TAFE-Net: Task-Aware Feature Embeddings for Low Shot Learning

Xin Wang   Fisher Yu   Ruth Wang   Trevor Darrell   Joseph E. Gonzalez

UC Berkeley

Abstract

Learning good feature embeddings for images often requires substantial training data. As a consequence, in settings where training data is limited (e.g., few-shot and zero-shot learning), we are typically forced to use a generic feature embedding across various tasks. Ideally, we want to construct feature embeddings that are tuned for the given task. In this work, we propose Task-Aware Feature Embedding Networks (TAFE-Nets[1]) to learn how to adapt the image representation to a new task in a meta learning fashion. Our network is composed of a meta learner and a prediction network. Based on a task input, the meta learner generates parameters for the feature layers in the prediction network so that the feature embedding can be accurately adjusted for that task. We show that TAFE-Net is highly effective in generalizing to new tasks or concepts and evaluate the TAFE-Net on a range of benchmarks in zero-shot and few-shot learning. Our model matches or exceeds the state-of-the-art on all tasks. In particular, our approach improves the prediction accuracy of unseen attribute-object pairs by 4 to 15 points on the challenging visual attribute-object composition task.

1. Introduction

Feature embeddings are central to computer vision. By mapping images into semantically rich vector spaces, feature embeddings extract key information that can be used for a wide range of prediction tasks. However, learning good feature embeddings typically requires substantial amounts of training data and computation. As a consequence, a common practice [8, 14, 53] is to re-use existing feature embeddings from convolutional networks (e.g., ResNet [18], VGG [37]) trained on large-scale labeled training datasets (e.g., ImageNet [36]); to achieve maximum accuracy, these generic feature embedding are often fine-tuned [8, 14, 53] or transformed [19] using additional task specific training data.

In many settings, the training data are insufficient to learn or even adapt generic feature embeddings to a given task. For example, in zero-shot and few-shot prediction tasks, the

---
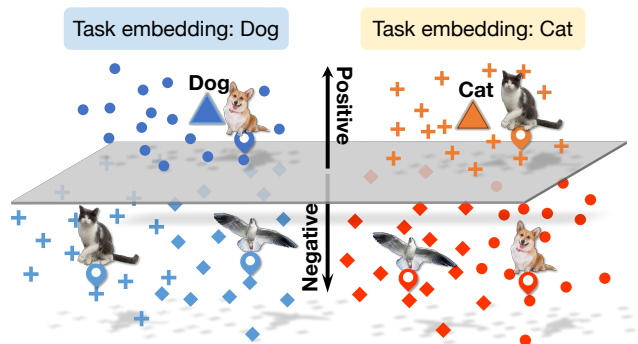
[1]Pronounced taffy-nets

Figure 1: A cartoon illustration of Task-aware Feature Embeddings (TAFEs). In this case there are two binary prediction tasks: `hasCat` and `hasDog`. Task-aware feature embeddings mean that the same image can have different embeddings for each task. As a consequence, we can adopt a single task independent classification boundary for all tasks.

scarcity of training data forces the use of generic feature embeddings [26, 49, 55]. As a consequence, in these situations, much of the research instead focuses on the design of joint task and data embeddings [4, 12, 55] that can be generalized to unseen tasks or tasks with fewer examples. Some have proposed treating the task embedding as linear separators and learning to generate them for new tasks [42, 29]. Others have proposed hallucinating additional training data [50, 17, 45]. However, in all cases, a common image embedding is shared across tasks. Therefore, the common image embedding may be out of the domain or sub-optimal for any individual prediction task and may be even worse for completely new tasks. This problem is exacerbated in settings where the number and diversity of training tasks is relatively small [11].

In this work, we explore the idea of dynamic feature representation by introducing the task-aware feature embedding network (TAFE-Net) with a meta-learning based parameter generator to transform generic image features to task-aware feature embeddings (TAFEs). As illustrated in Figure 1, the representation of TAFEs is adaptive to the given semantic task description, and thus able to accommodate the need of new tasks at testing time. The feature transformation is realized with a task-aware meta learner, which generates the parameters of feature embedding layers within the classi-
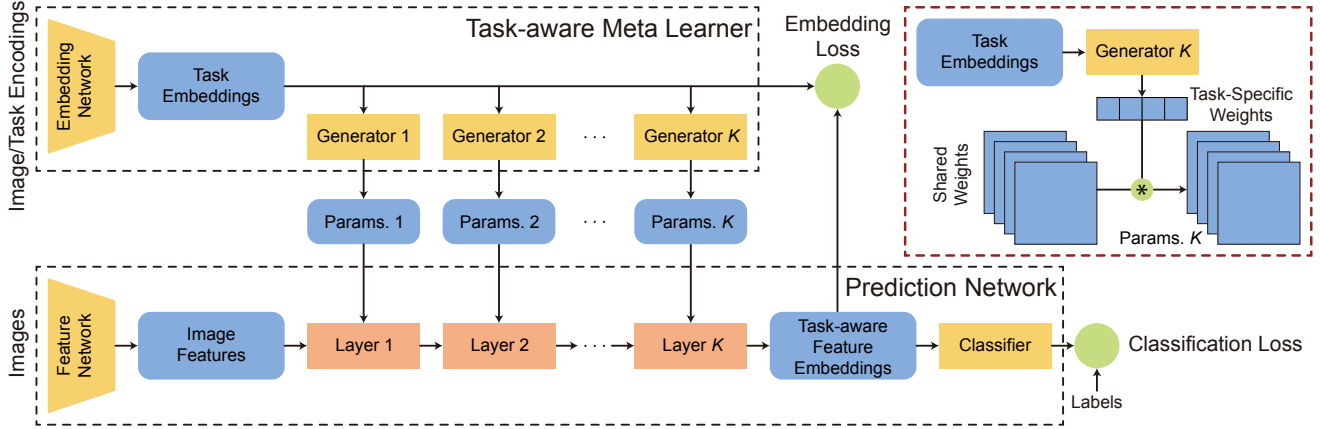
Figure 2: TAFE-Net architecture design. TAFE-Net has a task-aware meta learner that generates the parameters of the feature layers within the classification subnetwork to transform the generic image features to TAFEs. The generated weights are factorized into low-dimensional task-specific weights and high-dimensional shared weights across all tasks to reduce the complexity of the parameter generation. A single classifier is shared across all tasks taking the resulting TAFEs as inputs.

