Few-Shot Learning in Object Classification using Meta-Learning with Between-Class Attribute Transfer

Majed Alsadhan
mos@ksu.edu
Computer Science Department
Kansas State University
Manhattan, Kansas, USA

William H Hsu
bhsu@ksu.edu
Computer Science Department
Kansas State University
Manhattan, Kansas, USA

ABSTRACT

We present a novel framework for the problem of transfer learning between few-shot source and target domains, using synthetic attributes in addition to convolutional neural networks that are pre-trained on larger image corpora. In these corpora, no labeled instances of the target domains are present, though they may contain instances of their superclasses. Using probabilistic inference over predicted classes and inferred attributes, we developed a meta-learning ensemble method that builds upon that of [10]. This paper introduces the new framework **BCAT** (**B**etween-**C**lass **A**ttribute **T**ransfer), adapting inter-class attribute transfer designed for zero-shot learning (ZSL), combined with fusing transfer learning and probabilistic priors, and thereby extending and improving upon existing deep meta-learning models for FSL. We show how probabilistic learning architectures can be adapted to use state-of-the-field deep learning components in this framework. We applied our technique to four baseline convnet-based FSL ensembles and boosted accuracy by up to 6.24% for 1-shot learning and up to 4.11% for 5-shot learning on the mini-ImageNet dataset, the best result of which is competitive with the current state of the field; using the same technique, we improved accuracy by up to 7.83% for 1-shot learning and up to 3.67% for 5-shot learning on the tiered-ImageNet dataset.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; **Object recognition**; **Transfer learning**.

KEYWORDS

few-shot learning, transfer learning, meta-learning, deep learning, residual networks, object classification

ACM Reference Format:
Majed Alsadhan and William H Hsu. 2022. Few-Shot Learning in Object Classification using Meta-Learning with Between-Class Attribute Transfer. In *Proceeding of the 14th International Conference on Machine Learning and Computing, February 18–21, 2022, Shenzhen, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3529836.3529914

ICMLC '22, February 18–21, 2022, Shenzhen, China

ACM ISBN 978-1-4503-9570-0/22/02.
https://doi.org/10.1145/3529836.3529914

1 INTRODUCTION

Meta-learning of class-specific features embeddings in vision tasks like detecting and classifying objects achieved effective transfer learning for few-shot domains [4, 11, 18], but a problem remains: inter-class attribute transfer from source to rarely-seen target domains (both of which may be few-shot) [10]. In this paper, we present a novel synthesis of these two methods that incorporates inter-class attribute transfer into a state-of-the-field meta-learning convolutional neural network system. The purpose of this approach is to improve generalization accuracy across both metric-based classifiers (e.g., prototypical networks) and optimization-based regularized linear classifiers (e.g., SVM). Our method improves the accuracy of current deep learning classifiers based on residual networks with meta-learning for parametric feature embedding [11] and the probabilistic zero-shot attribute transfer methods that we extend [10], outperforming both and establishing a framework for further advances using between-class attribute transfer.

The performance of most state-of-the-field convolutional neural networks for basic vision tasks is achieved through large quantities of annotated samples such as labeled images [20]. While acquiring large quantities of labeled images might be feasible in some application domains, building such an image corpora in others [20] remains difficult. Labeled images might be unavailable or prohibitively expensive to annotate even for subject matter experts or crowdworkers. Few-shot learning (FSL) is significant to machine learning for many reasons, one of which is reducing the effort involved in collecting data. Because an FSL model requires less data for training, the costs associated with collecting and labeling data are greatly reduced. Because training models that can perform well in FSL tasks is difficult, meta-learning frameworks such as those of [2] were developed to transfer knowledge from the meta-training stage to the meta-testing stage. This transfer of knowledge can be further improved by adding knowledge obtained from attributes.

Meta-learning frameworks that incorporate existing convnets have been proposed for FSL [4]. In such frameworks, FSL classification tasks are formulated by sampling from base categories (meta-training), and a model is optimized to perform well on those tasks. Typically, a task T consists of K-way and N-shot (which means K classes and N support samples per K), and Q query samples for every class. The goal here is to classify all $Q * K$ samples into K classes based on only $N * K$ support samples using a learned model that represents the parametric inductive bias of an embedding. This model is then applied to tasks that have been sampled from novel categories (meta-testing). FSL techniques are evaluated for small values of N where $N \in \{1, 5\}$.

