


RESEARCH

Open Access



# Multimodal few-shot classification without attribute embedding

Jun Qing Chang<sup>1\*</sup> , Deepu Rajan<sup>1†</sup> and Nicholas Vun<sup>1†</sup>

<sup>†</sup>Deepu Rajan and Nicholas Vun contributed equally.

\*Correspondence:  
[junqing001@e.ntu.edu.sg](mailto:junqing001@e.ntu.edu.sg)

<sup>1</sup>Nanyang Technological University, 50 Nanyang Ave, Singapore, Singapore

## Abstract

Multimodal few-shot learning aims to exploit complementary information inherent in multiple modalities for vision tasks in low data scenarios. Most of the current research focuses on a suitable embedding space for the various modalities. While solutions based on embedding provide state-of-the-art results, they reduce the interpretability of the model. Separate visualization approaches enable the models to become more transparent. In this paper, a multimodal few-shot learning framework that is inherently interpretable is presented. This is achieved by using the textual modality in the form of attributes without embedding them. This enables the model to directly explain which attributes caused it to classify an image into a particular class. The model consists of a variational autoencoder to learn the visual latent representation, which is combined with a semantic latent representation that is learnt from a normal autoencoder, which calculates a semantic loss between the latent representation and a binary attribute vector. A decoder reconstructs the original image from concatenated latent vectors. The proposed model outperforms other multimodal methods when all test classes are used, e.g., 50 classes in a 50-way 1-shot setting, and is comparable for lesser number of ways. Since raw text attributes are used, the datasets for evaluation are CUB, SUN and AWA2. The effectiveness of interpretability provided by the model is evaluated by analyzing how well it has learnt to identify the attributes.

**Keywords:** Multimodal learning, Few-shot classification, Image classification

## 1 Introduction

Few-shot learning (FSL) enables a model to learn in low data scenarios [1–5]. A popular method for FSL is meta-learning in which the model learns a representation for base classes that have abundant data available, and which is then fine-tuned and tested on novel classes with less data [1, 3, 4]. The support set from novel classes are employed for fine-tuning and the query set is used for testing. When there are  $N$  novel classes and  $K$  images in the support set for each class, the testing regime is termed as  $N$ -way- $K$ -shot.

When multiple modalities are available, the complementary information contained in them could be exploited to enrich the few-shot learning model. Typically, the base classes and the support set of novel classes contain images and texts that describe them, and the query set in the novel classes contain only images [6]. Two main approaches for multimodal FSL involve learning a multimodal representation

in a joint embedding space of visual and semantic information [7], and leveraging textual descriptions to generate additional training images [8, 9]. In this paper, while the visual information is represented in an embedding space, the semantic information is used in its raw form, i.e., as image attributes without embedding. There are two advantages that follow. First, the model becomes interpretable in that the specific attributes that contribute to a particular classification is immediately evident. In fact, the model is inherently interpretable implying that there is no need of additional visualization steps such as layerwise relevance propagation [10] or Grad-CAM and its variants [11]. Although [12] visualize attention maps directly from the learned latent embedding of a variational autoencoder (VAE), it is not evident that the method would work in a few-shot setting. In [13], a separate language model is trained to produce an explanation for a given feature embedding and class label. The second advantage is that, in effect, the model is learning the composition of an image. Compositionality is integral to the human representation of a concept by way of decomposing it into parts and cognitive studies have demonstrated its critical role in human vision [14]. The semantic information in an image readily provided by the model identifies not only the parts in the image, but also their attributes.

The method proposed in this paper is based on a hybrid autoencoder framework, i.e., it contains a basic autoencoder as well as a variational autoencoder (VAE). First, image features are encoded into a semantic space by the former where the encoded semantic features are enforced to be close to the binary ground-truth attributes. The attributes contributing to a classification can be directly read off from the semantic space. The learnt encoder weights are retained while the VAE encoder learns the embedding of image features into the visual space. The visual and semantic features are concatenated and decoded for classification. Thus, the proposed framework performs multimodal few-shot classification while directly providing the attributes for a classified image, this enables interpretability of the model. An example of application for the model would be in the aerospace manufacturing industry which needs to identify defective parts. It is required by regulatory authorities for such critical industries that decisions taken by a machine is interpretable. By performing classification that are interpretable, the proposed framework can be extended to such real-world scenarios.

In [7], the authors have addressed the question of how expensive it is to label images with attributes and furthermore, how to define a vocabulary for the attributes. The authors state that labeling 159 category-level attributes for a subset of ImageNet images took only 3 days, noting that the novel classes did not need attribute annotation.

The main contributions of this paper are as follows: (1) a multimodal framework is proposed consisting of a basic autoencoder whose semantic features can directly provide the attributes for a classified image, and a VAE that provides the visual features that are concatenated with semantic features for few-shot classification. (2) The model outperforms other multimodal methods in 50-way- $K$ -shot on CUB dataset, and have comparable results in fewer number of ways, while simultaneously explaining the results without additional training. Note that neither zero-shot learning [15, 16] nor generalized few-shot learning [17] is addressed in this paper. (3) The model is shown to be interpretable using the attributes predicted as part of the model.

## 2 Related work

### 2.1 Few-shot learning

A comprehensive survey on few-shot learning is presented in [18], where the three main approaches are data augmentation, reducing the space of hypotheses that maps input features to labels and searching for parameters of the best hypothesis. Here, some state-of-the-art methods are picked and reviewed briefly.

