GAN survey 2

PRESENTED BY 薛铭龙


ICLR
2017
PART
ONE

1 LR-GAN: Layered Recursive Generative Adversarial Networks for Image Generation

Jianwei Yang*
Virginia Tech
Blacksburg, VA
jw2yang@vt.edu

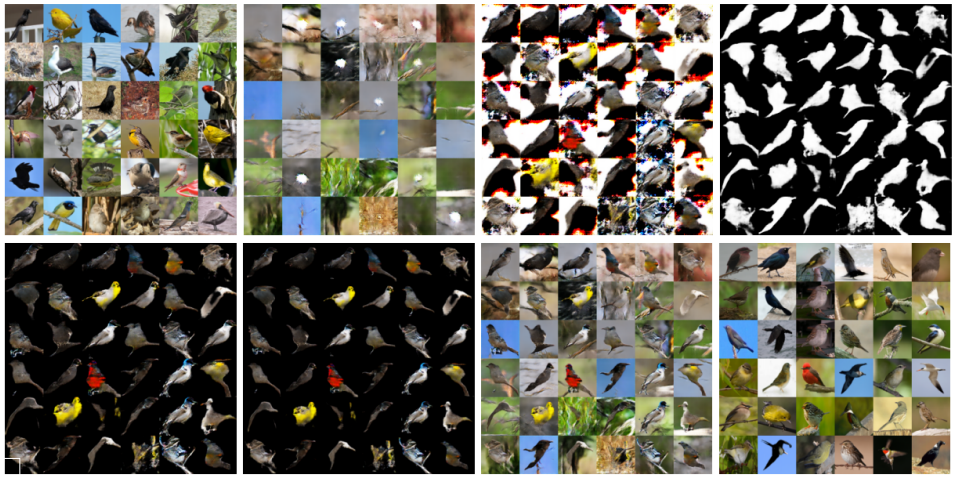
Anitha Kannan
Facebook AI Research
Menlo Park, CA
akannan@fb.com

Dhruv Batra* and Devi Parikh*
Georgia Institute of Technology
Atlanta, GA
{dbatra, parikh}@gatech.edu

主要内容

网络结构及实验结果

- **Problem** : While the holistic ‘gist’ of images generated by these approaches is beginning to look natural, there is clearly a long way to go. For instance, the foreground objects in these images tend to be deformed,blended into the background, and not look realistic or recognizable. One fundamental limitation of these methods is that they attempt to generate images without taking into account that images are 2D projections of a 3D visual world, which has a lot of structures in it. This manifests as structure in the 2D images that capture this world.



- **Method** : The key innovation of our work is the layered recursive generator, The generator in LR-GAN is recursive in that the image is constructed recursively using a recurrent network.

$$\boldsymbol{x}_t = \underbrace{ST(\boldsymbol{m}_t, \boldsymbol{a}_t)}_{\text{affine transformed mask}} \odot \underbrace{ST(\boldsymbol{f}_t, \boldsymbol{a}_t)}_{\text{affine transformed appearance}} + \underbrace{(1 - ST(\boldsymbol{m}_t, \boldsymbol{a}_t)) \odot \boldsymbol{x}_{t-1}}_{\text{pasting on image composed so far}}, \quad \forall t \in [1, T] \quad (4)$$

where $ST(\diamond, \boldsymbol{a}_t)$ is a spatial transformation operator that outputs the affine transformed version of $\diamond$ with $\boldsymbol{a}_t$ indicating parameters of the affine transformation.

Image = foreground(appearance f, shape m, pose a) + background

LR-GAN: Layered Recursive Generative Adversarial Networks for Image Generation

Jianwei Yang*
Virginia Tech
Blacksburg, VA
jw2yang@vt.edu

Anitha Kannan
Facebook AI Research
Menlo Park, CA
akannan@fb.com

Dhruv Batra* and Devi Parikh*
Georgia Institute of Technology
Atlanta, GA
{dbatra, parikh}@gatech.edu

网络结构及实验结果

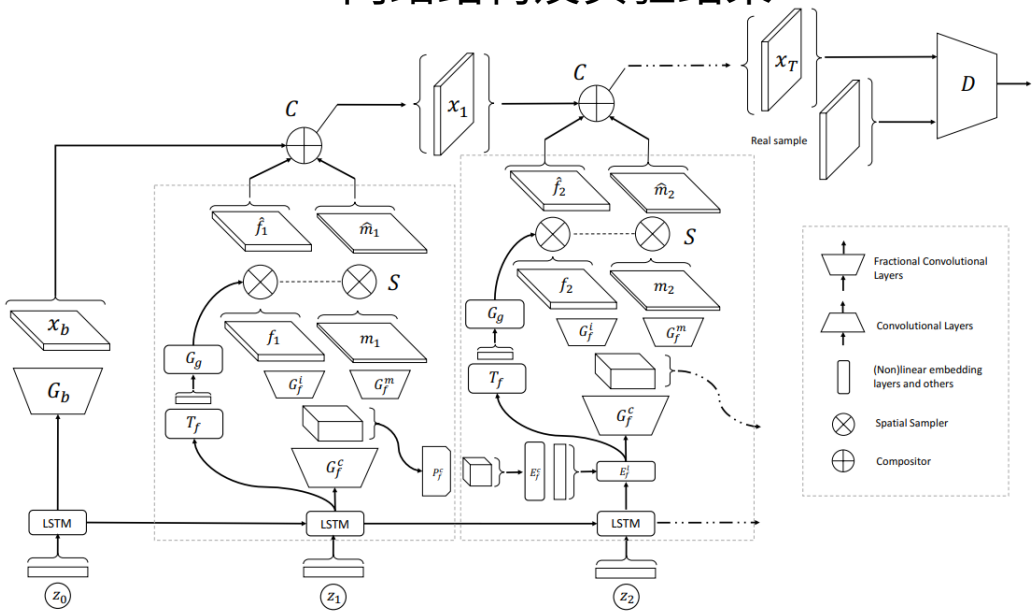


Figure 2: LR-GAN architecture unfolded to three timesteps. It mainly consists of one background generator, one foreground generator, temporal connections and one discriminator. The meaning of each component is explained in the legend.

$$\boldsymbol{x}_t = \underbrace{ST(\boldsymbol{m}_t, \boldsymbol{a}_t)}_{\text{affine transformed mask}} \odot \underbrace{ST(\boldsymbol{f}_t, \boldsymbol{a}_t)}_{\text{affine transformed appearance}} + \underbrace{(1 - ST(\boldsymbol{m}_t, \boldsymbol{a}_t)) \odot \boldsymbol{x}_{t-1}}_{\text{pasting on image composed so far}}, \quad \forall t \in [1, T] \quad (4)$$

where $ST(\diamond, \boldsymbol{a}_t)$ is a spatial transformation operator that outputs the affine transformed version of $\diamond$ with $\boldsymbol{a}_t$ indicating parameters of the affine transformation.

Image = foreground(appearance f, shape m, pose a) + background

RenderGAN: Generating Realistic Labeled Data

Leon Sixt, Benjamin Wild, & Tim Landgraf
Fachbereich Mathematik und Informatik
Freie Universität Berlin
Berlin, Germany
{leon.sixt, benjamin.wild, tim.landgraf}@fu-berlin.de

3D

主要内容

- **Problem** : While a GAN implicitly learns a meaningful latent embedding of the data, **there is no simple relationship between the latent dimensions and the labels of interest.** cGANs are an extension of GANs to sample from a conditional distribution given some labels. **However, training cGANs requires a labeled dataset.** Explicitly controlling the relationship between the latent space and generated samples without using labeled data is an open problem. i.e. sampling from p(x, l) without requiring labels for training.
- **Method** : Our contributions are as follows. We present an extension of the GAN framework that allows to sample from the joint distribution of data and labels. The generated samples are nearly indistinguishable from real data for a human observer and can be used to train a DCNN end-to-end to classify real samples. In the RenderGAN framework, we aim to solve the inverse problem to this regression task: generate data given the labels. **This is achieved by embedding a simple 3D model into the generator of a GAN.**

网络结构及实验结果

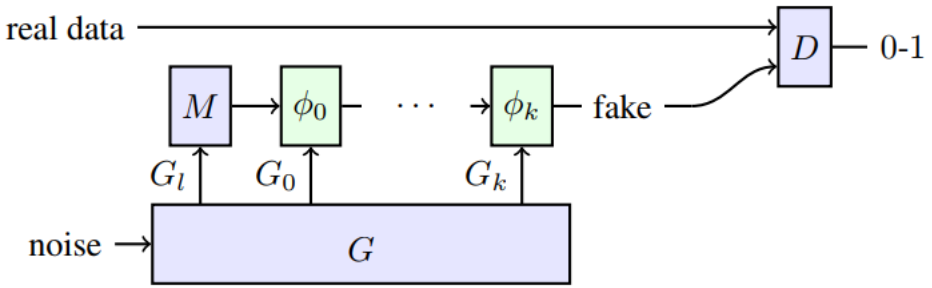
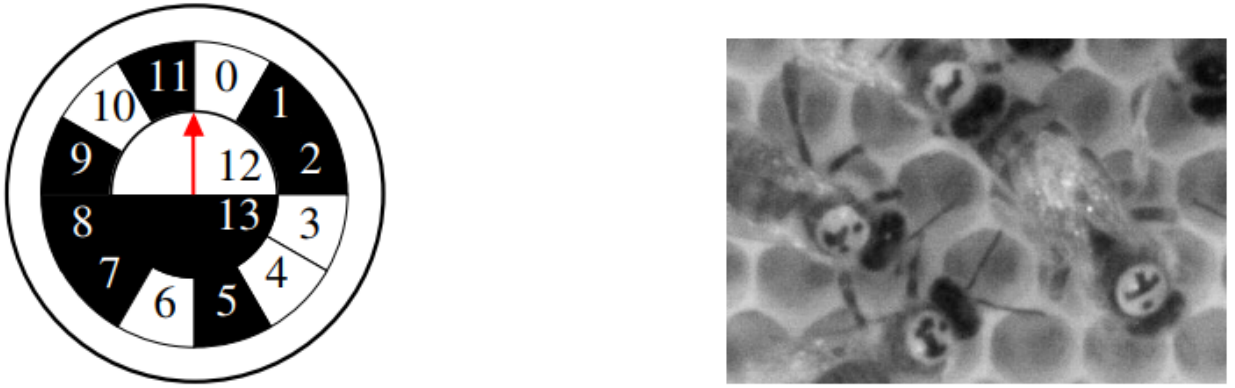


Figure 4: The generator G cannot directly produce samples. Instead, G has to predict parameters G_l for the 3D model M. The image generated by M is then modified through the augmentation functions φi parameterized by Gi to match the real data.

RenderGAN: Generating Realistic Labeled Data

Leon Sixt, Benjamin Wild, & Tim Landgraf
Fachbereich Mathematik und Informatik
Freie Universität Berlin
Berlin, Germany
{leon.sixt, benjamin.wild, tim.landgraf}@fu-berlin.de

Problem



(a) Tag structure

(b) Tagged bees in the hive

Figure 1: **(a)** The tag represents a unique binary code (cell 0 to 11) and encodes the orientation with the semicircles 12 and 13. The red arrow points towards the head of the bee. This tag encodes the id `100110100010`. **(b)** Cutout from a high-resolution image.

The RenderGAN framework was developed to solve the scarcity of labeled data in the BeesBook project (Wario et al., 2015) in which we analyze the social behavior of honeybees. A barcode-like marker is attached to the honeybees' backs for identification (see Fig. 1). Annotating this data is tedious, and therefore only a limited amount of labeled data exists. A 3D model (see the upper row of Fig. 2) generates a simple image of the tag based on position, orientation, and bit configuration. The RenderGAN then learns from unlabeled data to add lighting, background, and image details.

RenderGAN: Generating Realistic Labeled Data

Leon Sixt, Benjamin Wild, & Tim Landgraf
Fachbereich Mathematik und Informatik
Freie Universität Berlin
Berlin, Germany
{leon.sixt, benjamin.wild, tim.landgraf}@fu-berlin.de

3D

主要内容

- **Problem** : While a GAN implicitly learns a meaningful latent embedding of the data, **there is no simple relationship between the latent dimensions and the labels of interest.** cGANs are an extension of GANs to sample from a conditional distribution given some labels. **However, training cGANs requires a labeled dataset.** Explicitly controlling the relationship between the latent space and generated samples without using labeled data is an open problem. i.e. sampling from p(x, l) without requiring labels for training.
- **Method** : We present a novel framework called RenderGAN that can generate large amounts of realistic, labeled images by combining a 3D model and the Generative Adversarial Network framework. In our approach, image augmentations (e.g. lighting, background, and detail) are learned from unlabeled data such that the generated images are strikingly realistic while preserving the labels known from the 3D model.

网络结构及实验结果

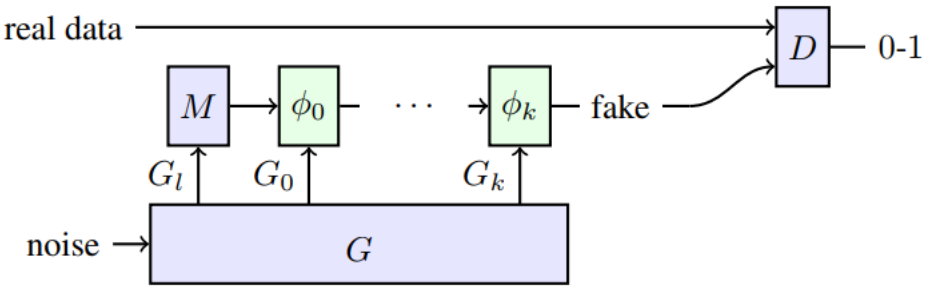


Figure 4: The generator G cannot directly produce samples. Instead, G has to predict parameters G_l for the 3D model M. The image generated by M is then modified through the augmentation functions φi parameterized by Gi to match the real data.

RenderGAN: Generating Realistic Labeled Data

3D

Leon Sixt, Benjamin Wild, & Tim Landgraf
Fachbereich Mathematik und Informatik
Freie Universität Berlin
Berlin, Germany
{leon.sixt, benjamin.wild, tim.landgraf}@fu-berlin.de

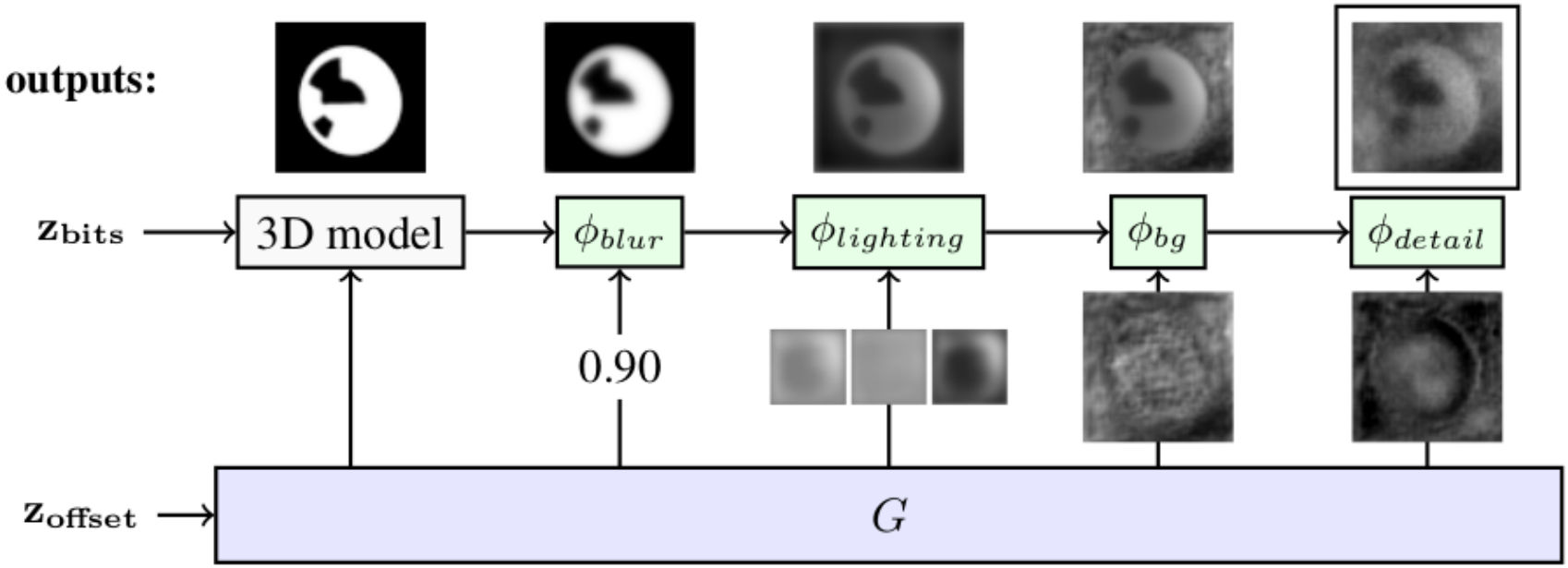


Figure 5: Augmentation functions of the RenderGAN applied to the BeesBook project. The arrows from G to the augmentation functions ϕ depict the inputs to the augmentation functions. The generator provides the position and orientations to the 3D model, whereas the bits are sampled uniformly. On top, the output of each stage is shown. The output of ϕ_{detail} is forwarded to the discriminator.

Published on ICLR in 2017

Towards Principled Methods for Training Generative Adversarial Networks

Week 2

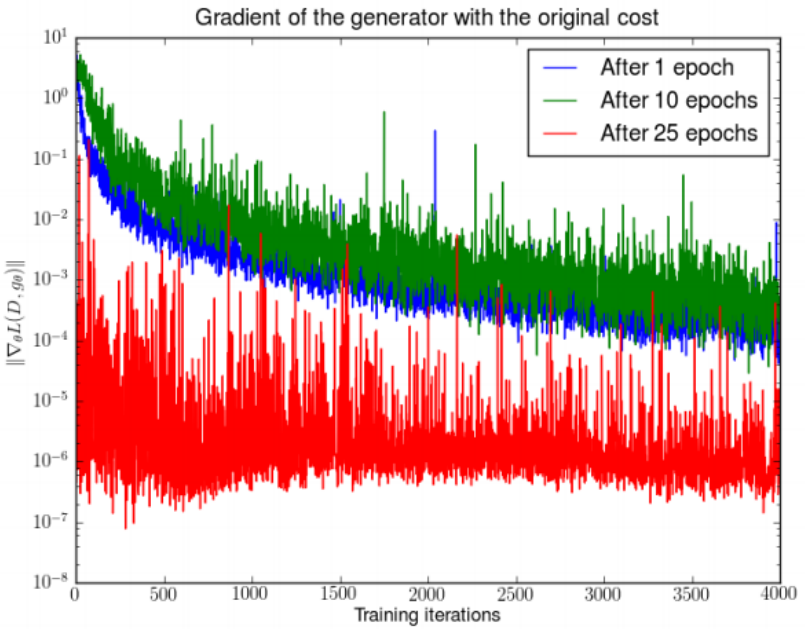
Martin Arjovsky
Courant Institute of Mathematical Sciences
martinarjovsky@gmail.com

Léon Bottou
Facebook AI Research
leonb@fb.com

主要内容

- **Problem** : Despite Generative adversarial networks (GANs) success, there is little to no theory explaining the unstable behaviour of GAN training. It is interesting to note that the architecture of the generator used by GANs doesn’ t differ significantly from other approaches like variational auto encoders. After all, at the core of it we first sample from a simple prior z ∼ p(z), and then output our final sample gθ(z).
- **Method** : The first section introduces the problem at hand. The second section is dedicated to studying and proving rigorously the problems including instability and saturation that arize when training generative adversarial networks. The third section examines a practical and theoretically grounded direction towards solving these problems, while introducing new tools to study them.

