`f-VAEGAN-D2`: A Feature Generating Framework for Any-Shot Learning

Yongqin Xian[1] Saurabh Sharma[1] Bernt Schiele[1] Zeynep Akata[1,2]

[1]Max Planck Institute for Informatics
Saarland Informatics Campus

[2]Amsterdam Machine Learning Lab
University of Amsterdam

Abstract

When labeled training data is scarce, a promising data augmentation approach is to generate visual features of unknown classes using their attributes. To learn the class conditional distribution of CNN features, these models rely on pairs of image features and class attributes. Hence, they can not make use of the abundance of unlabeled data samples. In this paper, we tackle any-shot learning problems i.e. zero-shot and few-shot, in a unified feature generating framework that operates in both inductive and transductive learning settings. We develop a conditional generative model that combines the strength of VAE and GANs and in addition, via an unconditional discriminator, learns the marginal feature distribution of unlabeled images. We empirically show that our model learns highly discriminative CNN features for CUB and FLO datasets, and establish a new state-of-the-art in any-shot learning, i.e. inductive and transductive generalized zero- and few-shot learning settings.

1. Introduction

Learning with limited labels has been an important topic of research as it is unrealistic to collect sufficient amounts of labeled data for every object. Recently, generating visual features of previously unseen classes [14, 3, 6, 4] has shown its potential to perform well on extremely imbalanced image collections. Our main focus in this work is a new model that generates visual features of any class, utilizing labeled samples when they are available and generalizing to unknown concepts whose labeled samples are not available. Prior work used GANs for this task [14, 4] as they directly optimize the divergence between real and generated data, but they suffer from mode collapse issues [2]. On the other hand, feature generation with VAE [6] is more stable. However, VAE optimizes the lower bound of log likelihood rather than the likelihood itself [5]. Our model combines the strengths of VAE and GANs by assembling them to a conditional feature generating model, called `f-VAEGAN-D2`, that synthesizes CNN image features from class embeddings, i.e. class-level attributes or word2vec [8]. Thanks to its additional discriminator that distinguishes real and generated features, our `f-VAEGAN-D2` is able to use unlabeled data from previously unseen classes without any condition. The features learned by our model are disciminative in that they boost the performance of any-shot learning as well as being visually and textually interpretable.

2. `f-VAEGAN-D2` Model

As shown in Figure 1, we proposes to enhance the feature generator by combining VAE and GANs with shared decoder and generator, and adding another discriminator (D_2) to distinguish real or generated features without applying any condition.

Setup. We are given a set of images $X = \{x_1, \ldots, x_l\} \cup \{x_{l+1}, \ldots, x_t\}$ encoded in the image feature space $\mathcal{X}$, a seen class label set Y^s, a novel label set Y^n, a.k.a unseen class label set Y^u in the zero-shot learning literature. The set of class embeddings $C = \{c(y) | \forall y \in Y^s \cup Y^n\}$ are encoded in the semantic embedding space $\mathcal{C}$ that defines high level semantic relationships between classes. The first l points $x_s (s \leq l)$ are labeled as one of the seen classes $y_s \in Y^s$ and the remaining points $x_n (l + 1 \leq n \leq t)$ are unlabeled, i.e. may come from seen or novel classes. In the inductive setting, the training set contains only labeled samples of seen class images, i.e. $\{x_1, \ldots, x_l\}$. On the other hand, in the transductive setting, the training set contains both labeled and unlabeled samples, i.e. $\{x_1, \ldots, x_l, x_{l+1}, \ldots, x_t\}$. In the generalized zero-shot learning, the goal is to classify those unlabeled points that can be either from seen or novel classes. Generalized few-shot learning is defined similarly when there are additional samples from novel classes available.

