Augmentation Network for Generalised Zero-Shot Learning

Rafael Felix[1,2,3][0000−0002−6186−9426], Michele Sasdelli[1,2,3][0000−0003−1021−6369],
Ian Reid[1,2,3][0000−0001−7790−6423], and Gustavo Carneiro[1,2,3][0000−0002−5571−6220]

[1] The University of Adelaide, Australia
[2] Australian Institute for Machine Learning (AIML),
[3] Australian Centre for Robotic Vision (ACRV),
{rafael.felixalves,michele.sasdelli,ian.reid,gustavo.carneiro}@adelaide.edu.au

Abstract. Generalised zero-shot learning (GZSL) is defined by a training process containing a set of visual samples from seen classes and a set of semantic samples from seen and unseen classes, while the testing process consists of the classification of visual samples from the seen and the unseen classes. Current approaches are based on inference processes that rely on the result of a single modality classifier (visual, semantic, or latent joint space) that balances the classification between the seen and unseen classes using gating mechanisms. There are a couple of problems with such approaches: 1) multi-modal classifiers are known to generally be more accurate than single modality classifiers, and 2) gating mechanisms rely on a complex one-class training of an external domain classifier that modulates the seen and unseen classifiers. In this paper, we mitigate these issues by proposing a novel GZSL method – augmentation network that tackles multi-modal and multi-domain inference for generalised zero-shot learning (AN-GZSL). The multi-modal inference combines visual and semantic classification and automatically balances the seen and unseen classification using temperature calibration, without requiring any gating mechanisms or external domain classifiers. Experiments show that our method produces the new state-of-the-art GZSL results for fine-grained benchmark data sets CUB and FLO and for the large-scale data set ImageNet. We also obtain competitive results for coarse-grained data sets SUN and AWA. We show an ablation study that justifies each stage of the proposed AN-GZSL.

Keywords: generalised zero-shot learning; multi-modal inference; multi-domain inference

1 Introduction

As computer vision systems start to be deployed in unstructured environments, they must have the ability to recognise not only the visual classes used during the training process (i.e., the seen classes) but also classes that are not available during training (i.e., unseen classes). The importance of such ability lies in the impracticality of collecting visual samples from all possible classes that will be shown to the system. In this context, approaches categorised as Generalised Zero-Shot Learning (GZSL) [1–3] play an important role due to their ability to classify visual samples from seen and unseen classes. In general, the training of GZSL methods involves the use of visual samples from seen classes and semantic samples (e.g., textual definition) from seen and unseen classes. The rationale behind the use of semantic samples is that they are readily available from various sources, such as Wikipedia, English dictionary [4], or manually annotated attributes [5]. Such training setup can potentially mitigate the issue of collecting visual samples from all possible unseen classes, and the success of GZSL lies in the effective transferring of knowledge between the semantic and visual modalities.

In recent years, we note three different approaches for solving GZSL. One type focuses on training a mapping function from the visual to the semantic space [6], and then inference relies on classification in the semantic space. Another type is based on training a conditional generative model for visual samples. The generated visual samples of unseen classes complement visual samples from the seen classes for a visual classifier [7, 2, 8–14]. Another type relies on an external domain classifier (seen vs unseen) trained with the visual samples from the seen classes via a one-class learning problem. The domain classification is combined with the classification models in each domain [15–19].

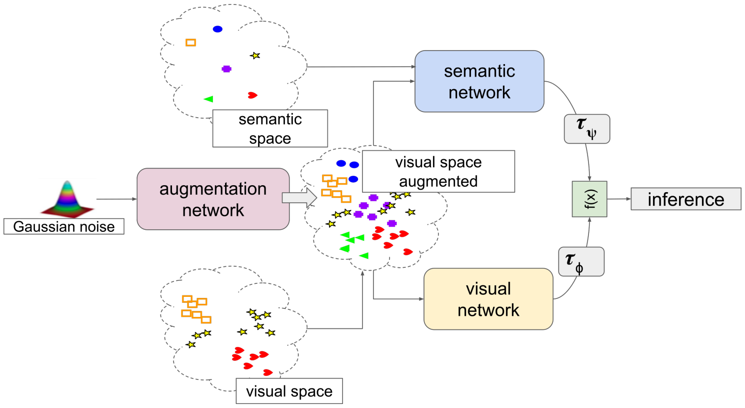


Fig. 1. Our proposed model Augmentation Network for multi-modal and multi-domain Generalised Zero-Shot Learning (AN-GZSL). AN-GZSL is composed of the augmentation network to generates visual samples, the visual and the semantic networks, a classification calibration (represented by τ_ψ and τ_ϕ in (2)) that enables multi-domain classification, and the multi-modal classification that combines the visual and semantic modules.

The GZSL methods above have a couple of issues: 1) even though the training process involves some sort of interaction between the visual and semantic modalities, the inference usually does not rely on a truly multi-modal classification (i.e., where both modalities are jointly used in the process) [15, 7, 2, 13] , which can be considered a weakness given the strong evidence that multi-modal inference can improve classification accuracy [20, 15, 19]; and 2) the one-class training of external domain classifiers that modulate the seen and unseen classification [15–19] is not a trivial process given the similarity between the seen and unseen class domains. In fact, it can be argued that samples from these domains are drawn from the same distribution, making it challenging to distinguish between them.

