



HAL
open science

Zero-Shot Classification by Generating Artificial Visual Features

Maxime Bucher, Stéphane Herbin, Frédéric Jurie

► **To cite this version:**

Maxime Bucher, Stéphane Herbin, Frédéric Jurie. Zero-Shot Classification by Generating Artificial Visual Features. RFIAP, Jun 2018, Paris, France. hal-01796440

HAL Id: hal-01796440

<https://hal.science/hal-01796440>

Submitted on 20 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Zero-Shot Classification by Generating Artificial Visual Features

Maxime Bucher^{1,2}

Stéphane Herbin¹

Frédéric Jurie²

¹ ONERA - The French Aerospace Lab, Palaiseau, France

² Normandie Univ, UNICAEN, ENSICAEN, CNRS

firstname.name @onera.fr @unicaen.fr

Abstract

This paper addresses the task of learning an image classifier when some categories are defined by semantic descriptions only (visual attributes) while the others are also defined by exemplar images. This task is often referred to as Zero-Shot classification (ZSC). Most of the previous methods rely on learning a common embedding space allowing to compare visual features of unknown categories with semantic descriptions. This paper argues that these approaches are limited as i) efficient discriminative classifiers can't be used and ii) classification tasks with seen and unseen categories (Generalized Zero-Shot Classification or GZSC) can't be addressed efficiently. This paper suggests to address ZSC and GZSC by i) learning a conditional generator using seen classes ii) generate artificial training examples for the categories without exemplars. ZSC is therefore turned into a standard supervised learning problem. Experiments with 4 generative models and 6 datasets experimentally validate the approach, giving state-of-the-art results on both ZSC and GZSC.

1 Introduction and related works

Zero-Shot Classification (ZSC) [26] addresses classification problems where not all the classes are represented in the training examples. ZSC can be made possible by defining a high-level description of the categories, relating the new classes (*the unseen classes*) to classes for which training examples are available (*seen classes*). Learning is usually done by leveraging an intermediate level of representation, the attributes, that provide semantic information about the categories to classify. As pointed out by [38] this paradigm can be compared to how human can identify a new object from a description of it, leveraging similarities between its description and previously learned concepts. Recent ZSC algorithms ([1, 7]) do the classification by defining a zero-shot prediction function that outputs the class y having the maximum compatibility score with the image x : $f(x) = \arg \max_y S(x, y)$. The compatibility function, for its part, is often defined as $S(x, y; W) = \theta(x)^t W \phi(y)$ where θ and ϕ are two projections and W is a bilinear function relating the two in a common embedding. There are different variants in the recent litera-

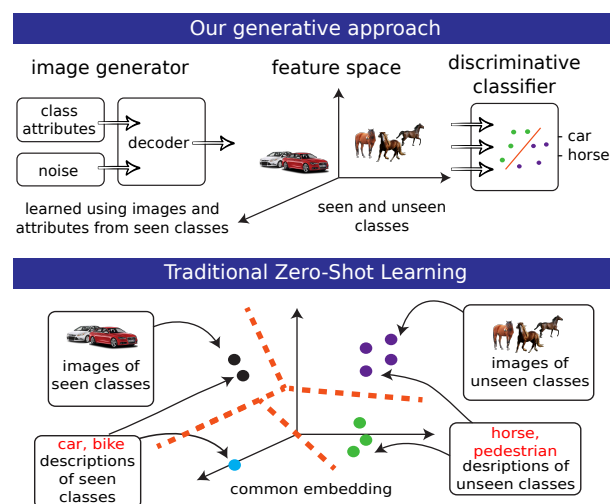


Figure 1: On opposite to traditional embedding approaches, our method consists in i) learning an image feature generator capable of generating artificial image representations from given attributes ii) learning a discriminative classifier from the artificially generated training data.

ture on how the projections or the similarity measure are computed [14, 11, 18, 34, 38, 45, 46, 49], but in all cases the class is chosen as the one maximizing the compatibility score. This embedding and maximal compatibility approach, however, does not exploit, in the learning phase, the information potentially contained in the semantic representation of the unseen categories. The only step where a discriminating capability is exploited is in the final label selection which uses an $\arg \max_y$ decision scheme, but not in the setting of the compatibility score itself.

A parallel can be easily done between the aforementioned approaches and *generative models* such as defined in the machine learning community. Generative models estimate the joint distribution $p(y, x)$ of images and classes, often by learning the class prior probability $p(y)$ and the class-conditional density $p(x|y)$ separately. However, as it has been observed for a long time [43], discriminative approaches trained for predicting directly the class label have better performance than model-based approaches as long

as the learning database reliably samples the target distribution.