fication subnetwork shown in Figure 2. Through the use of TAFEs, we can adopt a simple binary classifier to learn a task-independent linear boundary that can separate the positive and negative examples and generalize to new tasks.

We further propose two design innovations to address the challenges due to the limited number of training tasks [11] and the complexity of the parameter generation [3]. Dealing with the limited tasks, we couple the task embedding to the task aware feature embeddings with a novel embedding loss based on metric learning. The resulting coupling improves generalization across tasks by jointly clustering both images and tasks. Moreover, the parameter generation requires predicting a large number of weights from a low dimensional task embedding (e.g., a 300-dimensional vector extracted with GloVe [33]), which can be complicated and even infeasible to train in practice, we therefore introduce a novel decomposition to factorize the weights into a small set of task-specific weights needed for generation on the fly and a large set of static weights shared across all tasks.

We conduct an extensive experimental evaluation in Section 4. The proposed TAFE-Net exceeds the state-of-the-art zero-shot learning approaches on three out of five standard benchmarks (Section 4.1) without the need of additional data generation, a complementary approach that has shown boosted performance compared to mere discriminative models by the recent work [50]. On the newly proposed unseen attribute-object composition recognition task [31], we are able to achieve an improvement of 4 to 15 points over the state-of-the-art (Section 4.2). Furthermore, the proposed architecture can be naturally applied to few-shot learning (Section 4.3), achieving competitive results on the ImageNet based benchmark introduced by Hariharan *et al*. [17]. The code is available at https://github.com/ucbdrive/tafe-net.

2. Related Work

Our work is related to several lines of research in zero-shot learning as well as parameter generation, dynamic neural network designs, and feature modulation. Built on top of the rich prior works, to the best of our knowledge, we are the first to study dynamic image feature representation for zero-shot and few-shot learning.

Zero-shot learning falls into the multimodal learning regime which requires a proper leverage of multiple sources (e.g., image features and semantic embeddings of the tasks). Many [23, 52, 42, 55, 4, 12] have studied metric learning based objectives to jointly learn the task embeddings and image embeddings, resulting in a similarity or compatibility score that can later be used for classification [31, 42, 26, 1, 2, 12, 39]. Conceptually, our approach shares the *matching* spirit with the introduction of a binary classifier which predicts whether or not the input image matches the task description. In contrast to prior works, we transform the image features according to the task and thus we only need to learn a task-independent decision boundary to separate the positive and negative examples similar to the classic supervised learning. The proposed embedding loss in our work also adopts metric learning for joint embedding learning but with the main goal to address the limited number of training tasks in meta learning [11]. More recently, data hallucination has been used in the zero-shot [50, 57] and few-shot [17, 45] learning which indicate that the additional synthetic data of the unseen tasks are useful to learn the classifier and can be augmented with the discriminative models [50, 45]. Our (discriminative) model does not utilize additional data points and we show in experiments that our model can match or outperform the generative models on a wide range of benchmarks. We believe the approaches re-

quiring additional data generation can benefit from a stronger base discriminative model.

TAFE-Net uses a task-aware meta learner to generate parameters of the feature layers. Several efforts [3, 16, 7] have studied the idea of adopting one meta network to generate weights of another network. Our task-aware meta learner serves a similar role for the weight generation but in a more structured and constrained manner. We study different mechanisms to decompose the weights of the prediction network so that it can generate weights for multiple layers at once. In contrast, Bertinetton *et al*. [3] focus on generating weights for a single layer and Denil *et al*. [7] can generate only up to 95% parameters of a single layer due to the quadratic size of the output space.

The TAFE-Net design is also related to works on dynamic neural networks [44, 48, 43, 27] which focus on dynamic execution at runtime. SkipNet [44] proposed by Wang *et al*. introduces recurrent gating to dynamically control the network activations based on the input. In contrast, TAFE-Net dynamically re-configures the network parameters rather than the network structure as in the prior works [44, 48] aiming to learn adaptive image features for the given task.

In the domain of visual question answering, previous works [34, 6] explore the use of a question embedding network to modulate the features of the primary convolutional network. Our factorized weight generation scheme for convolutional layers can also be viewed as channel-wise feature modulation. However, the proposed parameter generation framework is more general than feature modulation which can host different factorization strategies [3].

3. Task-Aware Feature Embedding

As already widely recognized, feature embeddings are the fundamental building blocks for many applications [24, 28, 13] in computer vision. In this work, we introduce task-aware feature embeddings (TAFEs), a type of dynamic image feature representation that adapts to the given task. We demonstrate that such dynamic feature representation has applications in the zero-shot learning, few-shot learning and unseen attribute-object pair recognition.

