Multimodal Prototypical Networks for Few-shot Learning

Frederik Pahde[1, †], Mihai Puscas[2, ‡], Tassilo Klein[3], Moin Nabi[3]
[1]Amazon.com, Inc., [2]Huawei Research, Ireland, [3]SAP AI Research, Berlin, Germany
frederikpahde@gmail.com, mihai.puscas@huawei.com, {tassilo.klein, m.nabi}@sap.com

Abstract

Although providing exceptional results for many computer vision tasks, state-of-the-art deep learning algorithms catastrophically struggle in low data scenarios. However, if data in additional modalities exist (e.g. text) this can compensate for the lack of data and improve the classification results. To overcome this data scarcity, we design a cross-modal feature generation framework capable of enriching the low populated embedding space in few-shot scenarios, leveraging data from the auxiliary modality. Specifically, we train a generative model that maps text data into the visual feature space to obtain more reliable prototypes. This allows to exploit data from additional modalities (e.g. text) during training while the ultimate task at test time remains classification with exclusively visual data. We show that in such cases nearest neighbor classification is a viable approach and outperform state-of-the-art single-modal and multimodal few-shot learning methods on the CUB-200 and Oxford-102 datasets.

1. Introduction

Despite the great success of deep learning models, the necessity of large training sets for these models is often a limiting factor. Many applications come with the natural problem of limited data, making it too expensive or even impossible to collect a sufficient number of training samples and leading to poor model accuracies. This is in contrast to the human ability to quickly learn new concepts. Consequently, the study of few-shot classification, i.e. learning new concepts from a very limited amount of training data, has gathered focus in recent years [19, 39, 2, 30, 11, 37, 31]. Finetuning DNNs has been shown to be effective in a context where the big data assumption holds [25]. However, scenarios where access is limited to only very few samples of novel data are extremely susceptible to over-fitting.

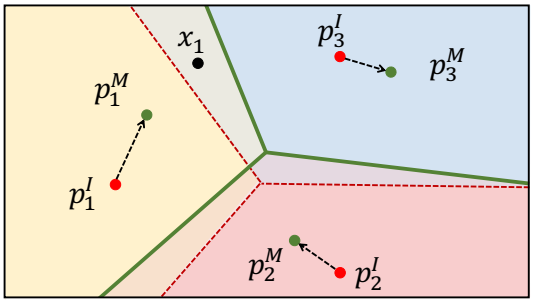

Figure 1: Multimodal Prototypical Networks: Cross-modal generated representations can condense the embedding space and move the visual prototypes p^I towards more reliable multimodal prototypes p^M and it improves the classification accuracy on unseen test samples (e.g., x_1).

Originally, few shot learning defined a scenario where the only very few samples per class were accessible [20, 18, 19]. With the advent of deep learning the assumption was broadened, into having large amounts of data accessible for a number of *base* classes, with *novel* classes bound by a scarce data regime. This more realistic scenario falls under a meta-learning context, where a representation is learned on the base classes to be employed later on the novel classes.

To leverage the powerful representations that can be learned on the base classes with a DNN, a wide variety of meta-learning methods have been proposed. Santoro et al. [34] make use of a memory network to better assimilate new data and make predictions using it. Edwards et al. [6] aim to make use of learned dataset statistics to better finetune on new samples.
In contrast to more *model-driven* methods, [39] learns an embedding of the labelled examples over which an attention mechanism can be utilized, while [37] learns a mapping from the input to an embedding for which its class is represented by a prototype. Upon learning an embedding, both methods make use of a simple k-nearest neighbor approach to infer the class membership of unseen samples, implying

[†]Work completed while at SAP AI Research, prior to joining Amazon.com, Inc.
[‡]Work completed while at SAP AI Research, prior to joining Huawei Research, Ireland

that they can leverage the representational power of DNNs in a low data regime. See also [41] for a discussion.
However, even when optimizing the learning process for the low-shot scenario, the lack of novel samples remains a hindrance. To mitigate this, a series of generative approaches have been developed, increasing the number of novel class samples that can be utilized during training. Hariharan et al. [11] facilitates training the classifier by generating features, disregarding realism or diversity criteria. While this approach provides a stable meta-learning process, and practically generates useful hallucinated samples, the diversity of generated samples is bound by the samples used to learn the generator. The key assumption is that incorporating multimodal data can provide the means to inject diversity into the generation process. This is achieved by learning a cross-modal mapping which in turn broadens the scope of the generated sample space. The most closely related work to us is [28], which makes use of additional textual data in an adversarial context, followed by a self-paced selection of the most discriminative samples.

Our method builds upon the observation that the representations learned through DNNs are powerful enough for the use of simple non-parametric classification techniques [1], and that multi-modal data can improve generation diversity. To this end, an image encoder is first trained on the available base classes, after which a text-conditional GAN learns a cross-modal mapping between the textual and visual embedding spaces. This mapping can then be used to generate feature representations that reside in the visual space, conditioned by textual data. Intuitively, our method makes use of the cross-modal feature mapping to shift single-modal prototypes p^I (representing visual data) to p^M, mimicking unseen samples of the novel classes. This process can be observed in Fig. 1, where a given sample x_i is classified differently though the shift in the prototypes. In a prototypical space, k-NN, a non-parametric classification technique is used, and thus only the representation learning stage requires multi-modal data, the inference stage requiring only visual data.