Our framework can be thought of as a data augmentation scheme where arbitrarily many synthetic features of sparsely populated classes aid in improving the disciminative power of classifiers. In the following, we only detail our feature generating network structure as the classifier is

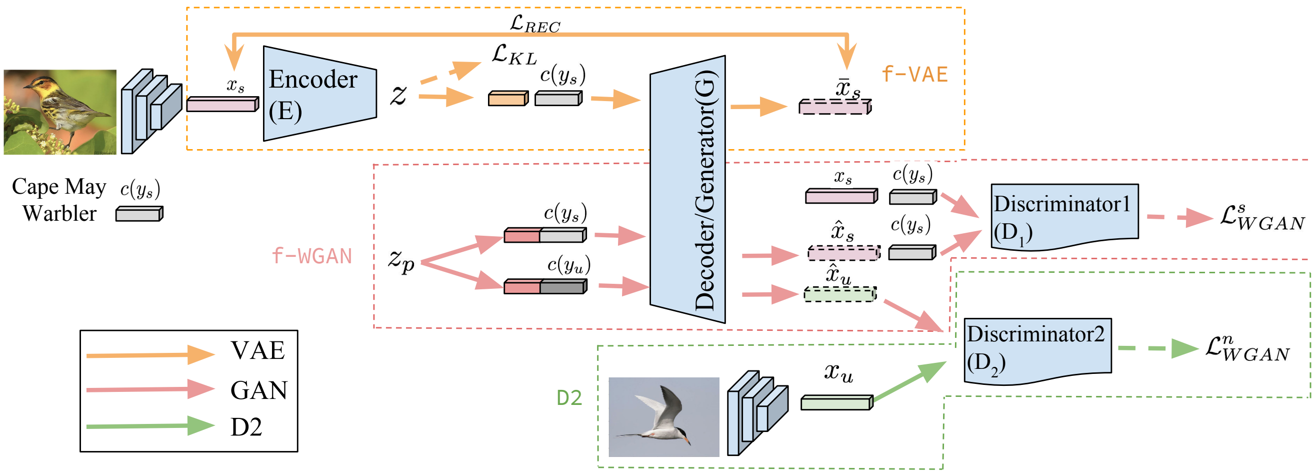

Figure 1: Our any-shot feature generating network (f-VAEGAN-D2) consist of a feature generating VAE (f-VAE), a feature generating WGAN (f-WGAN) with a conditional discriminator (D_1) and a transductive feature generator with a non-conditional discriminator (D_2) that learns from both labeled data of seen classes and unlabeled data of novel classes.

unconstrained (we use linear softmax classifiers).

2.1. Baseline Feature Generating Models

In feature generating networks (f-WGAN) [14] the generator $G(z,c)$ generates a CNN feature $\hat{x}$ from random noise z_p and a condition c, and the discriminator $D(x,c)$ takes as input a pair of input features x and a condition c and outputs a real value, optimizing:

$$\mathcal{L}^s_{WGAN} = \mathbb{E}[D(x,c)] - \mathbb{E}[D(\tilde{x},c)] - \lambda\mathbb{E}[(||\nabla_{\hat{x}} D(\hat{x},c)||_2 - 1)^2], \tag{1}$$

The feature generating VAE [5] (f-VAE) consists of an encoder $E(x,c)$, which encodes an input feature x and a condition c to a latent variable z, and a decoder $Dec(z,c)$, which reconstructs the input x from the latent z and condition c optimizing:

$$\mathcal{L}^s_{VAE} = KL(q(z|x,c)||p(z|c)) - \mathbb{E}_{q(z|x,c)}[\log p(x|z,c)], \tag{2}$$

2.2. Our f-VAEGAN-D2 Model

It has been shown that ensembling a VAE and a GAN leads to better image generation results [7]. We hypothesize that VAE and GAN learn complementary information for feature generation as well. This is likely when the target data follows a complicated multi-modal distribution where two losses are able to capture different modes of the data.