In this paper, we introduce the Augmentation Network for multi-modal and multi-domain Generalised Zero-Shot Learning (AN-GZSL) depicted in Fig. 1. The approach introduces a novel loss function that combines the training of three networks: augmentation network, visual network, and semantic network. **The proposed AN-GZSL represents the first method to perform inference with multiple modalities without the use of external domain classifiers for modulating the inference between seen and unseen classes.** The augmentation network is a generative model for visual samples conditioned on the semantic data, the generated visual samples are then used by the visual network (a visual classifier) and by the semantic network (a semantic classifier). The visual and semantic classifiers are then temperature calibrated to enable a modulation-free classification, which alleviates the burden of an external domain classifier. Then, the two calibrated classifiers are combined in a multi-modal classification. We show that the proposed approach produces state-of-the-art GZSL results on the fine-grained benchmark data sets CUB [21, 3] and FLO [22] and on the large-scale data set ImageNet [23, 24]. We also achieve competitive results for the coarse-grained data sets SUN [3] and AWA [5]. We finally show an ablation study that tests the importance of each component of the proposed model.

2 Literature Review

In this section we describe relevant literature that contextualises and motivates the proposed approach.

Generalised Zero-Shot Learning (GZSL). In recent years, we have observed a growing interest in GZSL. A catalyst for such interest was the paper by Xian et al. [3] that formalises the GZSL problem. Their work introduces a solid experimental setup and a robust evaluation metric based on the harmonic mean between the classification accuracy results of the seen and the unseen visual classes. Recently proposed GZSL methods can be roughly divided into three categories: **semantic attribute prediction**, **visual data augmentation**, and **domain balancing**. **Semantic attribute prediction** methods [25, 26, 5] tackle GZSL by training a regressor that maps visual samples from seen classes to their respective semantic samples. Hence, given a test visual sample (from a seen or unseen class), the regressor maps it into the semantic space, which is then used in a nearest neighbour semantic classification process. The main assumption of this approach is that the mapping from visual to semantic spaces learned from the seen class domain can be transferred to the unseen class domain. Unfortunately, such assumption is unwarranted, and a typical issue of this approach is that test visual samples from seen classes are classified correctly and samples from unseen classes are often incorrectly classified into one of the seen classes – this is referred to as a bias toward the seen classes [3]. Recent research exploring matching functions between visual and semantic samples can address the issue mentioned above, but they still show biased classification toward the seen classes [25].

Visual data augmentation relies on a generative model trained to produce visual samples from the corresponding semantic samples [7, 2, 8–10]. Such model allows the generation of visual samples for the unseen classes, which are then used in the modelling of a visual classifier that is trained with real visual samples from seen classes and generated visual samples from unseen classes. Methods based on this approach are effective because they solve, to a certain extent, the bias toward the seen classes [11–14]. Recently, the training process of this approach has been extended, forcing generated visual samples to regress to the corresponding semantic samples, in a multi-modal cycle consistent training [2, 14]. This extension represents the first attempt at a multi-modal training, which allowed further improvements in GZSL results. However, none of the methods above relies on a multi-modal inference process. It is interesting to note that the inference process of **semantic attribute prediction** focuses exclusively on the semantic space, while **visual data augmentation** works solely on the visual space. A multi-modal inference process that effectively merges the two spaces has yet to be proposed.

Domain balancing methods solve the bias toward the seen classes issue with a gating mechanism that modulates the classification of seen and unseen classes [15–19]. In particular, these methods consist of a (generally visual) classifier trained for the seen classes, a (usually semantic) classifier trained for the unseen classes, and a domain classifier for the modulation process [15, 19, 17].

Even though domain balancing approaches hold outstanding results [15, 19], they have the following challenges: 1) the training of multiple domain-specific classifiers, and 2) the non-trivial training of a gating mechanism that needs to classify between seen and unseen classes using a one-class classification process, which is a hard task considering that these classes arguably come from the same data distribution. In this paper, we also rely on visual data augmentation and domain balancing, but differently from the approaches above, our multi-modal classification relies on visual and semantic classifiers trained on all seen and unseen classes (i.e., they are not domain-specific). Furthermore, the balancing between seen and unseen domains is achieved with a classification calibration approach that does not need any gating mechanism.

3 Method

In the next sub-sections, we first formulate the GZSL problem. Then, we introduce our proposed augmentation network for multi-modal and multi-domain generalised zero-shot learning (AN-GZSL), with the explanation of the inference, architecture and training processes.

3.1 Problem Formulation

To formulate the GZSL problem [1, 3], we first define the visual data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$ denotes a visual sample (acquired from the second to last layer of a pre-trained deep residual nets [27]), and $y \in \mathcal{Y} = \{1, ..., C\}$ denotes the visual class, which can also be described with a one-hot vector $\mathbf{h} \in \{0, 1\}^C$, where the y-th position in $\mathbf{h}$ is assigned to 1, and all the others 0. The visual data set has N samples, denoting the number of images. We also need to define the semantic data set $\mathcal{R} = \{\mathbf{a}_y\}_{y \in \mathcal{Y}}$, which associates visual classes with semantic samples, where $\mathbf{a}_y \in \mathcal{A} \subseteq \mathbb{R}^L$ represents a semantic feature (e.g., *word2vec* features [3]). The semantic data set has as many elements as the number of classes. The set $\mathcal{Y}$ is split into the seen subset $\mathcal{Y}^S = \{1, ..., S\}$, and the unseen subset $\mathcal{Y}^U = \{(S+1), ..., (S+U)\}$. Therefore, $C = S + U$, with $\mathcal{Y} = \mathcal{Y}^S \cup \mathcal{Y}^U$, $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. Furthermore, $\mathcal{D}$ is also divided into mutually exclusive training and testing visual subsets $\mathcal{D}^{Tr}$ and $\mathcal{D}^{Te}$, respectively, where $\mathcal{D}^{Tr}$ contains a subset of the visual samples belonging to the seen classes, and $\mathcal{D}^{Te}$ has the visual samples from the seen classes held out from training and all samples from the unseen classes. The training data set comprises the semantic data set $\mathcal{R}$ and the training visual subset $\mathcal{D}^{Tr}$, while the testing data set consists of the testing visual subset $\mathcal{D}^{Te}$ and the same semantic data set $\mathcal{R}$.