We start with the TAFE-Net model design in Section 3.1 and then introduce the weight factorization (Section 3.2) and the embedding loss (Section 3.3) to address the challenges with the weight generation and the limited number of training tasks. We delay the specifications of different task descriptions and the setup of various applications to Section 3.4.

3.1. TAFE-Net Model

There are two sub-networks in TAFE-Net as shown in Figure 2: a task-aware meta leaner $\mathcal{G}$ and a prediction network $\mathcal{F}$. The task-aware meta learner takes a task description $\mathbf{t} \in \mathcal{T}$ (e.g., word2vec [30] encoding or example images,

detailed in Section 3.4) and generates the weights of the feature layers in the prediction network.

For an input image $\mathbf{x} \in \mathcal{X}$, the prediction network:

$$\mathcal{F}(\mathbf{x}; \theta_t) = \mathbf{y}, \qquad (1)$$

predicts a binary label $\mathbf{y} \in \mathcal{Y}$ indicating whether or not the input image x is compatible with the task description $\mathbf{t}$. More specifically, we adopt a pre-trained feature extractor on ImageNet (e.g., ResNet [18], VGG [37] whose parameters are frozen during training) to produce generic features of the input images and then feed the generic features to a sequence of *dynamic* feature layers whose parameters denoted by θ_t are generated by $\mathcal{G}(\mathbf{t})$. The output of the dynamic feature layers is named as *task-aware feature embedding* (TAFE) in the sense that the feature embedding of the same image can be different under different task descriptions. Though not directly used as the input to $\mathcal{F}$, the task description $\mathbf{t}$ controls the parameters of the feature layers in $\mathcal{F}$ and further injects the task information to the image feature embeddings.

We are now able to introduce a simple binary classifier in $\mathcal{F}$, which takes TAFEs as inputs, to learn a task-independent decision boundary. When multi-class predictions are needed, we can leverage the predictions of $\mathcal{F}(\mathbf{x})$ under different tasks descriptions and use them as probability scores. The objective formulation is presented in Section 3.3.

The task-aware meta learner $\mathcal{G}$ paramterized by η is composed of an embedding network $\mathcal{T}(\mathbf{t})$ to generate a task embedding $\mathbf{e}_t$ and a set of weight generators $\mathbf{g}^i, i = \{1...K\}$ that generate parameters for K dynamic feature layers in $\mathcal{F}$ conditioned on the same task embedding $\mathbf{e}_t$.

3.2. Weight Generation via Factorization

We now present the weight generation scheme for the feature layers in $\mathcal{F}$. The feature layers that produce the task aware feature embeddings (TAFE) can either be convolutional layers or fully-connected (FC) layers. To generate the feature layer weights, we will need the output dimension of $\mathbf{g}^i$ (usually a FC layer) to match the weight size of the i-th feature layer in $\mathcal{F}$. As noted by Bertinetto *et al*. [3], the number of weights required for the meta-learner estimation is often much greater than that of the task descriptions Therefore, it is difficult to learn weight generation from a small number of example tasks. Moreover, the parametrization of the weight generators $\mathbf{g}$ can consume a large amount of memory, which makes the training costly and even infeasible.

To make our meta learner generalize effectively, we propose a weight factorization scheme along the output dimension of each FC layer and the output channel dimension of a convolutional layer. This is distinct from the low-rank decomposition used in prior meta-learning works [3]. The channel-wise factorization builds on the intuition that chan-

nels of a convolutional layer may have different or even orthogonal functionality.

Weight factorization for convolutions. Given an input tensor $\mathbf{x}^i \in \mathbb{R}^{w \times h \times c_{\text{in}}}$ for the i-th feature layer in $\mathcal{F}$ whose weight is $\mathbf{W}^i \in \mathbb{R}^{k \times k \times c_{\text{in}} \times c_{\text{out}}}$ (k is the filter support size and c_{in} and c_{out} are the number of input and output channels) and bias is $\mathbf{b}^i \in \mathbb{R}^{c_{\text{out}}}$, the output $\mathbf{x}^{i+1} \in \mathbb{R}^{w' \times h' \times c_{\text{out}}}$ of the convolutional layer is given by

$$\mathbf{x}^{i+1} = \mathbf{W}^i * \mathbf{x}^i + \mathbf{b}^i, \tag{2}$$

where $*$ denotes convolution. Without loss of generality, we remove the bias term of the convolutional layer as it is often followed by batch normalization [20]. $\mathbf{W}^i = \mathbf{g}^i(\mathbf{t})$ is the output of the i-th weight generator in $\mathcal{G}$ in the full weight generation setting. We now decompose the weight $\mathbf{W}^i$ into

$$\mathbf{W}^i = \mathbf{W}^i_s *_{c_{\text{out}}} \mathbf{W}^i_t, \tag{3}$$

where $\mathbf{W}^i_s \in \mathbb{R}^{k \times k \times c_{\text{in}} \times c_{\text{out}}}$ is a shared parameter aggregating all tasks $\{\mathbf{t}_1, ... \mathbf{t}_T\}$ and $\mathbf{W}_t \in \mathbb{R}^{1 \times 1 \times c_{\text{out}}}$ is a task-specific parameter depending on the current task input. $*_{c_{\text{out}}}$ denotes the grouped convolution along the output channel dimension, i.e. each channel of $x *_{c_{\text{out}}} y$ is simply the convolution of the corresponding channels in x and y. The parameter generator $\mathbf{g}^i$ only needs to generate $\mathbf{W}^i_t$ which reduces the output dimension of $\mathbf{g}^i$ from $k \times k \times c_{\text{in}} \times c_{\text{out}}$ to c_{out}.