The main contributions of this work include the use of a cross-modal feature generation network in the context of few-shot learning. Furthermore, we suggest a strategy to combine real and generated features, allowing us to infer the class membership of unseen samples with a simple nearest neighbor approach. Our method outperforms our baselines and the state-of-the-art approaches for multimodal and image-only few-shot learning by a large margin for the CUB-200 and Oxford-102 datasets.

2. Related Work

In this section we briefly review the previous works related to few-shot learning and multimodal learning.

2.1. Few-Shot Learning

For learning deep networks using limited amounts of data, different approaches have been developed in recent years. Following Taigman et al. [38], Koch et al. [18] interpreted this task as a verification problem, i.e. given two samples, it has to be verified, whether both samples belong to the same class. Therefore, they employed siamese neural networks [3] to compute the distance between the two samples and perform nearest neighbor classification in the learned embedding space. Some recent works approach few-shot learning by striving to avoid overfitting by modifications to the loss function or the regularization term. Yoo et al. [46] proposed a clustering of neurons on each layer of the network and calculated a single gradient for all members of a cluster during the training to prevent overfitting. The optimal number of clusters per layer is determined by a reinforcement learning algorithm. A more intuitive strategy is to approach few-shot learning on data-level, meaning that the performance of the model can be improved by collecting additional related data. Douze et al. [5] proposed a semi-supervised approach in which a large unlabeled dataset containing similar images was included in addition to the original training set. This large collection of images was exploited to support label propagation in the few-shot learning scenario. Hariharan et al. [11] combined both strategies (data-level and algorithm-level) by defining the squared gradient magnitude loss, that forces models to generalize well from only a few samples, on the one hand and generating new images by hallucinating features on the other hand. For the latter, they trained a model to find common transformations between existing images that can be applied to new images to generate new training data [42]. Other recent approaches to few-shot learning have leveraged meta-learning strategies. Ravi et al. [30] trained a long short-term memory (LSTM) network as meta-learner that learns the exact optimization algorithm to train a learner neural network that performs the classification in a few-shot learning setting. This method was proposed due to the observation that the update function of standard optimization algorithms like SGD is similar to the update of the cell state of a LSTM. Similarly, Finn et al. [9] suggested a model-agnostic meta-learning approach (MAML) that learns a model on base classes during a meta learning phase optimized to perform well when finetuned on a small set of novel classes. Moreover, Bertinetto et al. [2] trained a meta-learner feed-forward neural network that predicts the parameters of another, discriminative feed-forward neural network in a few-shot learning scenario. Another technique that has been applied successfully to few-shot learning recently is attention. [39] introduced matching networks for one-shot learning tasks. This network is able to apply an attention mechanism over embeddings of labeled samples in order to classify unlabeled samples. One further outcome

of this work is that it is helpful to mimic the one-shot learning setting already during training by defining mini-batches, called few-shot episodes with subsampled classes. Snell et al. [37] generalize this approach by proposing prototypical networks. Prototypical networks search for a non-linear embedding space (the prototype) in which classes can be represented as the mean of all corresponding samples. Classification is then performed by finding the closest prototype in the embedding space. Other related works include [43, 21, 29].

2.2. Multimodal Learning

Kiros et al [17] propose to align visual and semantic information in a joint embedding space using a encoder-decoder pipeline to learn a multimodal representation. Building upon this, Faghri et al [8] improve the mixed representation by incorporating a triplet ranking loss.

Karpathy et al [14] generate textual image descriptions given the visual data. Their model infers latent alignments between regions of images and segments of sentences of their respective descriptions. Reed et al [32] focus on fine-grained visual descriptions. They present an end-to-end trainable deep structured joint embedding trained on two datasets containing fine-grained visual descriptions.

In addition to multimodal embeddings, another related field using data from different modalities is text-to-image generation. Reed et al [33] study image synthesis based on textual information. Zhang et al [47] greatly improve the quality of generated images to a photo-realistic high-resolution level by stacking multiple GANs (StackGANs). Extensions of StackGANs include an end-to-end trainable version [47] and considering an attention mechanism over the textual input [45]. Sharma et al. [36] extended the conditioning by involving dialogue data and further improved the image quality. Beside the usage of GANs for conditioned image generation, other work employed Variational Autoencoders [16] to generate images [22]. However, they conditioned on attribute vectors instead of text.

Some works have leveraged multimodal data to improve classification results. Elhoseiny et al [7] collect noisy text descriptions and train a model that is able to connect relevant terms to its corresponding visual parts. This allows zero-shot classification for unseen samples, i.e. visual samples for novel classes do not exist. Similarly, Zhu et al. [48] train a classifier with images generated by a GAN given noisy text descriptions and test their approach in a zero-shot setup. Xian et al [44] follow this notion, however, generating feature vectors instead of images. In the context of few-shot learnig, Pahde et al [28, 27, 26] have leveraged textual descriptions to generate additional training images, as opposed to visual feature embeddings generated in this work. Along with a self-paced learning strategy for sample selection this method improves few-shot learning accuracies.

3. Multimodal Prototypical Networks

To define our developed method we first introduce the necessary notation and then describe the architecture of our framework.

