6], and We used the prototype vector of the positive class as the task encoding for these two models. We train all episodic models on flexible few-shot learning tasks that are derived from the training set of attributes, and we refer to this approach as Flexible Few-Shot Episodic (**FFSE**).

Figure 2: Sample FFSL episodes using Celeb-A (left) and Zappos-50K (right) datasets. Positive and negative examples are sampled according to the context attributes, but the context information is not revealed to the model at test time.

Following the success of performing supervised pretraining [1, 8], we instead propose to adopt a two-stage approach to address this problem. The first stage is representation learning, which can be obtained from either supervised or unsupervised pretraining. Then, the second stage is few-shot learning, where we solve a few-shot episode by reading out the representation into binary classes.

3.1 Stage 1: Representation learning

Supervised Attribute prediction (SA): Since attributes are the building blocks from which the "classes" are ultimately defined in our flexible episodes, it is natural to consider a training objective that explicitly trains the base backbone to recognize the training attributes, for the purpose of representation learning. For this, we add a classification layer on top of the feature extractor and use it for the task of attribute prediction. We use a sigmoid activation, effectively solving independent binary tasks to predict the presence or absence of each attribute. We refer to this approach as Supervised Attribute prediction (**SA**). As an oracle, we also consider a variant that performs attribute prediction for the full set of both training and held-out attributes, referred to as (**SA***).

Unsupervised learning (U): We hypothesized that learning general-purpose features that capture varying aspects of objects would be helpful to enable this desired flexibility across episodes. We therefore also considered a self-supervised approach, for its ability to learn general semantic features. We chose SIMCLR as a representative from this category due to its empirical success. We refer to this unsupervised approach as (**U**).

Unsupervised Learning with Fine-tuning (UFT) Finally, we explored whether we can combine the merits of unsupervised representation learning and supervised attribute classification, by fine-tuning SIMCLR's unsupervised representation for the task of attribute prediction discussed (SA) above. To prevent SA from overriding the unsupervised features, we add another classifier decoder MLP before the sigmoid classification layer.

3.2 Stage 2: Few-shot learning

Once a representation is learned, it remains to be decided how to use the small support set of each given test episode in order to make predictions for the associated query set. MatchingNet [9] uses nearest neighbor classifier, whereas ProtoNet [7] uses the nearest centroid. Following [1], we propose to directly learn a linear classifier on top of the representation. This approach learns a weight coefficient for each feature dimension, thus performing some level of feature selection, unlike the neareset-centroid and nearest-neighbour variants. Still, the weights need to be properly regularized to encourage high-fidelity selection. For this, we apply an L1 regularizer on the weights to encourage sparsity. The overall objective of the classifier is:

$$\arg\min_{\mathbf{w},b} -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) + \lambda\|\mathbf{w}\|_1, \tag{1}$$

where $\hat{y} = \sigma(\mathbf{w}^\top \mathbf{h} + b)$, and $\mathbf{h}$ is the representation vector extracted from the CNN backbone. The learning of a classifier is essentially done at the same time as the selection of feature dimensions. We refer to this approach as Logistic Regression (**LR**).

4 Experiments

In this section we present our experimental evaluations on our FFSL benchmarks using the different representation learning and few-shot learning methods described in the previous section. For the Celeb-A dataset, we included an additional representation learning method: **ID**, where the backbone was trained for the objective of solving the auxiliary task of face identity classification. In

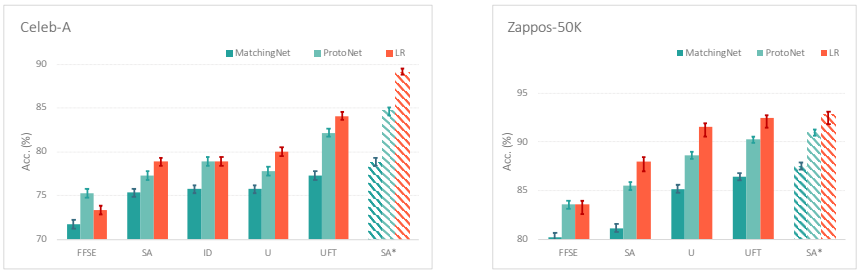

Figure 3: 20-shot FFSL results comparing different representation learning and FSL stage combinations. **FFSE**: Episodic learning using the flexible few-shot episodes. **SA**: Supervised attribute classification. **ID**: Auxiliary representation learning (for Celeb-A this is face ID classification). **U**: Unsupervised contrastive learning. **UFT**: Our proposed U pretraining followed by SA finetuning. **SA***: Supervised attribute binary classification on **all** attributes, which serves as an oracle (striped bars). A set of few-shot learners are evaluated: 1) MatchingNet 2) ProtoNet 3) Logistic Regression (LR). Chance is 50%.