Our foremost **novel contribution** in this research is to develop a new attribute-transfer technique for FSL in a meta-learning framework. The authors of [10] present an attribute-transfer technique for **zero**-shot learning (ZSL) which is nontrivial to apply to state-of-the-art FSL models with mini-ImageNet and tiered-ImageNet in a meta-learning framework because it pre-dates many convnet-based learning representations. We improve this ZSL architecture to meet these specific requirements: probabilistic inference on small samples to obtain inferred transferable attributes; fusion of probabilistic estimates from the attribute predictor of [10] with those based on labeled data; and metric scaling. In addition, we updated this entire ensemble to use a more modern pre-training methodology [4].

2 RELATED WORK: FEW-SHOT LEARNING

Research in FSL and meta-learning related to our own research can be broadly categorized into three approaches: metric-based, optimization-based, and semantics-based.

In the metric-based approach, the model extracts embeddings using a feature-extractor; the embeddings are then used along a distance-based prediction rule [4, 14, 18, 19]. One early work that uses this approach is matching networks [19], where attention and memory are used through an LSTM network. Another example of this approach is the prototypical network [18], where each class is represented by the mean embedding of the support set, and Euclidean distance is then used to calculate the similarity values between that mean (prototypes) and the images in the query set. Authors in [18] report that cosine similarity does not work as well as the Euclidean distance to calculate similarity between the prototypes and the query set. The authors in [14] proposed using metric scaling, which sufficed to close the gap between the Euclidean distance and cosine similarity as originally reported by [18]. One advantage of the approach taken by [4] is that, in contrast to complex models, it represents a simple yet effective model that achieves state of the art accuracy for FSL image classification tasks. Their approach was simply to pre-train the feature-extractor (with a classifier attached to it) on the base classes in a normal classification set up, then remove the classifier and carry out training on the feature-extractor using meta-learning with support and query sets, as well as a scaled cosine similarity measure between them [4]. Most models coming after the Meta-Baseline adopted the pre-training stage because it is simple and effective; we adopted pre-training as well.

Unlike the metric-based approach, recent optimization-based modeling approaches estimate parameters using a parameterized predictor together with a feature-extractor [2, 11]. Authors in [2] developed a deep neural network augmented with conventional learning components. Specifically, they used ridge regression in tandem with a feature-extractor (the deep neural network), which at the time achieved better performance than metric-based approaches. Following in the footsteps of [2], the authors of "MetaOptNetSVM" [11] used a support-vector machine instead of ridge regression for meta-learning. Their approach was to formulate and solve dual quadratic programming (QP) equations to learn linear classifiers like SVM and ridge regression. They thus could solve the dual QP equations using the differentiable GPU-based QP solver proposed by [1].

Our approach fits into the semantics-based branch of FSL work [5, 12, 21]. In semantics-based systems, textual semantic knowledge helps improve the performance of the model. The Dual TriNet Network [5] obtains feature representations from different layers of a ResNet-18 and treats them as different levels of abstract semantic information fed into an encoder to encode into the semantic space. These parameters are then decoded to implement feature augmentation using the Semantic Gaussian and Semantic Neighborhood methods [5]. The names of the base classes in the meta-training task were used in [12] as attributes and fed into a word-embedding model to extract semantic vectors for classes. An adaptive margin generator then obtained a margin penalty for each pair of classes. Finally, the classification loss was combined with this margin penalty to obtain the proposed adaptive margin loss [12]. Attributes extracted from WordNet [7] were used in [21] along with a prototype-completion network, "ProtoComNet". The model worked side-by-side with a feature-extractor, resulting in a hybrid system comprising their new model, ProtoComNet, used to complete prototypes using attributes, and a second model similar to the one in [4]. The outputs of the two models were then fused using their probability fusion strategy, "GaussFusion", based on the assumption that the estimated prototypes follow a Multivariate Gaussian Distribution (MGD) [21]. We used the same fusion strategy in our approach.