Despite one can expect discriminative methods to give better performance [43], they can't be used directly in the case of ZSC for obvious reasons: as no images are available for some categories, discriminative classifiers cannot be learned out-of-the-box.

This paper proposes to overcome this difficulty by generating training features for the unseen classes, in such a way that standard discriminative classifiers can be learned (Fig. 1). Generating data for machine learning tasks has been studied in the literature, [21] or [5] to compensate for imbalanced training sets. Generating novel training examples from the existing ones is also at the heart of the technique called *Data Augmentation*, frequently used for training deep neural networks [27]. When there is no training data at all for some categories, some underlying parametric transformation can be used to generate missing training data, assuming a mapping from the underlying transformation to the image space. [15] generated images by applying warping and other geometric / photometric transformations to prototypical logo exemplars. A similar idea was also presented in [23] for text spotting in images. [10] capture what they call *The Gist of a Gesture* by recording human gestures, representing them by a model and use this model to generate a large set of realistic gestures.

We build in this direction, in the context of ZSC, the underlying representation being an attribute or text based description of the unseen categories. The transformation from attributes to image features is learned from examples of seen classes. A relevant way to learn this transformation is to use generative models such as *denoising auto encoders* [6] and *generative adversarial nets* (GAN) [19] or their variants [13, 31]. GANs consist in estimating generative models via an adversarial process simultaneously learning two models, a generative model that captures the data distribution, and a discriminative model that estimates the probability that a sample came from the training data rather than the generator. The *Conditional Generative Adversarial Nets* of [33] is a very relevant variant adapted to our problem.

In addition to the advantage of using discriminative classifiers – which is expected to give better performance – our approach, by nature, can address the more realistic task of Generalized Zero-Shot Classification (GZSC). This problem, introduced in [12], assumes that both seen and unseen categories are present at test time, making the traditional approaches suffering from a decision bias towards seen classes, i.e. they are predominantly chosen. In contrast, the proposed approach uses (artificial) training examples of both seen and unseen classes during training, avoiding the aforementioned issues.

Another reason to perform classification inference directly in the visual feature space rather than in an abstract attribute or embedding space is that data are usually more easily separated in the former, especially when using dis-

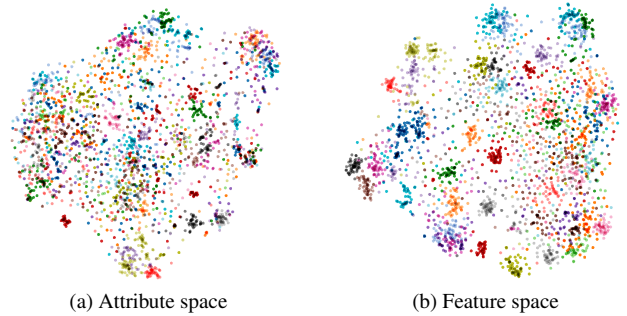


Figure 2: CUB dataset: T-SNE visualization of feature vectors in (a) attribute space and (b) visual feature space. Data seem visually better separated in visual feature space.

criminant deep features that are now commonly available. Fig. 2 displays a 2D projection of the two kinds of representations and justifies this intuition. It can therefore be expected that a reliable conditional visual feature generator will be able to translate ZSC into a well separated supervised classification problem.

This paper experimentally validates the proposed strategy on 5 standard Zero-Shot classification datasets (Animals with Attributes 1 (AWA1) [26], Animals with Attributes 2 (AWA2) [48], SUN attributes (SUN) [37], Apascal&Ayahoo (aP&Y) [17] and Caltech-UCSD Birds-200-2011 (CUB) [44]), and gives insight on how the approach scales on large datasets such as ImageNet [14]. It shows state-of-the-art performance on all datasets for both ZSC and GZSC.

This paper builds on [9] and updates the evaluation to comply with the de facto evaluation standard on ZSC described in [48]. We also compare our approach with papers that will be published soon [4, 47] and that exploit the same artificial feature generation idea that we previously proposed in [9].

2 Approach

2.1 Zero shot classification

As motivated in the introduction, we address in this paper the problem of learning a classifier capable of discriminating between a given set of classes where empirical data is only available for a subset of it, the so-called *seen* classes. In the vocabulary of zero-shot classification, the problem is usually qualified as inductive — we do not have access to any data from the unseen classes — as opposed to transductive where the *unseen* data is available but not the associated labels. We do not address in this paper the transductive setting, considering that the availability of target data is a big constraint in practice.