3.1. Preliminaries

Let $\mathcal{I}$ denote the image space, $\mathcal{T}$ the text space and $\mathcal{C} = \{1, ..., R\}$ be the discrete label space. Further, let $x_i \in \mathcal{I}$ be the i-th input data point, $t_i \in \mathcal{T}$ its corresponding textual description and $y_i \in \mathcal{C}$ its label. In the few-shot setting, we consider two disjunct subsets of the label space: $\mathcal{C}_{\text{base}}$ - labels for which we have access to sufficient data samples, and $\mathcal{C}_{\text{novel}}$ novel classes, which are underrepresented in the data. Note that both subsets exhaust the label space $\mathcal{C}$, i.e. $\mathcal{C} = \mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{novel}}$. We further assume that in general $|\mathcal{C}_{\text{novel}}| < |\mathcal{C}_{\text{base}}|$.

We organize the data set $\mathcal{S}$ as follows. Training data $\mathcal{S}_{\text{train}}$ consists of tuples $\{(x_i, t_i, y_i)\}_{i=1}^{n}$ taken from the whole data set and test data $\mathcal{S}_{\text{test}} = \{(x_i, y_i) : y_i \in \mathcal{C}_{\text{novel}}\}_{i=1}^{m}$ that belongs to novel classes such that $\mathcal{S} = \mathcal{S}_{\text{train}} \cup \mathcal{S}_{\text{test}}$, $\mathcal{S}_{\text{train}} \cap \mathcal{S}_{\text{test}} = \emptyset$. Naturally, we can also consider $\mathcal{S}_{\text{train}}^{\text{novel}} = \{(x_i, t_i, y_i) : (x_i, t_i, y_i) \in \mathcal{S}_{\text{train}}, y_i \in \mathcal{C}_{\text{novel}}\}_{i=1}^{k} \subset \mathcal{S}_{\text{train}}$, where in accordance with a few-shot scenario $k = \left|\mathcal{S}_{\text{train}}^{\text{novel}}\right| \ll \left|\mathcal{S}_{\text{train}}\right| = n$. Additionally, in a few-shot learning scenario, the number of samples per category of $\mathcal{C}_{\text{base}}$ may be limited to g, denoted by $\mathcal{S}_{\text{train}}^{\text{novel}}(g)$.

Note that contrary to the benchmark defined by Hariharan et al. [11], the few-shot learning scenario in this paper is multimodal in training. However, the testing phase is single-modal on image data of $\mathcal{C}_{\text{novel}}$.

3.2. Nearest Neighbor in Visual Embedding Space

The classification in the embedding space is performed with a simple nearest neighbor approach. The assumption is that given a powerful feature representation, such as ResNet-18 feature vectors, nearest neighbor is a viable choice as classification model and has proven to outperform more sophisticated few-shot learning approaches [1]. Therefore, we use the visual data from base classes $\mathcal{C}_{\text{base}}$ to train an image encoder Φ_I, providing a discriminative visual embedding space φ. For novel visual samples $x_i \in \mathcal{S}_{\text{train}}^{\text{novel}}$, $\Phi_I(x_i)$ then provides the embedding accordingly, featuring discriminativeness given by the pre-trained visual embedding space φ.

Following [37], for every novel class $k \in \mathcal{C}_{\text{novel}}$ we calculate a visual prototype p^k of all encoded training samples:

$$p^k = \frac{1}{|\mathcal{S}_{\text{train}}^{\text{k}}|} \sum_{(x_i, y_i) \in \mathcal{S}_{\text{train}}^{\text{k}}} \Phi_I(x_i), \tag{1}$$

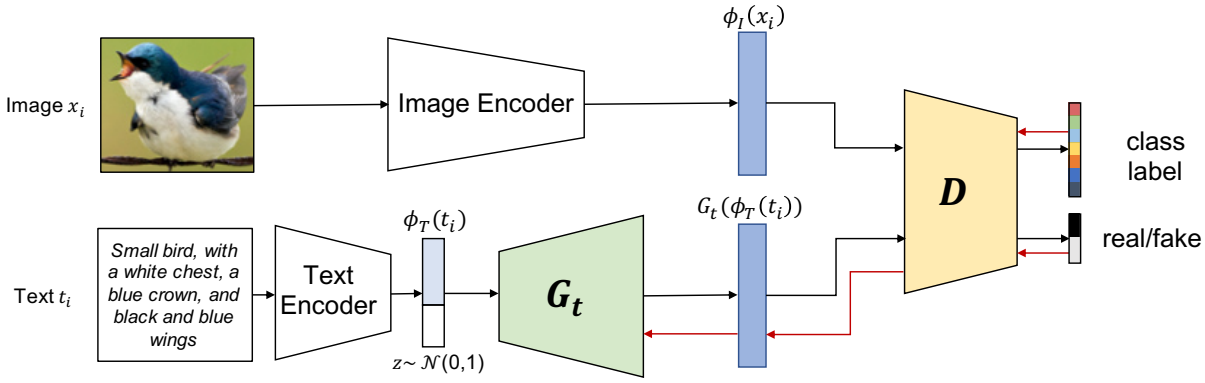

Figure 2: Architecture of our proposed feature-generating network for few-shot learning: The GAN framework containing a generator G_t and a discriminator D is optimized to transform a text embedding given a pre-trained text encoder into the visual embedding space φ yielded by a pre-trained image encoder. The discriminator computes a reconstruction loss (real/fake) and an auxiliary classification loss.