3.2 AN-GZSL Calibrated Inference

The inference procedure consists of estimating the class label of a test visual sample $\mathbf{x}$ that optimises

$$f(y|\mathbf{x}, \mathcal{R}) = \sigma(\phi(y|\mathbf{x}), \tau_\phi) + \sigma(\psi(y|\mathbf{x}, \mathcal{R}), \tau_\psi), \tag{1}$$

where $f(.)$ denotes the classification function, $\phi(.)$ and $\psi(.)$ represent the visual network (defined in Sec. 3.5) and the semantic network (Sec. 3.4) that return a logit, and $\sigma(.)$ represents the softmax activation function with temperature calibration [28], defined by

$$\sigma(l_y, \tau) = \frac{e^{(l_y/\tau)}}{\sum_{c=1}^{C} \mathrm{e}^{(l_c/\tau)}}, \tag{2}$$

where the logit $l_y \in \mathbb{R}$ represents the y^{th} output of a network (i.e., the visual or the semantic), and the temperature scaling τ represents a calibrating factor.

The multi-modal inference in (1) consists of a sum of the results from the visual and semantic classifiers, where the final classification is achieved by

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} f(y|\mathbf{x}, \mathcal{R}). \tag{3}$$

The GZSL inference in (3) balances the seen and unseen classes with a confidence calibrated by the temperature scaling, which is a much simpler strategy than to previously used gating mechanisms [15, 18, 19] that have to deal with complicated one-class domain classification problems. Furthermore, (3) creates a simple multi-modal inference without any hyper-parameter tuning to combine the classifiers.

3.3 Augmentation Network

The augmentation network relies on a generative model [13] trained to produce visual samples conditioned on their semantic samples. This model allows to generate visual samples for the unseen classes that together with real visual samples from the seen classes are used to train a classifier [7, 2, 8–14]. This approach has been recently extended with a cycle consistency loss that regularises the training process [2]. The augmentation network is optimised with a Wasserstein generative adversarial network (WGAN) [29] loss and cycle-consistent loss [2], defined by

$$\ell_{AN} = \ell_{WGAN} + \ell_{CYC}, \tag{4}$$

where ℓ_{WGAN} represents the WGAN loss [29] that optimises a conditional generator network $g(.)$ and discriminator network $d(.)$. The loss ℓ_{WGAN} is defined by

$$\begin{aligned} \ell_{WGAN} = & \mathbb{E}_{(\mathbf{x},\mathbf{a})\sim\mathbb{P}_s^{x,a}}[d(\mathbf{x},\mathbf{a};\theta_d)] - \mathbb{E}_{(\tilde{\mathbf{x}},\mathbf{a})\sim\mathbb{P}_g^{x,a}}[d(\tilde{\mathbf{x}},\mathbf{a};\theta_d)] \\ & - \kappa\mathbb{E}_{(\hat{\mathbf{x}},\mathbf{a})\sim\mathbb{P}_\alpha^{x,a}}[(||\nabla_{\hat{\mathbf{x}}} d(\hat{\mathbf{x}},\mathbf{a};\theta_d)||_2 - 1)^2], \end{aligned} \tag{5}$$

where $\mathbb{E}[.]$ represents the expected value operator. The joint distribution of visual and semantic samples from the seen classes is given by $\mathbb{P}_s^{x,a}$, and $\mathbb{P}_g^{x,a}$ represents the joint distribution of semantic and visual samples produced by the augmented network using the generator network, as follows: $\tilde{\mathbf{x}} \sim g(\mathbf{a}, \mathbf{z}; \theta_g)$, where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. The coefficient κ in (5) weights the contribution of the third term of the loss, and the joint distribution of the semantic and visual samples produced by $\hat{\mathbf{x}} \sim \alpha\mathbf{x} + (1 - \alpha)\tilde{\mathbf{x}}$ with $\alpha \sim \mathcal{U}(0, 1)$ (i.e., uniform distribution) is given by $\mathbb{P}_\alpha^{x,a}$. In this network, the generator receives a semantic sample $\mathbf{a}$ and a noise vector $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ to generate visual samples, $\nabla_{\hat{\mathbf{x}}}$ represents the gradient penalty [29]. Then, the discriminator network aims to differentiate the generated from the real visual samples [2]. The loss ℓ_{CYC} provides a cycle-consistent training regularisation which guarantees that generated visual samples can reconstruct the corresponding semantic samples. The loss ℓ_{CYC} is defined by

$$\begin{aligned} \ell_{CYC} = & \mathbb{E}_{\mathbf{a}\sim\mathbb{P}_s^a,\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\mathbf{a} - r(g(\mathbf{a},\mathbf{z};\theta_g);\theta_r)\|_2^2\right] \\ & + \mathbb{E}_{\mathbf{a}\sim\mathbb{P}_u^a,\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\mathbf{a} - r(g(\mathbf{a},\mathbf{z};\theta_g);\theta_r)\|_2^2\right], \end{aligned} \tag{6}$$

where the function $r(.)$ represents a regressor network parameterised by θ_r that estimates the original semantic samples from the visual samples generated by $g(.)$, the latent variable $\mathbf{z}$ represents Gaussian noise, and the distributions of the semantic samples of seen and unseen domains are represented by $\mathbb{P}_s^a$ and $\mathbb{P}_u^a$. In contrast to previous approaches [7, 2, 8–14], our proposed augmentation network feeds the visual *and* semantic networks with generated visual samples from both domains – this allows the visual and semantic classifiers to jointly learn an effective discriminating space for *all seen and unseen classes.*