While current semantics-based methods for FSL achieve leading-edge performance, they suffer from two main limitations:

(1) **Generalization.** Most of these models achieve FSL via direct optimization and/or using specific embeddings, not through inter-class attribute transfer, which could improve multiple models. Thus, generalizing the transfer mechanism remains a challenge.
(2) **Complexity.** Most of these models are based on complex artificial neural network architectures to infer and/or incorporate attributes. This incurs greater computational costs for training and inference and limits the semantic transparency of the model itself.

Our framework overcomes the first limitation via attribute transfer via a framework that imposes negligible overhead with respect to the second limitation. In the following two sections, we discuss related work on the attribute transfer subtasks of FSL for image classification with known ground attributes. We then present our system, which improves on existing meta-learning FSL systems by solving these subtasks in a recombinable fashion.

3 ATTRIBUTE TRANSFER

The authors in [10] addressed the **zero**-shot learning (ZSL) problem and its specific challenges in data-poor computer vision domains. ZSL differs from FSL in the explicit necessity of attribute transfer because ZSL has no instances of labeled training images (shots) from some classes.

Other researchers in [10] proposed two different ways of transferring attributes in a ZSL setup. One is Indirect Attribute Prediction (IAP) where attributes are inferred from known classes, and then the inferred attributes are used to infer the class labels of the unseen classes. When training an IAP under zero-shot learning, an ordinary multi-class classifier is used to predict known classes [10].

When using this classifier with unseen classes, zl, attributes are first inferred from seen classes, k, using the following equation [10] with the attributes a_m and samples x:

$$p(a_m|x) = \sum_{k=1}^{K} p(a_m|y_k)p(y_k|x) \quad (1)$$

where $p(y_k|x)$ is the probabilistic multi-class classifier's estimates. They set $p(a_m|y)$ assuming a deterministic dependence between attributes and classes. Then those attributes are transferred into labels using the following equation [10]:

$$f(x) = \underset{l=1,\ldots,L}{argmax} \prod_{m=1}^{M} = \frac{p(a_m^{z_l}|x)}{p(a_m^{z_l})} \quad (2)$$

where for the factor $p(a_m)$ they assume a factorial distribution $p(a) = \prod_{m=1}^{M} p(a_m)$, using the empirical means $p(a_m) = \frac{1}{k}\sum_{k=1}^{K} a_m^{yk}$ over the training classes as attribute priors.

We used an approach similar to the IAP method, hybridized with both meta-learning of marginal posteriors for FSL [4, 11, 18] and data-driven estimates of priors. In the following section, a fusion method [21] is presented that further improves this means of inferring transferable attributes.

4 PROBABILISTIC FUSION

The authors in [21] created two combined models using a probabilistic fusion process they call GaussFusion [21]. The motivation for this fusion strategy (refereed to as prototype fusing in [21]) because they observed that one model produced better estimates (p_k) with more shots; while their proposed ProtoComNet produced better estimates ($\hat{p}_k$) with fewer shots. They thus discovered that the two estimates complement each other and thereby identified a need for their probabilistic fusion strategy in such situations [21]. In fusing prototypes, the authors in [21] applied Bayesian estimation. They assumed that both prototypes p_k and $\hat{p}_k$ were sampled from a Multivariate Gaussian Distribution (MGD), $N(\mu_k, diag(\sigma_k^2))$ and $N(\hat{\mu}_k, diag(\hat{\sigma}_k^2))$, where μ_k and $\hat{\mu}_k$ are the means and $diag(\sigma_k^2)$ and $diag(\hat{\sigma}_k^2)$ are the diagonal covariances. The goal is to calculate a fused representation, $N(\hat{\mu'}_k, diag(\hat{\sigma'}_k^2))$, where $\hat{\mu'}_k$ is the MGD mean and $diag(\hat{\sigma'}_k^2)$ is the diagonal covariance [21]. They calculated the mean, μ_k, using the following equation [21]:

$$\mu_k = \frac{1}{\sum\limits_{x \in S \cup Q} P(k|x)} \sum_{x \in S \cup Q} P(k|x) f_{\theta_f}(x) \quad (3)$$

where $P(y = k|x)$ is the probability of a sample x belonging to a K class, S is the support set, Q is the query set, and $f_{\theta_f}(x)$ is the extracted embedding. $\hat{\mu}_k$ is calculated similarly using $\hat{P}(y = k|x)$. To calculate σ_k the following equation [21] is used:

$$\sigma_k = \sqrt{\frac{1}{\sum\limits_{x \in S \cup Q} P(k|x)} \sum_{x \in S \cup Q} P(k|x)(f_{\theta_f}(x) - \mu_k)^2} \quad (4)$$

$\hat{\sigma}_k$ can be calculated similarly using $\hat{\mu}_k$. Having calculated μ_k, $\hat{\mu}_k$, σ_k, and $\hat{\sigma}_k$, we can then calculate $\hat{\mu'}_k = \frac{\sigma_k \odot \hat{\mu}_k + \hat{\sigma}_k \odot \mu_k}{\hat{\sigma}_k \odot \sigma_k}$ and $diag(\hat{\sigma'}_k^2) = diag(\frac{\sigma_k^2 \odot \hat{\sigma}_k^2}{\hat{\sigma}_k^2 \odot \sigma_k^2})$ which provides a new set of estimates, i.e., *prototypes* $\hat{p'}_k$, that can be used to calculate new probabilities for the query samples $\hat{P'}(y = k|x)$, which represent the fused probabilities [21].

5 OUR APPROACH

This section provides details on our approach as illustrated in Figure 1. Subsection 5.1 proposes the pre-training step that we used before fine-tuning the model. Subsection 5.2 explains our own version of the Indirect Attribute Prediction (IAP) and how we incorporated the work in [10]. Subsection 5.3 explains how we used the fusion strategy, and finally, we discuss scaling in subsection 5.4.

5.1 Step 1: Pre-training

We used the pre-training stage, proposed by [4], in our approach. We found that pre-training is a very important step to boost the results of the model to which we apply it. Pre-training by itself does not always boost the performance of all models as we show in section 7. When pre-training, we used a Resnet-12 [9] with a linear classifier to classify the base classes. We then removed the classifer and used the ResNet-12 [9] as a feature-extractor that we fine tuned using steps 2-4.

5.2 Step 2: Indirect Attribute Prediction (IAP)

We previously introduced the IAP approach for zero-shot learning that was proposed by [10]. Here, we propose a very similar approach for few-shot learning. After pre-training the feature-extractor from the previous step, we began fine-tuning the model using the probabilistic multi-class classifier's estimates $p(y_k|x)$ to infer attributes with the following equation:

$$p(a_m|x) = \sum_{k=1}^{K} \frac{e(a_m|y_k)}{\frac{1}{k}\sum_{m=1}^{m} e(a_m|y_k)} p(y_k|x) \quad (5)$$

where $e(a_m|y_k)$ is a static encoding of attributes for every label. We found that scaling the static encoding by dividing by $\frac{1}{k}\sum_{m=1}^{m} e(a_m|y_k)$ gave better results because the static encoding was converted into a probability distribution. We calculated the attribute priors $p(a_m)$ using the scaled encoding of labels for every attribute, $p(y_k|a_m)$:

$$p(a_m) = \frac{p(y_k|a_m)}{\sum_{k=1}^{k} p(y_k|a_m)} \quad (6)$$

Finally, we inferred new probabilistic multi-class estimates using the following:

$$p(\hat{y}_k|x) = \sum_{m=1}^{m} p(a_m|x)p(a_m) \quad (7)$$

This was an estimator that we needed to combine in the next step.

5.3 Step 3: Fusion

In this step, we fused the new probabilistic multi-class estimates $p(\hat{y}_k|x)$ obtained from IAP with the original probabilistic estimates $p(y_k|x)$ of the multi-class classifier, using the fusion strategy proposed by [21]. Using Equation 3, we first calculated the multivariate Gaussian distribution means μ_k and $\hat{\mu}_k$ for $p(y_k|x)$ and $p(\hat{y}_k|x)$. Next, using Equation 4, we calculated the covariances σ_k^2 and $\hat{\sigma}_k^2$. We could then calculate new prototypes to use along a cosine function to calculate similar logits in the query set.