The learning dataset \mathcal{D}_s is defined by a series of triplets $\{x_i^s, a_i^s, y_i^s\}_{i=1}^{N_s}$ where $x_i^s \in \mathcal{X}$ is the raw data (image or features), $y_i^s \in \mathcal{Y}_s$ is the associated class label and a_i^s is a rich semantic representation of the class (attributes, word vector or text) belonging to \mathcal{A}_s . This semantic repre-

sentation is expected to i) contain enough information to discriminate between classes by itself, ii) be predictable from raw data and iii) infer unambiguously the class label $y = l(\mathbf{a})$.

In an inductive ZSC problem, all that is known regarding the new target domain is the set of semantic class representations \mathcal{A}_u of the *unseen* classes. The goal is to use this information and the structure of the semantic representation space to design a classification function f able to predict the class label $\hat{y} = f(\mathbf{x}; \mathcal{A}_u, \mathcal{D}_s)$. The classification function f is usually parametric and settled by the optimization of an empirical learning criterion.

2.2 Discriminative approach for ZSC

In ZSC, the main problem is precisely the fact that no data is available for the unseen classes. The approach taken in this paper is to artificially generate data for the unseen classes given that seen classes and their semantic representations provide enough information to do so, and then apply a discriminative approach to learn the class predictor.

The availability of data for the unseen classes has two main advantages: it can make the classification of seen *and* unseen classes as a single homogeneous process, allowing to address Generalized Zero Shot Classification as a single supervised classification problem; it potentially allows a larger number of unseen classes, which is for instance required for datasets such ImageNet [14].

Let $\hat{\mathcal{D}}_u = \{\hat{x}_i^u, a_i^u, y_i^u\}_{i=1}^{N_u}$ be a database generated to account for the unseen semantic class representation $a^u \in \mathcal{A}_u$. The ZSC classification function becomes: $\hat{y} = f_D(\mathbf{x}; \hat{\mathcal{D}}_u, \mathcal{D}_s)$ and can be used in association with the seen data \mathcal{D}_s , to learn a homogeneous supervised problem.

2.3 Generating unseen data

Our generators of unseen data build on the recently proposed approaches for conditional data generation as presented in section 1. The idea is to learn globally a parametric random generative process G using a differentiable criterion able to compare, as a whole, a target data distribution and a generated one.

Given \mathbf{z} a random sample from a fixed multivariate prior distribution, typically uniform or Gaussian, and \mathbf{w} the set of parameters, new sample data consistent with the semantic description \mathbf{a} are generated by applying the function: $\hat{\mathbf{x}} = G(\mathbf{a}, \mathbf{z}; \mathbf{w})$. A simple way to generate conditional $\hat{\mathbf{x}}$ data is to concatenate the semantic representation \mathbf{a} and the random prior \mathbf{z} as the input of a multi-layer network, as shown in Fig. 3.

We now present 4 different strategies to design such a conditional data generator, the functional structure of the generator being common to all the described approaches.

Generative Moment Matching Network A first approach is to adapt the Generative Moment Matching Network (GMMN) proposed in [28] to conditioning. The generative process will be considered as good if for each semantic description \mathbf{a} two random populations $\mathcal{X}(\mathbf{a})$ from

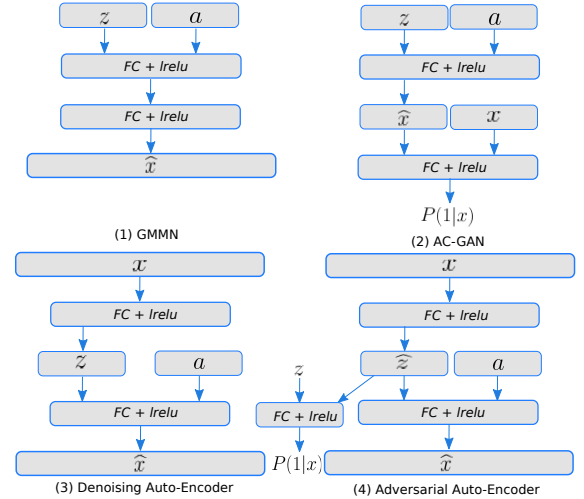


Figure 3: Architecture of the different generative models studied. z and a are respectively sample noise and semantic representation. $FC + lrelu$ represents fully connected layer with leaky-relu non-linearity. x stands for real feature sample, \hat{x} a generated one.

\mathcal{D}_s and $\hat{\mathcal{X}}(\mathbf{a}; \mathbf{w})$ sampled from the generator have low *maximum mean discrepancy* which is a probability divergence measure between two distributions. This divergence can be approximated using a Hilbert kernel based statistics [20] – typically a linear combination of Gaussian functions with various widths — which has the big advantage of being differentiable and may be thus exploited as a machine learning cost. Network parameters \mathbf{w} are then obtained by optimizing the differentiable statistics by stochastic gradient descent, using batches of generated and real data conditioned by the semantic description \mathbf{a} .