Weight factorization for FCs. Similar to the factorization of the convolution weights, the FC layer weights $\mathbf{W}^i \in \mathbb{R}^{m \times n}$ can be decomposed into

$$\mathbf{W}^i = \mathbf{W}^i_s \cdot \text{diag}(\mathbf{W}^i_t), \tag{4}$$

where $\mathbf{W}^i_s \in \mathbb{R}^{m \times n}$ is the shared parameters for all tasks and $\mathbf{W}^i_t \in \mathbb{R}^n$ is the task-specific parameter. Note that this factorization is equivalent to the feature activation modulation, that is, for an input $\mathbf{x} \in \mathbb{R}^{1 \times m}$,

$$\mathbf{x} \cdot (\mathbf{W}^i_s \cdot \text{diag}(\mathbf{W}^i_t)) = (\mathbf{x} \cdot \mathbf{W}^i_s) \odot \mathbf{W}^i_t, \tag{5}$$

where $\odot$ denotes element-wise multiplication.

As a consequence, the weight generators only need to generate low-dimensional task-specific parameters for each task in lower dimension and learn one set of high dimensional parameters shared across all tasks.

3.3. Embedding Loss for Meta Learner

The number of task descriptions used for training the task-aware meta learner is usually much smaller than the number of images available for training the prediction network. The data scarcity issue may lead to a degenerate meta learner. We, therefore, propose to add a secondary *embedding loss* $\mathcal{L}_{\text{emb}}$ for the meta learner alongside the classification loss $\mathcal{L}_{\text{cls}}$ used for the prediction network. Recall that we adopt a shared

binary classifier in $\mathcal{F}$ to predict the compatibility of the task description and the input image. To be able to distinguish which task (i.e., class) the image belong to, instead of using a binary cross-entropy loss directly, we adopt a calibrated multi-class cross-entropy loss [52] defined as

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left[\frac{\exp(\mathcal{F}(\mathbf{x}_i; \theta_t)) \cdot y^i_t}{\sum_{j=1}^{T} \exp(\mathcal{F}(\mathbf{x}_i; \theta_j))} \right], \quad (6)$$

where x_i is the i-th sample in the dataset with size N and $\mathbf{y}_i \in \{0, 1\}^T$ is the one-hot encoding of the ground-truth labels. T is the number of tasks either in the whole dataset or in the minibatch during training.

For the embedding loss, the idea is to project the latent task embedding $\mathbf{e}_t = \mathcal{T}(\mathbf{t})$ into a joint embedding space with the task-aware feature embedding (TAFE). We adopt a metric learning approach that for positive inputs of a given task, the corresponding TAFE is closer to the task embedding $\mathbf{e}_t$ while for negative inputs, the corresponding TAFE is far from the task embedding as illustrated in Figure 1. We use a hinged cosine similarity as the distance measurement (i.e. $\phi(p, q) = \max(\texttt{cosine\_sim}(p, q), 0)$) and the embedding loss is defined as

$$\mathcal{L}_{\text{emb}} = \frac{1}{NT} \sum_{i}^{N} \sum_{t}^{T} ||\phi(\text{TAFE}(\mathbf{x}_i; \theta_t), \mathbf{e}_t) - y^i_t||_2^2. \quad (7)$$

We find in experiments this additional supervision helps training the meta learner especially under the case where the number of training tasks is extremely limited. So far, we can define the overall objective as

$$\min_{\theta, \eta} \mathcal{L} = \min_{\theta, \eta} \mathcal{L}_{\text{cls}} + \beta \cdot \mathcal{L}_{\text{emb}}, \tag{8}$$

where β is the hyper-parameter to balance the two terms. We use β as 0.1 in our experiments if not specified.

3.4. Applications

We now describe how TAFE-Net design can be utilized in various applications (e.g., zero-shot learning, unseen attribute-object recognition and few shot learning) and specify the task descriptions adopted in this work.

Zero-shot learning. In the zero-shot learning (ZSL) setting, the set of classes seen during training and evaluated during testing are disjoint [26, 1]. Specifically, let the training set be $\mathcal{D}_s = \{(x, t, y)|x \in \mathcal{X}, t \in \mathcal{T}, y \in \mathcal{Y}\}$, and the testing set be $\mathcal{D}_u = \{(x, u, z)|x \in \mathcal{X}, u \in \mathcal{U}, z \in \mathcal{Z}\}$, where $\mathcal{T} \cap \mathcal{U} = \phi$, $|T| = |Y|$ and $|U| = |Z|$. In benchmark datasets (e.g., CUB [46], AWA [25]), each image category is associated with an attribute vector, which can be used as the task description in our work. The goal is to learn a classifier $f_{\text{zsl}} : \mathcal{X} \to \mathcal{Z}$. More recently, Xian *et al*. [49] proposed the generalized zero-shot learning (GZSL) setting which is

more realistic compared to ZSL. The GZSL setting involves classifying test examples from both seen and unseen classes, with no prior distinction between them. The classifier in GZSL maps $\mathcal{X}$ to $\mathcal{Y} \cup \mathcal{Z}$. We consider both the ZSL and GZSL settings in our work.

