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.
unconstrained (we use linear softmax classifiers).

2.1. Baseline Feature Generating Models

In feature generating networks \( \mathcal{f} \)-\( \mathcal{WGAN} \) [14] the generator \( G(z, c) \) generates a CNN feature \( \hat{x} \) from random noise \( z \) 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:

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

The feature generating VAE [5] \( \mathcal{f} \)-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:

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

2.2. Our \( \mathcal{f} \)-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 \( \mathcal{f} \)-VAE and \( \mathcal{f} \)-\( \mathcal{WGAN} \), 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:

\[
\mathcal{L}^s_{\mathcal{VAEGAN}} = \mathcal{L}^s_{\mathcal{VAE}} + \gamma \mathcal{L}_{\mathcal{WGAN}} \tag{3}
\]

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. \( \gamma \) 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 \( D_2 \) in \( \mathcal{f} \)-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 \( D_2 : \mathcal{X} \rightarrow \mathbb{R} \) distinguishes real and synthetic unlabeled samples using a WGAN loss:

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

where \( \tilde{x}_n = G(z, y_n) \) with \( y_n \in Y^n \), \( \tilde{x}_n = \alpha x_n + (1 - \alpha) x_n \) with \( \alpha \sim U(0, 1) \). Since \( \mathcal{L}_{\mathcal{WGAN}} \) 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, \( \mathcal{L}_{\mathcal{WGAN}} \) captures the marginal distribution of CNN features and provides useful signals of novel classes to generate transferable CNN features. Hence, our unified \( \mathcal{f} \)-VAEGAN-D2 model optimizes the following objective function:

\[
\min_{G,E,D_1,D_2} \max_{\mathcal{D}_1,\mathcal{D}_2} \mathcal{L}^s_{\mathcal{VAEGAN}} + \mathcal{L}_{\mathcal{WGAN}} \tag{5}
\]
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 (u) 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 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 GSFL 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 GSFL 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 back-propagates 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 (GSFL) tasks on CUB and FLO datasets.
References