Conditional Generative adversarial models Our second model builds on the principles of the generative adversarial networks (GAN), which is to learn a discrepancy measure between a true and a generated distributions — the *Discriminator* — simultaneously with the data generator. One extension allowing to produce conditional distributions is the AC-GAN [36] (Fig. 3) where the generated and the true distributions are compared using a binary classifier, and the quality of the conditional generation is controlled by the performance of this auxiliary task.

This model bears similarities with the GMMN model, the key difference being that in the GMMN distributions of true and generated data are compared using the kernel based empirical statistics while in the AC-GAN case it is measured by a learned discriminative parametric model.

Denoising Auto-Encoder Our third generator relies on the work presented in [6], where an encoder/decoder structure is proposed to design a data generator, the latent code playing the role of the random prior \mathbf{z} used to generate the data. A simple extension able to introduce a conditional data generation control has been developed by concatenating the semantic representation \mathbf{a} to the code that is fed to

the decoder (Fig. 3).

In practice, this model is learned as a standard auto-encoder, except that i) some noise is added to the input and ii) the semantic representation \mathbf{a} is concatenated to the code in the hidden layer. For generating novel examples, only the decoder part, the head of the network using \mathbf{z} and \mathbf{a} as input to produce $\hat{\mathbf{x}}$ is used.

Adversarial Auto-Encoder Our fourth generator is inspired by [31], which is an extension of the denoising auto-encoder. It introduces an adversarial criterion to control the latent code produced by the encoder part, so that the code distribution matches a fixed prior distribution. This extra constraint is expected to ensure that all parts of the sampling prior space will produce meaningful data.

During training, both the auto-encoder and the discriminator are learned simultaneously. For generating novel examples, as for the denoising auto-encoder, only the decoder part is used.

2.4 Implementing the generators

We implemented our 4 generative models with neural networks, whose architectures are illustrated Fig. 3. Hidden layers are fully connected (FC) with leaky-relu non-linearity [29] (leakage coefficient of 0.2). For the models using a classifier (the AC-GAN and the Adversarial auto-encoder), the classifier is a linear classifier (fully connected layer + Softmax activation function). The loss used to measure the quality of the reconstruction in the two auto-encoders is the L2 norm.

Regarding how to sample the noise \mathbf{z} , we did not observe any difference between sampling it from a Gaussian distribution or from a uniform distribution.

3 Experiments

This section presents the datasets and the experimental settings, compares the different generative models described in the previous section, shows how our approach can be used for the Generalized Zero-shot Classification Task—which is one of the key contributions of the paper –, provides some experiments on a large Zero-Shot classification task, and, finally, compares our approach with state-of-the-art Zero-Shot approaches on the regular Zero-shot Classification Task.

3.1 Datasets and Settings

The experimental evaluation is done on 5 standard publicly available ZSC benchmarks: Animals with Attributes 1 (AWA1) [26], Animals with Attributes 2 (AWA2) [48], SUN attributes (SUN) [37], Apascal&Ayahoo (aP&Y) [17] and Caltech-UCSD Birds-200-2011 (CUB) [44]. These benchmarks exhibit a great diversity of concepts: SUN and CUB are for fine-Grained categorization and include respectively birds and scenes images; AWA1 and AWA2 contain images of animals from 50 different categories; finally, aP&Y has broader concepts, from cars to animals. For each dataset, attributes descriptions are given, either at

the class level or at image level. aP&Y, CUB and SUN have per image binary attributes that we average to produce per class real valued representations. In order to make comparisons with other works possible, we follow the same training/testing splits as [48].

Image features are computed using the 2048-dim top-layer hidden unit activations of a 101-layered ResNet [22]. We keep the weights learned on ImageNet fixed and don't apply any fine-tuning.

The classifiers are obtained by adding a standard Fully Connected with Softmax layer to the pre-trained networks. We purposively chose a simple classifier to better observe the behavior of the generators. In all our experiments we generated 500 artificial image features per class, which we consider to be a reasonable trade-off between accuracy and training time; we have not observed any significant improvement when adding more images.

Each architecture has its own set of hyper-parameters (typically the number of units per layer, the number of hidden layers, the learning rate, *etc.*). They are obtained through a 'Zero-shot' cross-validation procedure. In this procedure, 20% of the seen classes are considered as unseen (hence used as validation set), allowing to choose the hyper-parameters maximizing the accuracy on this so-obtained validation set. In practice, typical values for the number of neurons (resp. the number of hidden layers) are in the range of [500-2000] (resp. 1 or 2).