3.4 Semantic Network

The semantic network consists of a bilinear neural network, novel in this application, which extends the ranking loss proposed by Akata et al. [25]. We define the semantic network as $\psi(y|\mathbf{x}, \mathcal{R}) = \mathbf{x}^T \theta_\psi \mathbf{a}_y$, represented by a bi-linear model parameterised by $\theta_\psi \in \mathbb{R}^{K \times L}$, with $\mathbf{a}_y \in \mathcal{R}$ and $\mathbf{x}$ being either a real sample from a seen class or a generated sample from an unseen class. *This approach is the first to use data augmentation to train a semantic classifier, when compared to the one in [25].* For training we optimise the semantic loss loss, defined by

$$\ell_{SN} = \sum_{i=1}^{M} \sum_{c=1}^{C} \lambda(\mathbf{x}_i, \theta_\psi, \mathbf{a}_{y_i}, \mathbf{h}_i) \Big[h_{i,c} + \mathbf{x}_i^T \theta_\psi \mathbf{a}_{y_i} - \mathbf{x}_i^T \theta_\psi \mathbf{a}_c \Big]_+, \tag{7}$$

where $[.]_+$ represents the hinge loss, $\mathbf{a}_{y_i}$ denotes the semantic vector associated with the class y_i of the i^{th} training sample, $h_{i,c}$ represents the c^{th} position of the one-hot vector for the i^{th} training sample $\mathbf{h}_i$, and M is the size of the training set (including the generated visual features), and $\mathbf{a}_c$ is a semantic vector associated with a class. In (7), the term inside of the hinge loss consists of a compatibility bi-linear loss,

$$\beta(\mathbf{x}_i, \theta_\psi, \mathbf{a}_c, \mathbf{a}_{y_i}, h_{i,c}) = h_{i,c} + \mathbf{x}_i^T \theta_\psi \mathbf{a}_{y_i} - \mathbf{x}_i^T \theta_\psi \mathbf{a}_c, \tag{8}$$

and $\lambda(.)$ represents a ranking regularization, defined by

$$\lambda(\mathbf{x}_i, \theta_\psi, \mathbf{a}_{y_i}, \mathbf{h}_i) = \Big(\textstyle\sum_{c=1}^{C} \mathbb{1}\Big(\beta(\mathbf{x}_i, \theta_\psi, \mathbf{a}_c, \mathbf{a}_{y_i}, h_{i,c})\Big)\Big)^{-1}, \tag{9}$$

where $\mathbb{1}(.)$ represents a Heaviside step function, with the divisor computing the ranking of the transformation according to the semantic data set.

The optimisation of (7) forces $\psi(y|\mathbf{x}, \mathcal{R})$ to be higher when $\mathbf{x}$ and $\mathbf{a}_y$ match correctly. This result is then calibrated by (2) to enable an effective multi-domain classification.

3.5 Visual Network

The visual network is a fully connected neural network represented by $\phi(y|\mathbf{x})$, parameterised by θ_ϕ, where $\mathbf{x}$ can be a real sample from a seen class or a generated sample from an unseen class. The network is trained with the usual cross-entropy loss defined by ℓ_{VN}. Similarly to the semantic classifier, this visual classifier is also calibrated with (2) for the multi-domain classification.

3.6 AN-GZSL Training

The loss function for our proposed AN-GZSL model is defined by

$$\ell_{AN-GZSL} = \ell_{AN} + \ell_{VN} + \ell_{SN}, \tag{10}$$

which is minimised to estimate the parameters $\theta_g, \theta_d, \theta_r, \theta_\phi, \theta_\psi$. For training, we use the visual samples produced by the augmentation network as input to the proposed visual and semantic networks. This approach not only augments the number of samples from the seen classes, but it also generates samples from the unseen classes. In practice, we perform an alternating training where we first optimise θ_g, θ_d and θ_r, then we optimise θ_ψ and θ_ϕ. Empirically, we have observed that the augmentation network tends to generate random samples at early stages of training [30]. Hence, the alternating strategy provides stronger gradients signal for the optimisation of θ_ψ and θ_ϕ, at late stages. After all the optimisation of (10) are completed, the temperatures (τ_ϕ and τ_ψ in Eq. 1) are quickly estimated by grid-search using the logits of a validation set held out from training [28].

4 Experiments

In this section, we describe the benchmark data sets, evaluation criteria and the setup adopted for the experiments. Then, we present a set of ablation studies and the results of the proposed method, which are compared with the state of the art (SOTA).

4.1 Data Sets

We assess the proposed method on publicly available benchmark GZSL data sets. More specifically, we perform experiments on CUB-200-2011 [21, 3], FLO [22], SUN [3], and AWA [5, 3] with the GZSL experimental setup described by Xian et al. [3]. We also perform GZSL experiments on ImageNet [23, 24]. The data sets CUB and FLO are generally regarded as fine-grained, while AWA and SUN are coarse-grained, and ImageNet is large-scale[4].