网络结构



First, we trained a DCGAN for 1, 10 and 25 epochs Then, with the generator fixed we train a discrimina from scratch and measure the gradients with the original cost function. We see the gradient norms decay quickly, in the best case 5 orders of magnitu after 4000 discriminator iterations. Note the logarithmic scale.

Improving Generative Adversarial Networks with Denoising Feature Matching

David Warde-Farley & Yoshua Bengio*
Montreal Institute for Learning Algorithms, * CIFAR Senior Fellow
Université de Montréal
Montreal, Quebec, Canada
{david.warde-farley,yoshua.bengio}@umontreal.ca

主要内容

- **Problem** : In practice, GANs are well known for being quite challenging to train effectively. The relative model capacities of the generator and discriminator must be carefully balanced in order for the generator to effectively learn. Compounding the problem is the lack of an **unambiguous and computable convergence criterion**. Nevertheless, particularly when trained on image collections from relatively narrow domains such as bedroom scenes (Yu et al., 2015) and human faces (Liu et al., 2015), GANs have been shown to produce very compelling results.
- **Method** : We propose to augment the generator’s training criterion with a second training objective which guides the generator towards samples more like those in the training set by explicitly modeling the data density in addition to the adversarial discriminator. Rather than deploy a second computationally expensive convolutional network for this task, the additional objective is computed in the space of **features learned by the discriminator**. In that space, we train a denoising auto-encoder, a family of models which is known to estimate the energy gradient of the data on which it is trained.

实验结果

We evaluate the denoising auto-encoder on samples drawn from the generator, and use the “denoised” features as targets – nearby feature configurations which are more likely than those of the generated sample, according to the distribution estimated by the denoiser.

Real data	Ours	GAN Baseline
$26.08 \pm .26$	8.51 ± 0.13	$7.84 \pm .07$

Table 2: Inception scores for models of the unlabeled set of STL-10.

- Higher Inception scores, as well as visual inspection, suggest that the procedure captures class-specific features of the training data in a manner superior to the original adversarial objective alone.

Published on ICLR in 2017

Improving Generative Adversarial Networks with Denoising Feature Matching

David Warde-Farley & Yoshua Bengio*
Montreal Institute for Learning Algorithms, * CIFAR Senior Fellow
Université de Montréal
Montreal, Quebec, Canada
{david.warde-farley,yoshua.bengio}@umontreal.ca

主要内容

- **Conclusion** : We have shown that training a denoising model on high-level discriminator activations in a GAN, and using the denoiser to propose high-level feature targets for the generator, can usefully improve GAN image models. Higher Inception scores, as well as visual inspection, suggest that the procedure captures class-specific features of the training data in a manner superior to the original adversarial objective alone. That being said, we do not believe we are yet making optimal use of the paradigm. The non-stationarity of the feature distribution on which the denoiser is trained could be limiting the ability of the denoiser to obtain a good fit, and the information backpropagated to the generator is always slightly stale. Steps to reduce this non-stationarity may be fruitful; we experimented briefly with historical averaging as explored in Salimans et al. (2016) but did not observe a clear benefit thus far. Structured denoisers, including denoisers that learn an energy function for multiple hidden layers at once, could conceivably aid in obtaining a better fit.

- **Conclusion** : Learning a partially stochastic transition operator rather than a deterministic denoiser could conceivably capture interesting multimodalities that are “blurred” by a unimodal denoising function. Our method is orthogonal and could conceivably be used in combination with several other GAN extensions. For example, methods incorporating an encoder component (Donahue et al., 2016; Dumoulin et al., 2016), various existing conditional architectures (Mirza & Osindero, 2014; Denton et al., 2015; Reed et al., 2016), or the semi-supervised variant employed in Salimans et al. (2016), could all be trained with an additional denoising feature matching objective.
- We have proposed a useful heuristic, but a better theoretical grounding regarding how GANs are trained in practice is a necessary direction for future work, including grounded criteria for assessing mode coverage and mass misassignment, and principled criteria for assessing convergence or performing early stopping.

Published on ICLR in 2017

Energy-based Generative Adversarial Networks

Junbo Zhao, Michael Mathieu and Yann LeCun
Department of Computer Science, New York University
Facebook Artificial Intelligence Research
{jakezhao, mathieu, yann}@cs.nyu.edu

- **Abstract** : We introduce the "Energy-based Generative Adversarial Network" model (EBGAN) which views the discriminator as an energy function that attributes low energies to the regions near the data manifold and higher energies to other regions. Similar to the probabilistic GANs, a generator is seen as being trained to produce contrastive samples with minimal energies, while the discriminator is trained to assign high energies to these generated samples. Viewing the discriminator as an energy function allows to use a wide variety of architectures and loss functionals in addition to the usual binary classifier with logistic output. Among them, we show one instantiation of EBGAN framework as using an auto-encoder architecture, with the energy being the reconstruction error, in place of the discriminator. We show that this form of EBGAN exhibits more **stable** behavior than regular GANs during training. We also show that a single-scale architecture canbe trained to generate high-resolution images.

- Objective Function :

$$\mathcal{L}_D(x, z) = D(x) + [m - D(G(z))]^+$$
$$\mathcal{L}_G(z) = D(G(z))$$

where $[\cdot]^+ = \max(0, \cdot)$.

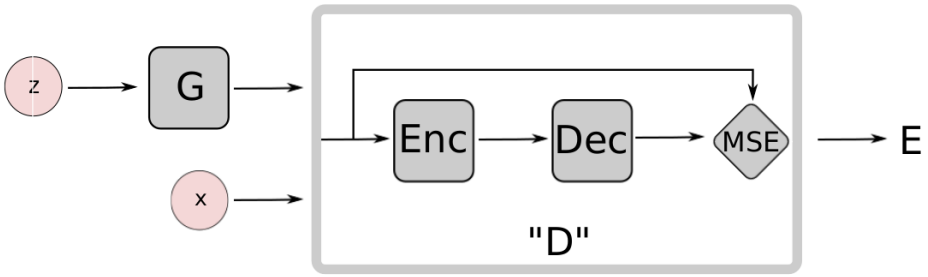


Figure 1: EBGAN architecture with an auto-encoder discriminator.

- Conclusion: EBGANs show better convergence pattern and scalability to generate high-resolution images. A family of energy-based loss functionals presented in LeCun et al. (2006) can easily be incorporated into the EBGAN framework. For the future work, the conditional setting (Denton et al., 2015; Mathieu et al., 2015) is a promising setup to explore. We hope the future research will raise more attention on a broader view of GANs from the energy-based perspective.

Energy-based Generative Adversarial Networks

- **Review** : This paper is a parallel work to Improving Generative Adversarial Networks with Denoising Feature Matching. The main solution of both papers is introducing autoencoder into discriminator to improve the stability and quality of GAN. Different to Denoising Feature Matching, EBGAN uses encoder-decoder instead of denoising only, and use hingle loss to replace original loss function.
- This paper proposes a novel extension of generative adversarial networks that replaces the traditional binary classifier discriminator with one that assigns a scalar energy to each point in the generator's output domain. The discriminator minimizes a hinge loss while the generator attempts to generate samples with low energy under the discriminator. The authors show that a Nash equilibrium under these conditions yields a generator that matches the data distribution (assuming infinite capacity). Experiments are conducted with the discriminator taking the form of an autoencoder, optionally including a regularizer that penalizes generated samples having a high cosine similarity to other samples in the minibatch.
- Use encoder-decoder model as D is an interesting and reasonable idea.

- Qustions:

1) **What’ s the difference between this paper and Improving Generative Adversarial Networks with Denoising Feature Matching?** However, we do project some possibility of relating GAN-DFM within the EBGAN framework. For instance, on top of the auto-encoder reconstruction energies in the EBGAN auto-encoder model, we can further add a binary classifier upon the top layer of the encoder, which results in a combinatorial energy function formulation: hierarchical reconstruction losses from all layers of encoder-decoder structures with a logistic loss. In other words, the discriminator of GAN-DFM can be seen as an energy function, constructed by a discriminative encoder-decoder architecture trained only with real data samples in a layer-wise manner.

Unrolled generative adversarial networks

Luke Metz*
Google Brain
lmetz@google.com

Ben Poole†
Stanford University
poole@cs.stanford.edu

David Pfau
Google DeepMind
pfau@google.com

Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

主要内容

- **Problem** : In practice, however, GANs suffer from many issues, particularly during training. **Mode collapse and unstable**.Explicitly solving for the optimal discriminator parameters θ D (θ G) for every update step of the generator G is computationally infeasible for discriminators based on neural networks. Therefore this minimax optimization problem is typically solved by alternating gradient descent on θ G and ascent on θ D .

- **Method** : We introduce a method to stabilize Generative Adversarial Networks (GANs) by defining the generator objective with respect to an unrolled optimization of the discriminator. This allows training to be adjusted between using the optimal discriminator in the generator’ s objective, which is ideal but infeasible in practice, and using the current value of the discriminator, which is often unstable and leads to poor solutions. We show how this technique solves the common problem of mode collapse, stabilizes training of GANs with complex recurrent generators,and increases diversity and coverage of the data distribution by the generator.

Unrolling GANs

A local optimum of the discriminator parameters θ D iterative optimization procedure:

$$\theta_D^0 = \theta_D$$

$$\theta_D^{k+1} = \theta_D^k + \eta^k \frac{\mathrm{d}f\left(\theta_G, \theta_D^k\right)}{\mathrm{d}\theta_D^k}$$

$$\theta_D^*\left(\theta_G\right) = \lim_{k\to\infty} \theta_D^k,$$

where η k is the learning rate schedule. For clarity, we have expressed Eq. 7 as a full batch steepest gradient ascent equation. More sophisticated optimizers can be similarly unrolled. In our experiments we unroll Adam (Kingma & Ba, 2014).

- By unrolling for K steps, we create a surrogate objective for the update of the generator,

$$f_K\left(\theta_G, \theta_D\right) = f\left(\theta_G, \theta_D^K\left(\theta_G, \theta_D\right)\right).$$

When K = 0 this objective corresponds exactly to the standard GAN objective, while as K → ∞∗it corresponds to the true generator objective function f (θ G , θ D(G)). By adjusting the number of unrolling steps K, we are thus able to interpolate between standard GAN training dynamics withtheir associated pathologies, and more costly gradient descent on the true generator loss.

Unrolled generative adversarial networks

Luke Metz*
Google Brain
lmetz@google.com

Ben Poole†
Stanford University
poole@cs.stanford.edu

David Pfau
Google DeepMind
pfau@google.com

Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

网络结构

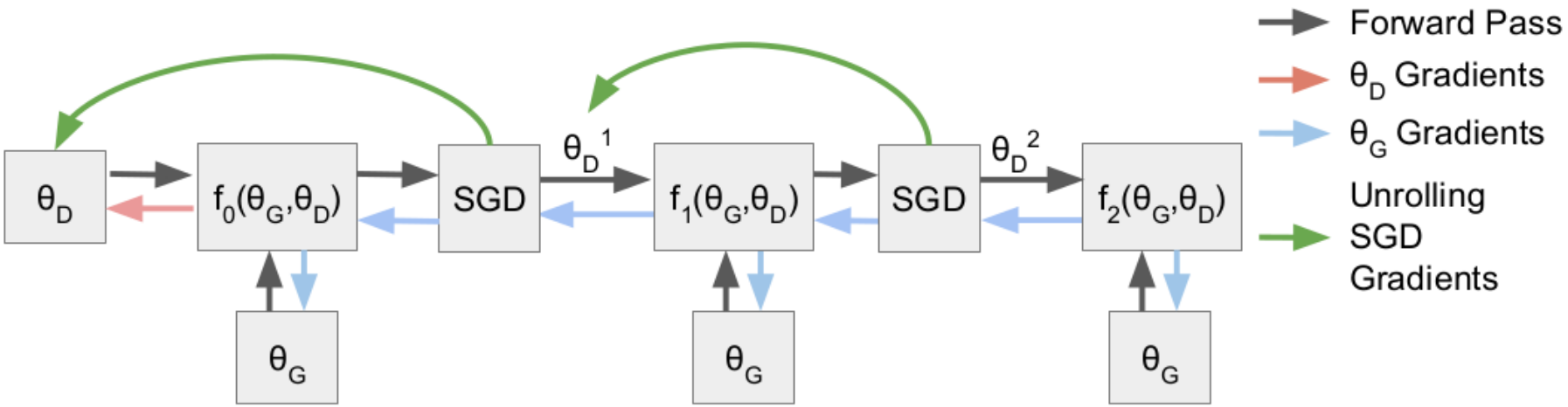


Figure 1: An illustration of the computation graph for an unrolled GAN with 3 unrolling steps. The generator update in Equation 10 involves backpropagating the generator gradient (blue arrows) through the unrolled optimization. Each step k in the unrolled optimization uses the gradients of f_k with respect to θ_D^k, as described in Equation 7 and indicated by the green arrows. The discriminator update in Equation 11 does not depend on the unrolled optimization (red arrow).

Unrolled generative adversarial networks

Luke Metz*
Google Brain
lmetz@google.com

Ben Poole†
Stanford University
poole@cs.stanford.edu

David Pfau
Google DeepMind
pfau@google.com

Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

主要内容

- **Discussion** : In this work we developed a method to stabilize GAN training and reduce mode collapse by defining the generator objective with respect to unrolled optimization of the discriminator. We then demonstrated the application of this method to several tasks, where it either rescued unstable training, or reduced the tendency of the model to drop regions of the data distribution.
- The main drawback to this method is computational cost of each training step, which increases linearly with the number of unrolling steps. There is a trade-off between better approximating the true generator loss and the computation required to make this estimate. Depending on the architecture, one unrolling step can be enough. In other more unstable models, such as the RNN case, more are needed to stabilize training. We have some initial positive results suggesting it may be sufficient to further perturb the training gradient in the same direction that a single unrolling step perturbs it. While this is more computationally efficient, further investigation is required.

The method presented here bridges some of the gap between theoretical and practical results for training of GANs. We believe developing better update rules for the generator and discriminator is an important line of work for GAN training. In this work we have only considered a small fraction of the design space. For instance, the approach could be extended to unroll G when updating D as well – letting the discriminator react to how the generator would move. It is also possible to unroll sequences of G and D updates. This would make updates that are recursive: G could react to maximize performance as if G and D had already updated.

- **Review**:The paper presents an approach for tackling the instability problem that is present in generative adversarial networks. The general idea is to allow the generator to "peek ahead" at how the discriminator will evolve its decision boundary over-time with the premise that this information should prevent the generator from collapsing to produce only samples from a single mode of the data distribution.

Mode Regularized Generative Adversarial Networks

†Tong Che,* ‡Yanran Li,* †,§Athul Paul Jacob, †Yoshua Bengio, ‡Wenjie Li
†Montreal Institute for Learning Algorithms, Université de Montréal, Montréal, QC H3T 1J4, Canada
‡Department of Computing, The Hong Kong Polytechnic University, Hong Kong
§David R. Cheriton School of Computer Science, University Of Waterloo, Waterloo, ON N2L 3G1, Canada
{tong.che,ap.jacob,yoshua.bengio}@umontreal.ca
{csyli,cswjli}@comp.polyu.edu.hk

主要内容

- **Abstract** : Although Generative Adversarial Networks achieve state-of-the-art results on a variety of generative tasks, they are regarded as highly unstable and prone to miss modes. We argue that these bad behaviors of GANs are due to the very particular functional shape of the trained discriminators in high dimensional spaces, which can easily make training stuck or push probability mass in the wrong direction, towards that of higher concentration than that of the data generating distribution. We introduce several ways of regularizing the objective, which can dramatically stabilize the training of GAN models. We also show that our regularizers can help the fair distribution of probability mass across the modes of the data generating distribution, during the early phases of training and thus providing a unified solution to the missing modes problem.

- **Conclusion** : We provide systematic ways to measure and avoid the missing modes problem and stabilize training with the proposed autoencoder-based regularizers. The key idea is that some geometric metrics can provide more stable gradients than trained discriminators, and when combined with the encoder, they can be used as regularizers for training. These regularizers can also penalize missing modes and encourage a fair distribution of probability mass on the generation manifold.

Mode Regularized Generative Adversarial Networks

- **GEOMETRIC METRICS REGULARIZER** : Compared with the objective for the GAN generator, the optimization targets for supervised learning are more stable from an optimization point of view. The difference is clear: the optimization target for the GAN generator is a learned discriminator. While in supervised models, the optimization targets are distance functions with nice geometric properties. The latter usually provides much easier training gradients than the former, especially at the early stages of training.
- Inspired by this observation, we propose to incorporate a supervised training signal as a regularizer on top of the discriminator target. Assume the generator $G(z) : Z \rightarrow X$ generates samples by sampling first from a fixed prior distribution in space Z followed by a deterministic trainable transformation G into the sample space X. Together with G, we also jointly train an encoder $E(x) : X \rightarrow Z$. Assume d is some similarity metric in the data space, we add $Ex\sim pd$ [d(x, $G \circ E(x)$)] as a regularizer, where p d is the data generating distribution. The encoder itself is trained by minimizing the same reconstruction error.

In practice, there are many options for the distance measure d. For instance, the pixel-wise L2 distance, or the distance of learned features by the discriminator (Dumoulin et al., 2016) or by other networks, such as a VGG classifier. (Ledig et al., 2016)

- The geometric intuition for this regularizer is straightforward. We are trying to move the generated manifold to the real data manifold using gradient descent. In addition to the gradient provided by the discriminator, we can also try to match the two manifolds by other geometric distances, say, L^s metric. The idea of adding an encoder is equivalent to first training a point to point mapping G(E(x)) between the two manifolds and then trying to minimize the expected distance between the points on these two manifold.

Mode Regularized Generative Adversarial Networks

- **MODE REGULARIZER :**

In addition to the metric regularizer, we propose a mode regularizer to further penalize missing modes. In traditional GANs, the optimization target for the generator is the empirical sum $\sum_i \nabla_\theta \log D(G_\theta(z_i))$. The missing mode problem is caused by the conjunction of two facts: (1) the areas near missing modes are rarely visited by the generator, by definition, thus providing very few examples to improve the generator around those areas, and (2) both missing modes and non-missing modes tend to correspond to a high value of D, because the generator is not perfect so that the discriminator can take strong decisions locally and obtain a high value of D even near non-missing modes.

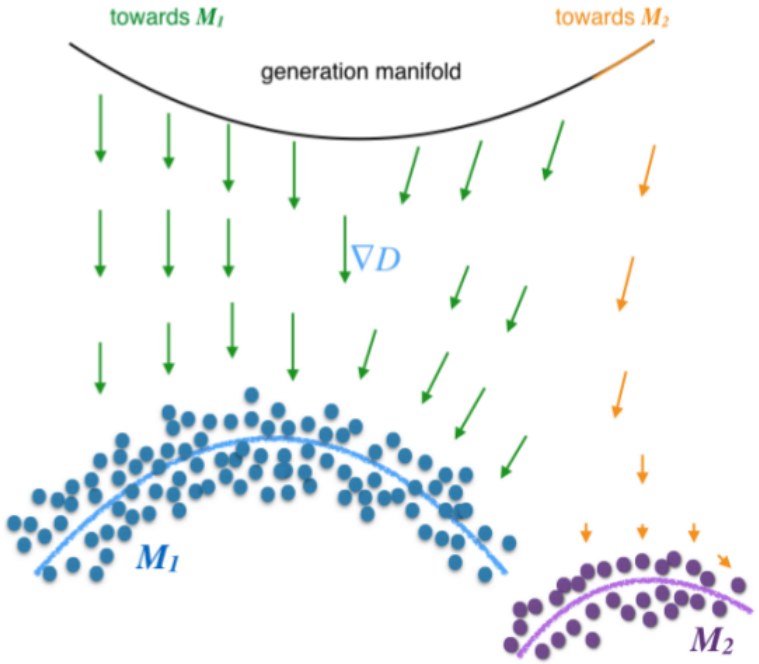


Figure 2: Illustration of missing modes problem.