where $\mathcal{S}^{\text{k}}_{\text{train}} = \{(x_i, y_i) \in \mathcal{S}^{\text{novel}}_{\text{train}} | y_i = k\}_{i=1}^{n}$ is the set of all training pairs (x_i, y_i) for class k. Classification of test samples is performed by finding the closest prototype given a distance function $d(\cdot)$. Thus, given a sample $x^{\text{test}} \in \mathcal{S}_{\text{test}}$ the class membership is predicted as follows:

$$c = \arg\min_k d(\Phi_I(x_{\text{test}}), p^k) \tag{2}$$

This assigns the class label of the closest prototype to an unseen test sample. Given the assumption that φ is a discriminative representation of visual data, Eq. 2 provides a powerful classification model. However, due to the few-shot scenario and the intrinsic feature sparsity in training space, $\mathcal{S}^{\text{novel}}_{\text{train}}$ is rather limited such that the computed class prototypes $\{p^k : k \in \mathcal{C}_{\text{novel}}\}$ consequentially yields merely a rough approximation of the true class mean.

3.3. Cross-modal Feature Generation

A viable solution to enrich the training space to enable the calculation of more reliable estimations of the class prototypes is to leverage the multimodality in $\mathcal{S}^{\text{novel}}_{\text{train}}$. Thus, the core idea of our method is to use textual descriptions provided in the training data to generate additional visual feature vectors compensating the few-shot feature sparsity. Therefore, we propose to train a text-conditional generative network G_t that learns a mapping from the encoded textual description into the pre-trained visual feature space φ for a given training tuple (x_i, t_i, y_i) according to

$$G_t(\Phi_T(t_i)) \approx \Phi_I(x_i). \tag{3}$$

For the purpose of cross-modal feature generation we use a modified version of text-condtional generative adversarial networks (tcGAN) [33, 47, 45]. The goal of tcGAN is to generate an image given its textual description in the GAN framework [10]. More specifically, the tcGAN is provided with an embedding $\phi_T(\cdot)$ of the textual description. Therefore, a common strategy is to define two agents G and D solving the adversarial game of generating images that cannot be distinguished from real samples (G) and detecting the generated images as fake (D). Because our strategy is to perform nearest-neighbor classification in a pre-trained embedding space φ, we slightly change the purpose of tcGAN. Instead of generating images $x_i \in \mathcal{I}$, we optimize G to generate its feature representation $\Phi_I(x_i)$ in the space φ. Generally, the representation vector in an embedding space has a significantly lower dimensionality than the original image. Consequentially, the generation of feature vectors is a computational cheaper task compared to the generation of images, can be trained more efficiently and is less error-prone.

To this end, using data from $\mathcal{S}_{\text{train}}$ our modified tcGAN can be trained by optimizing the following loss function,

$$\begin{aligned}\mathcal{L}_{tcGAN}(G_t, D) = \mathbb{E}_{x_i \sim p_{data}}\left[\log D\left(\Phi_I(x_i)\right)\right] \\ + \mathbb{E}_{t_i \sim p_{data}, z}\left[\log D\left(G_t\left(\Phi_T(t_i), z\right)\right)\right],\end{aligned} \tag{4}$$

which entails the reconstruction loss that is used for the traditional GAN implementation [10]. Moreover, following [24, 28, 48] we define the auxiliary task of class prediction during the training of the tcGAN. This entails augmenting the tcGAN loss given in Eq. 4 with a discriminative classification term, which is defined as

$$\mathcal{L}_{class}(D) = \mathbb{E}_{C,I}\left[\log p\left(C \mid I\right)\right] \tag{5}$$

$$\text{and} \quad \mathcal{L}_{class}(G_t) \triangleq \mathcal{L}_{class}(D). \tag{6}$$

Augmenting the original GAN loss with the defined auxiliary term, the optimization objectives for D and G_t can now

Dataset	Method	1-shot	2-shot	5-shot	10-shot	20-shot
CUB	Pahde et al. [28]	57.67	59.83	73.01	78.10	84.24
	Image Only Baseline(Resnet-18+NN)	62.65±0.22	73.52±0.15	82.44±0.09	85.64±0.08	87.27±0.08
	ZSL Baseline (Generated Resnet-18+NN)	58.28±0.22	65.62±0.19	71.79±0.14	74.15±0.11	75.32±0.13
	Our Method (Multimodal Resnet-18 + NN)	**70.39±0.19**	**78.62±0.12**	**84.32±0.06**	**86.23±0.08**	**87.47±0.09**
Oxford-102	Pahde et al. [28]	78.37	91.18	92.21	-	-
	Image Only Baseline (Resnet-18+NN)	84.18±0.48	90.25±0.20	94.18±0.13	95.63±0.14	96.25±0.10
	ZSL Baseline (Generated Resnet-18+NN)	73.35±0.52	77.52±0.34	81.14±0.25	82.95±0.28	83.97±0.21
	Our Method (Multimodal Resnet-18 + NN)	**86.52±0.36**	**91.31±0.18**	**94.57±0.13**	**95.74±0.13**	**96.38±0.10**

Table 1: Top-5 accuracy in comparison to other multimodal few-shot learning approaches and our baselines for CUB-200 (50-way classification) and Oxford-102 (20-way classification) datasets with $n \in \{1, 2, 5, 10, 20\}$

be defined as

$$\mathcal{L}(D) = \mathcal{L}_{tcGAN}(G_t, D) + \mathcal{L}_{class}(D) \tag{7}$$

$$\mathcal{L}(G_t) = \mathcal{L}_{tcGAN}(G_t, D) - \mathcal{L}_{class}(G_t), \tag{8}$$

which are optimized in an adversarial fashion. The adversarial nature of the task forces the generator to focus on the most class-discriminative feature elements. A visualization of our cross-modal feature generating method can be seen in Fig. 2.