Model parameters are initialized according to a centered Gaussian distribution ($\sigma = 0.02$). They are optimized with the Adam solver [24] with a cross-validated learning rate (typically of 10^{-4}), using mini-batches of size 128 except for the GMMN where each batch contains all the training images of one class, to make the estimation of the statistics more reliable. In order to avoid over-fitting, we used dropout [41] at every layer (probability of drop of 0.2 for the inputs layers and of 0.5 for the hidden layers). Input data (both image features and w2c vectors) are scaled to [0,1] by applying an affine transformation. With the TensorFlow framework [16] running on a Nvidia Titan X Pascal GPU, the learning stage takes around 10 minutes for a given set of hyper-parameters.

3.2 Comparing the different generative models

Our first round of experiments consists in comparing the performance of the 4 generative models described in Section 2.3, on the regular and Generalized Zero-shot classification tasks. Performance is reported in Table 1 and Table 3. We can see that the GMMN model outperforms the 3 other ones, on average. Its optimization is also computationally more stable than the adversarial versions. We consequently chose this generator for the large scale Zero-Shot classification case.

In order to have a better understanding of the behavior of the 4 models, we visualized the generated data, after projecting them in 2D with a T-SNE projection [30]. The pro-

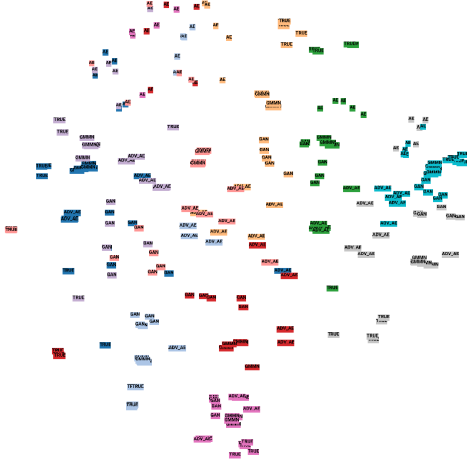


Figure 4: AwA1 dataset: T-SNE visualization of artificial image representations generated by the GMMN, AC-GAN, ADV AE (Adversarial Auto-encoder), AE (Denoising Auto-encoder) and TRUE (real data). Best viewed on a computer screen with strong zoom factor.

jections are given Fig. 4). For each model, 10 images per class are generated and compared to 10 real images taken from unseen classes. As we can see, the GMMN model produced clusters that are, on average, very similar to the real data clusters.

We explain the superiority of the GMMN model by the fact it aligns the distributions by using an explicit model of the divergence of the distributions while the adversarial autoencoder and the AC-GAN have to learn it. For its part, the denoising autoencoder doesn't have any guaranty that the distributions are aligned.

3.3 Generalized Zero-Shot Classification task

In this section, we follow the Generalized Zero-Shot Learning (GZSC) protocol introduced by Chao *et al* [12]. In this protocol, test data are from any classes, seen or unseen. This task is more realistic and harder, as the number of class candidates is larger.

We follow the notations of [48]: u is the classification accuracy of test images from unseen classes, s is the classification accuracy of test images from seen classes and H denotes the harmonic mean between u and s .

In the two cases, the classifier is learned with training data combining image feature generated for unseen classes and real feature for seen classes.

Most of the recent ZSC works, [3, 8, 7, 38] are focused on improving the embedding or the scoring function. However, [12] has shown that this type of approach is unpractical with GZSC. Indeed the scoring function is in this case biased toward seen classes, leading to very low accuracy on the unseen classes. This can be seen on Table 1 (u column), where the accuracy drops significantly compared to regular ZSC performance. The data distribution of the ZSC

datasets are strongly subject to this bias, as unseen classes are very similar to seen classes both in terms of visual appearance and attribute description. When seen and unseen classes are candidates, it becomes much harder to distinguish between them. For example, the horse (seen) and the zebra classes (unseen) of the AwA dataset cannot be distinguished by standard ZSC methods.

Table 1 shows that the generative approach outperforms by a large margin all the other approaches that rely on a common semantic embedding. This gain can be explained by the fact that the discrimination step doesn't suffer from the bias induced by only learning from seen data. Two very recent works [4, 47] (to be published at CVPR'2018) that describe very similar approaches to the one we previously described in [9] but with slight variations on the generative network, confirm the advantage of using generated data to solve the ZSC problem.