As an example, consider the situation in Figure 2. For most z, the gradient of the generator $\nabla_\theta \log D(G_\theta(z))$ pushes the generator towards the major mode M_1. Only when $G(z)$ is very close to the mode M_2 can the generator get gradients to push itself towards the minor mode M_2. However, it is possible that such z is of low or zero probability in the prior distribution p_0.

Given this observation, consider a regularized GAN model with the metric regularizer. Assume M_0 is a minor mode of the data generating distribution. For $x \in M_0$, we know that if $G \circ E$ is a good autoencoder, $G(E(x))$ will be located very close to mode M_0. Since there are sufficient training examples of mode M_0 in the training data, we add the mode regularizer $\mathbb{E}_{x \sim p_d}[\log D(G \circ E(x))]$ to our optimization target for the generator, to encourage $G(E(x))$ to move towards a nearby mode of the data generating distribution. In this way, we can achieve fair probability mass distribution across different modes.

Mode Regularized Generative Adversarial Networks

主要内容

Mode collapse

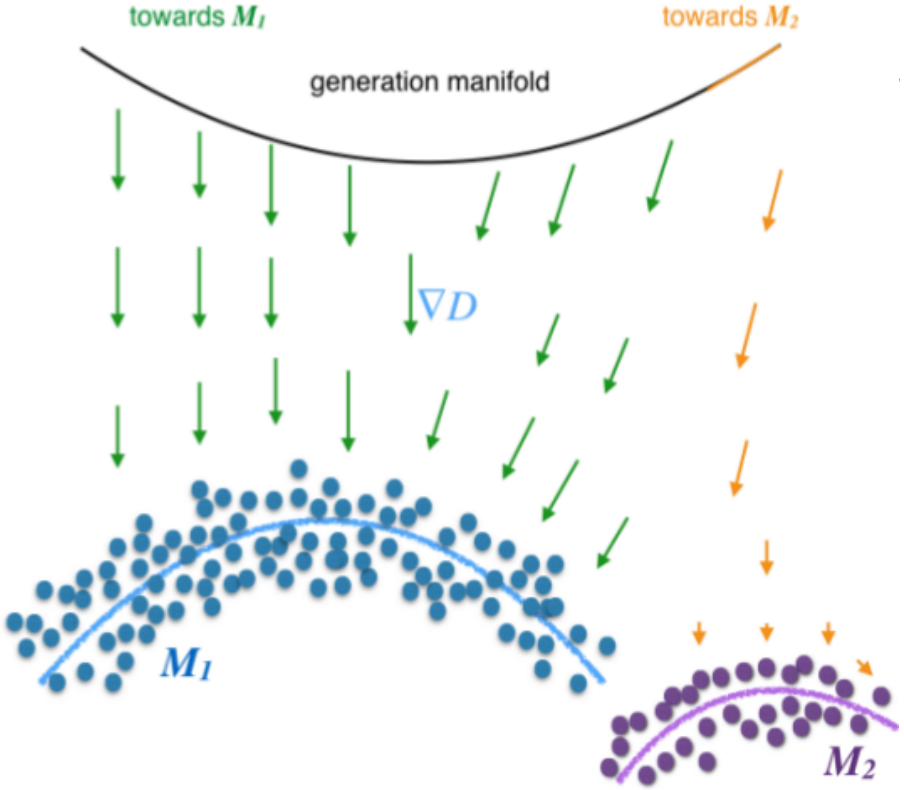


Figure 2: Illustration of missing modes problem.

2 regularized term

$$= -\mathbb{E}_z[\log D(G(z))] + \mathbb{E}_{x\sim p_d}[\lambda_1 d(x, G\circ E(x)) + \lambda_2 \log D(G\circ E(x))] \quad (1)$$

$$T_E = \mathbb{E}_{x\sim p_d}[\lambda_1 d(x, G\circ E(x)) + \lambda_2 \log D(G\circ E(x))] \quad (2)$$

Geometric Metrics Regularizer & Mode Regularizer

A new Metric -- Mode score

$$\exp\left(\mathbb{E}_x KL(p(y|\mathbf{x})||p(y)) - KL(p^*(y)||p(y))\right)$$

- where p(y) is the distribution of labels in the training data.

Mode Regularized Generative Adversarial Networks

Luke Metz*
Google Brain
lmetz@google.com

Ben Poole†
Stanford University
poole@cs.stanford.edu

David Pfau
Google DeepMind
pfau@google.com

主要内容

Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

- **Review** : The authors identify two very valid problem of mode-missing in Generative Adversarial Networks, explain their intuitions as to why these problems occur and propose ways to remedy it. The first problem is about the discriminator becoming too good (close to 0 on fake, and 1 on real data) and providing 0 gradients to the generator. The second problem is that GANs are prone to missing modes of the data generating distribution entirely. The authors propose two regularization techniques to address these problems: Geometric Metrics Regularizer and Mode Regularizer
- Detailed comments on the Geometric Metrics Regularizer: The motivation for this is to provide a way to measure and penalize distance between two degenerate probability distributions concentrated on non-overlapping manifolds, those of the generator and of the real data. There are different ways one could go about measuring difference between two manifolds or probability distributions concentrated on manifolds, for example:

- projection heuristic: measure the average distance between each point x on manifold A and the corresponding nearest point on manifold B (let’ s call it the projection of x onto B).
- earth mover’ s distance: establish a smooth mapping between the two manifolds that maps denser areas on manifold A to nearby denser areas of manifold B, and measure the average distance between corresponding pairs.

- The two heuristics are similar but while the earth mover distance is a divergence measure for distributions, the projection heuristic only measures the divergence of the manifolds, disregarding the distributions in question.

Mode Regularized Generative Adversarial Networks

- **Review** : The paper proposes two regularization approaches for training GAN, aiming to provide stronger gradient signal to move the generated distribution to data distribution and to avoid the generated distribution from getting trapped in only one or a few modes of the data distribution. The presented approaches are entirely based on some intuitive arguments. As such intuitions are interesting, likely useful, and deserve further exploration in a broader context, they stay as heuristics as this point. The paper will benefit from more rigorous theoretical justification of the presented approaches.
- think this is an interesting paper that discusses the mode-missing behavior of GANs and proposes new evaluation metric to evaluate this behavior. However, the core ideas of this paper are not very innovative to me. Specifically, there has been a lot of papers that combine GAN with an autoencoder and the settings of this paper is very similar to the other papers such as Larsen et al. As I pointed out in my pre-review comments, in the Larsen et al. both the geometric regularizer and model regularizer has been proposed in the context of VAEs and the way they are used is essentially the same as this paper.

I understand the argument of the authors that the VAEGAN is a VAE that is regularized by GAN and in this paper the main generative model is a GAN that is regularized by an autoencoder, but at the end of the day, both the models are combining the autoencoder and GAN in a pretty much same way, and to me the resulting model is not very different. I also understand the other argument of the authors that Larsen et al is using VAE while this paper is using an autoencoder, but I am still not convinced how this paper outperforms the VAEGAN by just removing the KL term of the VAE. I do like that this paper looks at the autoencoder objective as a way to alleviate the missing mode problem of GANs, but I think that alone does not have enough originality to carry the paper.

Published on ICLR in 2017

Adversarial Feature Learning

Jeff Donahue
jdonahue@cs.berkeley.edu
Computer Science Division
University of California, Berkeley

Philipp Krähenbühl
philkr@utexas.edu
Department of Computer Science
University of Texas, Austin

Trevor Darrell
trevor@eecs.berkeley.edu
Computer Science Division
University of California, 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.

- Objective Function :

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

where

$$V(D,E,G) := \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{X}}}\big[\underbrace{\mathbb{E}_{\mathbf{z}\sim p_E(\cdot|\mathbf{x})}\left[\log D(\mathbf{x},\mathbf{z})\right]}_{\log D(\mathbf{x},E(\mathbf{x}))}\big] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{Z}}}\big[\underbrace{\mathbb{E}_{\mathbf{x}\sim p_G(\cdot|\mathbf{z})}\left[\log\left(1-D(\mathbf{x},\mathbf{z})\right)\right]}_{\log(1-D(G(\mathbf{z}),\mathbf{z}))}\big].$$

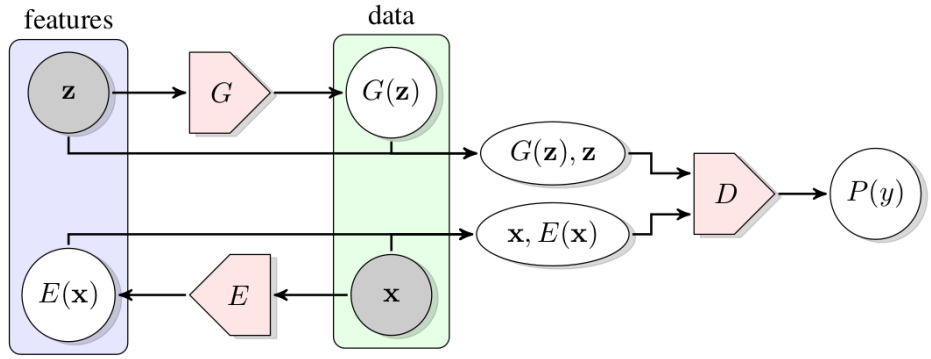


Figure 1: The structure of Bidirectional Generative Adversarial Networks (BiGAN).

- Conclusion: BiGAN and other unsupervised learning approaches are agnostic to the domain of the data.
- BiGAN and other unsupervised approaches needn't suffer from domain shift between the pre-training task and the transfer task, unlike self-supervised methods in which some aspect of the data is normally removed or corrupted in order to create a non-trivial prediction task.

Adversarial Feature Learning

Interpolations in the latent space of the generator produce smooth and plausible semantic variations, and certain directions in this space correspond to particular semantic attributes along which the data distribution varies. For example, Radford et al. (2016) showed that a GAN trained on a database of human faces learns to associate particular latent directions with gender and the presence of eyeglasses.

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* (BiGAN). The overall model is depicted in Figure 1. In short, in addition to the generator G from the standard GAN framework (Goodfellow et al., 2014), BiGAN includes an *encoder* E which maps data $\mathbf{x}$ to latent representations $\mathbf{z}$. The BiGAN discriminator D discriminates not only in data space ($\mathbf{x}$ versus $G(\mathbf{z})$), but jointly in data and latent space (tuples $(\mathbf{x}, E(\mathbf{x}))$ versus $(G(\mathbf{z}), \mathbf{z})$), where the latent component is either an encoder output $E(\mathbf{x})$ or a generator input $\mathbf{z}$.

Published on ICLR in 2017

Adversarially Learned Inference

Vincent Dumoulin[1], **Ishmael Belghazi**[1], **Ben Poole**[2]
Olivier Mastropietro[1], **Alex Lamb**[1], **Martin Arjovsky**[3]
Aaron Courville[1†]
[1] MILA, Université de Montréal, firstname.lastname@umontreal.ca.
[2] Neural Dynamics and Computation Lab, Stanford, poole@cs.stanford.edu.
[3] New York University, martinarjovsky@gmail.com.
[†]CIFAR Fellow.

- **Abstract** : We introduce the adversarially learned inference (ALI) model, which jointly learns a generation network and an inference network using an adversarial process. The generation network maps samples from stochastic latent variables to the data space while the inference network maps training examples in data space to the space of latent variables. An adversarial game is cast between these two networks anda discriminative network is trained to distinguish between joint latent/data-space samples from the generative network and joint samples from the inference network. We illustrate the ability of the model to learn mutually coherent inference and generation networks through the inspections of model samples and reconstructions and confirm the usefulness of the learned representations by obtaining a performance competitive with state-of-the-art on the semi-supervised SVHN and CIFAR10 tasks.

- Objective Function :

$$\min_G \max_D V(D,G) = \mathbb{E}_{q(\boldsymbol{x})}[\log(D(\boldsymbol{x}, G_z(\boldsymbol{x})))] + \mathbb{E}_{p(\boldsymbol{z})}[\log(1 - D(G_x(\boldsymbol{z}), \boldsymbol{z}))]$$
$$= \iint q(\boldsymbol{x})q(\boldsymbol{z} \mid \boldsymbol{x}) \log(D(\boldsymbol{x}, \boldsymbol{z}))d\boldsymbol{x}d\boldsymbol{z}$$
$$+ \iint p(\boldsymbol{z})p(\boldsymbol{x} \mid \boldsymbol{z}) \log(1 - D(\boldsymbol{x}, \boldsymbol{z}))d\boldsymbol{x}d\boldsymbol{z}.$$

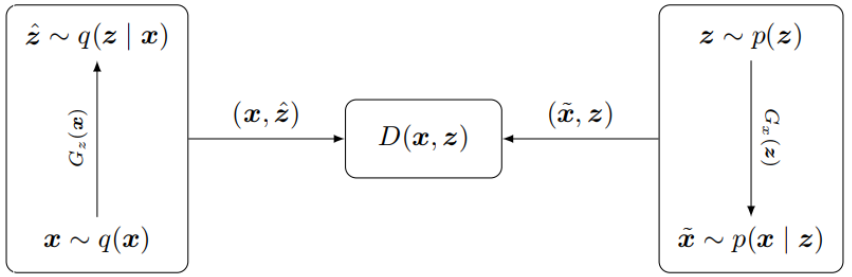


Figure 1: The adversarially learned inference (ALI) game.

- Conclusion: We introduced the adversarially learned inference (ALI) model, which jointly learns a generation network and an inference network using an adversarial process. The model learns mutually coherent inference and generation networks, as exhibited by its reconstructions. The induced latent variable mapping is shown to be useful, achieving results competitive with the state-of-the-art on the semi-supervised SVHN and CIFAR10 tasks.

Adversarially Learned Inference

- **Review** : The idea is using auto encoder to provide extra information for discriminator. This approach seems is promising from reported result.
- This paper proposes a new method for learning an inference network in the GAN framework. ALI's objective is to match the joint distribution of hidden and visible units imposed by an encoder and decoder network. ALI is trained on multiple datasets, and it seems to have a good reconstruction even though it does not have an explicit reconstruction term in the cost function. This shows it is learning a decent inference network for GAN.
- There are currently many ways to learn an inference network for GANs: One can learn an inference network after training the GAN by sampling from the GAN and learning a separate network to map X to Z. There is also the infoGAN approach (not cited) which trains the inference network at the same time with the generative path.
- This paper extends the GAN framework to allow for latent variables. The observed data set is expanded by drawing latent variables z from a conditional distribution q(z|x). The joint distribution on x,z is then modeled using a joint generator model p(x,z)=p(z)p(x|z).

Both q and p are then trained by trying to fool a discriminator. This constitutes a worthwhile extension of GANs: giving GANs the ability to do inference opens up many applications that could previously only be addressed by e.g. VAEs.

- **Question**:
- What is the difference of the ALI model to BiGAN
- Both methods were developed independently and published at roughly the same time (BiGAN released May 31 2016, ALI released 2 days later on June 2 2016). Both papers acknowledge that the methods were developed independently and are very similar.

There are differences in the experiments and ALI considered stochastic encoders, but the two proposed models are essentially the same.

Published on ICLR in 2017

Unsupervised Cross-Domain Image Generation

Vincent Dumoulin[1], **Ishmael Belghazi**[1], **Ben Poole**[2]
Olivier Mastropietro[1], **Alex Lamb**[1], **Martin Arjovsky**[3]
Aaron Courville[1†]
[1] MILA, Université de Montréal, firstname.lastname@umontreal.ca.
[2] Neural Dynamics and Computation Lab, Stanford, poole@cs.stanford.edu.
[3] New York University, martinarjovsky@gmail.com.
[†]CIFAR Fellow.

- **Abstract** :We study the problem of transferring a sample in one domain to an analog sample in another domain. Given two related domains, S and T, we would like to learn a generative function G that maps an input sample from S to the domain T, such that the output of a given representation function f, which accepts inputs in either domains, would remain unchanged. Other than f, the training data is unsupervised and consist of a set of samples from each domain, without any mapping between them. The Domain Transfer Network (DTN) we present employs a compound loss function that includes a multiclass GAN loss, an f preserving component, and a regularizing component that encourages G to map samples from T to themselves. We apply our method to visual domains including digits and face images and demonstrate its ability to generate convincing novel images of previously unseen entities, while preserving their identity

- Objective Function :

$$R_{\text{GAN}} = \max_D \mathbb{E}_{x \sim \mathcal{D}_S} \log[1 - D(G(x))] + \mathbb{E}_{x \sim \mathcal{D}_T} \log[D(x)],$$

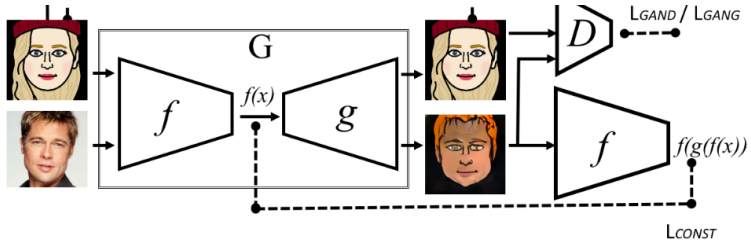


Figure 1: The Domain Transfer Network. Losses are drawn with dashed lines, input/output with solid lines. After training, the forward model G is used for the sample transfer.

- Conclusion: Asymmetry is central to our work. Not only does our solution handle the two domains S and T differently, the function f is unlikely to be equally effective in both domains since in most practical cases, f would be trained on samples from one domain. While an explicit domain adaptation step can be added in order to make f more effective on the second domain, we found it to be unnecessary. Adaptation of f occurs implicitly due to the application of D downstream.
- Using the same function f , we can replace the roles of the two domains, S and T .
- Domain transfer, as an unsupervised method, could prove useful across a wide variety of computational tasks.

Unsupervised Cross-Domain Image Generation

- **Review** : The authors propose a application of GANs to map images to new domains with no labels. E.g., an MNIST 3 is used to generate a SVHN 3. Ablation analysis is given to help understand the model. The results are (subjectively) impressive and the approach could be used for cross-domain transfer, an important problem. All in all, a strong paper.
- This paper presents an unsupervised domain transfer from the image of domain S to the image of domain T. It was really refreshing that this conversion was possible without any mapping data. For example, in the paper, the model can transfer the SVHN image '3' to the MNIST image '3' without the mapping data. The model can be roughly divided into GAN and Content Extractor (f in the paper).
- 1. GAN. During training, the discriminator sees the mnist image and learns to determine it as a real image. And with GAN loss, the generator learns to get the mnist image as output when it receives an svhn image as input to deceive the discriminator.
- 2. Content Extractor. If the model use only GAN loss, the content in the image may not be retained even if the domain is changed. For example, the generator may convert the svhn image '3' to the mnist image '2' to deceive the discriminator. In this paper, authors introduce a new function called 'f' to maintain the content. The generator includes f and generates a fake mnist image when it receives an svhn image as input. The original svhn image and the generated fake mnist image are put back into f. Then additional loss function is set so that the resulting values are the same. Here, f is learning to extract content regardless of domain.
- This paper presents an unsupervised image transformation method that maps a sample from source domain to target domain

Published on ICLR in 2017

Calibrating Energy-based Generative Adversarial Networks

Zihang Dai[1], Amjad Almahairi[2]*, Philip Bachman[3], Eduard Hovy[1] & Aaron Courville[2]
[1] Language Technologies Institute, Carnegie Mellon University.
[2] MILA, Université de Montréal.
[3] Maluuba Research.

- **Abstract** :In this paper, we propose to equip Generative Adversarial Networks with the ability to produce direct energy estimates for samples. Specifically, we propose a flexible adversarial trai framework, and prove this framework not only ensures the generator converges to the true data distribution, but also enables the discriminator to retain the density information at the global optim We derive the analytic form of the induced solution, and analyze the properties. In order to make the proposed framework trainable in practice, we introduce two effective approximation techniques. Empirically, the experiment results closely match our theoretical analysis, verifying the discriminator is able to recover the energy of data distribution.

