Adversarial Feature Learning

Jeff Donahue
http://jeffdonahue.com
Philipp Krähenbühl
http://philkr.net
Trevor Darrell
https://people.eecs.berkeley.edu/~trevor/

Computer Science Division,
UC Berkeley

Abstract

The ability of the Generative Adversarial Networks (GANs) framework to learn generative models mapping from simple latent distributions to arbitrarily complex data distributions has been demonstrated empirically, with compelling results showing generators learn to "linearize semantics" in the latent space of such models. Intuitively, such latent spaces may serve as useful feature representations for auxiliary problems where semantics are relevant. However, in their existing form, GANs have no means of learning the inverse mapping – projecting data back into the latent space. We propose Bidirectional Generative Adversarial Networks (BiGANs) as a means of learning this inverse mapping, and demonstrate that the resulting learned feature representation is useful for auxiliary supervised discrimination tasks, competitive with contemporary approaches to unsupervised and self-supervised feature learning.

1 Introduction

Deep convolutional networks (convnets) have become a staple of the modern computer vision pipeline. After training these models on a massive database of image-label pairs like ImageNet [17], the network easily adapts to a variety of similar visual tasks, achieving impressive results on image classification [5, 16, 20] or localization [9, 14] tasks. In other perceptual domains such as natural language processing or speech recognition, deep networks have proven highly effective as well [2, 11, 18]. However, all of these recent results rely on a supervisory signal from large-scale databases of hand-labeled data, ignoring much of the useful information present in the structure of the data itself.

Meanwhile, Generative Adversarial Networks (GANs) [10] have emerged as a powerful framework for learning generative models of arbitrarily complex data distributions. The GAN framework learns a *generator* mapping samples from an arbitrary latent distribution to data, as well as an adversarial *discriminator* which tries to distinguish between real and generated samples as accurately as possible. The generator's goal is to "fool" the discriminator by producing samples which are as close to real data as possible. GANs produce impressive results on databases of natural images [3, 15]. Interpolations in the latent space of the generator produce smooth and plausible semantic variations [15]. Based on these intuitions from observation of qualitative results, it appears that the generator learned by the GAN framework learns to "linearize the semantics" of the data distribution in the latent space.

A natural question arises from this ostensible "semantic juice" flowing through the weights of generators learned using the GAN framework: can GANs be used for unsupervised learning of rich feature representations for arbitrary data distributions? An obvious issue with doing so is that the generator maps latent samples to generated data, but the framework does not include an *inverse* mapping from data to latent representation.

Hence, we propose a novel unsupervised feature learning framework, *Bidirectional Generative Adversarial Networks* (BiGANs). The overall model is depicted in Figure 1. In short, in addition to the generator G and discriminator D from the standard GAN framework [10], we additionally learn an *encoder* E which maps data $\mathbf{x}$ to latent representations $\mathbf{z}$.

BiGANs are a robust and highly generic approach to unsupervised feature learning, making no assumptions about the structure or type of data to which they are applied, as our theoretical results will demonstrate. Our empirical studies of their feature learning abilities will show that despite their generality, BiGANs are competitive with contemporary approaches to unsupervised and weakly supervised feature learning tailor-made for a notoriously complex data distribution – natural images.

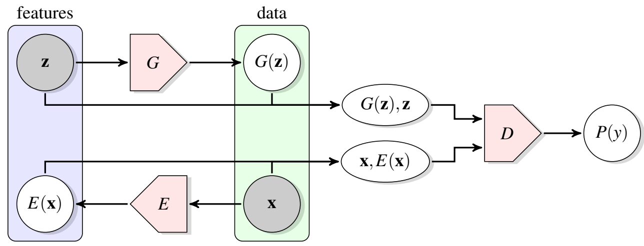

Figure 1: The structure of a Bidirectional Generative Adversarial Network (BiGAN).

Dumoulin *et al.* [6] independently proposed an identical model in their concurrent work, exploring the case of a stochastic encoder E and the ability of such models to learn in a semi-supervised setting.