One approach to data augmentation for FSL is to use hand-crafted rules such as [19] where rotation and translation of images are used to augment the dataset to train VAE for new samples generation. By generating new and additional samples, it circumvents the problem of having small amount of data in a few-shot learning setting. The main disadvantage in using hand-crafted rules is that it is impossible to enumerate all possible variations. In addition, it is costly and requires domain knowledge to apply these hand-crafted rules. To avoid using hand-crafted rules, Generative Adversarial Network (GAN) models have been used for generation of synthetic data [20]. The GAN model learns from other larger datasets first, before being used on few-shot datasets. This allows new samples to be generated without the need of hand-crafted rules. One disadvantage of this is that the larger datasets used to train the GAN model have to be related to the few-shot datasets. This in practice might not be possible all the time.

Storing knowledge from training data as external memory is a method that reduces the space of hypotheses. Instead of using embedding of samples directly, [21] uses these embeddings as a key to query the most similar memory. The values from the most similar memory are combined to form the representation of the sample. This allows the model to predict based on these representation instead of embeddings of few-shot examples that might not be sufficient. The downside of this is that manipulating the memory is expensive resulting in the memory being typically small. When the memory is full, a decision on which memory to replace has to be made which may result in worse performance if the wrong memory is chosen. More recently, by using embedding learning to reduce the search space, [2] suggests that training a linear classifier on top of a supervised or self supervised representation is sufficient for few-shot learning. By doing so, it is believed that only a good embedding is required for few-shot classification. This however raises the question on how to obtain a good embedding. In a scenario for classifying common objects, obtaining an embedding for that might be easy. However, in more complicated cases for example in industrial applications, such embeddings might not be easily obtainable.

FSL can also be approached by searching for parameters of the best hypothesis. This can be done by teaching an optimizer to find the optimal update parameters at every step [22]. This allows the step size or search direction to be determined by the learned optimizer instead of using hand-crafted update rules. However, this may raise issues on how to transfer the optimizer between different data sources or granularities.

### 2.2 Multimodal few-shot learning

Multimodal few-shot learning extends regular few-shot learning by including additional modalities. Common forms of modalities include attributes like those used in zero-shot learning such as shape and color of bird parts, objects in a scene, or the

color and behavior of animals. Song et al. [23] did a more recent survey on few-shot learning and included a section on multimodal few-shot learning. Here, some methods in multimodal few-shot learning are discussed.

CCAM [24] encodes context and visual information to the same embedding space, allowing the use of contextual prototypes to be used instead of real labels. Classification can be done by comparing the distance to these prototypes. This allows for contextual information to be used instead for few-shot learning. Schwartz et al. [25] refine visual prototypes by using Multi-Layer Perceptrons (MLPs) to generate semantic prototypes. The visual and semantic prototypes are then combined to form a final prototype. This allows the prototype to be more suitable for few-shot learning. Pahde et al. [26] introduce hallucinated samples conditioned on textual description as an augmented dataset. This enables additional samples to be generated for few-shot learning. Using prototypical networks, [8] propose a multimodal prototypical network model to map semantic information to the visual space for a better prototype. This allows additional visual features to be generated as prototypes for multimodal few-shot learning. Instead of using modality-alignment methods, [27] introduces an adaptive modality mixture mechanism for multimodal few-shot learning. By combining the visual and textual modality, it significantly improves performance on few-shot learning problems. Chen et al. [28] approaches the problem by mapping samples to the semantic space and augmenting them with noise. These augmented features can then be projected back to the visual space to generate new samples. By constraining image representations to predict natural language, [29] uses language as a bottleneck to reconstruct features used for classification. This use of natural language proved to help significantly with few-shot learning. Mu et al. [30] use the same constraints on image representation and classify with the learned visual representations, further improving the previous method without the need of the language model at test time. This makes the model simpler and more data efficient. Compositionality for multimodal few-shot learning is addressed in [7] by applying constraints to ensure that the similarities between the image and textual representation is maximized. Improved performance of multimodal few-shot learning is shown when learning compositions in images.

These methods achieve good performance in multimodal few-shot learning, but are not interpretable due to the need of embedding the modalities. Interpretable few-shot learning has been presented in [13], but it is only through learning a language model to generate captions from their feature space, which is a separate module to the basic framework. This requires an additional step on top of classification.

Multimodal learning has also been explored in a zero-shot learning setting. To account for problems in generation shifts such as semantic inconsistency, [31] introduce a generative flow framework using conditional affine coupling layers. Some generalized zero-shot learning methods introduce small amounts of visual information to their existing framework for generalized few-shot learning setting. By aligning embeddings of visual and other modalities using VAE, [17] perform generalized zero-shot learning using embeddings of other modalities as classification samples. Samuel et al. [32] address the zero-shot learning problem by introducing a module that address the long-tail problem by rebalancing class predictions across classes on

a sample-by-sample basis. In both methods, by introducing small samples of visual information, it is shown that these methods are able to perform generalized few-shot learning as well.

### 3 Variational autoencoders (VAE)

An autoencoder consists of a combination of an encoder and a decoder and aims to learn a latent representation of given data by constraining information flowing through a network with a bottleneck. The latent representation is learnt by minimizing a loss between the input  $x$  to the encoder and the output  $f(x)$  of the decoder. If the loss is the L1 distance, it is given by

$$\mathcal{L} = \mathbb{E}[|x - f(x)|], \quad (1)$$

where the expectation is taken over the training data.