- Objective Function :

$$\max_{c} \min_{p_{\text{gen}} \in \mathcal{P}} \quad \mathop{\mathbb{E}}_{x \sim p_{\text{gen}}} [c(x)] - \mathop{\mathbb{E}}_{x \sim p_{\text{data}}} [c(x)] + K(p_{\text{gen}}),$$

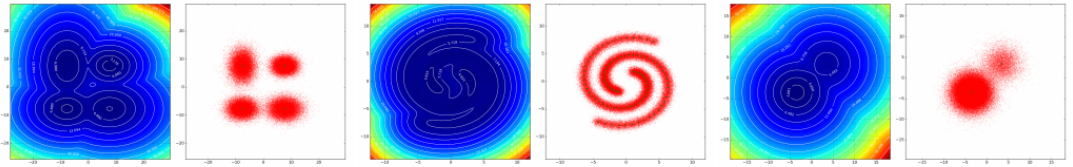


Figure 1: True energy functions and samples from synthetic distributions. Green dots in the sample plots indicate the mean of each Gaussian component.

- **Conclusion**: In this paper we have addressed a fundamental limitation in adversarial learning approaches, which is their inability of providing sensible energy estimates for samples. We proposed a novel adversarial learning formulation which results in a discriminator function that recovers the true data energy. We provided a rigorous characterization of the learned discriminator in the non-parametric setting, and proposed two methods for instantiating it in the typical parametric setting. Our experimental results verify our theoretical analysis about the discriminator properties, and show that we can also obtain samples of state-of-the-art quality.

Calibrating Energy-based Generative Adversarial Networks

- **Review** : his paper addresses one of the major shortcomings of generative adversarial networks - their lack of mechanism for evaluating held-out data. While other work such as BiGANs/ALI address this by learning a separate inference network, here the authors propose to change the GAN objective function such that the optimal discriminator is also an energy function, rather than becoming uninformative at the optimal solution. Training this new objective requires gradients of the entropy of the generated data, which are difficult to approximate, and the authors propose two methods to do so, one based on nearest neighbors and one based on a variational lower bound. The results presented show that on toy data the learned discriminator/energy function closely approximates the log probability of the data, and on more complex data the discriminator give a good measure of quality for held out data.
- Problem:

  1. **In the appendix, you compare your algorithm to f-GAN. However, It would be more natural that you compare it with GMMN because the formulation of maximum mean discrepancy is more similar to (1)**.

  Answer:Firstly, just for clarification, the reason we mention f-GAN in the appendix is to support the point that "adding the same calibrating term K(p_gen) to the original GAN formulation (or to the f-GAN family) will NOT enable the discriminator to recover the energy function while ensuring the generator matches the data distribution". Hopefully, this point is well conveyed.

  For the comparison among GMMN, f-GAN, and our proposed calibrating energy-based GAN (CEGAN for short), the difference lies in how they achieve the central target: matching p_gen and p_data. Specifically, GMMN tries to do the matching directly in the primal space (generator space) by employing the MMD as the divergence criterion between distributions. On the contrary, f-GAN and CEGAN choose to train a criterion (the discriminator) at the same time via a minimax game, where the discriminator corresponds to the dual variable in the dual space.


ICLR
2016
PART
ONE

Published on ICLR in 2016

Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks

Jost Tobias Springenberg
University of Freiburg
79110 Freiburg, Germany
springj@cs.uni-freiburg.de

- **Abstract** :In this paper we present a method for learning a discriminative classifier from unlabeled or partially labeled data. Our approach is based on an objective function that trades-off mutual information between observed examples and their predicted categorical class distribution, against robustness of the classifier to an adversarial generative model. The resulting algorithm can either be interpreted as a natural generalization of the generative adversarial networks (GAN) framework or as an extension of the regularized information maximization (RIM) framework to robust classification against an optimal adversary. We empirically evaluate our method - which we dub categorical generative adversarial networks (or CatGAN) - on synthetic data as well as on challenging image classification tasks, demonstrating the robustness of the learned classifiers. We further qualitatively assess the fidelity of samples generated by the adversarial generator that is learned alongside the discriminative classifier, and identify links between the CatGAN objective and discriminative clustering algorithms (such as RIM).

- Objective Function :

$$\max_{c} \min_{p_{\text{gen}} \in \mathcal{P}} \quad \mathop{\mathbb{E}}_{x \sim p_{\text{gen}}} \left[c(x) \right] - \mathop{\mathbb{E}}_{x \sim p_{\text{data}}} \left[c(x) \right] + K(p_{\text{gen}}),$$

- **Conclusion**: We have presented categorical generative adversarial networks, a framework for robust unsupervised and semi-supervised learning. Our method combines neural network classifiers with an adversarial generative model that regularizes a discriminatively trained classifier. We found the proposed method to yield classification performance that is competitive with state-of-the-art results for semi-supervised learning for image classification and further confirmed that the generator, which is learned alongside the classifier, is capable of generating images of high visual fidelity.

Published on ICLR in 2016

Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks

Jost Tobias Springenberg
University of Freiburg
79110 Freiburg, Germany
springj@cs.uni-freiburg.de

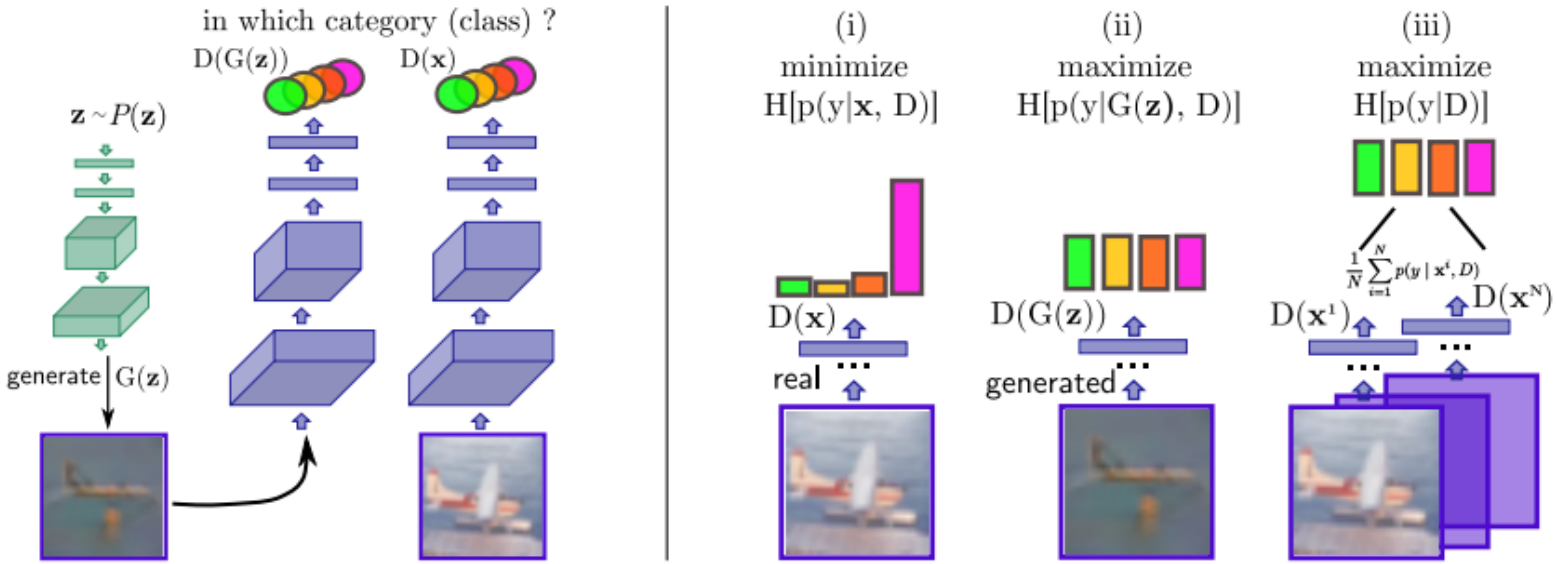


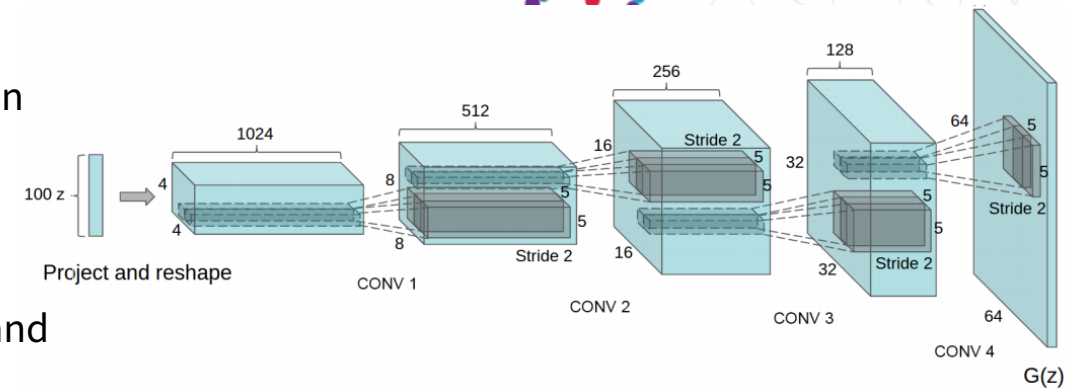
Figure 1: Visualization of the information flow through the generator (in green) and discriminator (in violet) neural networks (left). A sketch of the three parts (i) - (iii) of the objective function $\mathcal{L}_D$ for the discriminator (right). To obtain certain predictions the discriminator minimizes the entropy of $p(y|\mathbf{x}, D)$, leading to a peaked conditional class distribution. To obtain uncertain predictions for generated samples the the entropy of $p(y|G(\mathbf{z}), D)$ is maximized which, in the limit, would result in a uniform distribution. Finally, maximizing the marginal class entropy over all data-points leads to uniform usage of all classes

Published on ICLR in 2016

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

Jost Tobias Springenberg
University of Freiburg
79110 Freiburg, Germany
springj@cs.uni-freiburg.de

- **Abstract** :In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. In this work we hope to help bridge the gap between the success of CNNs for supervised learning and unsupervised learning. We introduce a class of CNNs called deep convolutional generative adversarial networks (DCGANs), that have certain architectural constraints, and demonstrate that they are a strong candidate for unsupervised learning. Training on various image datasets, we show convincing evidence that our deep convolutional adversarial pair learns a hierarchy of representations from object parts to scenes in both the generator and discriminator. Additionally, we use the learned features for novel tasks - demonstrating their applicability as general image representations.



- **Conclusion**: We propose a more stable set of architectures for training generative adversarial networks and we give evidence that adversarial networks learn good representations of images for supervised learning and generative modeling. There are still some forms of model instability remaining - we noticed as models are trained longer they sometimes collapse a subset of filters to a single oscillating mode.
- Further work is needed to tackle this from of instability. We think that extending this framework to other domains such as video (for frame prediction) and audio (pre-trained features for speech synthesis) should be very interesting. Further investigations into the properties of the learnt latent space would be interesting as well.

Published on ICLR in 2016

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

Jost Tobias Springenberg
University of Freiburg
79110 Freiburg, Germany
springj@cs.uni-freiburg.de

- **Contribution**
- We propose and evaluate a set of constraints on the architectural topology of Convolutional GANs that make them stable to train in most settings. We name this class of architectures Deep Convolutional GANs (DCGAN)
- We use the trained discriminators for image classification tasks, showing competitive performance with other unsupervised algorithms.
- We visualize the filters learnt by GANs and empirically show that specific filters have learned to draw specific objects.
- We show that the generators have interesting vector arithmetic properties allowing for easy manipulation of many semantic qualities of generated samples.

Architecture guidelines for stable Deep Convolutional GANs:

- Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).
- Use batchnorm in both the generator and the discriminator.
- Remove fully connected hidden layers for deeper architectures.
- Use ReLU activation in generator for all layers except for the output, which uses Tanh
- Use LeakyReLU activation in the discriminator for all layers.


NIPS
2017
PART
ONE

Published on NIPS in 2017

Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis

Jian Zhao[1,2*†] Lin Xiong[3] Karlekar Jayashree[3] Jianshu Li[1] Fang Zhao[1]
Zhecan Wang[4†] Sugiri Pranata[3] Shengmei Shen[3]
Shuicheng Yan[1,5] Jiashi Feng[1]
[1]National University of Singapore [2]National University of Defense Technology
[3] Panasonic R&D Center Singapore [4] Franklin. W. Olin College of Engineering
[5] Qihoo 360 AI Institute
{zhaojian90, jianshu}@u.nus.edu {lin.xiong, karlekar.jayashree, sugiri.pranata, shengmei.shen}@sg.panasonic.com
zhecan.wang@students.olin.edu {elezhf, eleyans, elefjia}@u.nus.edu

- **Abstract** :Synthesizing realistic profile faces is promising for more efficiently training deep pose-invariant models for large-scale unconstrained face recognition, by populating samples with extreme poses and avoiding tedious annotations. However, learning from synthetic faces may not achieve the desired performance due to the discrepancy between distributions of the synthetic and real face images. To narrow this gap, we propose a Dual-Agent Generative Adversarial Network (DA-GAN) model, which can improve the realism of a face simulator' s output using unlabeled real faces, while preserving the identity information during the realism refinement. The dual agents are specifically designed for distinguishing real v.s. fake and identities simultaneously. In particular, we employ an off-the-shelf 3D face model as a simulator to generate profile face images with varying poses. DA-GAN leverages a fully convolutional network as the generator to generate high-resolution images and an auto-encoder as the discriminator with the dual agents.

Besides the novel architecture, we make several key modifications to the standard GAN to preserve pose and texture, preserve identity and stabilize training process: (I) a pose perception loss; (ii) an identity perception loss; (iii) an adversarial loss with a boundary equilibrium regularization term. Experimental results show that DA-GAN not only presents compelling perceptual results but also significantly outperforms state-of-the-arts on the large-scale and challenging NIST IJB-A unconstrained face recognition benchmark. In addition, the proposed DA-GAN is also promising as a new approach for solving generic transfer learning problems more effectively. DA-GAN is the foundation of our submissions to NIST IJB-A 2017 face recognition competitions, where we won the 1 st places on the tracks of verification and identification..

Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis

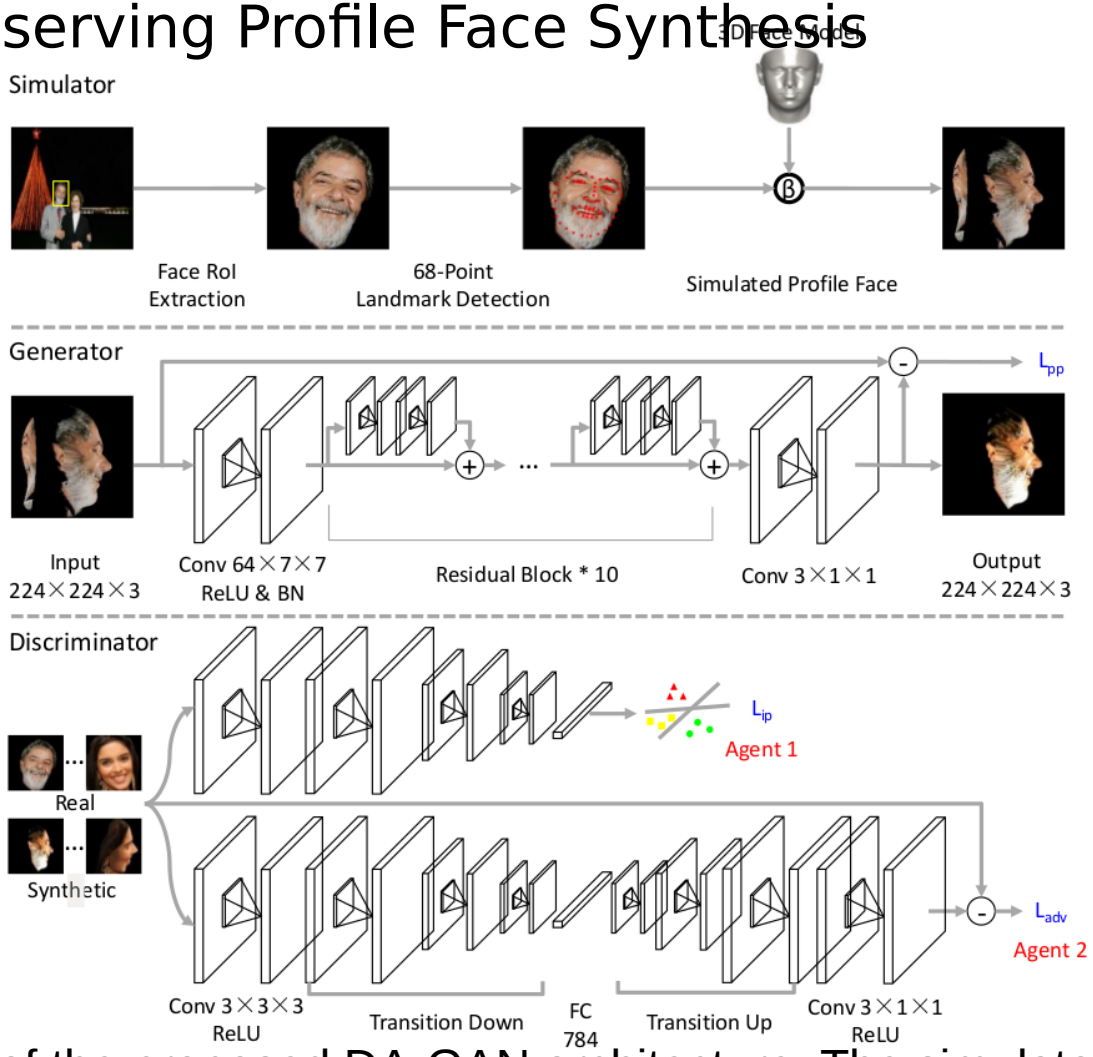


Figure 2: Overview of the proposed DA-GAN architecture. The simulator (upper panel) extracts face RoI, localizes landmark points and produces synthesis faces with arbitrary poses, which are fed
to DA-GAN for realism refinement. DA-GAN uses a fully convolutional skip-net as the generator (middle panel) and an auto-encoder as the discriminator (bottom panel). The dual agents focus on both discriminating real v.s. fake (minimizing the loss L adv) and preserving identity information (minimizing the loss L ip). Best viewed in color.

Published on NIPS in 2017

Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis

Jian Zhao[1,2*†] Lin Xiong[3] Karlekar Jayashree[3] Jianshu Li[1] Fang Zhao[1]
Zhecan Wang[4†] Sugiri Pranata[3] Shengmei Shen[3]
Shuicheng Yan[1,5] Jiashi Feng[1]
[1]National University of Singapore [2]National University of Defense Technology
[3] Panasonic R&D Center Singapore [4] Franklin. W. Olin College of Engineering
[5] Qihoo 360 AI Institute
{zhaojian90, jianshu}@u.nus.edu {lin.xiong, karlekar.jayashree, sugiri.pranata, shengmei.shen}@sg.panasonic.com
zhecan.wang@students.olin.edu {elezhf, eleyans, elefjia}@u.nus.edu

- **Review** :This work uses GANs to generate synthetic data to use for supervised training of facial recognition systems. More specifically, they use an image-to-image GAN to improve the quality of faces generated by a face simulator. The simulator is able to produce a wider range of face poses for a given face, and the GAN is able to refine the simulators output such that it is more closely aligned with the true distribution of faces (i.e. improve the realism of the generated face) while maintaining the facial identity and pose the simulator outputted. They show that by fine tuning a facial recognition system on this additional synthetic data they are able to improve performance and outperform previous state of the art methods.
- This paper presents a method for augmenting natural face data by 3D synthesis which does not suffer from overfitting on artifacts. The approach uses a GAN network to filter synthesized images so as to automatically remove artifacts.
- The paper shows that the approach provides a significant boost over a state-of-the-art model on the IJB 'faces in the wild' dataset: reducing the errors by about 25%. The idea of augmenting natural images using 3D models is not new. However, real gains over state-of-the-art performance have not materialized due to the models overfitting on the artifacts of the 3D synthesis process. The authors prove that argument by showing that adding unfiltered augmented data to the baseline model actually degrades performance. I believe this paper shows a promising approach to solve this issue that I have not seen elsewhere so far.
- In this paper the authors proposed a model-free imitation learning that 1) obtains significant performance gains over existing model-free methods in imitating learning/behavioral cloning, and 2) bypasses the indirect nature of inverse reinforcement learning where a cost function is first learned with expert's data and an optimal policy is obtained by RL afterwards.