To combine f-VAE and f-WGAN, we introduce an encoder $E(x,c): \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{Z}$, which encodes a pair of feature and class embedding to a latent representation, and a discriminator $D_1: \mathcal{X} \times \mathcal{C} \rightarrow \mathbb{R}$ maps this embedding pair to a compatibility score, optimizing:

$$\mathcal{L}^s_{VAEGAN} = \mathcal{L}^s_{VAE} + \gamma\mathcal{L}^s_{WGAN} \tag{3}$$

where the generator $G(z,c)$ of the GAN and decoder $Dec(z,c)$ of the VAE share the same parameters. The superscript s indicates that the loss is applied to feature and class embedding pair of seen classes. γ is a hyperparameter to control the weighting of VAE and GAN losses.

Furthermore, when unlabeled data of novel classes becomes available, we propose to add a non-conditional discriminator D_2 (D2 in f-VAEGAN-D2) which distinguishes between real and generated features of novel classes. This way D_2 learns the feature manifold of novel classes. Formally, our additional non-conditional discriminator $D_2: \mathcal{X} \rightarrow R$ distinguishes real and synthetic unlabeled samples using a WGAN loss:

$$\mathcal{L}^n_{WGAN} = \mathbb{E}[D_2(x_n)] - \mathbb{E}[D_2(\tilde{x}_n)] - \lambda\mathbb{E}[(||\nabla_{\hat{x}_n} D_2(\hat{x}_n)||_2 - 1)^2], \tag{4}$$

where $\tilde{x}_n = G(z, y_n)$ with $y_n \in Y^n$, $\hat{x}_n = \alpha x_n + (1-\alpha x_n)$ with $\alpha \sim U(0,1)$. Since $\mathcal{L}^s_{WGAN}$ is trained to learn CNN features using labeled data conditioned on class embeddings of seen classes and class embeddings encode shared properties across classes, we expect these CNN features to be transferable across seen and novel classes. However, this heavily relies on the quality of semantic embeddings and suffers from domain shift problems. Intuitively, $\mathcal{L}^n_{WGAN}$ captures the marginal distribution of CNN features and provides useful signals of novel classes to generate transferable CNN features. Hence, our unified f-VAEGAN-D2 model optimizes the following objective function:

$$\min_{G,E} \max_{D_1,D_2} \mathcal{L}^s_{VAEGAN} + \mathcal{L}^n_{WGAN} \tag{5}$$

	Method	CUB u	CUB s	CUB H	FLO u	FLO s	FLO H
IND	ALE [1]	23.7	62.8	34.4	13.3	61.6	21.9
	CLSWGAN [14]	43.7	57.7	49.7	59.0	73.8	65.6
	Cycle-CLSWGAN [4]	47.9	59.3	53.0	61.6	69.2	65.2
	Ours	48.4	60.1	53.6	56.8	74.9	64.6
	Ours-finetuned	**63.2**	**75.6**	**68.9**	**63.3**	**92.4**	**75.1**
TRAN	ALE-tran [13]	23.5	45.1	30.9	13.6	61.4	22.2
	GFZSL [11]	24.9	45.8	32.2	21.8	75.0	33.8
	DSRL [15]	17.3	39.0	24.0	26.9	64.3	37.9
	UE-finetune [10]	74.9	71.5	73.2	-	-	-
	Ours	61.4	65.1	63.2	78.7	87.2	82.7
	Ours-finetuned	**73.8**	**81.4**	**77.3**	**91.0**	**97.4**	**94.1**

Table 1: Comparing with the-state-of-the-art. Top: inductive methods (IND), Bottom: transductive methods (TRAN). Fine tuning is performed only on seen class images as this does not violate the zero-shot condition. We measure Top-1 accuracy on seen (**s**) and unseen (**s**) classes as well as their harmonic mean (**H**) in GZSL setting.

3. Experiments

Generalized Zero-shot Learning We validate our model on two widely-used datasets for zero-shot learning, i.e. Caltech-UCSD-Birds (CUB) [12] and Oxford Flowers (FLO) [9]. We follow the exact class splits as well as the evaluation protocol of [13] and for fair comparison we use the same image and class embeddings for all models.