Unseen attribute-object pair recognition. Motivated by the human capability to compose and recognize novel visual concepts, Misra *et al.* [31] recently proposed a new recognition task to predict unseen compositions of a given set of attributes (e.g., red, modern, ancient, etc) and objects (e.g., banana, city, car, etc) during testing and only a subset of attribute-object pairs are seen during training. This can be viewed as a zero-shot learning problem but requires more understanding of the contextuality of the attributes. In our work, the attribute-object pairs are used as the task descriptions.

Few-shot Learning. In few-shot learning, there are one or a few examples from the novel classes and plenty of examples in the base classes [17]. The goal is to learn a classifier that can classify examples from both the novel and base classes. The sample image features from different categories can be used as the task descriptions for TAFE-Nets.

4. Experiments

We evaluate our TAFE-Nets on three tasks: zero-shot learning (Section 4.1), unseen attribute-object composition (Section 4.2 and few-shot learning (Section 4.3). We observe that TAFE-Net is highly effective in generalizing to new tasks or concepts and is able to match or exceed the state-of-the-art on all the tasks.

Model configurations. We first describe the network configurations. The task embedding network $\mathcal{T}$ is a 3-layer FC network with the hidden unit size of 2048 except for the aPY dataset [9] where we choose $\mathcal{T}$ as a 2-layer FC network with the hidden size of 2048 to avoid overfitting. The weight generator $\mathbf{g}^i$ is a single FC layer with the output dimension matching the output dimension of the corresponding feature layer in $\mathcal{F}$. For the prediction network $\mathcal{F}$, the TAFE is generated through a 3-layer FC network with the hidden size of 2048 with input image features extracted from different pretrained backbones (e.g., ResNet-18, ResNet-101, VGG-16, VGG-19, etc.)

4.1. Zero-shot Learning

Datasets and evaluation metrics. We conduct our experiments on 5 benchmark datasets: SUN [51], CUB [47], AWA1 [25], AWA2 [49] and aPY [9], which have different numbers of categories and granularity. In particular, there are only 20 classes (i.e. tasks) available in the aPY dataset while 645 classes are available for training in the SUN dataset. The dataset statistics are shown in Table 1.

Table 1: Datasets used in GZSL

Dataset	SUN	CUB	AWA1	AWA2	aPY
No. of Images	14,340	11,788	30,475	37,322	15,339
Attributes Dim.	102	312	85	85	64
$\mathcal{Y}$	717	200	50	50	32
$\mathcal{Y}^{seen}$	645	150	40	40	20
$\mathcal{Y}^{unseen}$	72	50	10	10	12
Granularity	fine	fine	coarse	coarse	coarse

Following the settings proposed by Xian *et al.*, we consider both the generalized zero-shot learning (GZSL) and the conventional zero-shot learning (ZSL). For GZSL, we report the average per class top-1 accuracy of both unseen acc_u and seen classes acc_s and the harmonic mean $H = 2 \times (acc_u \times acc_s)/(acc_u + acc_s)$. For conventional ZSL, we report the average per-class top-1 accuracy of the unseen classes and adopt the new split provided by Xian *et al.* [49].

Training details. We set the batch size to 32 and use Adam [22] as the optimizer with the initial learning rate of 10^{-4} for the prediction network and weight generators, and 10^{-5} for the task embedding network. We reduce the learning rate by $10\times$ at epoch 30 and 45, and train the network for 60 epochs. For AWA1, we train the network for 10 epochs and reduce the learning rate by $10\times$ at epoch 5.

Baselines. We compare our model with two lines of prior works in our experiments. (1) Discriminative baselines which focus on mapping the images into a rich semantic embedding space. We include the recent competitive baselines: LATEM [55], ALE [1], DeViSE [12], SJE [2], SYNC [4], DEM [54] and the newly proposed RelationNet [52]. (2) Generative models that tackle the data scarcity problem by generating synthetic images for the unseen classes using a GAN [15, 56] based approach. The generative models can combine different discriminative models as base networks [50, 45]. We conduct comparison with f-CLSWGAN [50], SE [41], SP-AEN [5] in this category. Our model falls into the discriminative model category requiring no additional synthetic data.

Quantitative results. We compare the performance of TAFE-Net to the prior works in Table 2. Overall, our model outperforms existing approaches including the generative models on the AWA1, AWA2 and aPY datasets under the ZSL setting and on the AWA1 and aPY datasets under the GZSL setting. TAFE-Net outperforms the discriminative models (denoted in **blue** in Table 2) by a large margin (e.g., roughly 16 points improvement on AWA1 and 17 points on aPY) on the GZSL test. For the more challenging fine-grained SUN and CUB datasets, we are able to improve the results by 7 and 2 points. The results indicate that better embeddings can aid in model generalization.

Table 2: Evaluate TAFE-Net on five standard benchmarks under the ZSL and the GZSL settings. Models with † (f-CLSWGAN, SE and SP-AEC) generate additional data for training while the remaining models do not. **Red** denotes the best performing model on each dataset and **blue** denotes the prior art of discriminative models. Our model is better than all the other discriminative models and also competitive compared to models with additional synthetic data.