Published on NIPS in 2017

AdaGAN: Boosting Generative Models

Ilya Tolstikhin
MPI for Intelligent Systems
Tübingen, Germany
ilya@tue.mpg.de

Sylvain Gelly
Google Brain
Zürich, Switzerland
sylvaingelly@google.com

Olivier Bousquet
Google Brain
Zürich, Switzerland
obousquet@google.com

Carl-Johann Simon-Gabriel
MPI for Intelligent Systems
Tübingen, Germany
cjsimon@tue.mpg.de

Bernhard Schölkopf
MPI for Intelligent Systems
Tübingen, Germany
bs@tue.mpg.de

- **Abstract** :Generative Adversarial Networks (GAN) are an effective method for training generative models of complex data such as natural images. However, they are notoriously hard to train and can **suffer from the problem of missing modes where the model is not able to produce examples in certain regions of the space**. We propose an iterative procedure, called AdaGAN, where at every step we add a new component into a mixture model by running a GAN algorithm on a re-weighted sample. This is inspired by boosting algorithms, where many potentially weak individual predictors are greedily aggregated to form a strong composite predictor. We prove analytically that such an incremental procedure leads to convergence to the true distribution in a finite number of steps if each step is optimal, and convergence at an exponential rate otherwise. We also illustrate experimentally that this procedure addresses the problem of missing modes.

ALGORITHM 1 AdaGAN, a meta-algorithm to construct a "strong" mixture of T individual generative models (f.ex. GANs), trained sequentially.

Input: Training sample $S_N := \{X_1, \dots, X_N\}$.
Output: Mixture generative model $G = G_T$.

Train vanilla GAN $G_1 = \text{GAN}(S_N, W_1)$ with a uniform weight $W_1 = (1/N, \dots, 1/N)$ over the training points
for $t = 2, \dots, T$ **do**
#Choose the overall weight of the next mixture component
$\beta_t = \text{ChooseMixtureWeight}(t)$
#Update the weight of each training example
$W_t = \text{UpdateTrainingWeights}(G_{t-1}, S_N, \beta_t)$
#Train t-th "weak" component generator G_t^c
$G_t^c = \text{GAN}(S_N, W_t)$
#Update the overall generative model:
#Form a mixture of G_{t-1} *and* G_t^c.
$G_t = (1 - \beta_t)G_{t-1} + \beta_t G_t^c$
end for

Published on NIPS in 2017

AdaGAN: Boosting Generative Models

Ilya Tolstikhin
MPI for Intelligent Systems
Tübingen, Germany
ilya@tue.mpg.de

Sylvain Gelly
Google Brain
Zürich, Switzerland
sylvaingelly@google.com

Olivier Bousquet
Google Brain
Zürich, Switzerland
obousquet@google.com

Carl-Johann Simon-Gabriel
MPI for Intelligent Systems
Tübingen, Germany
cjsimon@tue.mpg.de

Bernhard Schölkopf
MPI for Intelligent Systems
Tübingen, Germany
bs@tue.mpg.de

f-Divergences In order to measure the agreement between the model distribution and the true distribution we will use an f-divergence defined in the following way:

$$D_f(Q\|P) := \int f\left(\frac{dQ}{dP}(x)\right) dP(x) \tag{1}$$

for any pair of distributions P, Q with densities dP, dQ with respect to some dominating reference measure μ (we refer to Appendix D for more details about such divergences and their domain of definition). Here we assume that f is convex, defined on $(0, \infty)$, and satisfies $f(1) = 0$. We will denote by $\mathcal{F}$ the set of such functions. [2]

As demonstrated in [16, 17], several commonly used symmetric f-divergences are *Hilbertian metrics*, which in particular means that their square root satisfies the triangle inequality. This is true for the Jensen-Shannon divergence[3], the Hellinger distance and the Total Variation among others. We will denote by $\mathcal{F}_H$ the set of functions f such that D_f is a Hilbertian metric.

- **Conclusion** :We studied the problem of minimizing general f -divergences with additive mixtures of distributions. The main contribution of this work is a detailed theoretical analysis, which naturally leads to an iterative greedy procedure. On every iteration the mixture is updated with a new component, which minimizes f -divergence with a re-weighted target distribution. We provided conditions under which this procedure is guaranteed to converge to the target distribution at an exponential rate. While our results can be combined with any generative modelling techniques, we focused on GANs and provided a boosting-style algorithm AdaGAN. Preliminary experiments show that AdaGAN successfully produces a mixture which iteratively covers the missing modes.

AdaGAN: Boosting Generative Models

- **Review** :This paper builds a mega-algorithm that can incorporate various sub-models to tackle with missing mode problem. It tactfully applied AdaBoost and other mixture model spirits in the context of GAN. The paper theoretically analyzed the optimal and suboptimal distributions that are added as mixture components, and get the corresponding convergence results. The paper is well organized and the details are clearly elaborated.
- The paper proposes a new method inspired by AdaBoost to address the missing mode problem often occurs in GAN training. In general, the paper is quite well written. The studied problem is interesting and important, and the theoretical analyses seem solid. The proposed algorithm is novel and shows good empirical results on synthetic dataset. Below are my minor comments:

- AdaGAN is a meta-algorithm proposed for GAN. The key idea of AdaGAN is: at each step reweight the samples and fits a generative model on the reweighted samples. The final model is a weighted addition of the learned generative models. The main motivation is to reduce the mode-missing problem of GAN by reweighting samples at each step.

https://www.youtube.com/watch?v=5EEaY_cVYkk

Published on NIPS in 2017

Gradient descent GAN optimization is locally stable

Ilya Tolstikhin
MPI for Intelligent Systems
Tübingen, Germany
ilya@tue.mpg.de

Sylvain Gelly
Google Brain
Zürich, Switzerland
sylvaingelly@google.com

Olivier Bousquet
Google Brain
Zürich, Switzerland
obousquet@google.com

Carl-Johann Simon-Gabriel
MPI for Intelligent Systems
Tübingen, Germany
cjsimon@tue.mpg.de

Bernhard Schölkopf
MPI for Intelligent Systems
Tübingen, Germany
bs@tue.mpg.de

- **Abstract** :Despite the growing prominence of generative adversarial networks (GANs), optimization in GANs is still a poorly understood topic. In this paper, we analyze the "gradient descent" form of GAN optimization, i.e., the natural setting where we simultaneously take small gradient steps in both generator and discriminator parameters. We show that even though GAN optimization does not correspond to a **convex-concave** game (even for simple parameterizations), under proper conditions, equilibrium points of this optimization procedure are still locally asymptotically stable for the traditional GAN formulation. On the other hand, we show that the recently proposed Wasserstein GAN can have **non-convergent limit cycles near equilibrium**. Motivated by this stability analysis, we propose an additional **regularization term** for gradient descent GAN updates, which is able to guarantee local stability for both the **WGAN** and the **traditional GAN**, and also shows practical promise in speeding up convergence and addressing **mode collapse.**

- **Conclusion** :In this paper, we presented a theoretical analysis of the local asymptotic stability of GAN optimization under proper conditions. We further showed that the recently proposed WGAN is not asymptotically stable under the same conditions, but we introduced a **gradient-based regularizer** which stabilizes both traditional GANs and the WGANs, and can improve convergence speed in practice. The results here provide substantial insight into the nature of GAN optimization, perhaps even offering some clues as to why these methods have worked so well despite **not** being **convex-concave**. However, we also emphasize that there are substantial limitations to the analysis, and directions for future work. Perhaps most notably, the analysis here only provides an understanding of what happens locally, close to an equilibrium point.

Published on NIPS in 2017

Gradient descent GAN optimization is locally stable

Ilya Tolstikhin
MPI for Intelligent Systems
Tübingen, Germany
ilya@tue.mpg.de

Sylvain Gelly
Google Brain
Zürich, Switzerland
sylvaingelly@google.com

Olivier Bousquet
Google Brain
Zürich, Switzerland
obousquet@google.com

Carl-Johann Simon-Gabriel
MPI for Intelligent Systems
Tübingen, Germany
cjsimon@tue.mpg.de

Bernhard Schölkopf
MPI for Intelligent Systems
Tübingen, Germany
bs@tue.mpg.de

- **Conclusion** : For non-convex architectures this may be all that is possible, but it seems plausible that much stronger global convergence results could hold for simple settings like the linear quadratic GAN (indeed, as the streamline plots show, we observe this in practice for simple domains). Second, the analysis here does not show the equilibrium points necessarily exist, but only illustrates convergence if there do exist points that satisfy certain criteria: the existence question has been addressed by previous work [Arora et al., 2017], but much more analysis remains to be done here. GANs are rapidly becoming a cornerstone of deep learning methods, and the theoretical and practical understanding of these methods will prove crucial in moving the field forward.

https://youtu.be/9gQt6Vqb854

Published on NIPS in 2017

f-GANs in an Information Geometric Nutshell

Richard Nock[†,‡,§] **Zac Cranko**[‡,†] **Aditya Krishna Menon**[†,‡]
Lizhen Qu[†,‡] **Robert C. Williamson**[‡,†]
[†]Data61, [‡]the Australian National University and [§]the University of Sydney
{firstname.lastname, aditya.menon, bob.williamson}@data61.csiro.au

- **Abstract** :Nowozin et al showed last year how to extend the GAN principle to all f divergences. The approach is elegant but falls short of a full description of the supervised game, and says little about the key player, the generator: for example, what does the generator actually converge to if solving the GAN game means convergence in some space of parameters? How does that provide hints on the generator' s design and compare to the flourishing but almost exclusively experimental literature on the subject? In this paper, we unveil a broad class of distributions for which such convergence happens — namely, deformed exponential families, a wide superset of exponential families —. We show that current deep architectures are able to factorize a very large number of such densities using an especially compact design, hence displaying the power of deep architectures and their concinnity in the f-GAN game. This result holds given a sufficient condition on activation functions — which turns out to be satisfied by popular choices. The key to our results is a variational generalization of an old theorem that relates the KL divergence between regular exponential families and divergences between their natural parameters. We complete this picture with additional results and experimental insights on how these results may be used to ground further improvements of GAN architectures, via (i) a principled design of the activation functions in the generator and (ii) an explicit integration of proper composite losses' link function in the discriminator.
- **Conclusion** :It is hard to exaggerate the success of GAN approaches in modelling complex domains, and with their success comes an increasing need for a rigorous theoretical understanding [34]. In this paper, we complete the supervised understanding of the generalization of GANs introduced in [30], and provide a theoretical background to understand its unsupervised part, showing in particular how deep architectures can be powerful at tackling the generative part of the game. Experiments display that the tools we develop may help to improve further the state of the art.

Published on NIPS in 2017

f-GANs in an Information Geometric Nutshell

Richard Nock[†,‡,§] **Zac Cranko**[‡,†] **Aditya Krishna Menon**[†,‡]
Lizhen Qu[†,‡] **Robert C. Williamson**[‡,†]
[†]Data61, [‡]the Australian National University and [§]the University of Sydney
{firstname.lastname, aditya.menon, bob.williamson}@data61.csiro.au

- **Review** :This paper considers the f-GAN principle generalized to f-divergences. Generalizing the connection between the KL divergence and regular exponential families by using the chi-logarithm, the authors prove the variational information geometric f-GAN (vig-f-GAN) identity, which provides an interpretation of the objective function of the f-GAN. The authors also prove a sufficient condition on activation functions to obtain factored generative distributions.
- This paper proposed an information geometric (IG) view of the f-GAN algorithm. It first showed that f-GAN converges in parameter space using the 1-1 mapping of f-divergence and chi-divergence and a Bregman divergence result. Then it also discussed a proper way to implement f-GAN. Finally in the main text it provided a factorisation result of the deep neural network representation and discussed the choice of activation functions which has 1-1 mapping to the chi (or f) function.

I didn't check every detail in the appendix but it seems to me that the proofs (except for Thm. 8 which I don't have time to read before due) are correct.

- he authors identify several interesting GAN-related questions, such as to what extend solving the GAN problem implies convergence in parameter space, what the generator is actually fitting when convergence occurs and (perhaps most relevant from a network architectural point of view) how to choose the output activation function of the discriminator so as to ensure proper compositeness of the loss function. The authors set out to address these questions within the (information theoretic) framework of deformed exponential distributions, from which they derive among other things the following theoretical results: They present a variational generalization (amenable to f-GAN formulation) of a known theorem that relates an f-divergence between distributions to a corresponding Bregman divergence between the parameters of such distributions.

Published on NIPS in 2017

f-GANs in an Information Geometric Nutshell

Richard Nock[†,‡,§] **Zac Cranko**[‡,†] **Aditya Krishna Menon**[†,‡]
Lizhen Qu[†,‡] **Robert C. Williamson**[‡,†]
[†]Data61, [‡]the Australian National University and [§]the University of Sydney
{firstname.lastname, aditya.menon, bob.williamson}@data61.csiro.au

- **Review** :As such, this theorem provides an interesting connection between the information-theoretic view point of measuring dissimilarity between probability distributions and the information-geometric perspective of measuring dissimilarity between the corresponding parameters. They show that under a reversibility assumption, deep generative networks factor as so called escorts of deformed exponential distributions. Their theoretical investigation furthermore suggests that a careful choice of the hidden activation functions of the generator as well as a proper selection of the output activation function of the discriminator could potentially help to further improve GANs. They also briefly mention an alternative interpretation of the GAN game in the context of expected utility theory.

- One of the author's main contributions, the information-geometric f-GAN identity in Eq.(7), relates the well-known variational f-divergence formulation over distributions to an information-geometric optimization problem over parameters. Can the authors explain what we gain from this parameter-based point of view? Is it possible to implement GANs in terms of this parameter-based optimization and what would be the benefits? I would really have liked to see experimental results comparing the two approaches to optimization of GANs. Somewhat surprising, the authors don't seem to make use of the right hand side of this identity, other than to assert that solving the GAN game implies convergence in parameter space (provided the residual J(Q) is small). And why is this implication not obvious? Can the authors give a realistic scenario where convergence in the variational f-divergence formulation over distributions does not imply convergence in parameter space?

https://www.facebook.com/nipsfoundation/videos/1554402331317667/

Published on NIPS in 2017

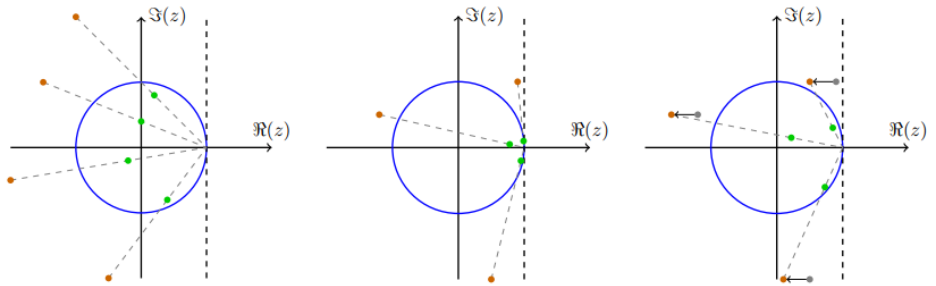
The Numerics of GANs

Lars Mescheder
Autonomous Vision Group
MPI Tübingen
lars.mescheder@tuebingen.mpg.de

Sebastian Nowozin
Machine Intelligence and Perception Group
Microsoft Research
sebastian.nowozin@microsoft.com

Andreas Geiger
Autonomous Vision Group
MPI Tübingen
andreas.geiger@tuebingen.mpg.de

- **Abstract** :In this paper, we analyze the numerics of common algorithms for training Generative Adversarial Networks (GANs). Using the formalism of smooth two-player games we analyze the associated gradient vector field of GAN training objectives. Our findings suggest that the convergence of current algorithms suffers due to two factors: i) presence of eigenvalues of the **Jacobian** of the gradient vector field with zero real-part, and ii) eigenvalues with big imaginary part. Using these findings, we design **a new algorithm** that overcomes some of these limitations and has better convergence properties. Experimentally, we demonstrate its superiority on training common GAN architectures and show convergence on GAN architectures that are known to be notoriously hard to train.



(a) Illustration how the eigenvalues are projected into unit ball. (b) Example where h has to be chosen extremely small. (c) Illustration how our method alleviates the problem.

Figure 1: Images showing how the eigenvalues of A are projected into the unit circle and what causes problems: when discretizing the gradient flow with step size h, the eigenvalues of the Jacobian at a fixed point are projected into the unit ball along rays from 1. However, this is only possible if the eigenvalues lie in the left half plane and requires extremely small step sizes h if the eigenvalues are close to the imaginary axis. The proposed method moves the eigenvalues to the left in order to make the problem better posed, thus allowing the algorithm to converge for reasonable step sizes.

- **Conclusion** :In this work, starting from GAN objective functions we analyzed the general difficulties of finding local Nash-equilibria in smooth two-player games. We pinpointed the major numerical difficulties that arise in the current state-of-the-art algorithms and, using our insights, we presented a new algorithm for training generative adversarial networks. Our novel algorithm has favorable properties in theory and practice: from the theoretical viewpoint, we showed that it is locally convergent to a Nash equilibrium even if the eigenvalues of the Jacobian are problematic.

Published on NIPS in 2017

The Numerics of GANs

Ilya Tolstikhin
MPI for Intelligent Systems
Tübingen, Germany
ilya@tue.mpg.de

Sylvain Gelly
Google Brain
Zürich, Switzerland
sylvaingelly@google.com

Olivier Bousquet
Google Brain
Zürich, Switzerland
obousquet@google.com

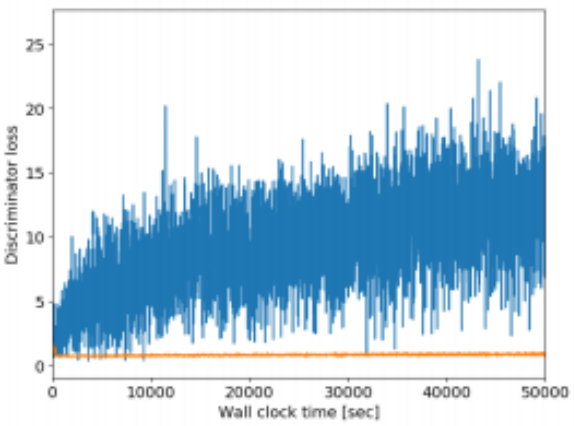
Carl-Johann Simon-Gabriel
MPI for Intelligent Systems
Tübingen, Germany
cjsimon@tue.mpg.de

Bernhard Schölkopf
MPI for Intelligent Systems
Tübingen, Germany
bs@tue.mpg.de

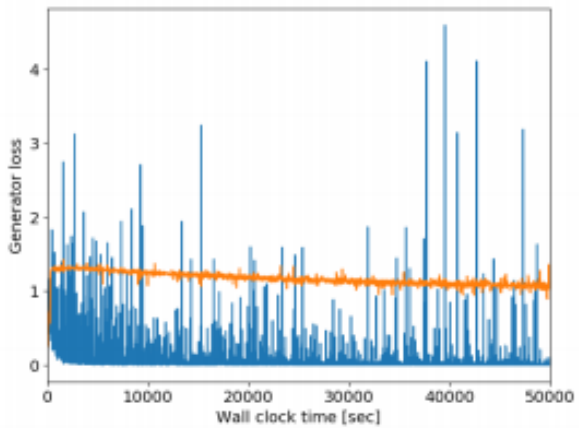
- **Conclusion** : This is particularly interesting for games that arise in the context of GANs where such problems are common. From the practical viewpoint, our algorithm can be used in combination with any GAN-architecture whose objective can be formulated as a two-player game to stabilize the training. We demonstrated experimentally that our algorithm stabilizes the training and successfully combats training issues like mode collapse. We believe our work is a first step towards an understanding of the numerics of GAN training and more general deep learning objective functions.