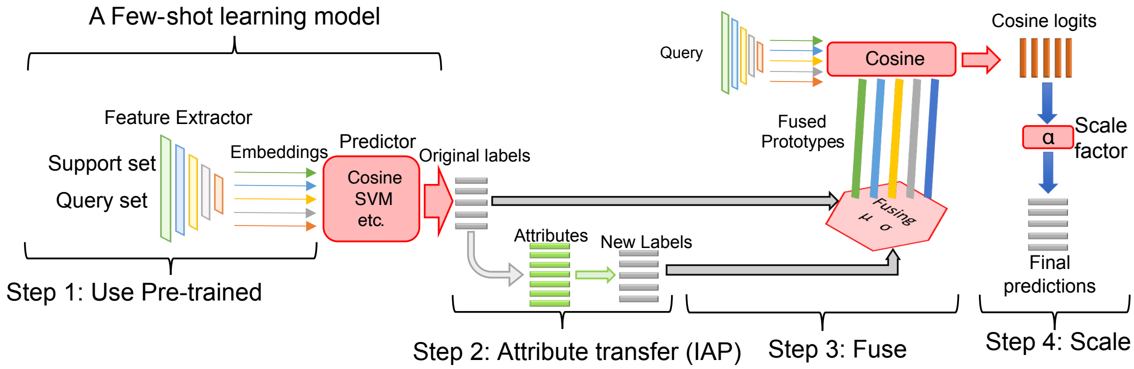

Figure 1: Overview of our approach. We first pre-trained the feature extractor using base class labels in a normal classification setting before using it in a few-shot learning model. Next, we inferred new probabilistic estimates for the query set by inferring attributes from the original probabilistic estimates for the query set. We then fused the inferred probabilistic estimates with the original ones to produce prototypes that we use to calculate cosine similarity for the query set. Finally, we scaled the similarities with a learnable scaling factor.

5.4 Step 4: Scaling

In this step, we used the metric scaling proposed in [14] because we concluded with the cosine function in the previous fusing step to calculate similarities. We made the scaling factor learnable by the network.

6 EXPERIMENT DESIGN

In Section 6.1, we introduce the baselines to which we applied our approach. We discuss the sampling protocol in Section 6.2. We also discuss the dataset used in these experiments in Section 6.3, and We explain our experimental setup in Section 6.4.

6.1 Baselines

Prototypical Networks [18] are a metric-based approach that we used as our first baseline. The authors [18] used a CNN network as a feature-extractor; we used a ResNet-12 [9]. We used the implementation provided by [11]

MetaOptNetSVM/Ridge [11] are optimization-based approaches that we used as our second and third baselines. We used the original implementation provided by [11]

Meta-baseline [4] is a metric-based approach and was our fourth baseline. Our implementation followed the implementation in [4].

6.2 Original Dataset

The **Mini-ImageNet** dataset was developed by [19] and has become one of the most popular benchmark datasets for FSL. This dataset is a subset of ImageNet [6]. It consists of 100 classes split by [16] into three sets: a training set with 64 classes, a validation set with 16 classes, and a testing set with 20 classes. Each category contains 600 images that each are 84 by 84 pixels. We used the same splits as specified by [16]. In addition to images, we used attributes that we extracted from WordNet [7].

Tiered-ImageNet dataset was proposed by [17]. This dataset is also a subset of ImageNet. It consists of 608 classes grouped into 34 high-level categories/clusters with a training set of 20 clusters, validation set with 6 clusters, and testing set with 8 clusters. We also extracted attributes for the images from WordNet [7].

6.3 Sampling Protocol for FSL With Attributes

In this research, we extended the sampler and data loaders developed and open-sourced by [8] and used by [11]. The sampling process begins by randomly sampling K classes from the base classes that are used to sample the meta-learning set for a task by sampling $S + Q$ images where S is the number of shots in the support set and Q is the number of query images. The same process was repeated for meta-validation and meta-testing. We also extended their sampler to look up the hierarchical labeling of the K classes from WordNet [7] by using the nltk [13] library. We used the hierarchical labeling of a category as the attributes for all samples in that category. Therefore, our attributes were per class and not per sample. Per-class attributes was used in the research in [10].