The irregularity in the latent space of an autoencoder arising due to overfitting is addressed by forcing the encoder to return a distribution over the latent space as opposed to a single point. This structure is called a variational autoencoder [33]. Consider the latent representation  $z$  to be sampled from a prior distribution  $p(z)$ . The encoder outputs parameters to the distribution of the encoded variable given input as  $q_\theta(z|x)$ . The decoder takes as input the latent representation and outputs the parameters to the distribution of the data, i.e.,  $p_\phi(x|z)$ . The loss function is given by

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)] + D_{\text{KL}}[q_\theta(z|x) || p(z)]. \quad (2)$$

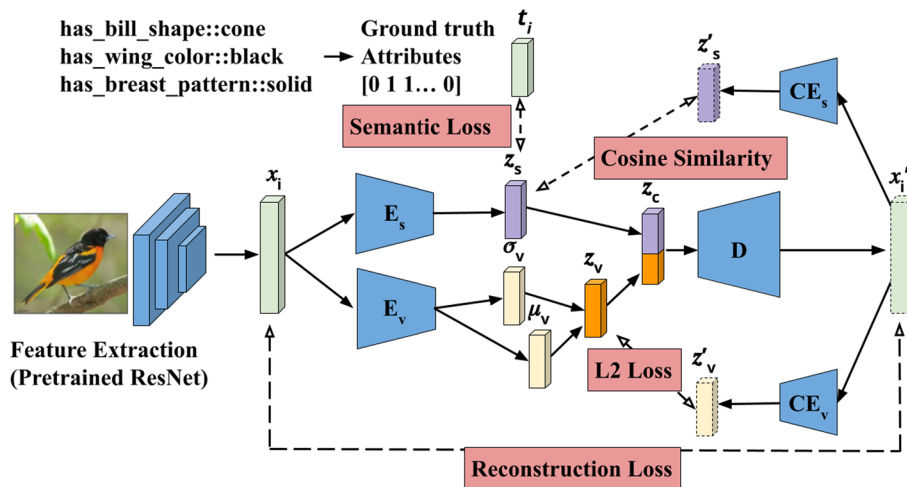
The first term is the reconstruction loss that forces the decoder to learn to reconstruct the data. The second term is a regularizer whose objective is to make the distributions returned by the encoder to be close to a standard Gaussian distribution. This enables the latent space to be organized such that encodings of similar datapoints are close together. It is implemented through the KL divergence between the encoder's distribution  $q_\theta(z|x)$  and the prior  $p(z)$ .

### 4 Proposed method

Let  $C$  be the discrete label space. For multimodal few-shot setting, the training set  $D_{\text{train}} = \{(x_i, y_i, m_i^{(u)})\}_{i=1}^n$  from  $C_{\text{base}}$  classes that contain sufficient number of data samples, where  $x_i$  is the image data with class label  $y_i$  and  $m_i^{(u)}$  are different modalities indexed by  $u$ . The support set  $D_{\text{support}} = \{(x_i, y_i, m_i^{(u)})\}_{i=1}^s$  consists of similar triplets drawn from  $C_{\text{novel}}$  classes that contain fewer data. The query set  $D_{\text{test}} = \{x_i\}_{i=1}^l$  is drawn from  $C_{\text{novel}}$  classes. Thus  $D_{\text{train}}$  is the meta training set and  $D_{\text{support}}$  and  $D_{\text{test}}$  together make up the meta testing set. In this case,  $u = 1$  for text modality.

Figure 1 shows the architecture of the proposed multimodal few-shot learning framework. It consists of (i) a basic autoencoder whose encoder is denoted as  $E_s$  and decoder as  $D$ , (ii) a VAE whose encoder is  $E_v$  and shares the same decoder  $D$  and (iii) two additional encoders denoted as  $CE_s$  and  $CE_v$  that ensure cyclic consistency for the semantic and visual components, respectively.

Following [8], CNN features from a pre-trained ResNet-18 model are used. These features are encoded into a semantic representation of the image as  $z_s$ , which is the latent variable that the model learns to represent the attributes of the image. It is from this representation



**Fig. 1** The proposed model for multimodal few-shot learning with losses between components shown. Reconstruction loss is enforced on the input  $x_i$  and reconstructed features  $x'_i$ . Semantic loss is enforced on the output of encoder  $E_s$  to ensure that it is as close as possible to the ground-truth attributes. To ensure cyclic consistency, two different losses are used for each modality: the cosine similarity loss between  $z_s$  and  $z'_s$  for the semantics, and L2 loss between  $z_v$  and  $z'_v$  for the visual

that the attributes contributing towards a certain classification can be read off. The learning of  $z_s$  is achieved by forcing the representation towards a binary attribute vector  $t$  whose element is 1 indicating the presence of an attribute and 0 otherwise. Eventually, the learnt latent representation consists of the probability of each attribute in the image enforced through a sigmoid function. Thus, the proposed framework is naturally interpretable without additional computation for visualization or other models for interpretation. A Bernoulli latent representation in an autoencoder has been studied in [34]. The formulation of the semantic loss can be considered as a multi-label problem where the targets are the attributes and is written in the form of a binary cross-entropy loss as

$$\mathcal{L}_S = -\frac{1}{N} \sum_{j=1}^N t_{ij} \cdot \log(p(t_{ij})) + (1 - t_{ij}) \cdot \log(1 - p(t_{ij})), \tag{3}$$

where  $t_{ij}$  is the  $j$ th attribute for sample  $i$  and  $p(t_{ij})$  is its estimated probability.