3.4. Multimodal Prototype

Having learned a strong text-to-image feature mapping G_t on data provided for the base classes $\mathcal{C}_{\text{base}}$ we can employ the conditional network to generate additional visual features $G_t(\Phi_T(t_i))$ given an textual description t_i and a pre-trained text encoder $\Phi_T(\cdot)$. This allows for computing a prototype from generated samples $G_t(t_i)$ according to

$$p_T^k = \frac{1}{|\mathcal{S}^k_{\text{train}}|} \sum_{(t_i, y_i) \in \mathcal{S}^k_{\text{train}}} G_t(\Phi_T(t_i)). \tag{9}$$

Next, having both the true visual prototype p^k from Eq. 1 and a prototype p_T^k computed from generated feature vectors conditioned on textual descriptions from Eq. 9 a new joint prototype can be computed using a weighted average of both representations:

$$p^k = \frac{p^k + \lambda * p_T^k}{1 + \lambda}, \tag{10}$$

where λ is a weighting factor and $k \in \mathcal{C}_{\text{novel}}$ represents the class label of the prototype. Note that the step in Eq. 10 can be repeated multiple times, because G_t allows for the generation of a potentially infinite number of visual feature vectors in φ. The prediction of the class membership of unseen test samples can now be performed with Eq. 2 using the updated prototypes.

4. Experiments

To confirm the general applicability of our method we perform several experiments using two datasets. These experiments include comparisons to state-of-the-art multimodal and single-modal approaches for few-shot learning.

4.1. Datasets

We test our method on two fine-grained multimodal classification datasets. Specifically, we use the CUB-200-2011 [40] with bird data and Oxford-102 [23] containing flower data for our evaluation. The CUB-200 dataset contains 11,788 images of 200 different bird species, with $\mathcal{I} \subset \mathbb{R}^{256 \times 256}$. The data is split equally into training and test data. As a consequence, samples are roughly equally distributed, with training and test set each containing ≈ 30 images per category. Additionally, 10 short textual descriptions per image are provided by [32]. Similar to [47], we use the text-encoder pre-trained by Reed et al. [32], yielding a text embedding space $\mathcal{T} \subset \mathbb{R}^{1024}$ with a CNN-RNN-based encoding function. Following [47], we split the data such that $|C_{base}| = 150$ and $|C_{novel}| = 50$. To simulate few-shot learning, $n \in \{1, 2, 5, 10, 20\}$ images of C_{novel} are used for training, as proposed by [11]. We perform 50-way classification, such that during test time, all classes are considered for the classification task. In contrast, the Oxford-102 dataset contains images of 102 different categories of flowers. Similar to the CUB-200 dataset, 10 short textual descriptions per image are available. As for the CUB-200 dataset, we use the text-encoder pre-trained by Reed et al. [32], yielding a text embedding space $\mathcal{T} \subset \mathbb{R}^{1024}$. Following Zhang et al. [47], we split the data such that $|C_{base}| = 82$ and $|C_{novel}| = 20$. To simulate few-shot learning, $n \in \{1, 2, 5, 10, 20\}$ images of C_{novel} are used for training. Again, we perform classification among all available novel classes, yielding a 20-way classification task.

4.2. Implementation Details

Image Encoding For image encoding we utilize a slightly modified version of the ResNet-18 architecture [12].

Method	1-shot	5-shot
MAML [9]	38.43	59.15
Meta-Learn LSTM [30]	40.43	49.65
Matching Networks [39]	49.34	59.31
Prototypical Networks [37]	45.27	56.35
Metric-Agnostic Conditional Embeddings [13]	60.76	74.96
ResNet-18 [4]	66.54 ± 0.53	82.38 ± 0.43
ResNet-18 + Gaussian [4]	65.02 ± 0.60	80.79 ± 0.49
ResNet-18 + Dual TriNet [4]	69.61 ± 0.46	84.10 ± 0.35
Image Only Baseline (ResNet-18 + NN)	68.85 ± 0.86	83.93 ± 0.57
Our Full Method (Multimodal ResNet-18 + NN)	$\mathbf{75.01 \pm 0.81}$	$\mathbf{85.30 \pm 0.54}$

Table 2: Top-1 accuracies for the 5-way classification task on the CUB-200 dataset of our approach compared with single-modal state-of-the-art few-shot learning methods. We report the average accuracy of 600 randomly samples few-shot episodes including 95% confidence intervals.

Specifically, we halve the dimensionality of every layer and add two 256-dimensional fully connected layers with LeakyRelu activation after the last pooling layer, followed by a softmax classification layer with $|\mathcal{C}_{\text{base}}|$ units. This network is trained on base classes using the Adam optimizer [15] for 200 iterations with learning rate 10^{-3}, which is decreased to 5×10^{-4} after 20 iterations. The last fully connected layer is employed as embedding space φ.