6.4 Experimental Setup

In this work, we used PyTorch [15] as our primary development framework. To apply attributes to the support set, we modified the data loader and sampler provided by [8]. We extended the open-source library developed by [11], which included implementations of prototypical networks [18], MetaOptNetSVM [11], and MetaOptNetRidge [11]. We implemented the open-source meta-baseline by [4]. We replaced the Resnet-12 [9] implementation provided by [11] with the implementation provided by [4].

For performing experiments, we also pre-trained ResNet-12 [9] for the baselines for fair comparisons. In pre-training, we followed the same pre-training protocol used by [4]. During fine-tuning, authors in [11] used a label smoothing of 0.5 with MetaOptNetSVM/Ridge baselines, but we omitted this in our approach for better results. The cosine scaling of the meta-baseline was set to 10, while the scaling factor in our approach was set to 50. We fine-tuned the baselines and our approach and report the best results of both in Section 7.

7 RESULTS AND ANALYSIS

7.1 Meta-Learning/Fusion Ensemble

Table 1 lists, in columns from left to right, the original baseline results that cited authors reported in their respective papers, followed by our results from running the baselines again using pre-trained feature extractors, and finally our BCAT results and the commensurate increases in accuracy over **the better of** pre-trained and published baselines. This shows the marginal gain for our approach relative to all baselines. For brevity, we refer the reader to [21] for results of other semantic models that they compare their system to and which our system transitively outperforms.

Model	Published	Our pre-trained	BCAT	Increase
5-shot mini-ImageNet				
Prototypical	68.20 ± 0.66	**78.11 ± 0.14**	**82.22 ± 0.14**	4.11
MetaOptNetSVM	**78.63 ±0.46**	77.21 ± 0.16	**80.60 ± 0.14**	1.97
MetaOptNetRidge	**77.88 ±0.46**	76.73 ± 0.15	**81.14 ± 0.14**	3.26
Meta-baseline	**79.26 ±0.17**	78.55 ± 0.15	**81.85 ± 0.15**	2.59
1-shot mini-ImageNet				
Prototypical	49.42 ±0.78	**60.46 ± 0.21**	**66.70 ± 0.22**	6.24
MetaOptNetSVM	**62.64 ±0.61**	61.49 ± 0.22	**65.19 ± 0.21**	2.55
MetaOptNetRidge	61.41 ±0.61	**61.61 ± 0.20**	**66.32 ± 0.21**	4.71
Meta-baseline	**63.17 ±0.23**	62.30 ± 0.21	**69.40 ± 0.22**	6.23
5-shot tiered-ImageNet				
Prototypical	72.69 ±0.74	**83.09 ± 0.16**	**85.37 ± 0.16**	2.28
MetaOptNetSVM	**81.56 ±0.53**	79.75 ± 0.17	**84.77 ± 0.16**	3.21
MetaOptNetRidge	81.34 ±0.52	**81.82 ± 0.16**	**85.49 ± 0.16**	3.67
Meta-baseline	83.74 ±0.18	**83.84 ± 0.16**	**85.50 ± 0.16**	1.66
1-shot tiered-ImageNet				
Prototypical	53.31 ±0.89	**65.49 ± 0.24**	**71.96 ± 0.24**	6.47
MetaOptNetSVM	**65.99 ±0.72**	61.37 ± 0.24	**70.90 ± 0.23**	4.91
MetaOptNetRidge	65.36 ±0.71	**65.61 ± 0.23**	**72.69 ± 0.24**	7.08
Meta-baseline	**68.62 ±0.27**	67.78 ± 0.23	**76.45 ± 0.23**	7.83

Table 1: Analysis of mini-ImageNet and tiered-ImageNet. Average 5-way accuracy (%) with 95% confidence interval