The VAE uses encoder  $E_v$  to learn a latent representation of the visual feature  $x_i$  parameterized by a Gaussian distribution with mean  $\mu_v$  and standard deviation  $\sigma_v$ . These parameters are employed to randomly generate the latent visual representation  $z_v$  through the reparameterization trick [33]. The output of both encoders is then concatenated into a latent variable  $z_c$  that fully describes the image in terms of the visual features as well as its attributes. A single decoder  $D$  is used to reconstruct the image features from  $z_c$ . Thus, the two encoders and a decoder form what is named in this paper as a hybrid autoencoder that consists of a basic autoencoder and a VAE. The reconstruction loss for the basic autoencoder is taken as

$$\mathcal{L}_R = \mathbb{E}[|x'_i - x_i|], \tag{4}$$

where  $x'_i$  is the reconstructed image feature. As described in Sect. 3, the loss function for a VAE consists of the reconstruction loss and a regularizer. Since the semantic information is encapsulated in the basic autoencoder, the expected log-likelihood term of the VAE reconstruction loss is replaced with  $\mathcal{L}_R$ . The regularizer term is retained as the KL divergence between the distribution of encoder  $E_v$ ,  $q_\phi(z_v | x_i)$ , and  $p_\theta(z_v)$ . Taken together, the loss function of the hybrid autoencoder is

$$\mathcal{L}_{\text{HAE}} = \alpha \mathcal{L}_R - \beta D_{\text{KL}}[q_\phi(z | x_i) || p_\theta(z_v)], \quad (5)$$

where  $\alpha$  and  $\beta$  are the weights for each component.

Next, cyclic consistency is considered to ensure that the reconstructed feature  $x'_i$  generated by the hybrid autoencoder can fully encode both the semantic as well as visual information in the image. To this end,  $x'_i$  is converted back into latent semantic and visual representations,  $z'_s$  and  $z'_v$ , respectively, through encoders  $CE_s$  and  $CE_v$  that have the same structure as  $E_s$  or  $E_v$ . The semantic similarity between the encoded semantic representation is the cosine distance. For visual similarity, the output of  $CE_v$  is compared with the output of  $E_v$  using L2 loss.

Applying the visual cyclic constraint to  $\mu_v$  results in a softer constraint compared to applying it to  $z_v$ , since in the latter case, the similarity of a specific sample to its reconstruction is maximized as opposed to maximizing to the mean of the distribution. From the experiments shown in Table 1, it is observed that a more restrictive constraint results in a better performing model, specifically at lower number of shots. Maximizing the similarities corresponds to minimizing the representation consistency loss [35]:

$$\mathcal{L}_{\text{cyclic}} = \frac{\|\mu_v - \mu'_v\|^2}{\cos(z_s, z'_s) + \epsilon}, \quad (6)$$

where  $\cos$  is the cosine similarity and  $\epsilon = 0.1$  is a constant to avoid division by zero.

The overall loss for the model combines the semantic loss, the hybrid autoencoder loss and the cyclic loss as

$$\mathcal{L} = \mathcal{L}_{\text{HAE}} + \gamma \mathcal{L}_S + \delta \mathcal{L}_{\text{cyclic}}, \quad (7)$$

where  $\gamma$  and  $\delta$  are weights for semantic and cyclic loss, respectively.

**Table 1** Comparing accuracy when cyclic consistency is considered with respect to  $z_v$  (hard constraint) and to  $\mu_v$  (soft constraint) in 50-way classification on CUB

Method	Metric	1-shot	2-shot	5-shot	10-shot	20-shot
$z_v$	Top-1	<b>29.00</b>	<b>33.83</b>	<b>49.10</b>	57.60	64.57
	Top-3	<b>49.74</b>	53.89	<b>70.96</b>	<b>80.83</b>	84.70
	Top-5	<b>59.89</b>	64.74	<b>80.79</b>	<b>87.92</b>	90.59
$\mu_v$	Top-1	26.50	33.55	45.60	59.02	<b>65.94</b>
	Top-3	44.60	<b>54.59</b>	69.19	79.41	<b>86.83</b>
	Top-5	54.61	<b>65.16</b>	79.09	86.66	<b>92.58</b>

Bold values represent best performing scores in the individual categories

## 5 Experiments

First, the datasets used to evaluate the proposed hybrid autoencoder framework are described. Next, some implementation details are discussed and then comparison of the model with other state-of-the-art methods for multimodal few-shot image classification are provided. Finally, the effectiveness of the inherent interpretability of the model is demonstrated.

### 5.1 Datasets

The model is evaluated on three datasets: Caltech-UCSD Birds-200-2011 (CUB) [36], SUN [37], and Animals with Attributes 2 (AWA2) [15]. CUB dataset is an image dataset of 200 bird species and their attributes. The image features used were obtained from the final pooling layer of a ResNet-18 similar to [8]. In addition, to ensure that support and test classes are disjoint from the classes in ResNet, the proposed training splits in [15] were used. In this split,  $|C_{\text{base}}| = 150$  and  $|C_{\text{novel}}| = 50$ . Following few-shot learning methods, this results in  $K \in \{1, 2, 5, 10, 20\}$  images of  $C_{\text{novel}}$  in the support set.  $N \in \{5, 10, 20, 50\}$  way classification were performed. The image features and attributes generated are also provided in the dataset.

SUN is a scene dataset with 717 classes split into  $|C_{\text{base}}| = 645$  and  $|C_{\text{novel}}| = 72$  with  $N \in \{5, 10, 20, 50, 72\}$  and  $K \in \{1, 2, 5, 10\}$ . Unfortunately, there are no results reported for few-shot learning on this dataset; instead it is used for zero-shot learning and generalized few-shot learning, which the model is not intended for.

AWA2 is an animal dataset consisting of 50 classes that are split into  $|C_{\text{base}}| = 40$  and  $|C_{\text{novel}}| = 10$  with  $N \in \{5, 10\}$  and  $K \in \{1, 2, 5, 10, 20\}$ . Similar to the SUN dataset, there are no results reported for this dataset for multimodal few-shot classification.