For the semantic features, we use the 1024-dimensional vector produced by CNN-RNN [31] for CUB-200-2011 [3] and FLO [22]. These semantic features are extracted from a set of textual description of 10 sentences per image. To define a unique semantic sample per-class, the semantic features of all images belonging to each class are averaged [3]. For the SUN and AWA data sets, we use manually annotated semantic features (attributes) containing 102 and 85 dimensions, respectively [3]. For the visual samples, we follow the protocol by Xian et al. [3], where the features are represented by the activation of the 2048-dimensional top pooling layer of ResNet-101 [27], obtained for the image. To guarantee reproducible and consistent results, we follow the data set split proposed by Xian et al. [3], which prevents the model to violate the zero-shot conditions.

[4] See supplementary material for more information on data sets.

For the ImageNet experiment [23], there can be several testing splits for GZSL (e.g., 2-hop, 3-hop), which rely on the training set of 1K classes and testing set on 22K classes. However, recent studies reported that such splits show overlap between seen and unseen classes for GZSL [2]. To demonstrate the robustness of the proposed approach to large data sets, we experiment with ImageNet [23] for a split containing 100 classes for testing [24] and the standard 1K classes for training [24], without any overlap between seen and unseen classes. For ImageNet, we used 500-dimensional semantic samples [24] and 2048-dimensional ResNet-features, where images are resized to 256×256 pixels, cropped to 224×224 pixels, normalised with means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$ per RGB channel [3].

4.2 Evaluation Protocol

The evaluation protocol is based on computing the average per-class top-1 accuracy measured independently for each class before dividing their cumulative sum by the number of classes [3]. For GZSL, after computing the average per-class top-1 accuracy on seen classes $\mathcal{Y}^S$ and unseen classes $\mathcal{Y}^U$, we compute the harmonic mean of the seen and the unseen classification accuracy [3]. We also show results using the receiver operating characteristics (ROC) curve that measures the seen and the unseen classification accuracy over many operating points of the classifier [1].

4.3 Implementation Details

In this section, we describe the implementation details for the augmentation network, visual network and semantic networks that compose the model AN-GZSL, in terms of the model architecture and hyper-parameters (e.g. *number of epochs, batch size, number of layers, learning rate, weight decay, and learning rate decay*). Firstly, the augmentation network (composed of a generator (θ_g), a discriminator (θ_d), and a regressor (θ_r)) is defined in terms of a generative adversarial network (GAN) with cycle-consistency loss [2]. The generator consists of a single hidden layer with 4096 nodes and LeakyReLU activation [32] with an output layer of 2048 nodes (same dimension as ResNet [27] feature layer). The discriminator consists of a single hidden layer with 4096 nodes with a LeakyReLU activation function, the output layer has no activation. Secondly, the visual network (θ_ϕ) consists of a model parameterised with one fully connected layer from the 2048-dimensional visual space into label space $\mathcal{Y}$. Thirdly, the semantic network (θ_ψ) is defined as a bi-linear model [25] that matches the 2048-dimensional visual space with the semantic space. We introduce a dropout layer (rate equal to 0.2) for the visual and the semantic networks for regularisation during training. On all the benchmark data sets [13] we generate 300 visual samples per class for the training of the visual and the semantic networks. Temperature calibration (2) is done after the training finding the parameters τ_ψ and τ_ϕ with grid search minimization of the losses for the visual and semantic networks, using the validation set [3](this procedure does not require re-training of the whole model). Finally, we perform a Bayesian inference using Monte-Carlo dropout [33] because recent results suggest that such Bayesian inference can improve classification calibration and accuracy [34]. All hyper-parameters of the proposed AN-GZSL model are estimated with standard model selection methods on the validation sets proposed by Xian et. al. [3].

4.4 Ablation Study

In Table 1, we report the ablation study for the proposed method AN-GZSL. First, we report the results for inference computed by the visual network ($AN-GZSL^{\phi}$). Second, $AN-GZSL^{\psi}$ reports the results for our semantic network. Then, $AN-GZSL^{\tau=1}$ shows the combination of the visual and semantic network without the temperature calibration. $AN-GZSL^{T=1}$ shows the results without MC dropout inference, and the last row shows the results with MC dropout using the calibrated (i.e., multi-domain) multi-modal networks.

Table 1. GZSL results using per-class average top-1 accuracy on the test sets of unseen classes $\mathcal{Y}^U$, seen classes $\mathcal{Y}^S$, and H-mean result H – all results shown in percentage. The highlighted values represent the best ones for each column.

	CUB			**FLO**			**SUN**			**AWA**		
Classifier	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H
$AN-GZSL^{\phi}$	46.2	**61.5**	52.8	60.0	70.8	65.0	48.7	33.1	39.4	55.4	64.8	59.7
$AN-GZSL^{\psi}$	**77.7**	41.8	54.4	**84.9**	36.6	51.2	47.2	21.6	29.6	46.4	**67.3**	54.9
$AN-GZSL^{\tau=1}$	46.2	**61.5**	52.8	60.1	70.9	65.1	**53.3**	32.8	40.6	55.6	65.0	60.0
$AN-GZSL^{T=1}$	62.3	54.7	58.2	66.5	**84.4**	74.4	48.8	33.1	39.5	55.2	69.4	61.5
$AN-GZSL$	60.5	56.6	**58.5**	80.7	69.3	**74.5**	41.7	**37.1**	**41.7**	**58.2**	66.1	**61.9**

4.5 Results

In Table 2, we compare the GZSL results on CUB, FLO, SUN and AWA, produced by the proposed model AN-GZSL and several other methods previously proposed in the field. These methods are split into three groups: semantic approach, generative approach and domain balancing. We report the following metrics in Table 2: the accuracy for the unseen domain ($\mathcal{Y}^U$), the seen domain ($\mathcal{Y}^S$) and the harmonic-mean (H) between the two. Table 3 shows the top-1 accuracy on ImageNet for the proposed AN-GZSL and the results reported by previous methods on the same experimental setup.