Cross-Modal Generation For the text-to-image feature mapping we use a tcGAN architecture inspired by StackGAN++ [47]. In the generator G_t, following [47] the text embedding $\Phi_T(t_i)$ is first passed into a conditioning augmentation layer to condense the input space during training. This is followed by some upsampling convolutions, yielding a 256-dim output vector, equivalent to the dimensionality of the image feature space φ. Given the calculated feature vector $G_t(\Phi_T(t_i)$ and the original text embedding $\Phi_T(t_i)$, the discriminator D outputs a conditional and an unconditional loss (see [47]) along with the auxiliary classification loss. Adam is used to optimize both networks with learning rate 2×10^{-4} for 500 iterations. Having trained a feature generating network G_t, we compute $G_t(\Phi_T(t_i))$ for all 10 available textual descriptions per image and take the average in φ as its feature representation.

Classification We predict the class membership of test samples by calculating the nearest prototype in the embedding space φ (see Eq. 2). As distance function we use cosine distance. To average visual and textual prototypes we set $\lambda = 1$ (see Eq. 10) and repeat this step 10 times, updating G_t in every iteration. Hence, in each iteration we reuse real samples from $\mathcal{S}^{\text{k}}_{\text{train}}$, combined with novel generated samples given an updated generator G_t.

4.3. Results

For the evaluation, we test our approach in the 50-way classification task for CUB-200, and 20-way classification for Oxford-102. We designed a strong baseline, in which we predict the class label of unseen test samples by finding the nearest prototype in the the embedding space φ, where the prototype p_I^k is computed exclusively using the limited visual samples (**image only**). Note that nearest neighbor classification is a powerful baseline in the context of few-shot learning, as similarly suggested by other works [1].

Furthermore, we evaluate our method in a zero-shot setting, in which we generate feature vectors given the textual descriptions. The class-label of unseen test samples is predicted by computing the nearest prototype p_T^k containing exclusively generated features conditioned on the textual descriptions (**ZSL**). Our full method calculates the average of both prototypes (**multimodal**).

We compare our method with [28], which to the best of our knowledge is the only existing work leveraging multimodal data in the context of few-shot learning. Because the classification results highly depend on the choice of samples available in a few-shot scenarios, we run the experiments 600 times following [37] and sample a random few-shot episode, i.e. a random choice of n samples per class in every iteration to cover randomness. We report the average top-5 accuracy including 95% confidence intervals in Tab. 1.

It can be observed that in every n-shot scenario we outperform our strong baselines and the other existing approach for multimodal few-shot learning. In the CUB-200 dataset, we outperform the baselines by a large margin, confirming our assumption that multimodal data in training is beneficial. For Oxford-102 the margins are lower, however, we still increase the classification results and outperform state-of-the-art results. Interestingly, our approach also stabilizes the results as the confidence intervals decrease com-

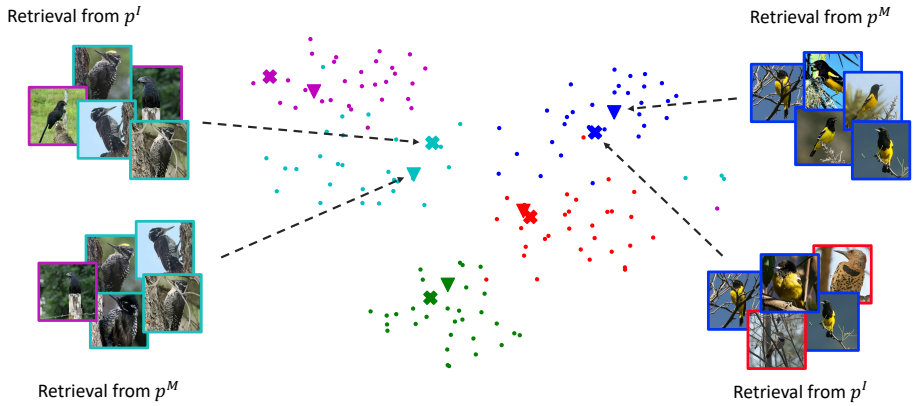

Figure 3: tSNE visualization of test samples (dots), prototypes p_I computed from only real image features (crosses) and multimodal prototypes p_M computed from real image features and generated features conditioned on text (triangles) in 5-way 1-shot scenario for CUB-200. The color indicates the class membership. Furthermore, we show the top-5 results for an image retrieval task for unseen images given the image-only prototype p_I and the multimodal prototype p_M. The color of the border indicates the class membership.

pared to the baselines.

4.4. Comparison to Single-modal Methods

Due to the lack of existing approaches leveraging multimodal data for few-shot learning, we additionally compare our approach to existing methods using only image data during training. Outperforming these state-of-the-art image-only few-shot learning proves the beneficial impact of additional text data during training. Specifically, we compare our method with MAML [9], meta-learning LSTM [30], matching networks [39], prototypical networks [37] and metric-agnostic conditional embeddings [13]. The results for CUB-200 for these methods are provided in [4]. We also include theirs results in our comparison. However, their experimental setup differs slightly from our evaluation protocol. Instead of performing 50-way classification, the results in [4] are reported for 5-way classification in the 1- and 5-shot scenarios. This implies that in every few-shot learning episode, 5 random classes are sampled for which a classification task has to be solved, followed by the choice of n samples that are available per class. For the sake of comparability, we also evaluated our approach in the same experimental setup. We repeat our experiment for 600 episodes and report average top-1 accuracy and 95% confidence intervals in Tab. 2.