### 5.2 Implementation details

Image features are embeddings of 512 dimensions obtained from an ImageNet pre-trained ResNet-18 from PyTorch. Semantic features are the class-level attributes provided with the dataset whose values range from 0.0 to 1.0. A binary attribute vector is created by assigning an attribute as 1 if its value is greater than zero and 0 otherwise. For the encoders and decoders, the sizes of the hidden layers are 1560 and 1660, respectively. The size of the latent space  $z_v$  is 64, and the size of  $z_s$  follows the number of attributes in a dataset. The optimizer that is used is an Adam optimizer with a learning rate of 0.00015. The class of the test samples are predicted by training a simple single-layer linear classifier on the concatenated  $z$ . Here cross-entropy loss is used. The Adam optimizer has a learning rate of 0.001.

### 5.3 Performance evaluation

First, the results are compared with [8], which has the best performing 50-way classification accuracy on CUB in the multimodal few-shot learning scenario. Table 2 compares the performance for 50-way classification on CUB with [26] and [8] including Top-1, Top-3 and Top-5 accuracies. The proposed method outperforms [26] in 5- or more shots for all metrics. It also outperforms [8] at higher number of shots for all metrics. Note that both [26] and [8] embed attributes into a semantic space unlike the



**Table 2** 50-way classification accuracy on CUB

Method	Metric	1-shot	2-shot	5-shot	10-shot	20-shot
<i>With attribute embedding</i>						
Pahde et al. [26]	Top-1	24.90	25.17	34.66	44.00	53.70
	Top-3	37.59	39.75	49.86	59.62	67.99
	Top-5	57.67	59.83	73.01	78.10	84.24
Multimodal prototypical Network [8]	Top-1	<b>34.16</b>	<b>41.43</b>	48.84	53.01	55.58
	Top-3	<b>58.56</b>	<b>67.44</b>	<b>74.65</b>	77.60	79.30
	Top-5	<b>70.39</b>	<b>78.62</b>	<b>84.32</b>	86.23	87.47
<i>Without attribute embedding</i>						
Proposed method (ResNet-18)	Top-1	29.00	33.83	<b>49.10</b>	<b>57.60</b>	<b>64.57</b>
	Top-3	49.74	53.89	70.96	<b>80.83</b>	<b>84.70</b>
	Top-5	59.89	64.74	80.79	<b>87.92</b>	<b>90.59</b>
Proposed method (ResNet-101)	Top-1	<b>42.61</b>	<b>52.95</b>	<b>63.67</b>	<b>71.79</b>	<b>74.73</b>
	Top-3	<b>66.30</b>	<b>75.76</b>	<b>85.57</b>	<b>88.16</b>	<b>90.70</b>
	Top-5	<b>73.57</b>	<b>84.90</b>	<b>90.98</b>	<b>94.00</b>	<b>94.36</b>

Bold values represent best performing scores in the individual categories

**Table 3** 5-way classification accuracy of Top-1 on CUB

Method	1-shot	5-shot
<i>With attribute embedding</i>		
AM3-TADAM [27]	74.10	79.70
Multimodal prototypical network [8]	75.01	85.30
Dual TriNet [28]	69.61	84.10
f-VAEGAN-D2 [13]	<b>84.00</b>	85.00
RS-FSL [38]	65.66	–
L3 [29]	53.96	–
LSL [30]	61.24	–
Multiple-semantics [25]	76.10	82.90
<i>Without attribute embedding</i>		
Proposed method (ResNet-18)	64.68	<b>85.36</b>
Proposed method (ResNet-101)	74.67	<b>87.22</b>

Bold values represent best performing scores in the individual categories

proposed model that uses raw textual attributes. The proposed framework performs better as the number of shot increases. The lower performance for lesser number of shots is believed to be due to the size of the latent variable and the amount of data, in this case number of shots, that is available to train the latent variable. Further experiments that follow help substantiate this claim.

Table 3 compares with other multimodal models that report 5-way classification on CUB for 1- and 2-shots. For 5-way-1-shot, the best performing model uses a combination of VAE and GAN. For 5-way 1-shot, the proposed model performs average across all methods. For 5-way 5-shot, the model performs better than the rest. The lower performance for 1-shot is believed to be due to the increased size of the latent vector as a consequence of using the raw attributes since for 1-shot, a low dimensional input could prove beneficial. Specifically, for CUB there are 312 attributes and together with the visual representation  $z_v$  of 64 dimensions, the total input dimension is 376. As noted earlier,

**Table 4** 72-way classification accuracy on SUN of proposed model

Backbone	Metric	1-shot	2-shot	5-shot	10-shot
ResNet-18	Top-1	10.70	23.52	49.91	62.95
	Top-3	17.61	39.69	70.58	83.29
	Top-5	20.21	47.63	80.30	90.03
ResNet-101	Top-1	36.40	46.14	59.26	66.67
	Top-3	55.41	67.28	80.56	87.08
	Top-5	65.20	76.23	87.31	94.31

**Table 5** 10-way classification accuracy on AWA2 of proposed model

Backbone	Metric	1-shot	2-shot	5-shot	10-shot	20-shot
ResNet-18	Top-1	59.67	74.79	79.12	87.19	91.47
	Top-3	84.73	90.88	94.64	96.78	98.41
	Top-5	93.13	96.67	98.08	98.91	99.60
ResNet-101	Top-1	65.37	78.13	83.56	87.12	89.50
	Top-3	87.98	92.12	95.32	97.22	98.61
	Top-5	95.01	96.73	97.61	98.77	99.64

the benefit of using the attributes directly is that it allows the model to be more interpretable and it provides a means for learning the compositionality of an image.