Method	Zero-shot Learning					Generalized Zero-shot Learning														
	SUN	CUB	AWA1	AWA2	aPY	SUN			CUB			AWA1			AWA2			aPY		
	T1	T1	T1	T1	T1	u	s	H	u	s	H	u	s	H	u	s	H	u	s	H
LATEM [55]	55.3	49.3	55.1	55.8	35.2	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
ALE [1]	58.1	54.9	59.9	62.5	39.7	21.8	33.1	**26.3**	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DeViSE[12]	56.5	52	54.2	59.7	**39.8**	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
SJE [2]	53.7	53.9	65.6	61.9	32.9	14.7	80.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL [35]	54.5	53.9	58.2	58.6	38.3	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [4]	56.3	**55.6**	54.0	46.6	23.9	7.9	43.3	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
RelationNet [52]	-	**55.6**	68.2	64.2	-	-	-	-	38.1	61.1	**47.0**	31.4	91.3	46.7	30.0	93.4	**45.3**	-	-	-
DEM [54]	**61.9**	51.7	**68.4**	**67.1**	35.0	20.5	34.3	25.6	19.6	57.9	29.2	32.8	84.7	**47.3**	30.5	86.4	45.1	11.1	75.1	**19.4**
f-CLSWGAN† [50]	60.8	57.3	68.2	-	-	42.6	36.6	**39.4**	57.7	43.7	**49.7**	61.4	57.9	59.6	-	-	-	-	-	-
SE† [41]	**63.4**	**59.6**	69.5	69.2	-	40.9	30.5	34.9	53.3	41.5	46.7	67.8	56.3	61.5	58.3	68.1	**62.8**	-	-	-
SP-AEN† [5]	59.2	55.4	-	58.5	24.1	24.9	38.6	30.3	34.7	70.6	46.6	-	-	-	23.3	90.9	37.1	13.7	63.4	22.6
TAFE-Net	60.9	56.9	**70.8**	**69.3**	**42.2**	27.9	40.2	33.0	41.0	61.4	49.2	50.5	84.4	**63.2**	36.7	90.6	52.2	24.3	75.4	**36.8**

Table 3: Ablation of the embedding loss on the five benchmarks under GZSL. Harmonic mean (H) is reported.

Method	SUN	CUB	AWA1	AWA2	aPY
TAFE-Net w/o EmbLoss	**33.1**	45.4	58.8	47.2	30.5
TAFE-Net	33.0	**49.2**	**63.2**	**52.2**	**36.8**

Embedding loss ablation. We provide the harmonic mean of our models with and without the embedding loss under the GZSL setting on five benchmark datasets in Table 3. In general, models with the embedding loss outperform those without the embedding loss except for the SUN dataset whose number of categories is about 3 to $22\times$ larger than the other datasets. This observation matches our assumption that the additional supervision on the joint embedding better addresses the data scarcity (i.e. fewer class descriptions than the visual inputs) of training the controller model.

Embedding visualization. In Figure 3, we visualize the task-aware feature embeddings of images from the aPY dataset under different task descriptions. As we can see, image embeddings of the same image are projected into different clusters conditioned on the task descriptions.

4.2. Unseen Visual-attribute Composition

Besides the standard zero-shot learning benchmarks, we evaluate our model on the visual-attribute composition task proposed by Misra *et al.* [31]. The goal is to compose a set of visual concept primitives like attributes and objects (e.g. `large elephant`, `old building`, etc.) to obtain new visual concepts for a given image. This is a more challenging "zero-shot" learning task, which requires the model not only to predict unseen visual concept compositions but also to model the contextuality of the concepts.

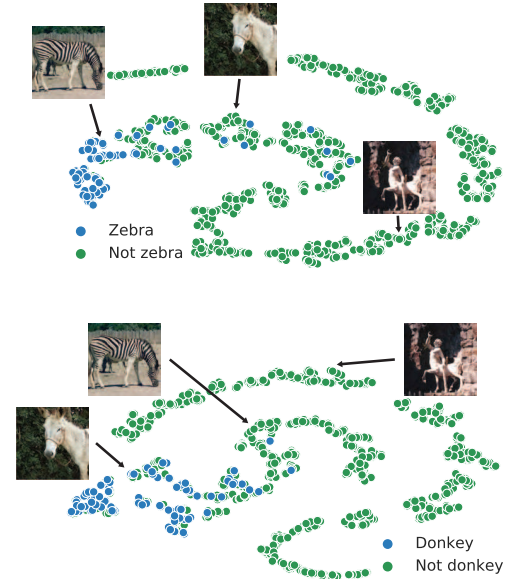**Datasets and evaluation metrics**. We conduct the experi-

Figure 3: Task-aware Image Feature Embedding projected into two dimensions using t-SNE [40] for two tasks (Zebra and Donkey). Note that changing the task produces different embeddings for the same data.

ments on two datasets: MITStates [21] (image samples in Figure 5) and the modified StanfordVRD [29] (image samples in Figure 4). The setup is the same as Misra *et al*. [31]. Each image in the MITStates dataset is assigned a pair of (attribute, object) as its label. The model is trained on 34K images with 1,292 label pairs and tested on 19K images with 700 unseen pairs. The second dataset is constructed based on the bounding box annotations of the StanfordVRD dataset. Each sample has an SPO (subject, predicate, object) tuple as the ground truth label. The dataset has 7,701 SPO triplets and 1,029 of them are seen only in the test split. We

Figure 4: Samples in StanfordVRD. Each image is described by a Subject-Verb-Object triplet. From top left to the bottom right: (elephant, on, grass), (giraffe, in, street), (person, walk, dog), (pillow, behind, person), (person, wears, jeans), (dog, has, shirt).