addition to the **SA*** oracle, we provided another oracle **GT-LR**, where the representations used in the test episodes are the ground-truth binary attribute values of the given examples, and the readout is performed using a linear classifier approach.

4.1 Results and discussion

Main results: Figure 3 shows our main results on Celeb-A and Zappos-50K with 20-shot FFSL episodes. On both benchmarks, training on flexible few-shot episodes based on training attributes (FFSE) performed worst. Similarly, supervised attribute (SA) learning and learning via the auxiliary task of class facial identification (ID) were not helpful for representation learning either. Interestingly, U attained relatively better test performance, suggesting that the training objective in contrastive learning indeed preserves more general features—not just shown for semantic classification tasks in prior literature, but also for the flexible class definitions present here. Our proposed UFT approach contributed further gains in performance, suggesting that a combination of unsupervised features with some supervised attribute information is indeed beneficial for this task. We also tried to finetune SIMCLR's representation using FFSE but this did not perform well. We conclude that episodic learning may not help learn higher-level features about the FFSL task itself. Lastly, we confirmed that UFT is able to reduce the generalization gap between SA and SA*, in fact almost closing it entirely in the case of Zappos-50K. These results were consistent across our benchmarks.

Discussion We hypothesize that the weak performance of FFSE and SA on our benchmarks is due to the fact that their training objectives essentially encourage ignoring features that aren't useful at training time, but may still be useful at test time, due to the shift in similarity contexts between the training and testing phases. In the supplementary materials, we study a toy FFSL problem which further illustrates these generalization issues. We explore training a ProtoNet model on data from a linear generative model, where each episode presents significant ambiguity in resolving the correct context. We show that in this setting, unlike in standard FSL tasks, the prototypical network is forced to discard information on the test attributes in order to solve the training tasks effectively, and thus fails to generalize.

Across our benchmarks, we found that UFT was the most effective representation learning algorithm we explored for FFSL. Interestingly, this result contrasts with standard FSL literature, where unsupervised representation learning still lags behind supervised pretraining [5]. On the other hand, our flexible few-shot learning results confirms a significant and complementary gain brought by unsupervised representation learning.

5 Conclusion

The notion of a class often changes depending on the context, yet existing few-shot classification relies on a fixed semantic class definition. In this paper, we propose a flexible few-shot learning paradigm where the similarity criteria change based on the episode context. We found that supervised representation learning generalizes poorly on the test set, due to the partitioning of training & test attributes. Unsupervised contrastive learning on the other hand preserved more generalizable features, and further fine tuning on supervised attribute classification yielded the best results.

Acknowledgments

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute (`www.vectorinstitute.ai/#partners`). This project is supported by NSERC and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

References

[1] W. Chen, Y. Liu, Z. Kira, Y. F. Wang, and J. Huang. A closer look at few-shot classification. In *Proceedings of the 7th International Conference on Learning Representations, ICLR*, 2019.

[2] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, 2017.

[3] B. M. Lake, R. Salakhutdinov, J. Gross, and J. B. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the 33th Annual Meeting of the Cognitive Science Society, CogSci*, 2011.

[4] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision, ICCV*, 2015.

[5] C. Medina, A. Devos, and M. Grossglauser. Self-supervised prototypical transfer learning for few-shot classification. *CoRR*, abs/2006.11325, 2020.

[6] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *8th International Conference on Learning Representations, ICLR*, 2020.

[7] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30, NIPS*, 2017.

[8] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola. Rethinking few-shot image classification: a good embedding is all you need? *CoRR*, abs/2003.11539, 2020.

[9] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29, NIPS*, 2016.

[10] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2014.