Algorithm 2 Consensus optimization

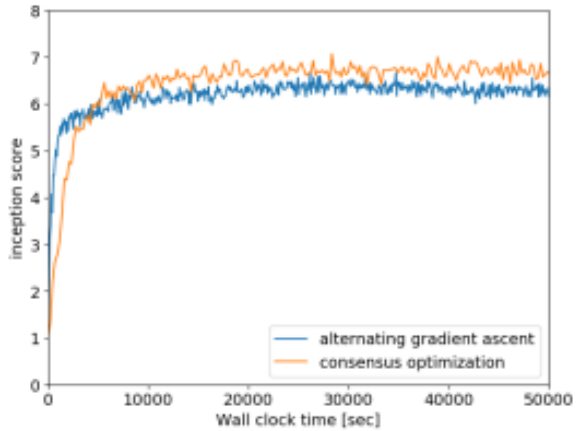
while not converged **do**
  $v_\phi \leftarrow \nabla_\phi(f(\theta,\phi) - \gamma L(\theta,\phi))$
  $v_\theta \leftarrow \nabla_\theta(g(\theta,\phi) - \gamma L(\theta,\phi))$
  $\phi \leftarrow \phi + h v_\phi$
  $\theta \leftarrow \theta + h v_\theta$
end while



(a) Discriminator loss



(b) Generator loss



(c) Inception score

Published on NIPS in 2017

The Numerics of GANs

Ilya Tolstikhin
MPI for Intelligent Systems
Tübingen, Germany
ilya@tue.mpg.de

Sylvain Gelly
Google Brain
Zürich, Switzerland
sylvaingelly@google.com

Olivier Bousquet
Google Brain
Zürich, Switzerland
obousquet@google.com

Carl-Johann Simon-Gabriel
MPI for Intelligent Systems
Tübingen, Germany
cjsimon@tue.mpg.de

Bernhard Schölkopf
MPI for Intelligent Systems
Tübingen, Germany
bs@tue.mpg.de

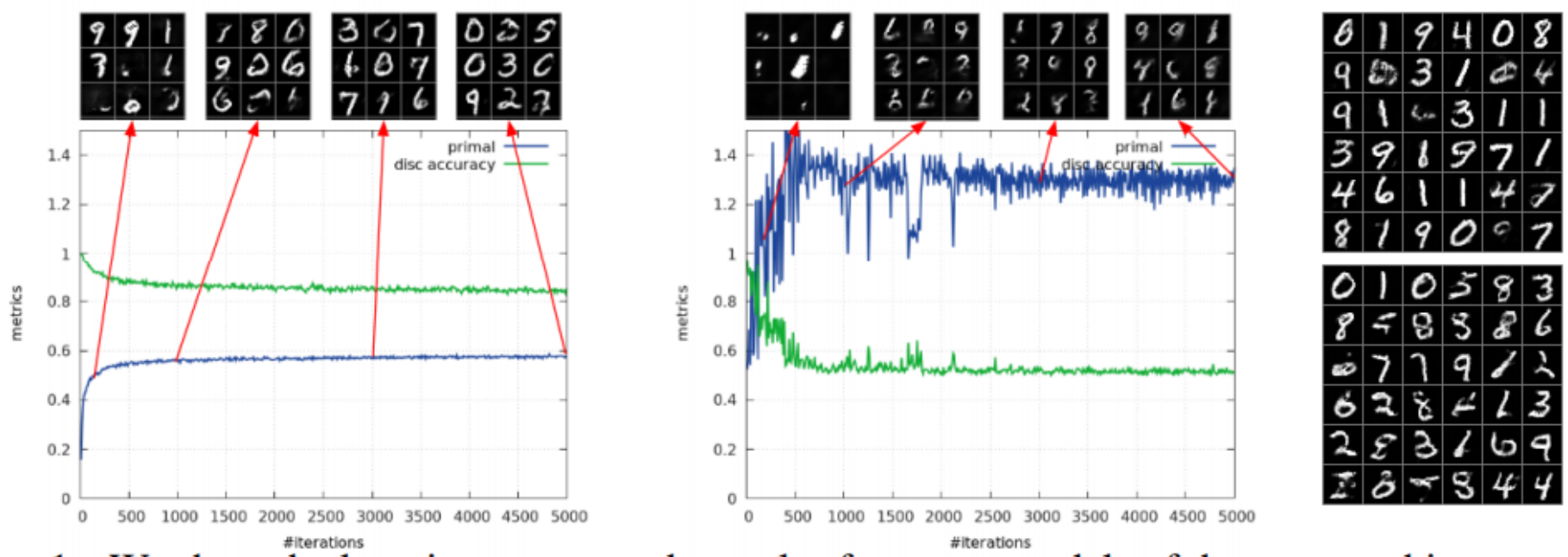


Figure 1: We show the learning curves and samples from two models of the same architecture, one optimized in dual space (left), and one in the primal space (*i.e.*, typical GAN) up to 5000 iterations. Samples are shown at different points during training, as well as at the very end (right top - dual, right bottom - primal). Despite having similar sample qualities in the end, they demonstrate drastically different training behavior. In the typical GAN setup, loss oscillates and has no clear trend, whereas in the dual setup, loss monotonically increases and shows much smaller oscillation. Sample quality is nicely correlated with the dual objective during training.

The Numerics of GANs

- **Review** :This paper presents a novel analysis of the typical optimization algorithm used in GANs (simultaneous gradient ascent) and identifies problematic failures when the Jacobian has large imaginary components or zero real components. Motivated by these failures, they present a novel consensus optimization algorithm for training GANs. The consensus optimization is validated on a toy MoG dataset as well as CIFAR-10 and CelebA in terms of sample quality and inception score.
- This is a very nice, elegant paper on the optimization challenges of GANs and a simple, well motivated fix for those problems. The continuous-time perspective on optimization makes it very clear why minimax problems suffer from oscillations that regular optimization problems do not, and the proposed solution is both theoretically motivated and simple enough to be practical. I anticipate this will be useful for a wider variety of nested optimization problems than just GANs and should get a wide viewing.

- The paper presents an analysis of the optimization of GANS by analyzing the jacobian of the vector field of the zero sum game. A saddle point exists if this jacobian is negative definite . The paper proposes to penalize the norms of gradients with respect to the parameters of the generator and the discriminator in order to ensure such conditions.

Published on NIPS in 2017

Dualing GANs

Yujia Li[1*] **Alexander Schwing**[3] **Kuan-Chieh Wang**[1,2] **Richard Zemel**[1,2]
[1]Department of Computer Science, University of Toronto [2]Vector Institute
[3]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
{yujiali, wangkua1, zemel}@cs.toronto.edu aschwing@illinois.edu

- **Abstract** :Generative adversarial nets (GANs) are a promising technique for modeling a distribution from samples. It is however well known that GAN training suffers from instability due to the nature of its saddle point formulation. In this paper, we explore ways to tackle the instability problem by dualizing the discriminator. We start from linear discriminators in which case conjugate duality provides a mechanism to reformulate the saddle point objective into a maximization problem, such that both the generator and the discriminator of this ‘dualing GAN’ act in concert. We then demonstrate how to extend this intuition to non-linear formulations. For GANs with linear discriminators our approach is able to remove the instability in training,while for GANs with nonlinear discriminators our approach provides an alternative to the commonly used GAN training algorithm.

- **Conclusion** :To conclude, we introduced ‘Dualing GANs,’ a framework which considers duality based formulations for the duel between the discriminator and the generator. Using the dual formulation provides opportunities to better train the discriminator. This helps remove the instability in training for linear discriminators, and we also adapted this framework to non-linear discriminators. The dual formulation also provides connections to other techniques. In particular, we discussed a close link to moment matching techniques, and showed that the cost function linearization for non-linear discriminators recovers the original gradient direction in standard GANs. We hope that our results spur further research in this direction to obtain a better understanding of the GAN objective and its intricacies.

Dualing GANs

- **Review** :To avoid the instability of the training curves present when training GANs generally, the paper proposes to solve the dual optimization problem for the discriminator rather than the primal problem, which is the standard GAN's formulation. The results show that dualizing the discriminator yields to monotonous training curves, which is not the case when considering the standard formulation where the training curves are spiky.
-
- Adopting the standard GAN approach, which is well-known in the literature, the paper proposes a new interesting way to solve the GAN's learning problem. The idea of dualizing the optimization problem is well motivated since it's fairly natural. Moreover, the proposed method is mathematically relevant and clearly validated by experiments. The paper is well written and the adopted formalism is consistent. Furthermore, the supplementary materials detail the experiments and the proofs of the main results, which are important to the paper's understanding.

- This submission proposes an alternative view on GAN training by replacing the inner learning of the discriminator by its dual program. In the case of linear discriminators this leads to a stable and simpler maximization problem. The authors propose strategies to cope with non-linear discriminators, by either taking linear approximations to the cost function, or to the scoring function (typically the CNN in the case of image GANs).
-
- This paper proposes a new method to train GAN: firstly it assumes the discriminator is linear, then the GAN model is solved by dual optimization, secondly since the discriminator is non-linear, to train the GAN model using dual optimization, local linearization for the cost or score function is repeated used. Experiments show the effectiveness of the proposed approach.

Published on NIPS in 2017

Fisher GAN

Youssef Mroueh*, **Tom Sercu***
mroueh@us.ibm.com, tom.sercu1@ibm.com
∗ Equal Contribution
AI Foundations, IBM Research AI
IBM T.J Watson Research Center

- **Abstract** :Generative Adversarial Networks (GANs) are powerful models for learning complex distributions. Stable training of GANs has been addressed in many recent works which explore different metrics between distributions. In this paper we introduce Fisher GAN which fits within the Integral Probability Metrics (IPM) framework for training GANs. Fisher GAN defines a critic with a data dependent constraint on its second order moments. We show in this paper that Fisher GAN allows for stable and time efficient training that does not compromise the capacity of the critic, and does not need data independent constraints such as weight clipping. We analyze our Fisher IPM theoretically and provide an algorithm based on Augmented Lagrangian for Fisher GAN. We validate our claims on both image sample generation and semi-supervised classification using Fisher GAN.

Table 1: Comparison between Fisher GAN and recent related approaches.

	Stability	Unconstrained capacity	Efficient Computation	Representation power (SSL)
Standard GAN [1, 9]	✗	✓	✓	✓
WGAN, McGan [6, 8]	✓	✗	✓	✗
WGAN-GP [7]	✓	✓	✗	?
Fisher Gan (Ours)	✓	✓	✓	✓

- **Conclusion** :We have defined Fisher GAN, which provide a stable and fast way of training GANs. The Fisher GAN is based on a scale invariant IPM, by constraining the second order moments of the critic. We provide an interpretation as whitened (Mahalanobis) mean feature matching and 2 distance. We show graceful theoretical and empirical advantages of our proposed Fisher GAN.

Published on NIPS in 2017

Fisher GAN

Youssef Mroueh*, **Tom Sercu***
mroueh@us.ibm.com, tom.sercu1@ibm.com
* Equal Contribution
AI Foundations, IBM Research AI
IBM T.J Watson Research Center

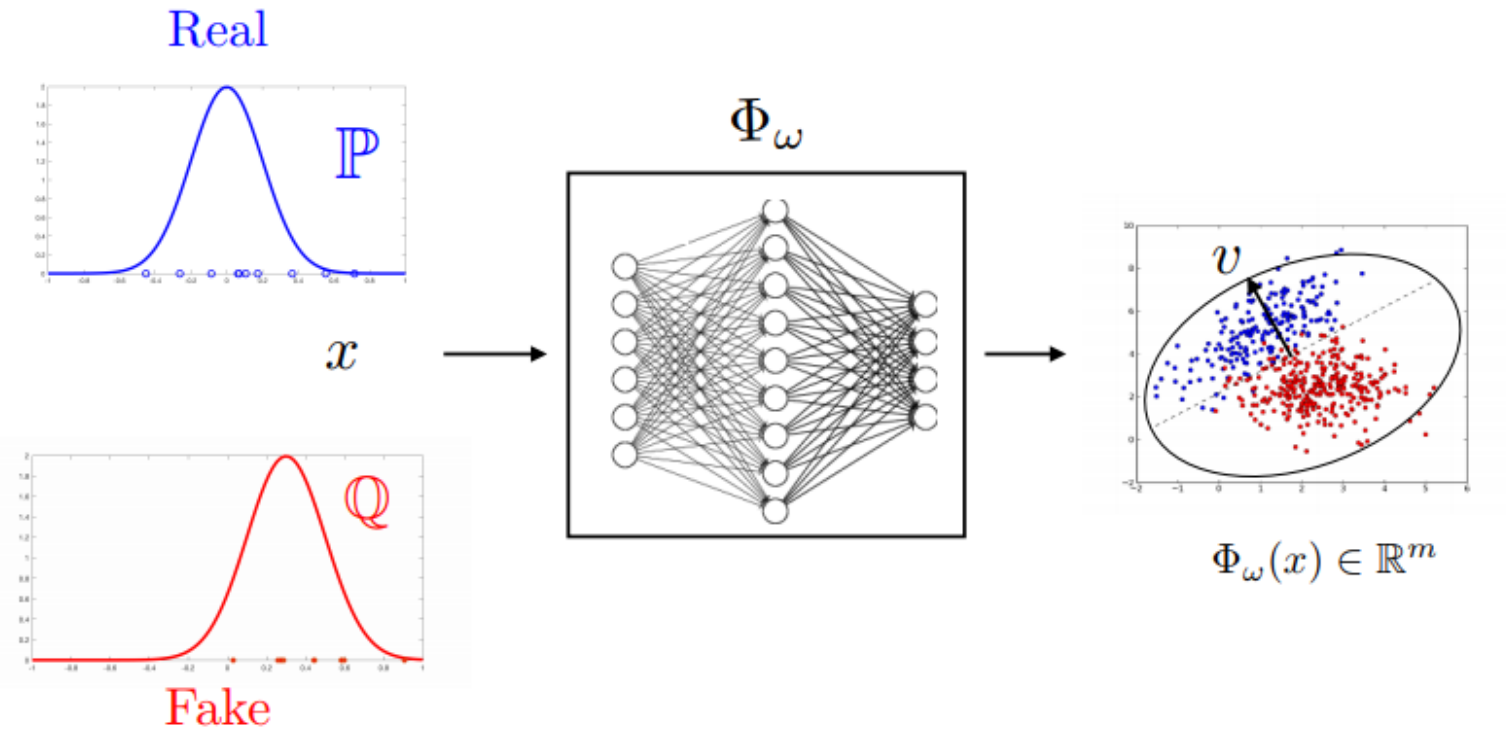


Figure 1: Illustration of Fisher IPM with Neural Networks. Φ_ω is a convolutional neural network which defines the embedding space. v is the direction in this embedding space with maximal mean separation $\langle v, \mu_\omega(\mathbb{P}) - \mu_\omega(\mathbb{Q})\rangle$, constrained by the hyperellipsoid $v^\top \Sigma_\omega(\mathbb{P};\mathbb{Q})\, v = 1$.

Published on NIPS in 2017

Fisher GAN

Youssef Mroueh*, Tom Sercu*
mroueh@us.ibm.com, tom.sercu1@ibm.com
* Equal Contribution
AI Foundations, IBM Research AI
IBM T.J Watson Research Center

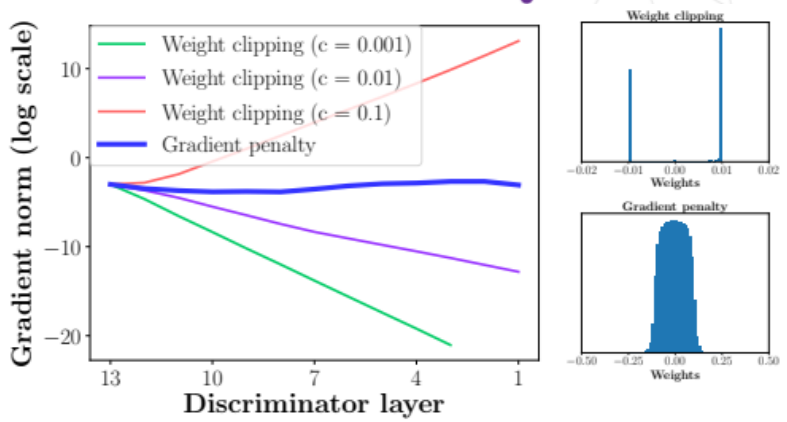
- **Review** :In this work, authors present a novel framework for training GAN based on Integral Probability Metrics (IPM). The different notions are introduced gradually which makes the article easy to read. Proof are provided in appendix. Experiments are well managed and complete.
- Overall the paper is mathematically strong and provides a good and well-founded basis for constraining the discriminator in an adversarial framework trained on IPM. The method is indeed simpler than WGAN-gp and provides a way of constraining the discriminator without the sort of clipping in the original WGAN paper
- 1) The constraining factors do indeed have computational advantages over WGAP-gp, where you have to compute the norm of the gradients backpropped onto the input space. Instead we only need to compute the covariance, which is in the simple case just the expectation of the square of the outputs.
- However, looking at this it seems possible that E[f^2] over samples from the generator go to 0 and 1 over the real samples. In this case, it seems like the gradients in the neighborhood of the samples could still go to 0 with this constraint and a neural network discriminator while the distance given the optimal discriminator are non-zero. WGAN-gp, on the other hand, actually uses a combination of generated and real samples in its regularization term, instead of something over the 2 distributions independently. This likely breaks a symmetry implicit in the Fisher IPM-based method that keeps the gradients from vanishing in the neighborhood of the generated samples. Could the authors comment on this?
- The stability experiments are lacking. Notably, the authors show stability when the discriminator batch norm is off, but I'm fairly certain that turning the discriminator batch norm off is a strong source of instability. Rather, turning the batch norm off on the generator, for instance, is much stronger. It is interesting that WGAN/gp doesn't perform as well, but it's always unclear if this is due to hyperparameter choices. Why not run the same exps as they did in the WGAN to compare?

Published on NIPS in 2017

Improved Training of Wasserstein GANs

Ishaan Gulrajani[1]*, **Faruk Ahmed**[1], **Martin Arjovsky**[2], **Vincent Dumoulin**[1], **Aaron Courville**[1,3]
[1] Montreal Institute for Learning Algorithms
[2] Courant Institute of Mathematical Sciences
[3] CIFAR Fellow
igul222@gmail.com
{faruk.ahmed,vincent.dumoulin,aaron.courville}@umontreal.ca
ma4371@nyu.edu

- **Abstract** :Generative Adversarial Networks (GANs) are powerful generative models, but suffer from training instability. The recently proposed Wasserstein GAN (WGAN) makes progress toward stable training of GANs, but sometimes can still generate only poor samples or fail to converge. We find that these problems are often due to the use of weight clipping in WGAN to enforce a Lipschitz constraint on the critic, which can lead to undesired behavior. We propose an alternative to clipping weights: penalize the norm of gradient of the critic with respect to its input. Our proposed method performs better than standard WGAN and enables stable training of a wide variety of GAN architectures with almost no hyperparameter tuning, including 101-layer ResNets and language models with continuous generators. We also achieve high quality generations on CIFAR-10 and LSUN bedrooms.