Table 4: Evaluation on 700 unseen (attribute, object) pairs on 19K images of the MITStates Dataset and 1029 unseen SPO triplets on 1000 images of the StanfordVRD Dataset. TAFE-Net improves over the baselines by a large margin.

| Method | MITStates | | | | StanfordVRD | | | |
| | AP | Top-k Accuracy | | | AP | Top-k Accuracy | | |
		1	2	3		1	2	3
Visual Product [31]	8.8	9.8	16.1	20.6	4.9	3.2	5.6	7.6
Label Embed (LE) [31]	7.9	11.2	17.6	22.4	4.3	4.1	7.2	10.6
LEOR [31]	4.1	4.5	6.2	11.8	0.9	1.1	1.3	1.3
LE + R [31]	6.7	9.3	16.3	20.8	3.9	3.9	7.1	10.4
Red Wine [31]	10.4	13.1	21.2	27.6	5.7	6.3	9.2	12.7
TAFE-Net	**16.3**	**16.4**	**26.4**	**33.0**	**12.2**	**12.3**	**19.7**	**27.5**

evaluate our models only on examples with unseen labels. We extract the image features with pre-trained models on ImageNet. We use VGG-16 and ResNet-101 as our main feature extractors and also test features extracted with VGG-19 and ResNet-18 for ablation. For the task descriptions, we concatenate the word embeddings of the attributes and objects with word2vec [30] trained with GoogleNews. We also consider one-hot encoding for the task ID in the ablation.

For evaluation metrics, we report the mean Average Precision (mAP) of images with unseen labels in the test set together with the top-k accuracy where $k = 1, 2, 3$. We follow the same training schedule as that used in the zero shot learning experiments.

Quantitative results. We compare our model with several baselines provided by Misra *et al.* [31] and summarize the results in Table 4 on both the MITStates and StanfordVRD datasets. Our model surpasses the state-of-the-art models with an improvement of more than 6 points in mAP and 4 to 15 points in top-k accuracy. Nagarajan and Grauman [32] recently proposed an embedding learning framework for visual-attribute composition. They report the top-1 accuracy of 12.0% on the MITStates dataset with ResNet-18 features.

Table 5: Ablation study with different task encoding and base network features. The variance of performance of TAFE-Net under different settings is minimal.

| Task Encoding | Features | AP | Top-k Accuracy | | |
			1	2	3
Word2vec	ResNet-101	16.2	17.2	27.8	35.7
Onehot	ResNet-101	16.1	16.1	26.8	33.8
Word2vec	VGG16	16.3	16.4	26.4	33.0
Onehot	VGG16	16.3	16.4	25.9	32.5
Word2vec	VGG19	15.6	16.2	26.0	32.4
Onehot	VGG19	16.3	16.4	26.0	33.1

For fair comparison, we use the same ResNet-18 features and obtain the top-1 accuracy of 15.1%.

Ablation on the feature extractor and task description. We consider different feature extractors (ResNet-101, VGG-16 and 19) and task encodings (word2vec and one-hot encoding) for ablation and summarize the results in Table 5. The average precision difference between different feature extractors are very minimal (within 0.1%) and the largest gap in Top-3 accuracy is within 2%. This indicates that TAFE-Net is robust in transforming the generic features into task-aware feature embeddings. For the task encoding, the one-hot encoding is comparable to the word2vec encoding and even stronger when using VGG-19 features. This shows that the task transformer network $\mathcal{T}$ is very expressive to extract rich semantic information simply from the task IDs.

Visualization. In Figure 5, we show the top retrievals of unseen attribute-object pairs from the MITStates dataset. Our model can learn to compose new concepts from the existing attributes and objects while respecting their context.

4.3. Few-shot Image Classification

Our model naturally fits the few-shot learning setting where one or few images of a certain category are used as the task descriptions. Unlike prior work on meta-learning which experiments with few classes and low resolution images [42, 38, 10], we evaluate our model on the challenging benchmark proposed by Hariharan and Girshick [17]. The benchmark is based on the ImageNet images and contains hundreds of classes that are divided into base classes and novel classes. At inference time, the model is provided with one or a few examples from the novel classes and hundreds of examples from the base classes. The goal is to obtain high accuracy on the novel classes without sacrificing the performance on the base classes.

Baselines. In our experiments, the baselines we consider are the state-of-the-art meta learning models: Matching Network (MN) [42] and Prototypical Network (PN) [38]. We also
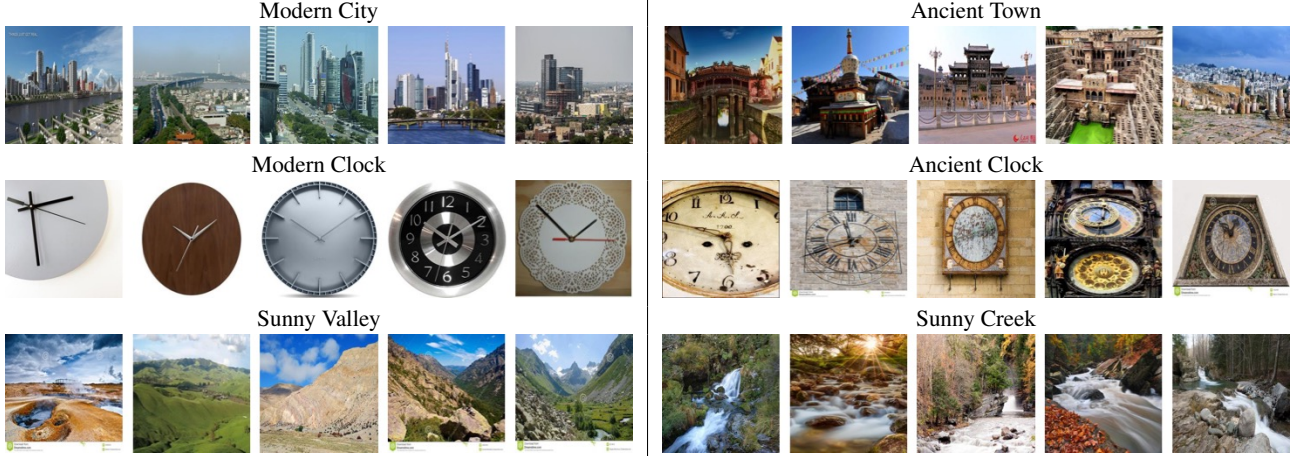
Figure 5: Top retrievals on the unseen pairs of the MITStates dataset. Our model can learn to compose new concepts from the existing attributes and objects while respecting their context. The second row shows some of the failure cases.