Tables 4 and 5 show the results for the proposed model on the SUN and AWA2 datasets. There are no comparisons with other models because these datasets are used for zero-shot or generalized few-shot learning, which are not considered here. When using all the test categories for classification, there is a continuous increase in accuracy as the number of shots is increased for SUN. The same is true for AWA2 although with 1-shot itself the accuracy of Top-5 reaches 93%. These results provide a sanity check that ensures that the framework works for other datasets as well. By using raw attributes in the proposed framework, the model is able to perform multimodal few-shot learning on these datasets. As the number of shots increases, the performance of the model increases as well.

In addition to results with a ResNet-18 backbone, results using a ResNet-101 backbone are also presented in all metrics. Results with the ResNet-18 backbone enable us to show a direct comparison with [8] and [26], whose works are closest, as the same backbone is used. ResNet-101 shows the effects on the model when using a directly comparable but stronger feature extraction. Results with the ResNet-101 backbone shows that when a better feature extraction is used with the model, results improve significantly.






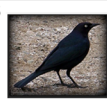
#### 5.4 Compositionality

Compositionality refers to the idea of representing a whole through a representation of its parts. Human knowledge representation is largely compositional and is applicable for spatial as well as temporal phenomena, e.g., a scene as composed of objects, an object as composed of parts or an activity as composed of events. Here, the attributes are viewed as representing the composition of an image; in fact it is more than that since the attributes not only describe the parts that an object is made up of but also describe

**Table 6** 100-way classification of Top-5 accuracy on CUB

Method	1-shot	2-shot	5-shot	10-shot	20-shot
PN [39] + [7]	42.20	55.70	68.70	–	–
MN [40] + [7]	51.60	59.90	72.00	–	–
Compositional [7]	<b>53.60</b>	<b>64.80</b>	<b>74.60</b>	78.70	–
Proposed method (ResNet-18)	43.49	58.03	72.13	<b>79.56</b>	<b>83.74</b>
Proposed method (ResNet-101)	52.57	<b>65.81</b>	<b>77.84</b>	<b>82.35</b>	<b>86.31</b>

Bold values represent best performing scores in the individual categories

ResNet-18				ResNet-101			
Image	Attribute	Probability Score	Ground Truth	Image	Attribute	Probability Score	Ground Truth
1 	has_upper_tail_color:black has_back_pattern:solid has_wing_color:green has_back_color:blue has_nape_color:white	0.6782 0.6526 0.6425 0.6409 0.6376	1 1 1 1 0	4 	has_eye_color:black has_belly_color:white has_underparts_color:white has_bill_length:shorter_than_head has_wing_color:brown	0.9975 0.9908 0.9890 0.9720 0.9680	1 1 1 1 1
2 	has_belly_pattern:striped has_back_color:brown has_underparts_color:rufous has_wing_shape:rounded-wings has_upper_tail_color:black	0.6649 0.6524 0.6426 0.6266 0.6266	1 1 0 1 1	5 	has_bill_length:shorter_than_head has_forehead_color:black has_eye_color:white has_throat_color:black has_crown_color:black	0.9954 0.9859 0.9674 0.9674 0.9431	1 1 1 1 1
3 	has_forehead_color:yellow has_upper_tail_color:black has_nape_color:white has_back_pattern:solid has_breast_color:grey	0.7166 0.7122 0.6739 0.6710 0.6567	1 1 0 1 1	6 	has_upperparts_color:black has_primary_color:black has_wing_color:black has_belly_color:black has_underparts_color:black	0.9999 0.9990 0.9989 0.9988 0.9975	1 1 1 1 1

**Fig. 2** Predicted probability of attributes versus ground truth

their characteristics. Tokmakov et al. [7] describe a model to learn compositional representations for few-shot learning by disentangling the feature space of a CNN into sub-spaces corresponding to category-level attributes. The performance of the proposed model without attribute embedding is compared to [7] in Table 6 for 100-way classification on CUB. The authors also applied their compositional algorithm on two few-shot recognition methods (described in their supplementary material), viz., Prototypical networks (PN) and matching network (MN). As seen in the table, the proposed method is comparable for 1-, 2- and 5-shot, but starts to performs better for 10-shot. The gap between performance of the proposed method and the other models is observed to decrease as the number of shots increases. This is likely due to the increased number of shots improving how well the model learns compositionality. The results of the proposed method with a ResNet-101 backbone is also presented. This improves the model results by 3 to 9%, similar to the behavior as observed in Sect. 5.3. With a stronger feature extraction, it results in a better representation of attributes for the model.

### 5.5 Interpretability

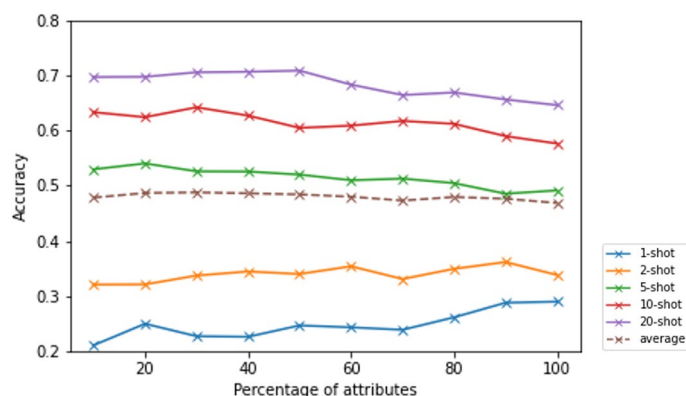
The proposed model is inherently interpretable because the probability of the attributes that contributed to a particular classification is readily available from the latent semantic representation  $z_s$ . In order to evaluate interpretability, the estimated probability of attributes is compared to the ground truth labels represented as the binary attribute vector. In Fig. 2, examples of images from the CUB dataset for which the Top-5 attributes are identified by the model is presented along with the ground truth