3.4 Large Scale Zero-Shot Classification

Table 2: Zero-shot and Generalized ZSC on ImageNet. Our model is evaluated on three different scenarios: i) 2-hop: 1,509 classes ii) 3-hop: 7,678 classes, iii) All: all unseen categories. (+1K) stands for the generalized case.

Scenario	Method	Flat Hit @K				
		1	2	5	10	20
2-hop	Frome [18]	6.0	10.0	18.1	26.4	36.4
	Norouzi [34]	9.4	15.1	24.7	32.7	41.8
	Changpinyo [11]	10.5	16.7	28.6	40.1	52.0
	Ours.	13.05	21.52	33.71	43.91	57.31
2-hop (+1K)	Frome [18]	0.8	2.7	7.9	14.2	22.7
	Norouzi [34]	0.3	7.1	17.2	24.9	33.5
	Ours.	4.93	13.02	20.81	31.48	45.31
3-hop	Frome [18]	1.7	2.9	5.3	8.2	12.5
	Norouzi [34]	2.7	4.4	7.8	11.5	16.1
	Changpinyo [11]	2.9	4.9	9.2	14.2	20.9
	Ours.	3.58	5.97	11.03	16.51	23.88
3-hop (+1K)	Frome [18]	0.5	1.4	3.4	5.9	9.7
	Norouzi [34]	0.2	2.4	5.9	9.7	14.3
	Ours.	1.99	4.01	6.74	11.72	16.34
All	Frome [18]	0.8	1.4	2.5	3.9	6.0
	Norouzi [34]	1.4	2.2	3.9	5.8	8.3
	Changpinyo [11]	1.5	2.4	4.5	7.1	10.9
	Ours.	1.90	3.03	5.67	8.31	13.14
All (+1K)	Frome [18]	0.3	0.8	1.9	3.2	5.3
	Norouzi [34]	0.2	1.2	3.0	5.0	7.5
	Ours.	1.03	1.93	4.98	6.23	10.26

We compared our approach with state-of-the-art methods on a large-scale Zero-Shot classification task. These experiences mirror those presented in [18]: 1000 classes from those of the ImageNet 2012 1K set [39] are chosen for training (seen classes) while 20,345 others are considered to be unseen classes with no image available. Image features are computed with the GoogLeNet network [42]. In contrast with ZSC datasets, no attributes are provided for defining unseen classes. We represent those categories using a skip-gram language model [32]. This model is learned on a dump of the Wikipedia corpus (≈ 3 billion words). Skip-gram is a language model learned to pre-

Table 1: Generalized Zero-Shot classification accuracy. Image features are obtained with the 101-layered ResNet [22] network. We follow the notations of [48]: u is the classification accuracy of test images from unseen classes, s is the classification accuracy of test images from seen classes and H denotes the harmonic mean between u and s .

Method	Generalized Zero-Shot Learning														
	CUB			P&Y			SUN			AWA1			AWA2		
	u	s	H	u	s	H	u	s	H	u	s	H	u	s	H
DAP [26]	1.7	67.9	3.3	4.8	78.3	9.0	4.2	25.1	7.2	0.0	88.7	0.0	0.0	84.7	0.0
IAP [26]	0.2	72.8	0.4	5.7	65.6	10.4	1.0	37.8	1.8	2.1	78.2	4.1	0.9	87.6	1.8
CONSE [35]	1.6	72.2	3.1	0.0	91.2	0.0	6.8	39.9	11.6	0.4	88.6	0.8	0.5	90.6	1.0
CMT [40]	7.2	49.8	12.6	1.4	85.2	2.8	8.1	21.8	11.8	0.9	87.6	1.8	0.5	90.0	1.0
CMT* [40]	4.7	60.1	8.7	10.9	74.2	19.0	8.7	28.0	13.3	8.4	86.9	15.3	8.7	89.0	15.6
SSE [49]	8.5	46.9	14.4	0.2	78.9	0.4	2.1	36.4	4.0	7.0	80.5	12.9	8.1	82.5	14.8
LATEM [46]	15.2	57.3	24.0	0.1	73.0	0.2	14.7	28.8	19.5	7.3	71.7	13.3	11.5	77.3	20.0
ALE [2]	23.7	62.8	34.4	4.6	73.7	8.7	21.8	33.1	26.3	16.8	76.1	27.5	14.0	81.8	23.9
DEVISE [18]	23.8	53.0	32.8	4.9	76.9	9.2	16.9	27.4	20.9	13.4	68.7	22.4	17.1	74.7	27.8
SJE [3]	23.5	59.2	33.6	3.7	55.7	6.9	14.7	30.5	19.8	11.3	74.6	19.6	8.0	73.9	14.4
ESZSL [38]	12.6	63.8	21.0	2.4	70.1	4.6	11.0	27.9	15.8	6.6	75.6	12.1	5.9	77.8	11.0
SYNC [11]	11.5	70.9	19.8	7.4	66.3	13.3	7.9	43.3	13.4	8.9	87.3	16.2	10.0	90.5	18.0
MLZSL [7, 8]	22.9	60.0	33.2	3.4	77.9	6.5	19.6	45.0	27.3	16.5	78.6	27.3	15.8	79.2	26.3
SAE [25]	7.8	54.0	13.6	0.4	80.9	0.9	8.8	18.0	11.8	1.8	77.1	3.5	1.1	82.2	2.2
f-CLSWGAN [47]	43.7	57.7	49.7	-	-	-	42.6	36.6	39.4	57.9	61.4	59.6	-	-	-
RG-GZSL [4]	31.9	-	40.0	-	-	-	32.9	-	25.5	42.2	-	54.1	56.7	-	65.1
GMMN	49.1	55.9	52.3	28.5	64.4	39.5	39.7	37.7	38.7	51.5	70.1	59.3	46.3	77.3	57.9
AE	16.3	31.6	21.5	24.8	38.9	29.6	9.1	22.0	12.9	49.3	52.8	51.0	49.5	63.3	55.7
Ad. AE	17.3	51.9	25.9	28.0	60.4	38.2	11.2	27.1	15.8	56.6	67.0	61.3	50.8	76.6	61.0
AC-GAN	18.3	37.2	24.5	6.9	75.7	12.6	22.9	20.6	21.7	0.45	81.9	0.9	2.4	89.8	4.7