2 Bidirectional Generative Adversarial Networks

In Bidirectional Generative Adversarial Networks (BiGANs) we not only train a generator, but additionally train an encoder $E : \Omega_{\mathbf{X}} \rightarrow \Omega_{\mathbf{Z}}$. The encoder induces a distribution $p_E(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{z} - E(\mathbf{x}))$ mapping data point $\mathbf{x}$ into the latent feature space of the generative model. The discriminator is also modified to take input from the latent space, predicting $P_D(Y|\mathbf{x},\mathbf{z})$, where $Y = 1$ if $\mathbf{x}$ is real (sampled from the real data distribution $p_{\mathbf{X}}$), and $Y = 0$ if $\mathbf{x}$ is generated (the output of $G(\mathbf{z}), \mathbf{z} \sim p_{\mathbf{Z}}$).

The BiGAN training objective is defined as a minimax objective

$$\min_{G,E} \max_{D} V(D,E,G) \tag{1}$$

where

$$V(D,E,G) = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{X}}}\left[\log D(\mathbf{x},E(\mathbf{x}))\right] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{Z}}}\left[\log\left(1 - D(G(\mathbf{z}),\mathbf{z})\right)\right]. \tag{2}$$

We optimize this minimax objective using the same alternating gradient based optimization as Goodfellow *et al.* [10].

BiGANs share many of the theoretical properties of GANs [10], while additionally guaranteeing that at the global optimum, both G and E are bijective functions and are each other's inverse. BiGANs are also closely related to autoencoders with an ℓ_0 loss function. In particular, the encoder and generator objective given an optimal discriminator $C(E,G) := \max_D V(D,E,G)$ can be rewritten as an ℓ_0 autoencoder loss function

$$C(E,G) = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{X}}}\left[\mathbf{1}_{\left[E(\mathbf{x})\in\hat{\Omega}_{\mathbf{Z}} \wedge G(E(\mathbf{x}))=\mathbf{x}\right]} \log f_{EG}(\mathbf{x},E(\mathbf{x}))\right] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{Z}}}\left[\mathbf{1}_{\left[G(\mathbf{z})\in\hat{\Omega}_{\mathbf{X}} \wedge E(G(\mathbf{z}))=\mathbf{z}\right]} \log\left(1 - f_{EG}(G(\mathbf{z}),\mathbf{z})\right)\right]$$

with $\log f_{EG} \in (-\infty, 0)$ and $\log(1 - f_{EG}) \in (-\infty, 0)$ almost everywhere on both $P_{E\mathbf{X}}$ and $P_{G\mathbf{Z}}$.

Here the indicator function $\mathbf{1}_{[G(E(\mathbf{x}))=\mathbf{x}]}$ is equivalent to an autoencoder with ℓ_0 loss, while the objective further encourages the functions $E(\mathbf{x})$ and $G(\mathbf{z})$ to produce valid outputs in the support of $P_{\mathbf{Z}}$ and $P_{\mathbf{X}}$ respectively. Unlike regular autoencoders, the ℓ_0 loss function does not make any assumptions about the structure or distribution of the data itself; in fact, all the structural properties of BiGAN are learned as part of the discriminator.

3 Evaluation

We evaluate the feature learning capabilities of BiGANs by first training them unsupervised, then transferring the encoder's learned feature representations for use in auxiliary supervised learning tasks. We evaluate

$G(\mathbf{z})$

$\mathbf{x}$

$G(E(\mathbf{x}))$

Figure 2: Qualitative results for ImageNet BiGAN training, including generator samples $G(\mathbf{z})$, real data $\mathbf{x}$, and corresponding reconstructions $G(E(\mathbf{x}))$.

trained layers	Classification (% mAP) fc8	fc6-8	all	*FRCN* [8] Detection (% mAP) all	*FCN* [14] Segmentation (% mIU) all
ImageNet [13]	77.0	78.8	78.3	56.8	48.0
Random (k-means) [12]	32.0	39.2	56.6	45.6	32.6
Agrawal *et al.* [1]	31.2	31.0	54.2	43.9	-
Wang & Gupta [19]	27.4	51.4	58.4	44.0	-
Doersch *et al.* [4]	44.7	55.1	65.3	51.1	-
Discriminator (D)	30.7	40.5	56.4	-	-
Latent Regressor (LR)	36.9	47.9	57.1	-	-
Joint LR	37.1	47.9	56.5	-	-
Autoencoder (ℓ_2)	24.8	16.0	53.8	41.9	-
BiGAN (ours)	37.5	48.7	58.9	46.2	34.9
BiGAN, 112×112 E (ours)	40.7	52.3	60.1	-	-

Table 1: Classification and detection results for the PASCAL VOC 2007 [7] test set, and segmentation results on the PASCAL VOC 2012 [7] validation set, under the standard mean average precision (mAP) or mean intersection over union (mIU) metrics for each task. Classification models are trained with various portions of the *AlexNet* [13] model frozen. The *fc8*, *fc6-8*, and *all* column headers signify which layers are "fine-tuned" using the VOC classification supervision.