in a 50-way setting. On the left (images 1, 2 and 3) shows results obtained from a ResNet-18 backbone and on the right (images 4, 5 and 6), results from a ResNet-101 backbone. In both setups, the model is able to predict the presence of attributes with high accuracy. For attributes that the prediction is wrong, potential reasons can be seen from the images. For example, from the Top-5 predicted attributes in image 1, the model detects the attribute “has\_nape\_color::white” with a confidence of 0.6376, however the ground truth indicates that the prediction is wrong. However, it can be observed from the image that the bird is surrounded by an object that is white near the nape area. This is likely the reason the model has a higher confidence for the presence of that attribute. In addition to showing results from a ResNet-18 backbone, the results for samples predicted with a ResNet-101 backbone shows significant improvement in predictions of these attributes when using a model with stronger representation power. The confidence score of each attribute rises to close to 1. From this, it can be inferred that the stronger the feature extraction is, the higher the confidence, and the more interpretable the results will be. This further shows the interpretability of the proposed model.

For a quantitative evaluation of interpretability, the L1 distance between the ground truth and the estimated probability score of the attribute is computed. Table 7 shows the L1 distance averaged across all attributes over the entire training and test dataset for a ResNet-18 and ResNet-101 backbone. The numbers in the table can be directly interpreted as the number of errors per attribute. On both training and test data, the average distance over all shots is around 0.5 for a ResNet-18 backbone, which is about a correct prediction for every other attribute. For the distance calculated from the results of a ResNet-101 backbone, the distance decreases to as much as about 0.000250 for training data and 0.2 for test data. This amounts to about 1 prediction error in every 4000 attributes for training data and 1 in 5 for test data. Both results suggest that the model has in some way learned the attributes from the data, and can detect the presence of attributes. Similar to the attribute prediction shown above, the use of a stronger feature extraction results in the model becoming more interpretable.

**Table 7** L1 distance of predicted attribute score to ground truth labels for 50-way CUB

Backbone	Number of shots	Training data	Test data
ResNet-18	1-shot	0.497340	0.497224
	2-shot	0.498417	0.498723
	5-shot	0.498584	0.498149
	10-shot	0.499778	0.499468
	20-shot	0.500094	0.499941
ResNet-101	1-shot	0.000291	0.182606
	2-shot	0.000276	0.187820
	5-shot	0.000275	0.192094
	10-shot	0.000298	0.190431
	20-shot	0.000293	0.194535



**Fig. 3** Top-1 accuracy for different shots as number of attributes increases for CUB data set

### 5.6 Effect of number of attributes

In this section, the effects of the number of attributes on the results is analyzed; in other words, should all the available attributes be used thereby increasing the size of the concatenated latent representation  $z_c$ ?

The number of attributes are increased by picking the first  $n\%$  of the attributes as indicated by the dataset. For example, for a dataset with 100 attributes, if 10% is chosen, only the first 10 attributes of the dataset will be used. Figure 3 shows the accuracy for 1, 2, 5, 10 and 20 shots as the percentage of attributes is increased from 10 to 100%. For 1- and 2-shot, it is observed that there is an improvement in accuracy of about 5 to 10% as the number of attributes increases. For 5-, 10- and 20-shot, the accuracy decreases about 2 to 5%. This phenomenon is believed to be caused by the size of the image features. When using a ResNet-18 backbone, the extracted image features has a size of 512. When there are low number of shots, each additional sample helps improve the mapping of image features to the  $n\% + 64$  sized  $z_c$  and back to the reconstructed features. As there are limited samples, each sample improves the mapping. However, as the number of shots increases, the model has to learn from more samples but still in small amounts that makes it difficult to learn a proper mapping between the different latent spaces. In Sect. 5.3, the lower performance is attributed to the size of the latent vector. The results for higher number of shots here mirrors this, as the percentage of attributes decreases, the learned mapping becomes easier for the model as the size of the latent space decreases. The same cannot be said for lower number of shots as due to a smaller number of samples, any form of additional information provided to the model improves the results. Reducing the size of the latent space is still believed to result in better performance, however not in the case when binary value of attributes are used.

## 6 Conclusion

In this work, a multimodal few-shot learning method that uses image attributes directly, without the need for an embedding space, is proposed. Embedded attributes prevent the model from being interpretable. Raw attributes also help determine the composition of an image. An inherently interpretable model is proposed using a hybrid autoencoder that has both a normal autoencoder and a variational autoencoder with a semantic

loss and cyclic consistency loss. This method outperforms existing methods in higher number of ways and shots on the CUB dataset with comparable results in fewer number of ways. The interpretability of the model is also evaluated by comparing the predicted attribute scores with the ground truth attribute labels, as well as show how with stronger feature extraction, the model becomes even more interpretable.

#### Acknowledgements

Not applicable.

#### Author contributions

JQC made contributions to the conception and design of this work, as well as the analysis and interpretation of the data. DR made contributions to the conception and design of this work, as well as the analysis and interpretation of the data. VN made contributions to the conception the project of which this research is a part of.

#### Funding

The authors declare that there is no funding body that aided in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

#### Availability of data and materials

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

#### Declarations

##### Competing interests

The authors declare that they have no competing interests.