- **Conclusion** :In this work, we demonstrated problems with weight clipping in WGAN and introduced an alternative in the form of a penalty term in the critic loss which does not exhibit the same problems. Using our method, we demonstrated strong modeling performance and stability across a variety of architectures. Now that we have a more stable algorithm for training GANs, we hope our work opens the path for stronger modeling performance on large-scale image datasets and language. Another interesting direction is adapting our penalty term to the standard GAN objective function, where it might stabilize training by encouraging the discriminator to learn smoother decision boundaries.

Published on NIPS in 2017

MMD GAN: Towards Deeper Understanding of Moment Matching Network

Chun-Liang Li[1,*] **Wei-Cheng Chang**[1,*] **Yu Cheng**[2] **Yiming Yang**[1] **Barnabás Póczos**[1]
[1] Carnegie Mellon University, [2]AI Foundations, IBM Research
{chunlial,wchang2,yiming,bapoczos}@cs.cmu.edu chengyu@us.ibm.com
(* denotes equal contribution)

- **Abstract** :Generative moment matching network (GMMN) is a deep generative model that differs from Generative Adversarial Network (GAN) by replacing the discriminator in GAN with a two-sample test based on kernel maximum mean discrepancy (MMD). Although some theoretical guarantees of MMD have been studied, the empirical performance of GMMN is still not as competitive as that of GAN on challenging and large benchmark datasets. The computational efficiency of GMMN is also less desirable in comparison with GAN, partially due to its requirement for a rather large batch size during the training. In this paper, we propose to improve both the model expressiveness of GMMN and its computational efficiency by introducing adversarial kernel learning techniques, as the replacement of a fixed Gaussian kernel in the original GMMN. The new approach combines the key ideas in both GMMN and GAN, hence we name it MMD GAN.

The new distance measure in MMD GAN is a meaningful loss that enjoys the advantage of weak← topology and can be optimized via gradient descent with relatively small batch sizes. In our evaluation on multiple benchmark datasets, including MNIST, CIFAR-10, CelebA and LSUN, the performance of MMD GAN significantly outperforms GMMN, and is competitive with other representative GAN works.

$$\min_{\theta} \max_{\phi} M_{f_{\phi_e}}(\mathbb{P}(\mathcal{X}), \mathbb{P}(g_\theta(\mathcal{Z}))) - \lambda \mathbb{E}_{y \in \mathcal{X} \cup g(\mathcal{Z})} \|y - f_{\phi_d}(f_{\phi_e}(y))\|^2.$$

Algorithm 1: MMD GAN, our proposed algorithm.

input : α the learning rate, c the clipping parameter, B the batch size, n_c the number of iterations o discriminator per generator update.

initialize generator parameter θ and discriminator parameter ϕ;

while θ *has not converged* **do**

  for $t = 1, \ldots, n_c$ **do**

    Sample a minibatches $\{x_i\}_{i=1}^B \sim \mathbb{P}(\mathcal{X})$ and $\{z_j\}_{j=1}^B \sim \mathbb{P}(\mathcal{Z})$

    $g_\phi \leftarrow \nabla_\phi M_{f_{\phi_e}}(\mathbb{P}(\mathcal{X}), \mathbb{P}(g_\theta(\mathcal{Z}))) - \lambda \mathbb{E}_{y \in \mathcal{X} \cup g(\mathcal{Z})} \|y - f_{\phi_d}(f_{\phi_e}(y))\|^2$

    $\phi \leftarrow \phi + \alpha \cdot \text{RMSProp}(\phi, g_\phi)$

    $\phi \leftarrow \text{clip}(\phi, -c, c)$

  Sample a minibatches $\{x_i\}_{i=1}^B \sim \mathbb{P}(\mathcal{X})$ and $\{z_j\}_{j=1}^B \sim \mathbb{P}(\mathcal{Z})$

  $g_\theta \leftarrow \nabla_\theta M_{f_{\phi_e}}(\mathbb{P}(\mathcal{X}), \mathbb{P}(g_\theta(\mathcal{Z})))$

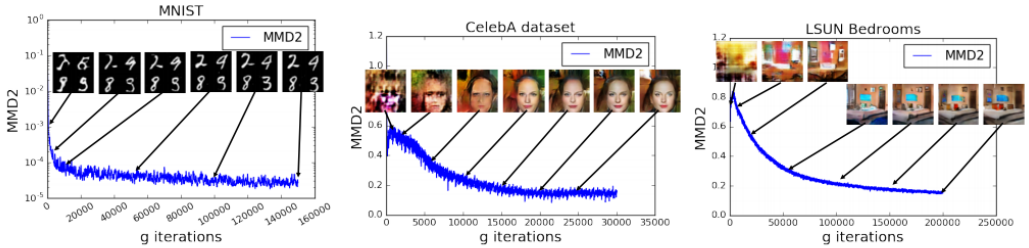
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$

MMD GAN: Towards Deeper Understanding of Moment Matching Network

Chun-Liang Li[1,*] **Wei-Cheng Chang**[1,*] **Yu Cheng**[2] **Yiming Yang**[1] **Barnabás Póczos**[1]
[1] Carnegie Mellon University, [2]AI Foundations, IBM Research
{chunlial,wchang2,yiming,bapoczos}@cs.cmu.edu chengyu@us.ibm.com
(* denotes equal contribution)

- **Conclusion** :We introduce a new deep generative model trained via MMD with adversarially learned kernels. We further study its theoretical properties and propose a practical realization MMD GAN, which can be trained with much smaller batch size than GMMN and has competitive performances with state-of-the art GANs. We can view MMD GAN as the first practical step forward connecting moment matching network and GAN. One important direction is applying developed tools in moment matching [15] on general GAN works based the connections shown by MMD GAN. Also, in Section 4, we connect WGAN and MMD GAN by first-order and infinite-order moment matching. [24] shows finite-order moment matching (↞ 5) achieves the best performance on domain adaption. One could extend MMD GAN to this by using polynomial kernels. Last, in theory, an injective mapping f is necessary for the theoretical guarantees.

However, we observe that it is not mandatory in practice as we show in Section 5.5. One conjecture is it usually learns the injective mapping with high probability by parameterizing with neural networks, which worth more study as a future work.



(a) MNIST (b) CelebA (c) LSUN Bedrooms

Figure 4: Training curves and generative samples at different stages of training. We can see a clear correlation between lower distance and better sample quality.

Method	Scores ± std.
Real data	11.95 ± .20
DFM [36]	7.72
ALI [37]	5.34
Improved GANs [28]	4.36
MMD GAN	**6.17** ± .07
WGAN	5.88 ± .07
GMMN-C	3.94 ± .04
GMMN-D	3.47 ± .03

Table 1: Inception scores

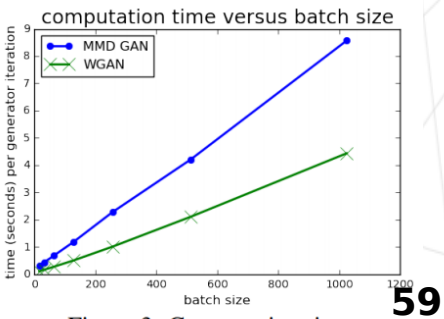


Figure 3: Computation time

Published on NIPS in 2017

GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium

Martin Heusel **Hubert Ramsauer** **Thomas Unterthiner** **Bernhard Nessler**

Sepp Hochreiter

LIT AI Lab & Institute of Bioinformatics,
Johannes Kepler University Linz
A-4040 Linz, Austria
{mhe,ramsauer,unterthiner,nessler,hochreit}@bioinf.jku.at

- **Abstract** :Generative Adversarial Networks (GANs) excel at creating realistic images with complex models for which maximum likelihood is infeasible. However, the convergence of GAN training has still not been proved. We propose a two time-scale update rule (TTUR) for training GANs with stochastic gradient descent on arbitrary GAN loss functions. TTUR has an individual learning rate for both the discriminator and the generator. Using the theory of stochastic approximation, we prove that the TTUR converges under mild assumptions to a stationary local Nash equilibrium. The convergence carries over to the popular Adam optimization, for which we prove that it follows the dynamics of a heavy ball with friction and thus prefers flat minima in the objective landscape. For the evaluation of the performance of GANs at image generation, we introduce the "Fréchet Inception Distance" (FID) which captures the similarity of generated images to real ones better than the Inception Score. In experiments, TTUR improves learning for DCGANs and Improved Wasserstein GANs (WGAN-GP) outperforming conventional GAN training on CelebA, CIFAR-10, SVHN, LSUN Bedrooms, and the One Billion Word Benchmark.

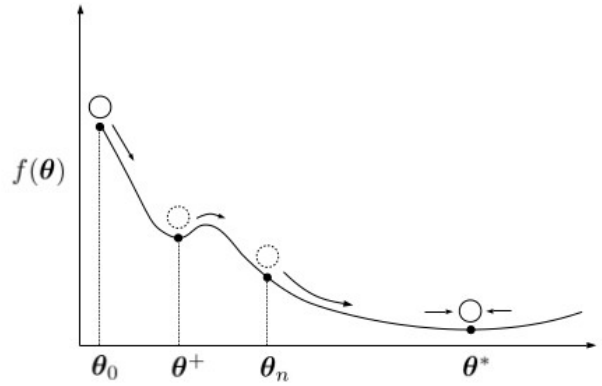


Figure 2: Heavy Ball with Friction, where the ball with mass overshoots the local minimum θ^+ and settles at the flat minimum θ^*.

Published on NIPS in 2017

GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium

Martin Heusel **Hubert Ramsauer** **Thomas Unterthiner** **Bernhard Nessler**

Sepp Hochreiter

LIT AI Lab & Institute of Bioinformatics,
Johannes Kepler University Linz
A-4040 Linz, Austria
{mhe,ramsauer,unterthiner,nessler,hochreit}@bioinf.jku.at

- **Conclusion** :For learning GANs, we have introduced the two time-scale update rule (TTUR), which we have proved to converge to a stationary local Nash equilibrium. Then we described Adam stochastic optimization as a heavy ball with friction (HBF) dynamics, which shows that Adam converges and that Adam tends to find flat minima while avoiding small local minima. A second order differential equation describes the learning dynamics of Adam as an HBF system. Via this differential equation, the convergence of GANs trained with TTUR to a stationary local Nash equilibrium can be extended to Adam. Finally, to evaluate GANs, we introduced the ‘Fréchet Inception Distance” (FID) which captures the similarity of generated images to real ones better than the Inception Score.

- In experiments we have compared GANs trained with TTUR to conventional GAN training with a one time-scale update rule on CelebA, CIFAR-10, SVHN, LSUN Bedrooms, and the One Billion Word Benchmark. TTUR outperforms conventional GAN training consistently in all experiments.

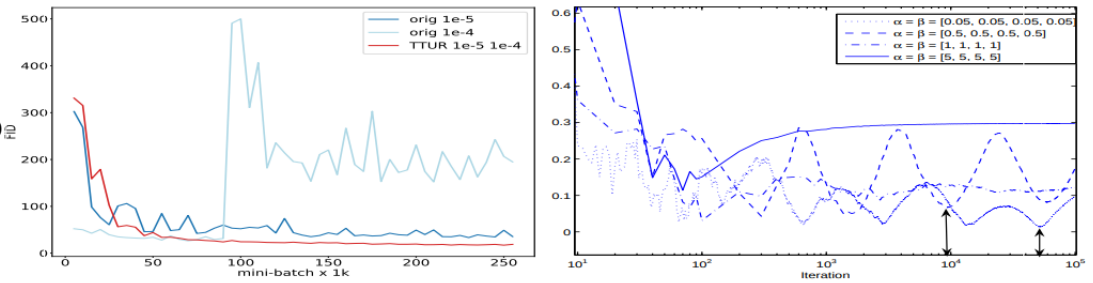


Figure 1: Left: Original vs. TTUR GAN training on CelebA. Right: Figure from Zhang 2007 [50] which shows the distance of the parameter from the optimum for a one time-scale update of a 4 node network flow problem. When the upper bounds on the errors (α, β) are small, the iterates oscillate and repeatedly return to a neighborhood of the optimal solution (cf. Supplement Section 2.3). However, when the upper bounds on the errors are large, the iterates typically diverge.

Published on NIPS in 2017

GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium

- **Review** :Generative adversarial networks (GANs) are turning out to be a very important advance in machine learning. Algorithms for training GANs have difficulties with convergence. The paper proposes a two time-scale update rule (TTUR) which is shown (proven) to converge under certain assumptions. Specifically, it shows that GAN Adam updates with TTUR can be expressed as ordinary differential equations, and therefore can be proved to converge using a similar approach as in Borkar' 1997 work. The recommendation is to use two different update rules for generator and discriminator, with the latter being faster, in order to have convergence guarantees.
- The authors of the submission propose a two time-scale update rule for GAN training, which is essentially about using two different learning rates (or learning rate schedules) for generator and discriminator. I see the two main contributions of the submission as follows: 1) a formal proof of GAN convergence under TTUR assumptions, and 2) introduction of FID (Frechet Inception Distance) score for more meaningful measurement of performance. Contribution 1) alone is providing a significant step toward the theoretical understanding of GAN training. Contribution 2) provides a useful tool for future evaluation of generative model performance, and its motivation is clear from Section 4.
- The paper proposes a two time-scale update rule (TTUR) for GAN training, which is to use different learning rates for the discriminator and the generator. Under certain assumptions, TTUR allows the proof of convergence of GANs, and the results carry to Adam. To evaluate the effectiveness of TTUR, the paper additionally proposed Frechet Inception Distance (FID), i.e., the approximate Frechet distance in the code space defined by the Inception model. Empirically, by tuning the two learning rates carefully, TTUR is able to achieve faster and more stable training as measured by FID, with comparable sample quality given the same FID score.

Published on NIPS in 2017

VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning

Akash Srivastava
School of Informatics
University of Edinburgh
akash.srivastava@ed.ac.uk

Lazar Valkov
School of Informatics
University of Edinburgh
L.Valkov@sms.ed.ac.uk

- **Abstract** :Deep generative models provide powerful tools for distributions over complicated manifolds, such as those of natural images. But many of these methods, including generative adversarial networks (GANs), can be difficult to train, in part because they are prone to **mode collapse**, which means that they characterize only a few modes of the true distribution. To address this, we introduce VEEGAN, which features a **reconstructor network,** reversing the action of the generator by mapping from data to noise. Our training objective retains the original asymptotic consistency guarantee of GANs, and can be interpreted as **a novel autoencoder** loss over the noise. In sharp contrast to a traditional autoencoder over data points, VEEGAN does not require specifying a loss function over the data, but rather only over the representations, which are standard normal by assumption. On an extensive set of synthetic and real world image datasets, VEEGAN indeed resists mode collapsing to a far greater extent than other recent GAN variants, and produces more realistic samples.

- **Objective function:**

$$\mathcal{O}(\gamma, \theta) = \mathrm{KL}\left[q_\gamma(x|z)p_0(z) \,\|\, p_\theta(z|x)p(x)\right] - E\left[\log p_0(z)\right] + E\left[d(z, F_\theta(x))\right].$$

	Stacked-MNIST		CIFAR-10
	Modes (Max 1000)	KL	IvOM
DCGAN	99	3.4	0.00844 ± 0.002
ALI	16	5.4	$\mathbf{0.0067} \pm 0.004$
Unrolled GAN	48.7	4.32	0.013 ± 0.0009
VEEGAN	**150**	**2.95**	$\mathbf{0.0068} \pm 0.0001$

Table 2: Degree of mode collapse, measured by modes captured and the inference via optimization measure (IvOM), and sample quality (as measured by KL) on Stacked-MNIST and CIFAR. VEEGAN captures the most modes and also achieves the highest quality.

- **Conclusion**:We have presented VEEGAN, a new training principle for GANs that combines a KL divergence in the joint space of representation and data points with an autoencoder over the representation space, motivated by a variational argument. Experimental results on synthetic data and real images show that our approach is much more effective than several state-of-the art GAN methods at avoiding mode collapse while still generating good quality samples.

Published on NIPS in 2017

VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning

- **Review** :This paper adds a reconstructor netowrk to reverse the action of the generator by mapping from data to noise, to solve the mode collapse problem of gan training. It obtains this reconstruction network by introducing an implicit variational principle to get an upper bound on the cross entropy between the reconstructor network of F_theta(X) and the original noise distribution of z. The paper also shows empirical results on MNIST, CIFAR10 and a synthetic dataset.
- GANs have received a lot of attention lately. They are very good at generating sharp samples in an unsupervised way. However this good sample generation performance comes with a cost: the optimization procedure can be unstable and might lead to mode collapsing -- which happens when the generator can only generate samples that lie in a subspace of the data manifold. There has been some works about providing solutions to this mode collapsing issue. This paper lies within that line of research. Several remarks on the paper below:
- This paper proposes a variation of the GAN algorithm aimed at improving the mode coverage of the generative distribution. The main idea is to use three networks: a generator, an encoder and a discriminator where the discriminator takes pairs (x,z) as input. The criterion to optimize is a GAN style criterion for the discriminator and a reconstruction in latent space for the encoder (unlike auto-encoders where the reconstruction is in input space).
- The proposed criterion is properly motivated and experimental evidence is provided to support the claim that the new criterion allows to produce good quality samples (similarly to other GAN approaches) but with more diversity and coverage of the target distribution (i.e. more modes are covered) than other approaches. The results are fairly convincing and the algorithm is evaluated with several of the metrics used in previous literature on missing modes in generative models.

Published on NIPS in 2017

Generating steganographic images via adversarial training

Jamie Hayes
University College London
j.hayes@cs.ucl.ac.uk

George Danezis
University College London
The Alan Turing Institute
g.danezis@ucl.ac.uk

- **Abstract** :Adversarial training has proved to be competitive against supervised learning methods on computer vision tasks. However, studies have mainly been confined to generative tasks such as image synthesis. In this paper, we apply adversarial training techniques to the discriminative task of learning a steganographic algorithm. Steganography is a collection of techniques for **concealing the existence of information** by **embedding it within a non-secret medium, such as cover texts or images**. We show that adversarial training can produce robust steganographic techniques: our unsupervised training scheme produces a steganographic algorithm that competes with state-of-the-art steganographic techniques. We also show that supervised training of our adversarial model produces a robust steganalyzer, which performs the discriminative task of deciding if an image contains secret information.

- We define a game between three parties, Alice, Bob and Eve, in order to simultaneously train both a steganographic algorithm and a steganalyzer. Alice and Bob attempt to communicate a secret message contained within an image, while Eve eavesdrops on their conversation and attempts to determine if secret information is embedded within the image. We represent Alice, Bob and Eve by neural networks, and validate our scheme on two independent image datasets, showing our novel method of studying steganographic problems is surprisingly competitive against established steganographic techniques.

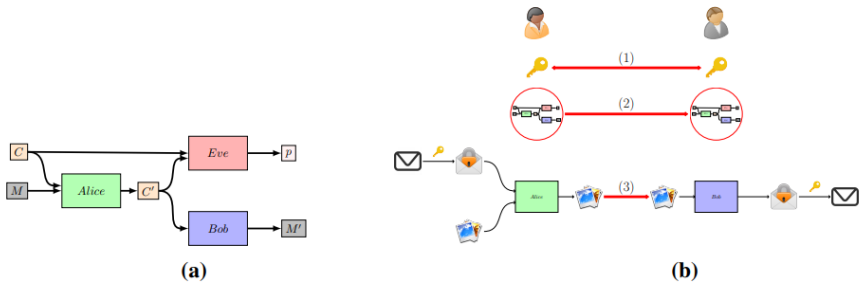


Figure 1: **(a)** Diagram of the training game. **(b)** How two parties, Carol and David, use the scheme in practice: (1) Two parties establish a shared key. (2) Carol trains the scheme on a set of images. Information about model weights, architecture and the set of images used for training is encrypted under the shared key and sent to David, who decrypts to create a local copy of the models. (3) Carol then uses the *Alice* model to embed a secret encrypted message, creating a steganographic image. This is sent to David, who uses the *Bob* model to decode the encrypted message and subsequently decrypt.