In Fig. 2, we show the ROC results of the proposed method AN-GZSL, and the cycle-WGAN [2], which has code available online and represents the SOTA for the measure, to the best of our knowledge. Furthermore, Figure 2 shows seen and unseen classification results for previously published GZSL methods (please refer to Tab. 2 for the original references). We represent previous methods [13] by single (diamond-shaped) points denoting the results for seen and unseen classification accuracies – this is because previous methods only report a single operating point for the classification of seen and unseen classes).

Table 2. GZSL results using per-class average top-1 accuracy on the test sets of unseen classes ($\mathcal{Y}^U$), seen classes ($\mathcal{Y}^S$), and H-mean result (H); – all results shown in percentage. The highlighted values represent the best for each column.

	CUB			**FLO**			**SUN**			**AWA**		
Classifier	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H
Semantic approach												
DAP [5]	4.2	25.1	7.2	–	–	–	1.7	67.9	3.3	0.0	**88.7**	0.0
IAP [5]	1.0	37.8	1.8	–	–	–	0.2	72.8	0.4	2.1	78.2	4.1
DEVISE [35]	23.8	53.0	32.8	9.9	44.2	16.2	16.9	27.4	20.9	13.4	68.7	22.4
SJE [36]	23.5	59.2	33.6	13.9	47.6	21.5	14.7	30.5	19.8	11.3	74.6	19.6
LATEM [37]	15.2	57.3	24.0	6.6	47.6	11.5	14.7	28.8	19.5	7.3	71.7	13.3
ESZSL [38]	12.6	63.8	21.0	11.4	56.8	19.0	11.0	27.9	15.8	6.6	75.6	12.1
ALE [25]	23.7	62.8	34.4	13.3	61.6	21.9	21.8	33.1	26.3	16.8	76.1	27.5
PQZSL [39]	43.2	51.4	46.9	–	–	–	35.1	35.3	35.2	31.7	70.9	43.8
AREN [40]	38.9	78.7	52.1	–	–	–	19.0	38.8	25.5	–	–	–
MLSE [41]	22.3	71.6	34.0	–	–	–	20.7	36.4	26.4	–	–	–
Generative approach												
SAE [42]	8.8	18.0	11.8	–	–	–	7.8	54.0	13.6	1.8	77.1	3.5
f-CLSWGAN [43]	43.8	60.6	50.8	58.8	70.0	63.9	47.9	32.4	38.7	56.0	62.8	59.2
cycle-WGAN [2]	46.0	60.3	52.2	59.1	71.1	64.5	48.3	33.1	39.2	56.4	63.5	59.7
CADA-VAE [12]	51.6	53.5	52.4	–	–	–	47.2	35.7	40.6	57.3	72.8	64.1
GDAN [8]	39.3	66.7	49.5	–	–	–	38.1	**89.9**	**53.4**	–	–	–
GMN [11]	56.1	54.3	55.2	–	–	–	**53.2**	33.0	40.7	61.1	71.3	**65.8**
Zhu et.al.[44]	33.4	**87.5**	48.4	–	–	–	–	–	–	–	–	–
LisGAN [9]	46.5	57.9	51.6	57.7	**83.8**	68.3	42.9	37.8	40.2	52.6	76.3	62.3
External Domain Classifier												
CMT [18]	7.2	49.8	12.6	–	–	–	8.1	21.8	11.8	0.9	87.6	1.8
DAZSL [15]	41.0	60.5	48.9	59.6	81.4	68.8	35.3	40.2	37.6	**64.8**	51.7	57.5
Ours												
$AN-GZSL$	**60.5**	56.6	**58.5**	**80.7**	69.3	**74.5**	41.7	37.1	41.7	58.2	66.1	61.9

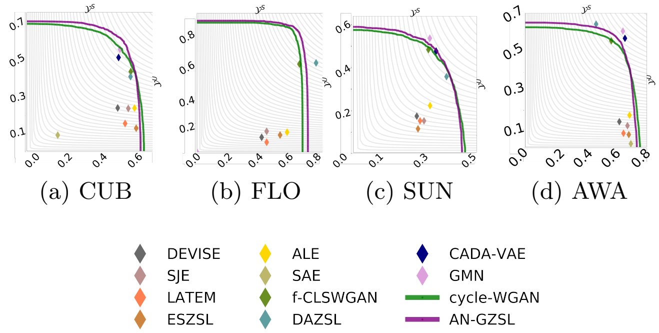
Table 3. GZSL ImageNet results – all results shown in percentage. Please see caption of Table 2 for details on each measure. The highlighted values represent the best ones.

Classifier	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H
f-CLSWGAN [13]	0.7	–	–
cycle-WGAN [2]	1.5	**66.5**	2.8
$AN-GZSL$	**2.5**	47.4	**4.8**

Using the graph in Fig. 2, we compute the AUSUC on each data set for AN-GZSL – results are shown in the supplementary material. Moreover, we added the results reported by the previous methods EZSL [38], fCLSWGAN [13], cycle-WGAN [2] and DAZSL [15]. We were able to compute the AUSUC results for AN-GZSL and cycle-WGAN, but the other AUSUC results were extracted from [15].