BiGANs on the high-resolution natural images of ImageNet [17]. GANs trained on ImageNet cannot perfectly reconstruct the data, but often capture some interesting aspects.

In these experiments, each module D, G, and E is a deep convnet. The BiGAN discriminator $D(\mathbf{x},\mathbf{z})$ takes data $\mathbf{x}$ as its initial input, and at each linear layer thereafter, the latent representation $\mathbf{z}$ is transformed using a learned linear transformation to the hidden layer dimension and added to the non-linearity input. In all experiments, the encoder E architecture follows AlexNet [13] through the fifth and last convolution layer (*conv5*). We also experiment with an AlexNet-based discriminator D as a baseline feature learning approach. We set the latent distribution $p_{\mathbf{Z}} = [\mathrm{U}(-1,1)]^{200}$.

Baseline methods Besides the BiGAN framework presented above, we considered alternative approaches to learning feature representations using different GAN variants. The discriminator D in a standard GAN takes data samples $\mathbf{x} \sim p_{\mathbf{X}}$ as input, making its learned intermediate representations natural candidates as feature representations for related tasks. We also consider an alternative encoder training by minimizing a reconstruction loss $\mathcal{L}(\mathbf{z}, E(G(\mathbf{z})))$, after or jointly during a regular GAN training, called latent regressor or joint latent regressor respectively. We use a sigmoid cross entropy loss $\mathcal{L}$.

Qualitative results In Figure 2 we present sample generations $G(\mathbf{z})$, as well as real data samples $\mathbf{x}$ and their BiGAN reconstructions $G(E(\mathbf{x}))$. The reconstructions, while certainly imperfect, demonstrate empirically that the BiGAN encoder E and generator G learn approximate inverse mappings.

VOC classification, detection, and segmentation We evaluate the transferability of BiGAN representations to the PASCAL VOC [7] computer vision benchmark tasks, including classification, object detection, and semantic segmentation. We report results on each of these tasks in Table 1, comparing BiGANs with contemporary approaches to unsupervised [4, 12] and weakly supervised [1, 19] feature learning in the visual domain, as well as the baselines discussed in Section 3. For best results, we also evaluate a BiGAN in which the encoder takes inputs at higher resolution 112×112.

Discussion Despite making no assumptions about the underlying structure of the data, the BiGAN unsupervised feature learning framework offers a representation competitive with existing self-supervised and even weakly supervised feature learning approaches for visual feature learning, while still being a purely generative model with the ability to sample data $\mathbf{x}$ and predict latent representation $\mathbf{z}$. Furthermore, BiGANs outperform the discriminator (D) and latent regressor (LR) baselines discussed in Section 3, confirming our intuition that these approaches may not perform well in the regime of highly complex data distributions such as that of natural images. We finally note that the results presented here constitute only a preliminary exploration of the space of model architectures possible under the BiGAN framework, and we expect results to improve significantly with advancements in generative image models and discriminative convolutional networks alike.

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *ICCV*, 2015.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[3] Emily L. Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a Laplacian pyramid of adversarial networks. In *NIPS*, 2015.

[4] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[5] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[6] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv:1606.00704*, 2016.

[7] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes challenge: A retrospective. *IJCV*, 2014.

[8] Ross Girshick. Fast R-CNN. In *ICCV*, 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[11] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, 2013.

[12] Philipp Krähenbühl, Carl Doersch, Jeff Donahue, and Trevor Darrell. Data-dependent initializations of convolutional neural networks. In *ICLR*, 2016.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[16] Ali Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *CVPR Workshops*, 2014.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and F. Li. ImageNet large scale visual recognition challenge. *IJCV*, 2015.

[18] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[19] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015.

[20] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.