dict context from words. The neural network has 1 input layer, 1 hidden layer and 1 output layer having the size of the vocabulary (same size as the input layer). The hidden layer has 500 neurons in our implementation. In the literature, the hidden layer has been reported to be an interesting embedding space for representing word. Consequently, We use this hidden layer to describe each class label by embedding the class name into this 500-dimensional space. Some classes cannot be represented as their name is not contained in the vocabulary established by parsing the Wikipedia corpus. Such classes are ignored, bringing the number of classes from 20,842 to 20,345 classes. For fair comparison, we take the same language model as [11] with the same classes excluded.

As in [11, 18] our model is evaluated on three different scenarios, with an increasing number of unseen classes: i) 2-hop: 1,509 classes ii) 3-hop: 7,678 classes, iii) All: all unseen categories.

For this task we use the Flat-Hit@K metric, the percentage of test images for which the model returns the true labels in the top K prediction scores.

Table 2 summarizes the performance on the 3 hops. As one can see, our model gets state-of the art performance for each configuration. As it can be observed from these experiments, our generative model is very suitable for this large scale GZSC problem, our approach improves by 5% best competitors for the Flat-Hit 1 metric on the 2-hop sce-

nario.

3.5 Classical Zero-Shot Classification task

In this last section, we follow the protocol of the standard ZSC task: during training, only data from seen classes are available while at test time new images (from unseen classes only) have to be assigned to one of the unseen classes.

As explained in the introduction, the recent ZSC literature ([3, 8, 7, 38], etc) mostly focuses on developing a good embedding for comparing attributes and images. One of our motivations for generating training images was to make the training of discriminative classifiers possible, assuming it would result in better performance. This section aims at validating this hypothesis on the regular ZSC task.

Table 3 summarizes our experiments, reporting the accuracy obtained by state of the art methods on the 5 ZSC datasets, with image features extracted from the 101-layered ResNet [22] network. Each entry is the mean accuracy computed on 5 different runs. Our method exhibits performances above state-of-the-art on each dataset (compared to embedding methods), with a noticeable mean improvement of more than 5%.

4 Conclusions

This paper introduces a novel way to address Zero-Shot Classification and Generalized Zero-Shot Classification

Table 3: Zero-shot classification accuracy. We report results image features extracted from the 101-layered ResNet [22] network.