It can be observed that even our image-only baseline, which performs nearest neighbor classification using prototypes in our modified ResNet-18 feature representation reaches state-of-the-art accuracies. Note that out image-only baseline can be interpreted as ResNet-version of Prototypical Network [37], which uses a simpler model network architecture in its original version. Including multimodal data during training outperforms the other approaches in both 1- and 5-shot learning scenarios. This proves the strength of our nearest neighbor baseline and shows that enriching the embedding space φ with generated features conditioned on data from other modalites further improves the classification accuracies. In Fig. 3 we show a tSNE visualization of the embedding space φ including the image-only and multimodal prototypes p_I and p_M respectively in the 5-way classification task. The graph clearly shows some clusters indicating the class membership. It can be observed that the generated feature vectors shift the prototypes into regions where more unseen test samples can be classified correctly. Moreover, Fig. 3 shows retrieval results of unseen classes for p_I and p_M. See further retrieval results in the Supplementary Materials.

5. Analysis

In order to get a further in-depth understanding of certain aspects of our method, we performed some additional experiments analyzing its behavior. To this end, we use the CUB-200 dataset for the experiments in this section.

5.1. Reducing Textual Data

In a first experiment we want to analyze the importance of the amount of available textual descriptions. Note that for the experiments in Tab. 1 we used all 10 textual descriptions per image to generate a feature vector $G_t(\Phi(t_i)$. In this experiment we want to understand how the model behaves at reduced text availability. Therefore, in addition to limiting the amount of available images per novel class to n, we limit the amount of textual descriptions per image to $k \in \{1, 2, 5, 10\}$. We evaluate the classification accuracy for $n \in \{1, 2, 5\}$ with reduced number of textual descrip-

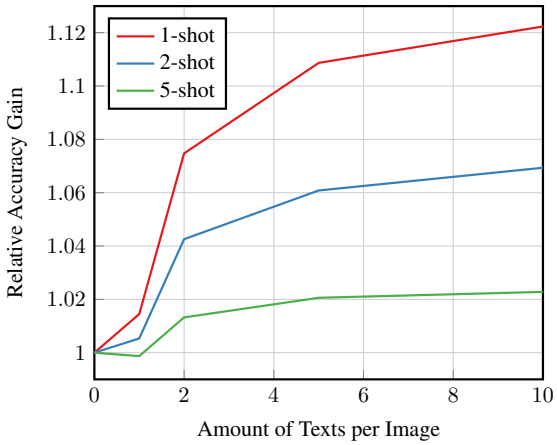

Figure 4: Relative top-5-accuracy gain for different amounts of available texts. The y-axis shows the accuracy gain in relation to the image-only baseline and the x-axis the amount of available texts per image k.

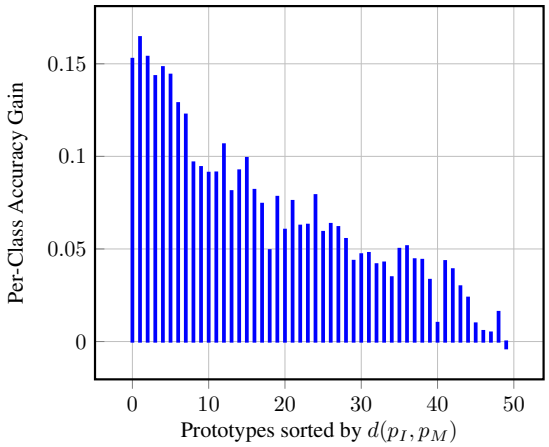

Figure 5: Per-class accuracy gain for prototypes after the adjustment with generated feature vectors. The x-axis shows the rank of the prototype sorted by $d(p_I, p_M)$ and the y-axis the top-5-accuracy gain for that particular class.

tions. In Fig. 4 we show the relative accuracy gains for the different amount of texts compared to the image-only baseline. The x-axis shows the amount of texts and the y-axis the relative accuracy gain. It can be observed that the lower the amount of images n the higher is the accuracy gain given the text. The graphs show an increasing trend which indicates that the more texts are available the more the classification accuracies can be increased. This proves our assumption that enriching the embedding space φ is crucial to reach high classification results. Interestingly, in every n-shot scenario the second text leads to the highest accuracy gain. However, adding more text constantly improves the results and is never harmful to the model.

