Ensembling Off-the-shelf Models for GAN Training

Nupur Kumari [1] Richard Zhang[2] Eli Shechtman[2] Jun-Yan Zhu[1]
[1]Carnegie Mellon University [2]Adobe

Abstract

The advent of large-scale training has produced a cornucopia of powerful visual recognition models. However, generative models, such as GANs, have traditionally been trained from scratch in an unsupervised manner. Can the collective "knowledge" from a large bank of pretrained vision models be leveraged to improve GAN training? If so, with so many models to choose from, which one(s) should be selected, and in what manner are they most effective? We find that pretrained computer vision models can significantly improve performance when used in an ensemble of discriminators. Notably, the particular subset of selected models greatly affects performance. We propose an effective selection mechanism, by probing the linear separability between real and fake samples in pretrained model embeddings, choosing the most accurate model, and progressively adding it to the discriminator ensemble. Interestingly, our method can improve GAN training in both limited data and large-scale settings. Given only 10k training samples, our FID *on* LSUN CAT *matches the StyleGAN2 trained on 1.6M images. On the full dataset, our method improves* FID *by* 1.5 *to* 2× *on cat, church, and horse categories of* LSUN.

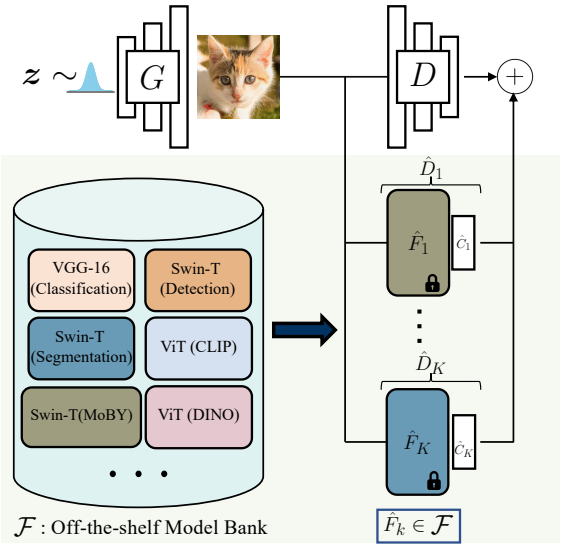

Figure 1. **Vision-aided-GAN training**. The model bank $\mathcal{F}$ consists of widely used and state-of-the-art pretrained networks. We automatically select a subset $\{\hat{F}\}_{k=1}^{K}$ from $\mathcal{F}$, which can best distinguish between real and fake distribution. Our training procedure consists of creating an ensemble of the original discriminator D and discriminators $\hat{D}_k = \hat{C}_k \circ \hat{F}_k$ based on the feature space of selected off-the-shelf models. $\hat{C}_k$ is a shallow trainable network over the frozen pretrained features.

1. Introduction

Image generation inherently requires being able to capture and model complex statistics in real-world visual phenomenon. Computer vision models, driven by the success of supervised and self-supervised learning techniques [15, 17, 30, 62, 74], have proven effective at capturing useful representations when trained on large-scale data [65, 88, 100]. What potential implications does this have on generative modeling? If one day, perfect computer vision systems could answer any question about any image, could this capability be leveraged to improve image synthesis models?

Surprisingly, despite the aforementioned connection between synthesis and analysis, current state-of-the-art generative adversarial networks (GANs) [9, 37, 38, 98] are trained in an unsupervised manner without the aid of such pretrained networks. With a plethora of useful models easily available in the research ecosystem, this presents a missed opportunity to explore. Can the knowledge of pretrained visual representations actually benefit GAN training? If so, with so many models, tasks, and datasets to choose from, which models should be used, and in what manner are they most effective?

In this work, we study the use of a "bank" of pretrained deep feature extractors to aid in generative model training. Specifically, GANs are trained with a discriminator, aimed at continuously learning the relevant statistics differentiating real and generated samples, and a generator, which aims to reduce this gap. Naïvely using such strong, pretrained networks as a discriminator leads to the overfitting and overwhelming the generator, especially in limited data settings. We show that freezing the pretrained network (with a small, lightweight learned classifier on top as shown in Figure 1) provides stable training when used with the original, learned discriminator. In addition, ensembling multiple pretrained networks encourages the generator to match the real distribution in different, complementary feature spaces.