In Table 1 we compare our model with the best performing recent methods on two zero-shot learning datasets in GZSL setting. We observe that feature generating methods, i.e. our model, CLSWGAN [14], Cycle-CLSWGAN [4] achieve better results than others. This is due to the fact that data augmentation through feature generation leads to a more balanced data distribution such that the learned classifier is not biased to seen classes.Note that although UE [10] is not a feature generating method, it leads to strong results as this model uses additional information, i.e. it assumes that unlabeled test samples always come from unseen classes. Our model with fine-tuning leads to 77.3% harmonic mean (H) on CUB, 94.1% H on FLO, achieving significantly higher results than all the prior works.

3.1. Generalized Few-shot Learning

In few-shot or low-shot learning scenarios, classes are divided into base classes that have a large number of labeled training samples and novel classes that contain only few labeled samples per category. We use the class splits from the standard ZSL setting, i.e. 150 base and 50 novel. For FLO we also follow the same class splits as in ZSL.

As shown in Figure 3 for both datasets both our inductive and transductive models have a significant edge over all

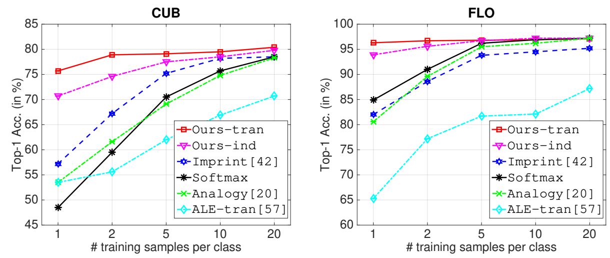

Figure 2: Generalized Few-Shot Learning (GFSL)

Figure 3: GFSL results on CUB and FLO with increasing number of training samples per novel class. Plots show the top-1 accuracy on all classes.

the competing methods when the number of samples from novel classes is small, e.g. 1,2 and 5. This shows that our model generates highly discriminative features even with only few real samples are present. In fact, only with one real sample per class, our model achieves almost the full accuracy obtained with 20 samples per class. Going towards the full supervised learning, e.g. with 10 or 20 samples per class, all methods perform similarly. This is expected since in the setting where a large number of labeled samples per class is available, then a simple softmax classifier that uses real ResNet-101 features achieves the state-of-the-art.

In inductive GFSL setting, our model with two samples per class achieves the same accuracy as softmax trained with ten samples per class on CUB. In the transductive GFSL setting, for FLO, for our model only one labeled sample is enough to reach the accuracy obtained with 20 labeled samples with softmax.

4. Conclusion

In this work, we develop a transductive feature generating framework that synthesizes CNN image features from a class embedding. Our generated features circumvent the scarceness of the labeled training data issues and allow us to effectively train softmax classifiers. Our framework combines conditional VAE and GAN architectures to obtain a more robust generative model. We further improve VAE-GAN by adding a non-conditional discriminator that handles unlabeled data from unseen classes. The second discriminator learns the manifold of unseen classes and backpropagates the WGAN loss to feature generator such that it generalizes better to generate CNN image features for unseen classes. Our feature generating framework is effective across generalized zero-shot (GZSL), and generalized few-shot learning (GFSL) tasks on CUB and FLO datasets.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *TPAMI*, 2016.

[2] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *ICLR*, 2017.

[3] M. Bucher, S. Herbin, and F. Jurie. Generating visual representations for zero-shot classification. *ICCV Workshop*, 2017.

[4] R. Felix, V. K. B. G, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018.

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

[6] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, 2018.

[7] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[9] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICCVGI*, 2008.

[10] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song. Transductive unbiased embedding for zero-shot learning. In *CVPR*, 2018.

[11] V. K. Verma and P. Rai. A simple exponential family framework for zero-shot learning. In *ECML*, 2017.

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

[13] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018.

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

[15] M. Ye and Y. Guo. Zero-shot classification with discriminative semantic representation learning. In *CVPR*, 2017.