5 Discussions

Ablation study. Table 1 shows the importance of each component of AN-GZSL, where the H-mean tends to be higher for the multi-modal approach, compared to each individual modality. The multi-domain multi-modal method that relies on Bayesian inference (last row) shows the highest H-mean on all data sets. The similarity between the results of the un-calibrated ($AN-GZSL^{\tau=1}$) and the visual network $AN-GZSL^{\phi}$ suggests that un-calibrated multi-modal classifiers rely entirely on the visual classifiers. This is explained by the fact that



(a) CUB (b) FLO (c) SUN (d) AWA

Fig. 2. ROC curves for the proposed method AN-GZSL, and several baseline and state-of-the-art methods (please see text and Table 2 for details about the methods). Note that these graphs are used to compute the AUSUC in Table **??**. (best seen on the digital format with colors)

the classification results produced by the un-calibrated semantic classifier show classification probabilities close to a uniform distribution, in contrast to the un-calibrated visual classifier that shows more non-uniform distributions. However, when calibration is applied, the classification probabilities produced by both classifiers are pushed further away from the uniform distribution, which means that the sum of calibrated classifiers can produce results that are different from the original visual and semantic classifiers. In fact, Table 1 shows that the AN-GZSL classification accuracy is always higher than single-modality classification results. This multi-modal calibrated classifier also produces the most balanced classification results between the seen and unseen domains for all data sets. These results suggest that our proposed method provides a way to correct some of the mistakes made using an individual modality. For example, this can happen when the classification probabilities of the correct class are relatively high for both modalities, but not the highest in any modality, and when summed, the correct class receives the highest confidence.

Another important point to notice from Table 1 is that our proposed AN-GZSL seems to be more advantageous in fine-grained (i.e., CUB and FLO) than in coarse-grained (i.e., SUN and AWA) data sets, where the key to explain such discrepancy lies in the effectiveness of temperature calibration. In coarse-grained data sets, the results from the calibrated visual classifier are almost binary, with the highest classification probability close to one and all other probability values close to zero. The calibrated semantic classifier shows a more uniform distribution, which when combined with the almost binary results of the visual classifier is less effective (than in fine-grained problems) to change a possibly incorrect visual classifier result for the multi-domain multi-modal model. On the other hand, in fine-grained data sets, the results from the calibrated visual classifier are farther from binary, which when combined with the results from the semantic classifier can be more effective to change an incorrect visual classifier result for the multi-domain multi-modal model. We speculate that this different performance between visual classifiers can be explained by the cluttered or the scattered nature of visual class distributions in fine-grained or coarse-grained data sets – that is, more cluttered distributions provide more space for improvement with an effective temperature calibration.

A final point from Table 1 is the apparent more accurate classification results for the unseen classes than for the seen classes for most of the data sets. We studied this issue by running an unpaired t-test to check the significance of these results, and for CUB, FLO and AWA the p-values are larger than 0.05, implying that we cannot reject the null hypothesis (i.e., the hypothesis that there is no significant difference between seen and unseen classification accuracies). For SUN, the p-value is smaller than 0.05, which we believe is due to the large size of the data set. It is important to note that previous methods have also reported similar classification results for SUN [2, 13]. Nevertheless, it is worth mentioning that the seen and unseen classification results represent the performance of a particular adjustable operating point of the methods, as shown in Fig. 2. Hence, measures that summarise the seen vs unseen classification, like H-mean or AUSUC, can characterise better the method performance, but the dependence of H-mean on an operating point makes it less reliable than AUSUC, so we advocate the use of AUSUC as a more general measure for GZSL approaches.

Comparison with SOTA. In Table 2, we notice a clear trend of the proposed AN-GZSL to perform substantially better than the SOTA in terms of H-mean and classification accuracy on unseen classes for fine-grained (CUB and FLO) data sets, and competitively for coarse-grained data sets (SUN and AWA). This result shows that the more challenging classification problem offered by the fine-grained data sets represents an ideal situation for exploring multi-modal and multi-domain classification. We discuss in the ablation study above, the reasons behind the superior performance in fine-grained data sets of our proposed AN-GZSL method.

Another interesting point to observe from Table 2 is that none of the competing methods stand out as a clear SOTA approach for all data sets since one method can be better in one data set, but worse in others. In fact, out of the four data sets studied, AN-GZSL is better in two, GDAN is better in one and GMN is better in another. It is also worth comparing the performance of previous semantic approaches in Table 3, and our proposed semantic network, represented by $AN - GZSL^{\psi}$ in Table 1. This comparison is important because our proposed semantic network introduces one significant novelty, which is the use of visual data augmentation for training the semantic classifier. Our proposed $AN - GZSL^{\psi}$ produces substantially better results in terms of H-mean and classification accuracy on unseen classes for CUB, FLO and AWA.

In terms of the large-scale data set ImageNet, we show in Table 3 that the proposed method establishes a new SOTA in terms of the H-mean result. More specifically, the proposed method achieves around 80% of relative H-mean improvement. We speculate that these results can be explained by the similar challenges present in fine-grained and large-scale data sets. Also, the proposed approach scales as well as f-CLSWGAN [13] and cycle-WGAN [2] with respect to the number of classes and samples.