Table 6: Few-shot ImageNet Classification on ImageNet. Our model is competitive compared to the state-of-the-art meta learning model without hallucinator.

Method	Novel Top-5 Acc		All Top-5 Acc	
	n=1	n=2	n=1	n=2
LogReg [17]	38.4	51.1	40.8	49.9
PN [38]	39.3	54.4	49.5	61.0
MN [42]	43.6	54.0	54.4	61.0
TAFE-Net	43.0	53.9	**55.7**	**61.9**
LogReg w/ Analogies [17]	40.7	50.8	52.2	59.4
PN w/ G [45]	45.0	55.9	56.9	63.2

compare the logistic regression (LogReg) baseline provided by Hariharan and Girshick [17]. Another line of research [45, 17] for few-shot learning is to combine the meta-learner with a "hallucinator" to generate additional training data. We regard these works as complementary approaches to our meta-learning model.

Experiment details. We follow the prior works [17, 45] to run five trials for each setting of n (the number of examples per novel class, n = 1 and 2 in our experiments) on the five different data splits and report the average top-5 accuracy of both the novel and all classes. We use the features trained with ResNet-10 using SGM loss provided by Hariharan and Girshick [17] as inputs. For training, we sample 100 classes in each iteration and use SGD with momentum of 0.9 as the optimizer. The initial learning rate is set to 0.1 except for the task embedding network (set to 0.01) and the learning rate is reduced by 10× every 8k iterations. The model is trained for 30k iterations in total. Other hyper-paramters are set to the same as Hariharan and Girshick [17] if not mentioned.

Quantitative results. As shown in Table 6, our model is on par with state-of-the-art meta learning models on the novel classes while outperforming them on all categories. Attaching a "hallucinator" to the meta learning model improves performance in general. Our model can be easily attached with a hallucinator and we leave the detailed study as future work due to the time constraint.

5. Conclusion

In this work, we explored a meta learning based approach to generate task aware feature embeddings for settings with little or no training data. We proposed TAFE-Net, a network that generates task aware feature embeddings (TAFE) conditioned on the given task descriptions. TAFE-Net has a task-aware meta learner that generates weights for the feature embedding layers in a standard prediction network. To address the challenges in training the meta learner, we introduced two key innovations: (1) adding an additional embedding loss to improve the generalization of the meta learner; (2) a novel weight factorization scheme to generate parameters of the prediction network more effectively. We demonstrated the general applicability of the proposed network design on a range of benchmarks in zero-/few- shot learning, and matched or exceeded the state-of-the-art.

Acknowledgments

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2016. 2, 4, 5, 6

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015. 2, 5, 6

[3] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *Advances in Neural Information Processing Systems*, pages 523–531, 2016. 2, 3

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 5, 6

[5] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1043–1052, 2018. 5, 6

[6] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017. 3

[7] Misha Denil, Babak Shakibi, Laurent Dinh, Nando De Freitas, et al. Predicting parameters in deep learning. In *Advances in neural information processing systems*, pages 2148–2156, 2013. 3

[8] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014. 1

[9] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. 5

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017. 7

[11] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Conference on Robot Learning*, pages 357–368, 2017. 1, 2

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. DeViSE: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 1, 2, 5, 6

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 5

[16] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 3

[17] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3037–3046. IEEE, 2017. 1, 2, 5, 7, 8

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018. 1

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[21] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 6

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Gregory Koch. Siamese neural networks for one-shot image recognition. 2015. 2

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[25] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009. 4, 5

[26] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 1, 2, 4

[27] Ji Lin, Yongming Rao, Jiwen Lu, and Jie Zhou. Runtime neural pruning. In *Advances in Neural Information Processing Systems*, pages 2178–2188, 2017. 3

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[29] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 1, 6

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013. 3, 7

[31] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, volume 2, page 6, 2017. 2, 5, 6, 7

[32] Tushar Nagarajan and Kristen Grauman. Attributes as operators. *ECCV*, 2018. 7

[33] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[34] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 3

[35] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015. 6

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 3

[38] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 7, 8

[39] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013. 2

[40] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 6

[41] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. 5, 6

[42] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016. 1, 2, 7, 8

[43] Xin Wang, Yujia Luo, Daniel Crankshaw, Alexey Tumanov, Fisher Yu, and Joseph E Gonzalez. Idk cascades: Fast deep learning by learning not to overthink. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018. 3

[44] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 409–424, 2018. 3

[45] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. *CVPR*, 2018. 1, 2, 5, 8

[46] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 4

[47] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 5

[48] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. 2018. 3

[49] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1, 4, 5

[50] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA*, 2018. 1, 2, 5, 6

[51] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010. 5

[52] Flood Sung Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA*, 2018. 2, 4, 5, 6

[53] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1

[54] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5, 6

[55] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016. 1, 2, 5, 6

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017. 5

[57] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1004–1013, 2018. 2