For the 5-shot/5-way mini-ImageNet experiments, pre-training improved the accuracy of prototypical networks (in the first row of Table 1) by 9.91%, while the accuracy of MetaOptNetSVM decreased by 1.4%, MetaOptNetRidge decreased by 1.15%, and Meta-baseline (which also relies on pre-training independently performed by the authors) decreased by 1%. As the last column shows, our attribute transfer approach augments the better of published and our pre-trained prototypical networks by 4.11%, MetaOptNetSVM by 1.97%, MetaOptNetRidge by 3.26%, and Meta-baseline by 2.59%. (Note that this marginal improvement is a conservative underestimate based on the slightly better accuracy of 79.26% that the authors of [4] achieved and can only increase if our pre-trained model improves.) In the case of 1-shot/5-way mini-ImageNet experiments, pre-training improved The accuracy of prototypical networks by 11.05%, MetaOptNetRidge increased by 0.20%, but the accuracy of MetaOptNetSVM decreased by 1.15%, and Meta-baseline decreased by 0.87%. Our attribute transfer approach augments the better of published and our pre-trained prototypical networks by 6.24%, MetaOptNetSVM by 2.55%, MetaOptNetRidge by 4.71%, and Meta-baseline by 6.23%. In the 5-shot/5-way tiered-ImageNet experiments, pre-training improved the accuracy of prototypical networks by 10.4%, MetaOptNetRidge increased by 0.48%, and Meta-baseline increased by 0.1%, but the accuracy of MetaOptNetSVM decreased by 1.81%. Our attribute transfer approach augmented the better of published and our pre-trained prototypical networks by 2.28%, MetaOptNetSVM by 2.21%, MetaOptNetRidge by 3.67%, and Meta-baseline by 1.66%. For 1-shot/5-way tiered-ImageNet experiments, pre-training improved The accuracy of prototypical networks by 12.18%, and MetaOptNetRidge increased by 0.25%. The accuracy of MetaOptNetSVM, however, decreased by 4.62%, and Meta-baseline decreased by 0.84%. Our attribute transfer approach augments the better of published and our pre-trained prototypical networks by 6.47%, MetaOptNetSVM by 4.01%, MetaOptNetRidge by 7.08%, and Meta-baseline by 7.83%.

For 5-shot learning on mini-ImageNet, our Prototypical + BCAT ensemble, which achieved 82.22% accuracy, was competitive with the state-of-the-field system of [21], ProtoComNet, which achieved 82.06% accuracy. The difference is not significant and largely attributable to our probabilistic fusion implementation. However, this result establishes that attribute transfer can be improved within BCAT, as our preliminary experiments showed.

7.2 Attribute Transfer

To test and then explore the potential benefits of attribute transfer learning using the approach we describe in Section 3, we used the hierarchical WordNet ontology as our attribute source. Figure 2 represents this hierarchy showing the label level L0 and internal levels Li for $i > 0$. We experimented with including attributes from different internal levels of the hierarchy, which correspond to the union of all interior attributes in the subtree rooted at a node.

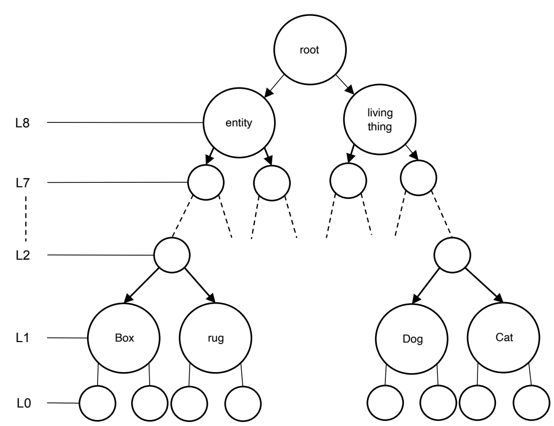

Figure 2: A hierarchy example obtained from WordNet

Evidence of transfer learning. In Figure 3 we show the marginal improvements obtained at different hierarchical levels of abstraction for our scheme added to the transfer of attributes. The local maxima of these improvements was observed at L7 for Meta-baseline + BCAT (L8 is listed in Table 1). While these improvements are incremental, this figure demonstrates that accuracy gain attributable to transfer is achieved when inferring new labels from inferred attributes.