5.2. Impact of Prototype Shift

We investigate how the adjustment of a certain prototype impacts the classification performance. Therefore, we analyze the per-class accuracy gain in correlation with the shift of the prototype when exposed to multimodal data. The assumption we want to confirm whether large adjustments to the prototype go along with higher accuracy gain compared to classes for which the prototype remains almost unchanged. To this end, we measure change in prototype between the original image-only prototype p_I and the updated multimodal prototype p_M using the cosine distance denoted by $d(p_I, p_M)$. For every novel class, we analyze the correlation of the prototype update to the accuracy gain compared to the image-only baseline. In Fig. 5 we show the per-class accuracy gain for all prototypes in the 1-shot scenario. The x-axis shows the rank of the prototypes for all 50 novel classes of the CUB-200 dataset sorted by $d(p_M, p_I)$ in a descending order. The y-axis represents the accuracy gain for the certain prototype. We report top-5 accuracy and show the average of the result for 100 few-shot episodes. It can be observed that the more the prototype is changed (low rank) the higher is the accuracy gain for this particular class. On average, the most changed prototype leads to a per-class top-5 accuracy gain of ca. 15%. Smaller changes have a smaller impact on the classification performance and on average, adjusting the prototype with multimodal data is not harmful for the accuracy. This suggests that the multimodal features carry complimentary information that is used to simulate unseen novel class samples. At the same time it shows that the text-to-image feature mapping is well learned, as the most diverse, or farthest multimodal features net the largest performance gains.

6. Conclusion and Future Work

In this paper we tackled the few-shot learning problem from a multimodal perspective. Therefore, we proposed to leverage a nearest neighbor classifier in a powerful representation space. To mitigate the low population problem caused by the few-shot scenario we developed a cross-modal generation framework that is capable of enriching the visual feature space given data in another modality (e.g. text). Classification can now be performed by finding the nearest multimodal class prototype to an unseen test sample. We evaluated our proposed methods on the two multimodal datasets CUB-200 and Oxford-102 and showed the applicability of our approach. We outperformed our strong baselines and state-of-the-art single-modal and multimodal methods by a large margin. For future work we plan to follow the notion of [35] which entails optimizing jointly the learned representation on multimodal data.

References

[1] M. Bauer, M. Rojas-Carulla, J. B. Świątkowski, B. Schölkopf, and R. E. Turner. Discriminative k-shot learning using probabilistic models. *Second Workshop on Bayesian Deep Learning at the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[2] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi. Learning feed-forward one-shot learners. In *NIPS*, pages 523–531, 2016.

[3] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. In *NIPS*, pages 737–744, 1994.

[4] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, and L. Sigal. Semantic feature augmentation in few-shot learning. *arXiv preprint arXiv:1804.05298*, 2018.

[5] M. Douze, A. Szlam, B. Hariharan, and H. Jégou. Low-shot learning with large-scale diffusion. *CoRR*, 2017.

[6] H. Edwards and A. Storkey. Towards a neural statistician. *ICLR*, 2017.

[7] M. Elhoseiny, Y. Zhu, H. Zhang, and A. Elgammal. Link the head to the "beak": Zero shot learning from noisy text description at part precision. In *CVPR*, July 2017.

[8] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. *arXiv:1707.05612 [cs]*, July 2017. arXiv: 1707.05612.

[9] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[11] B. Hariharan and R. Girshick. Low-shot Visual Recognition by Shrinking and Hallucinating Features. In *ICCV*, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] N. Hilliard, L. Phillips, S. Howland, A. Yankov, C. D. Corley, and N. O. Hodas. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018.

[14] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[17] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv:1411.2539 [cs]*, Nov. 2014. arXiv: 1411.2539.

[18] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

[19] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Cognitive Science Society*, volume 33, 2011.

[20] F.-F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006.

[21] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. Hwang, and Y. Yang. LEARNING TO PROPAGATE LABELS: TRANSDUCTIVE PROPAGATION NETWORK FOR FEW-SHOT LEARNING. In *International Conference on Learning Representations*, 2019.

[22] A. Mishra, M. Reddy, A. Mittal, and H. A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. *arXiv preprint arXiv:1709.00663*, 2017.

[23] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE, 2008.

[24] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724, 2014.

[26] F. Pahde, P. Jahnichen, T. Klein, and M. Nabi. Cross-modal hallucination for few-shot fine-grained recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR 2018), Workshop on Fine-grained Visual Categorization, Salt Lake City, USA*, 2018.

[27] F. Pahde, M. Nabi, T. Klein, and P. Jahnichen. Discriminative hallucination for multi-modal few-shot learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 156–160. IEEE, 2018.

[28] F. Pahde, O. Ostapenko, P. Jähnichen, T. Klein, and M. Nabi. Self-paced adversarial training for multimodal few-shot learning. *WACV*, 2019.

[29] F. Pahde, M. Puscas, J. Wolff, T. Klein, N. Sebe, and M. Nabi. Low-shot learning from imaginary 3d model. In *WACV*, 2019.

[30] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *InternationalConference on Learning Representations*, 2017.

[31] H. Raza, M. Ravanbakhsh, T. Klein, and M. Nabi. Weakly supervised one shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[32] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016.

[33] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069. PMLR, 2016.

[34] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

[35] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019.

[36] S. Sharma, D. Suhubdy, V. Michalski, S. E. Kahou, and Y. Bengio. Chatpainter: Improving text to image generation using dialogue. *ICLR Workshop*, 2018.

[37] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4080–4090. 2017.

[38] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface:

Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[39] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016.

[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[41] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

[42] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan. Low-Shot Learning from Imaginary Data. In *CVPR*, 2018.

[43] D. Wertheimer and B. Hariharan. Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2019.

[44] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018.

[45] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CVPR*, 2018.

[46] D. Yoo, H. Fan, V. N. Boddeti, and K. M. Kitani. Efficient K-Shot Learning with Regularized Deep Networks. In *AAAI*, 2018.

[47] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[48] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018.