Method	Standard Zero-Shot Learning					Avg.
	CUB	P&Y	SUN	AWA1	AWA2	
DAP [26]	40.0	33.8	39.9	44.1	46.1	40.8
IAP [26]	24.0	36.6	19.4	35.9	35.9	30.4
CONSE [35]	34.3	26.9	38.8	45.6	44.5	38.0
CMT [40]	34.6	28.0	39.9	39.5	37.9	36.0
SSE [49]	43.9	34.0	51.5	60.1	61.0	50.1
LATEM [46]	49.3	35.2	55.3	55.1	55.8	50.1
ALE [2]	54.9	39.7	58.1	59.9	62.5	55.0
DEVISE [18]	52.0	39.8	56.5	54.2	59.7	52.4
SJE [3]	53.9	32.9	53.7	65.6	61.9	53.6
ESZSL [38]	53.9	38.3	54.5	58.2	58.6	52.7
SYNC [11]	55.6	23.9	56.3	54.0	46.6	47.3
MLZSL [7, 8]	53.5	39.5	57.6	58.1	61.1	54.0
SAE [25]	33.3	8.3	40.3	53.0	54.1	37.8
f-CLSWGAN [47]	57.3	-	60.8	68.2	-	-
RG-GZSL [4]	59.6	-	63.3	69.5	69.2	-
GMMN	59.4	41.7	60.1	70.7	69.9	60.4
AE	53.7	33.9	54.2	69.4	69.4	56.1
Ad. AE	53.4	37.9	58.5	68.2	70.3	57.7
AC-GAN	38.7	38.1	47.2	59.4	57.7	48.2

tasks by learning a conditional generator from seen data and generating artificial training examples for the categories without exemplars, turning ZSC into a standard supervised learning problem. This novel formulation addresses the two main limitation of previous ZSC method, their intrinsic bias for Generalized Zero-Shot Classification tasks and their limitations in using discriminative classifiers in the deep image feature space. Our experiments with 4 generative models and 6 datasets experimentally validate the approach and give state-of-the-art performance.

Acknowledgements

Maxime Bucher is supported by a scholarship from the Région Normandie and from ONERA.

References

- [1] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue Zero-Shot Learning with Strong Supervision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 59–68. IEEE, June 2016.
- [2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-Embedding for Image Classification. *IEEE PAMI*, 2015.
- [3] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of Output Embeddings for Fine-Grained Image Classification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [4] Gundeep Arora, Vinay Kumar Verma, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.
- [5] Mohamed Bahy Bader-El-Den, Eleman Teitei, and Mo Adda. Hierarchical classification for dealing with the Class imbalance problem. *IJCNN*, 2016.
- [6] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013.
- [7] M Bucher, S Herbin, and F Jurie. Improving semantic embedding consistency by metric learning for zero-shot classification. In *European Conference on Computer Vision*, 2016.
- [8] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Hard negative mining for metric learning based zero-shot classification. In *Computer Vision—ECCV 2016 Workshops*, pages 524–531. Springer, 2016.
- [9] Maxime Bucher, Stéphane Herbin, and Frederic Jurie. Generating visual representations for zero-shot classification. In *International Conference on Computer Vision (ICCV) Workshops*, 2017.
- [10] Maria E Cabrera and Juan P Wachs. Embodied gesture learning from one-shot. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1092–1097. IEEE, 2016.
- [11] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 5327–5336. IEEE, 2016.
- [12] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, pages 52–68. Springer, 2016.
- [13] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode Regularized Generative Adversarial Networks. *arXiv*, December 2016.
- [14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [15] Christian Eggert, Anton Winschel, and Rainer Lienhart. On the Benefit of Synthetic Data for Company Logo Detection. In *ACM Multimedia*, 2015.
- [16] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [17] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [18] Andrea Frome, Gregory S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *Conference on Neural Information Processing Systems (NIPS)*, 2013.
- [19] I Goodfellow, J Pouget-Abadie, and M Mirza. Generative adversarial nets. In *NIPS*, 2014.

- [20] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.
- [21] Hongyu Guo and Herna L Viktor. Learning from imbalanced data sets with boosting and data generation - the DataBoost-IM approach. *SIGKDD Explorations*, 2004.
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [23] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading Text in the Wild with Convolutional Neural Networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.
- [24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [25] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017.
- [26] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Trans Pattern Anal Mach Intell*, 36(3):453–465, 2014.
- [27] Y LeCun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.
- [28] Yujia Li, Kevin Swersky, and Richard S Zemel. Generative moment matching networks. In *ICML*, pages 1718–1727, 2015.
- [29] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. Citeseer, 2013.
- [30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.
- [31] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.
- [32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [33] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv*, November 2014.
- [34] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.
- [35] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*, December 2013.
- [36] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.
- [37] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012.
- [38] Bernardino Romera-Paredes and Philip HS Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015.
- [39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [40] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In *Conference on Neural Information Processing Systems (NIPS)*, 2013.
- [41] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [43] Ilkay Ulusoy and Christopher M Bishop. Generative versus Discriminative Methods for Object Recognition. In *CVPR*, 2005.
- [44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [45] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 2764–2770. AAAI Press, 2011.
- [46] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.
- [47] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.
- [48] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [49] Ziming Zhang and Venkatesh Saligrama. Zero-Shot Learning via Semantic Similarity Embedding. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.