Seen and unseen classification graphs. Figure 2 shows the trade-off between the classification of seen and unseen classes for GZSL methods. In particular, it is interesting to notice a fact that is prevalent in GZSL methods, which is the classification imbalance that usually favours the seen classes – the figure illustrates that the majority of the previous methods (represented by diamonds) lie at the bottom-right part of the graphs, indicating the preference for seen classes. In terms of seen and unseen curves, the more balanced methods (see Table 2) usually lies close to the elbow of the curve, located at the top-right part of the graph.

AUSUC. Figure 2 shows that the proposed approach, AN-GZSL, outperforms previous methods on data sets CUB, SUN and FLO. For AWA, we achieve competitive performance, where the proposed method is the second best. It is worth emphasising that the AUSUC measure provides a more complete assessment of GZSL methods, where it is no longer necessary to commit to a particular operating point of the classification of seen and unseen classes.

6 Conclusions and Future Work

In this paper, we introduce a new approach to perform GZSL using a multi-modal multi-domain augmentation network. The proposed approach is the first to explore visual data augmentation for training visual *and* semantic classifiers, enabling a truly and novel multi-modal training and inference for GZSL. In addition, we show that the calibration of those visual and semantic classifiers provide an effective multi-domain classification, where the classification of seen and unseen classes are accurate and well balanced. The experimental results show that the proposed approach has established new state-of-the-art GZSL harmonic mean results for three benchmark data sets (CUB, FLO, and Imagenet). In particular, we report results that are substantially better than the previous methods on CUB and FLO, which are fine-grained data sets, and competitive on SUN and AWA, which are coarse-grained data sets. Moreover, the results of the proposed approach outperform previous methods on Imagenet data set by a large margin. Also, our proposed AN-GZSL achieves the best performance in terms of AUSUC for three benchmark data sets.

In the future, we intend to study more thoroughly the reason behind the performance difference observed between fine-grained and coarse-grained data sets. We will also investigate why it is challenging to obtain high classification accuracy on the unseen classes of the large scale ImageNet data set.

7 Acknowledge

This work was partially supported by Australian Research Council grants (FT190100525 and CE140100016).

References

1. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: European Conference on Computer Vision, Springer (2016) 52–68
2. Felix, R., Kumar, B.V., Reid, I., Carneiro, G.: Multi-modal cycle-consistent generalized zero-shot learning. In: European Conference on Computer Vision, Springer (2018) 21–37
3. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. CoRR **abs/1707.00600** (2017)
4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013) 3111–3119
5. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. (2009) 951–958
6. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **36** (2014) 453–465
7. Bucher, M., Herbin, S., Jurie, F.: Generating visual representations for zero-shot classification. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 2666–2673
8. Huang, H., Wang, C., Yu, P.S., Wang, C.D.: Generative dual adversarial network for generalized zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 801–810
9. Li, J., Jin, M., Lu, K., Ding, Z., Zhu, L., Huang, Z.: Leveraging the invariant side of generative zero-shot learning. arXiv preprint arXiv:1904.04092 (2019)
10. Paul, A., Krishnan, N.C., Munjal, P.: Semantically aligned bias reducing zero shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7056–7065
11. Sariyildiz, M.B., Cinbis, R.G.: Gradient matching generative networks for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2168–2178
12. Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: Generalized zero- and few-shot learning via aligned variational autoencoders. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 8247–8255
13. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. arXiv (2017)
14. Verma, V.K., Arora, G., Mishra, A., Rai, P.: Generalized zero-shot learning via synthesized examples. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
15. Atzmon, Y., Chechik, G.: Adaptive confidence smoothing for generalized zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 11671–11680
16. Bhattacharjee, S., Mandal, D., Biswas, S.: Autoencoder based novelty detection for generalized zero shot learning. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 3646–3650
17. Felix, R., Harwood, B., Sasdelli, M., Carneiro, G.: Generalised zero-shot learning with domain classification in a joint semantic and visual space. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), IEEE (2019) –

18. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: Advances in Neural Information Processing Systems. (2013) 935–943
19. Zhang, H., Koniusz, P.: Model selection for generalized zero-shot learning. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 0–0
20. Zhou, Z.H.: Ensemble methods: foundations and algorithms. Chapman and Hall/CRC (2012)
21. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. (2010)
22. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on, IEEE (2008) 722–729
23. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 248–255
24. Wang, P., Liu, L., Shen, C., Huang, Z., van den Hengel, A., Shen, H.T.: Multi-attention network for one shot learning. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR. (2017) 22–25
25. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE transactions on pattern analysis and machine intelligence **38** (2016) 1425–1438
26. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: The IEEE International Conference on Computer Vision (ICCV). (2015)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
28. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org (2017) 1321–1330
29. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv (2017)
30. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
31. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 49–58
32. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. Volume 30. (2013) 3
33. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Insights and applications. In: Deep Learning Workshop, ICML. (2015)
34. Gal, Y., Hron, J., Kendall, A.: Concrete dropout. In: Advances in Neural Information Processing Systems. (2017) 3581–3590
35. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems. (2013) 2121–2129
36. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2927–2936

37. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 69–77
38. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning. (2015) 2152–2161
39. Li, J., Lan, X., Liu, Y., Wang, L., Zheng, N.: Compressing unknown images with product quantizer for efficient zero-shot classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5463–5472
40. Xie, G.S., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., Shao, L.: Attentive region embedding network for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9384–9393
41. Ding, Z., Liu, H.: Marginalized latent semantic encoder for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6191–6199
42. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 3174–3183
43. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA (2018)
44. Zhu, P., Wang, H., Saligrama, V.: Generalized zero-shot recognition based on visually semantic embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2995–3003