Received: 27 February 2023 Accepted: 31 December 2023

Published online: 10 January 2024

#### References

1. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135 (2017)
2. Y. Tian, Y. Wang, D. Krishnan, J.B. Tenenbaum, P. Isola, Rethinking few-shot image classification: a good embedding is all you need? In: Proceedings of European Conference on Computer Vision, pp. 266–282 (2020)
3. A. Antoniou, H. Edwards, A. Storkey, How to train your MAML. In: International Conference on Learning Representations (2019)
4. C. Finn, K. Xu, S. Levine, Probabilistic model-agnostic meta-learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS'18, pp. 9537–9548 (2018)
5. F. Pahde, M.M. Puscas, J. Wolff, T. Klein, N. Sebe, M. Nabi, Low-shot learning from imaginary 3d model. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 978–985 (2019)
6. E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evcı, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, H. Larochelle, Meta-dataset: a dataset of datasets for learning to learn from few examples. In: International Conference on Learning Representations (2020)
7. P. Tokmakov, Y.-X. Wang, M. Hebert, Learning compositional representations for few-shot recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
8. F. Pahde, M. Puscas, T. Klein, M. Nabi, Multimodal prototypical networks for few-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2644–2653 (2021)
9. Y.-X. Wang, R. Girshick, M. Hebert, B. Hariharan, Low-shot learning from imaginary data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
10. G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed by Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. Layer-wise relevance propagation: an overview. (2019) pp. 193–209
11. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**(2), 336–359 (2019)
12. W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R.J. Radke, O. Camps, Towards visually explaining variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
13. Y. Xian, S. Sharma, B. Schiele, Z. Akata, F-vaegan-d2: a feature generating framework for any-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
14. I. Biederman, R.J. Mezzanotte, J.C. Rabinowitz, Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* **14**(2), 143–177 (1982)
15. Y. Xian, C. Lampert, B. Schiele, Z. Akata, Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (2017)
16. F. Qi, X. Yang, C. Xu, Zero-shot video emotion recognition via multimodal protagonist-aware transformer network. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1074–1083 (2021)

17. E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata, Generalized zero- and few-shot learning via aligned variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
18. Y. Wang, Q. Yao, J.T. Kwok, L.M. Ni, Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv.* **1**(1), 1–1134 (2002)
19. S. Benaim, L. Wolf, One-shot unsupervised cross domain translation. *Adv. Neural Inf. Process. Syst.* **31** (2018)
20. H. Gao, Z. Shou, A. Zareian, H. Zhang, S.-F. Chang, Low-shot learning via covariance-preserving adversarial augmentation networks. *Adv. Neural Inf. Process. Syst.* **31** (2018)
21. A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, J. Weston, Key-value memory networks for directly reading documents. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1400–1409 (2016)
22. M. Andrychowicz, M. Denil, S. Gómez, M.W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, N. de Freitas, Learning to learn by gradient descent by gradient descent. In Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS'16)
23. Y. Song, T. Wang, P. Cai, S.K. Mondal, J.P. Sahoo, A comprehensive survey of few-shot learning: evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* (2023). Just Accepted
24. M.P. Fortin, B. Chaib-draa, Towards contextual learning in few-shot object classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3279–3288 (2021)
25. E. Schwartz, L. Karlinsky, R. Feris, R. Giryes, A. Bronstein, Baby steps towards few-shot learning with multiple semantics. *Pattern Recogn. Lett.* **160**, 142–147 (2022)
26. F. Pahde, O. Ostapenko, P. Hnichen, T. Klein, M. Nabi, Self-paced adversarial training for multimodal few-shot learning. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 218–226 (2019)
27. C. Xing, N. Rostamzadeh, B. Oreshkin, P.O. Pinheiro, Adaptive cross-modal few-shot learning. *Adv. Neural Inf. Process. Syst.* **32** (2019)
28. Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, L. Sigal, Multi-level semantic feature augmentation for one-shot learning. *IEEE Trans. Image Process.* **28**(9), 4594–4605 (2019)
29. J. Andreas, D. Klein, S. Levine, Learning with latent language. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2166–2179 (2018)
30. J. Mu, P. Liang, N. Goodman, Shaping visual representations with language for few-shot classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4823–4830 (2020)
31. Z. Chen, Y. Luo, S. Wang, R. Qiu, J. Li, Z. Huang, Mitigating generation shifts for generalized zero-shot learning. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 844–852 (2021)
32. D. Samuel, Y. Atzmon, G. Chechik, From generalized zero-shot learning to long-tail with class descriptors. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 286–295 (2021)
33. D.P. Kingma, M. Welling, Auto-encoding variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014 (2014)
34. J. Fajtl, V. Argyriou, D. Monekosso, P. Remagnino, Latent bernoulli autoencoder. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 2964–2974 (2020)
35. Y. Zhang, S. Huang, X. Peng, D. Yang, Dizygotic conditional variational autoencoder for multi-modal and partial modality absent few-shot learning (2021). [arXiv:2106.14467](https://arxiv.org/abs/2106.14467)
36. C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology (2011)
37. G. Patterson, C. Xu, H. Su, J. Hays, The sun attribute database: beyond categories for deeper scene understanding. *Int. J. Comput. Vis.* **108**(1–2), 59–81 (2014)
38. M. Afham, S. Khan, M.H. Khan, M. Naseer, F.S. Khan, Rich semantics improve few-shot learning. 32nd British Machine Vision Conference (2021)
39. O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, D. Wierstra, Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* **29** (2016)
40. J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **30** (2017)

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.