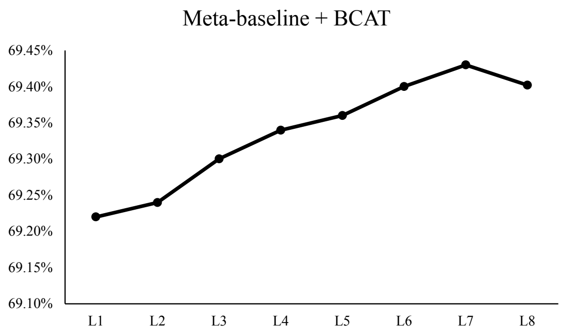

Figure 3: Chart showing incremental increases in accuracy when adding more attributes by going up in the hierarchy. This chart is produced by running Meta-Baseline + BCAT with 1-shot/5-way experiments on the mini-ImageNet dataset.

8 CONCLUSION

Through experiments on mini-ImageNet and tiered-ImageNet, we show that our technique of attribute transfer derived from zero-shot learning can be applied to few-shot learning when used with the meta-learning framework. Authors in [10] developed an attribute-transfer technique for zero-shot learning before recent advances using convnets. This technique does not work out of the box when applied to state-of-the-art FSL models with mini-ImageNet and tiered-ImageNet in a meta-learning framework. We successfully developed a technique for attribute transfer derived from [10] and adapted to FSL by using pre-training [4], probabilistic fusion [21], and metric scaling [14].

Our results show our proposed attribute transfer framework not only outperformed the baselines we applied it to, but also outperformed pre-trained versions in a completely supervised setting. Moreover, when we applied the probabilistic fusion mechanism of [21], we achieved results superior to meta-learning baselines and competitive with the state of the art with the possibility of further improvements using attribute transfer. As mentioned in the previous section, these results are comparable to the extant state of the field on 5-shot mini-ImageNet task [21] using our streamlined, independent implementation, while admitting further improvements to BCAT using both semi-supervised methods and tuning of hierarchical attribute ontologies and other representations as discussed in Section 7.

Given our preliminary findings on choosing transferable attributes from the WordNet hierarchy, a promising direction of future research lies in leveraging unlabeled samples within our framework. We are pursuing at least two main ways to explore this. First, we can take advantage of unlabeled data using semi-supervised and unsupervised methods, which may boost the performance of attribute transfer, because generally speaking, semi-supervised approaches provide better performance in this setting than their supervised counterparts. Second, we can use this framework to experiment with natural language processing methods like ontology extraction from corpora or analogical learning cf. [3]. Here, mini-ImageNet and tiered-ImageNet provide a data-driven basis for learning or inferring transferable attributes, and WordNet itself can serve as a starting point, because our existing attributes consist of words that are hierarchical labelings obtained using WordNet [7].

REFERENCES

[1] Brandon Amos and J Zico Kolter. 2017. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*. PMLR, 136–145.

[2] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. 2018. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136* (2018).

[3] Chao-Yeh Chen and Kristen Grauman. 2014. Inferring analogous attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*. IEEE, 200–207.

[4] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. 2020. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390* (2020).

[5] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. 2019. Multi-Level Semantic Feature Augmentation for One-Shot Learning. *IEEE Transactions on Image Processing* 28, 9 (Sept. 2019), 4594–4605. https://doi.org/10.1109/TIP.2019.2910052

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

[7] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

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

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 951–958.

[11] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10657–10665.

[12] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. 2020. Boosting Few-Shot Learning With Adaptive Margin Loss. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 12573–12581. https://doi.org/10.1109/CVPR42600.2020.01259

[13] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002).

[14] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123* (2018).

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035.

[16] Sachin Ravi and H. Larochelle. 2017. Optimization as a Model for Few-Shot Learning. In *ICLR*.

[17] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676* (2018).

[18] Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175* (2017).

[19] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems* 29 (2016), 3630–3638.

[20] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–34.

[21] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. 2021. Prototype Completion with Primitive Knowledge for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3754–3762.