Published on NIPS in 2017

Generating steganographic images via adversarial training

Jamie Hayes
University College London
j.hayes@cs.ucl.ac.uk

George Danezis
University College London
The Alan Turing Institute
g.danezis@ucl.ac.uk

- **Conclusion** :We have offered substantial evidence that our hypothesis is correct and machine learning can be used effectively for both steganalysis and steganographic algorithm design. In particular, it is **competitive against** designs using human-based rules. By leveraging adversarial training games, we confirm that neural networks are able to discover steganographic algorithms, and furthermore, these steganographic algorithms perform well against state-of-the-art techniques. Our scheme does not require domain knowledge for designing steganographic schemes. We model the attacker as another neural network and show that this attacker has enough expressivity to perform well against a state-of-the-art steganalyzer. We expect this work to lead to fruitful avenues of further research. Finding the balance between cover image reconstruction loss, Bob’ s loss and Eve’ s loss to discover an effective embedding scheme is currently done via grid search, which is a time consuming process.

- Discovering a more refined method would greatly improve the efficiency of the training process. Indeed, discovering a method to quickly check whether the cover image has the capacity to accept a secret message would be a great improvement over the trial-and-error approach currently implemented. It also became clear that Alice and Bob learn their tasks after a relatively small number of training steps, further research is needed to explore if Alice and Bob fail to improve due to limitations in the model or because of shortcomings in the training scheme.

Published on NIPS in 2017

Generating steganographic images via adversarial training

Jamie Hayes
University College London
j.hayes@cs.ucl.ac.uk

George Danezis
University College London
The Alan Turing Institute
g.danezis@ucl.ac.uk

- **Review** :The authors studied how to use adversarial training to learn the encoder, the steganalyzer and the decoder at the same time using unsupervised learning. Specifically, the authors designed an adversarial game of three parties. The encoder generates images based on the cover and the message. Then the generated image can be received to both decoder and steganalyzer. The goal of the encoder and decoder is to correctly encode and decode the message, and the goal of the steganalyzer is to determine whether the image is encrypted. Thus it is a minimax game.
- This work presents a straightforward and exciting application of adversarial training to steganography: One network is trained to conceal a message in an image while a second one (the steganalyser) is trained to tell whether a given image contains a secret message or not. The framework is well suited to adversarial training, the paper is well written, and the necessary material relative to security is carefully introduced.
- The adversarial steganography based technique is compared to standard handcrafted methods using public datasets. The experiments show the superiority of the presented approach. My primary concern with this paper is the experimental setting: In practice, one always uses several steganalysers rather than just one. Besides, the quality of the steganographer is tied to the capacity of the steganalyser. Therefore, a stronger steganalyser will likely be able to tell legitimate images apart from altered ones. I understand this problem is more general to adversarial training and goes beyond this particular application. Nonetheless, its importance is particularly emphasized in this case and calls for further investigation. Despite this remark, I believe this is an exciting contribution that opens a new avenue for application of adversarial training to steganography. The originality of the work, the importance of the problem and the clarity of the presentation outweighs the previously mentioned point. I vote for acceptance.

Published on NIPS in 2017

Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference

Abhishek Kumar* IBM Research AI, Yorktown Heights, NY, abhishk@us.ibm.com

Prasanna Sattigeri* IBM Research AI, Yorktown Heights, NY, psattig@us.ibm.com

P. Thomas Fletcher University of Utah, Salt Lake City, UT, fletcher@sci.utah.edu

- **Abstract** :Semi-supervised learning methods using Generative adversarial networks (GANs) have shown promising empirical success recently. Most of these methods use a shared discriminator/classifier which discriminates real examples from fake while also predicting the class label. Motivated by the ability of the GANs generator to capture the data manifold well, we propose to estimate the tangent space to the data manifold using GANs and employ it to inject invariances into the classifier. In the process, we propose enhancements over existing methods for learning the inverse mapping (i.e., the **encoder**) which greatly improves in terms of semantic similarity of the reconstructed sample with the input sample. We observe considerable empirical gains in semi-supervised learning over baselines, particularly in the cases when the number of labeled examples is low. We also provide insights into how fake examples influence the semi-supervised learning procedure.

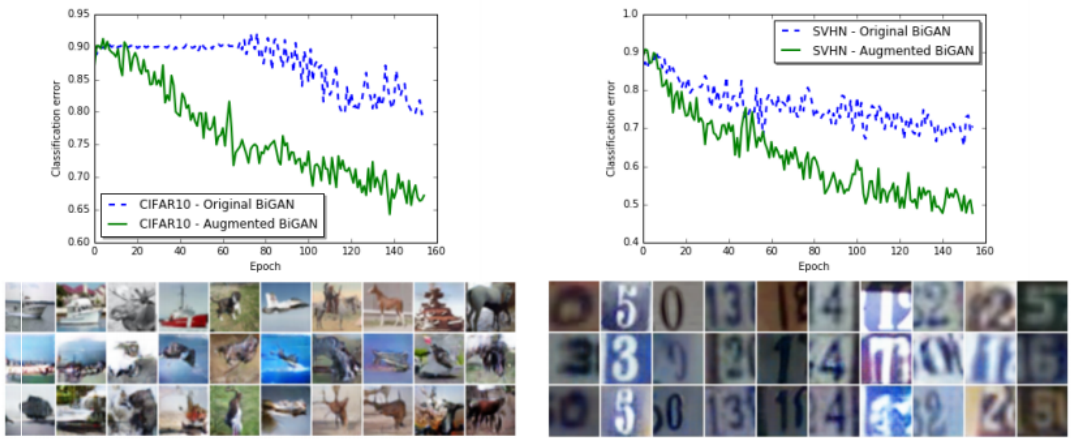


Figure 2: Comparing BiGAN with Augmented BiGAN based on the classification error on the reconstructed test images. *Left column:* CIFAR10, *Right column:* SVHN. In the images, the top row corresponds to the original images followed by BiGAN reconstructions in the middle row and the Augmented BiGAN reconstructions in the bottom row. More images can be found in the appendix.

$$L^f_{\text{unsup}} = \mathbb{E}_{x_g \sim p_g} \log \left(1 + \sum_{i=1}^{\kappa} e^{l_i(x_g)}\right) - \mathbb{E}_{x \sim p_d} \left[\log \sum_{i=1}^{\kappa} e^{l_i(x)} - \log \left(1 + \sum_{i=1}^{\kappa} e^{l_i(x)}\right)\right] \quad (5)$$

Taking the derivative w.r.t. discriminator's parameters θ followed by some basic algebra, we get $\nabla_\theta L^f_{\text{unsup}} =$

$$\mathbb{E}_{x_g \sim p_g} \sum_{i=1}^{k} p_f(y = i|x_g) \nabla l_i(x_g) - \mathbb{E}_{x \sim p_d} \left[\sum_{i=1}^{k} p_f(y = i|x, y \leq k) \nabla l_i(x) - \sum_{i=1}^{k} p_f(y = i|x) \nabla l_i(x)\right]$$

$$= \mathbb{E}_{x_g \sim p_g} \sum_{i=1}^{k} \underbrace{p_f(y = i|x_g)}_{a_i(x_g)} \nabla l_i(x_g) - \mathbb{E}_{x \sim p_d} \sum_{i=1}^{k} \underbrace{p_f(y = i|x, y \leq k) p_f(y = k+1|x)}_{b_i(x)} \nabla l_i(x) \quad (6)$$

$$L^f = L^f_{\text{sup}} + L^f_{\text{unsup}} + \lambda_1 \mathbb{E}_{x \sim p_d(x)} \sum_{v \in T_x} \|(J_x f)\, v\|_2^2 + \lambda_2 \mathbb{E}_{x \sim p_d(x)} \|J_x f\|_F^2.$$

Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference

- **Conclusion** :Our empirical results show that using the tangents of the data manifold (as estimated by the generator of the GAN) to inject invariances in the classifier improves the performance on semi-supevised learning tasks. In particular we observe impressive accuracy gains on SVHN (more so for the case of 500 labeled examples) for which the tangents obtained are good quality. We also observe improvements on CIFAR10 though not as impressive as SVHN. We think that improving on the quality of tangents for CIFAR10 has the potential for further improving the results there, which is a direction for future explorations. We also shed light on the effect of fake examples in the common framework used for semi-supervised learning with GANs where the discriminator predicts real class labels along with the fake label.

- Explicitly controlling the difficulty level of fake examples (i.e., pf (y = k + 1|xg) and hence indirectly pf (y = k + 1|x) in Eq. (6)) to do more effective semi-supervised learning is another direction for future work. One possible way to do this is to have a distortion model for the real examples (i.e., replace the generator with a distorter that takes as input the real examples) whose strength is controlled for more effective semi-supervised learning.

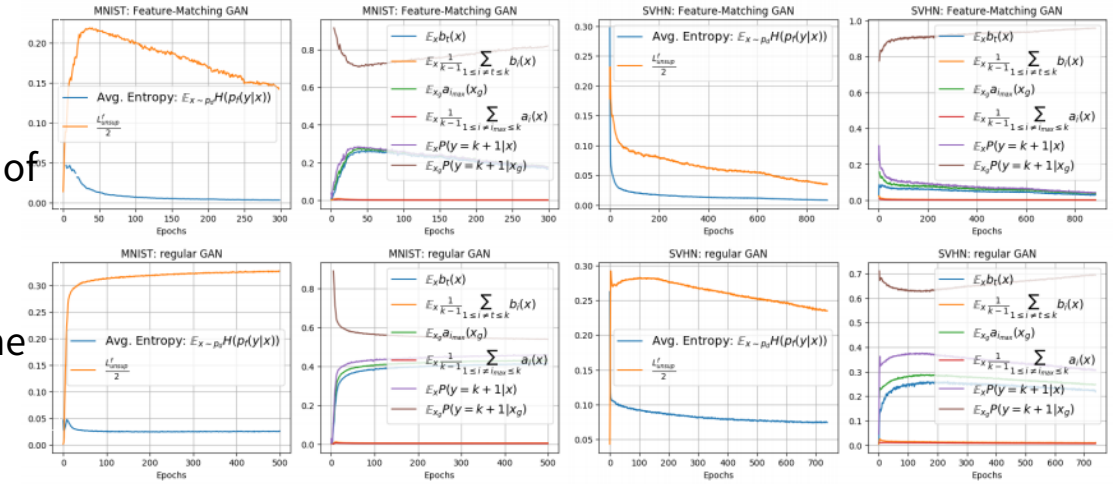


Figure 1: Plots of Entropy, L^f_{unsup} (Eq. (4)), $a_i(x_g)$, $b_i(x)$ and other probabilities (Eq. (6)) for regular GAN generator loss and feature-matching GAN generator loss.

Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference

- **Review** :he author(s) extend the idea of regularizing classifiers to be invariant to the tangent space of the learned manifold of the data to use GAN based architectures. This is a worthwhile idea to revisit as significant advances have been made in generative modeling in the intervening time since the last major paper in the area, the CAE was published.
- Crucial to the idea is the existence of an encoder learning an inverse mapping of the standard generator of GAN training. This is still an area of active research in the GAN literature that as of yet has no completely satisfactory approach.
- As current inference techniques for GANs are still quite poor, the authors propose two improvements to one technique, BiGAN, which are worthwhile contributions. 1) They adopt the feature matching loss proposed in "Improved techniques for training gans" and 2) they augment the BiGAN objective with another term that evaluates how the generator maps the inferred latent code for a given real example. Figure 2 provides good evidence of the usefulness of these modifications.
- Chiefly theoretical work with some empirical results on SVHN and CIFAR10. This paper proposes using a trained GAN to estimate mappings from and to the true data distribution around a data point, and use a kind of neural PCA to estimate tangents to those estimates, then used for training a manifold-invariant classifier. Some additional work investigating regular GAN vs feature matching GAN is briefly presented, and an augmentation to BiGAN.
- This paper describes a method for semi-supervised learning which is both adversarial and promotes the classifier's robustness to input variations. The GAN-based semi-supervised framework adopted in the paper is standard, treating the generated samples as an additional class to the regular classes that the classifier aims to label. What is new is the Jacobian-based regularizations that are introduced to encourage the classifier to be robust to local variations in the tangent space of the input manifold.

Published on NIPS in 2017

Bayesian GAN

Yunus Saatchi
Uber AI Labs

Andrew Gordon Wilson
Cornell University

- **Abstract** :Generative adversarial networks (GANs) can implicitly learn rich distributions over images, audio, and data which are hard to model with an explicit likelihood. We present a practical Bayesian formulation for **unsupervised and semi-supervised** learning with GANs. Within this framework, we use stochastic gradient Hamiltonian Monte Carlo to marginalize the weights of the generator and discriminator networks. The resulting approach is straightforward and obtains good performance without any standard interventions such as label smoothing or mini-batch discrimination. **By exploring an expressive posterior over the parameters of the generator, the Bayesian GAN avoids mode-collapse,** produces interpretable and diverse candidate samples, and provides state-of-the art quantitative results for semi-supervised learning on benchmarks including SVHN, CelebA, and CIFAR-10, outperforming DCGAN, Wasserstein GANs, and DCGAN ensembles.

Algorithm 1 One iteration of sampling for the Bayesian GAN. α is the friction term for SGHMC, η is the learning rate. We assume that the stochastic gradient discretization noise term $\hat{\beta}$ is dominated by the main friction term (this assumption constrains us to use small step sizes). We take J_g and J_d simple MC samples for the generator and discriminator respectively, and M SGHMC samples for each simple MC sample. We rescale to accommodate minibatches as in the supplementary material.

- Represent posteriors with samples $\{\theta_g^{j,m}\}_{j=1,m=1}^{J_g,M}$ and $\{\theta_d^{j,m}\}_{j=1,m=1}^{J_d,M}$ from previous iteration

for number of MC iterations J_g **do**

- Sample J_g noise samples $\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(J_g)}\}$ from noise prior $p(\mathbf{z})$. Each $\mathbf{z}^{(i)}$ has n_g samples.
- Update sample set representing $p(\theta_g|\theta_d)$ by running SGHMC updates for M iterations:

$$\theta_g^{j,m} \leftarrow \theta_g^{j,m} + \mathbf{v};\ \mathbf{v} \leftarrow (1-\alpha)\mathbf{v} + \eta \frac{\partial \log\left(\sum_i \sum_k p(\theta_g|\mathbf{z}^{(i)}, \theta_d^{k,m})\right)}{\partial \theta_g} + \mathbf{n};\ \mathbf{n} \sim \mathcal{N}(0, 2\alpha\eta I)$$

- Append $\theta_g^{j,m}$ to sample set.

end for

for number of MC iterations J_d **do**

- Sample minibatch of J_d noise samples $\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(J_d)}\}$ from noise prior $p(\mathbf{z})$.
- Sample minibatch of n_d data samples $\mathbf{x}$.
- Update sample set representing $p(\theta_d|\mathbf{z}, \theta_g)$ by running SGHMC updates for M iterations:

$$\theta_d^{j,m} \leftarrow \theta_d^{j,m} + \mathbf{v};\ \mathbf{v} \leftarrow (1-\alpha)\mathbf{v} + \eta \frac{\partial \log\left(\sum_i \sum_k p(\theta_d|\mathbf{z}^{(i)}, \mathbf{x}, \theta_g^{k,m})\right)}{\partial \theta_d} + \mathbf{n};\ \mathbf{n} \sim \mathcal{N}(0, 2\alpha\eta I)$$

- Append $\theta_d^{j,m}$ to sample set.

end for

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \int p(y_*|\mathbf{x}_*, \theta_d) p(\theta_d|\mathcal{D}) d\theta_d \approx \frac{1}{T} \sum_{k-1}^{T} p(y_*|\mathbf{x}_*, \theta_d^{(k)}), \theta_d^{(k)} \sim p(\theta_d|\mathcal{D}).$$

Bayesian GAN

- **Discussion** :By exploring rich multimodal distributions over the weight parameters of the generator, the Bayesian GAN can capture a diverse set of complementary and interpretable representations of data. We have shown that such representations can enable state of the art performance on semi-supervised problems, using a simple inference procedure.
- Effective semi-supervised learning of natural high dimensional data is crucial for reducing the dependency of deep learning on large labelled datasets. Often labeling data is not an option, or it comes at a high cost – be it through human labour or through expensive instrumentation (such as LIDAR for autonomous driving). Moreover, semi-supervised learning provides a practical and quantifiable mechanism to benchmark the many recent advances in unsupervised learning.
- Although we use MCMC, in recent years variational approximations have been favoured for inference in Bayesian neural networks.
- However, the likelihood of a deep neural network can be broad with many shallow local optima. This is exactly the type of density which is amenable to a sampling based approach, which can explore a full posterior. Variational methods, by contrast, typically centre their approximation along a single mode and also provide an overly compact representation of that mode. Therefore in the future we may generally see advantages in following a sampling based approach in Bayesian deep learning. Aside from sampling, one could try to better accommodate the likelihood functions common to deep learning using more general divergence measures (for example based on the family of α-divergences) instead of the KL divergence in variational methods; alongside more flexible proposal distributions.
- In the future, one could also estimate the marginal likelihood of a probabilistic GAN, having integrated away distributions over the parameters. The marginal likelihood provides a natural utility function for automatically learning hyperparameters, and for performing principled quantifiable model comparison between different GAN architectures.

Bayesian GAN

- **Review** :The paper introduces a Bayesian type of GAN algorithms, where the generator G and discriminator D do not have any fixed initial set of weights that gets gradually optimised. Instead, the weights for G and for D get sampled from two distributions (one for each), and it is those distributions that get iteratively updated. Different weight realisations of G may thus generate images with different styles, corresponding to different modes in the dataset. This, and the regularisation effect of the priors on the weights, promotes diversity and alleviates the mode collapse issue. The many experiments conducted in the paper support these claims.
- This paper provides a Bayesian formulation for GAN. The BayesGAN marginalizes the posteriors over the weights of the generator and discriminator using SGHMC. The quantitive experiment result is state-of-art, although the sample quality is not impressive.
- This work proposes a fully Bayesian approach to Generative Adversarial Nets (GANs) by defining priors over network weights (both the discriminator and the generator) and uses stochastic Gradient Hamiltonian Monte Carlo to sample from their posteriors. A fully Bayesian approach could allow for more extensive exploration of the state space and potentially avoid the mode collapse problem afflicting other GANs. Authors study this approach within the scope of semi-supervised learning and show that the proposed technique can achieve state of the art classification performance on several benchmark data sets but more importantly, can generate a more diversified artificial samples implying multimodal sampling behaviour for network weights.

Published on NIPS in 2017

Good Semi-supervised Learning That Requires a Bad GAN

Zihang Dai*, Zhilin Yang*, Fan Yang, William W. Cohen, Ruslan Salakhutdinov
School of Computer Science
Carnegie Melon University
dzihang,zhiliny,fanyang1,wcohen,rsalakhu@cs.cmu.edu

- **Abstract** :Semi-supervised learning methods based on generative adversarial networks (GANs) obtained strong empirical results, but it is not clear 1) how the discriminator benefits from joint training with a generator, and 2) why good semi-supervised classification performance and a good generator cannot be obtained at the same time. Theoretically we show that given the discriminator objective, good semi-supervised learning indeed requires a bad generator, and propose the definitionof a preferred generator. Empirically, we derive a novel formulation based on our analysis that substantially improves over feature matching GANs, obtaining state-of-the-art results on multiple benchmark datasets2.

Combining our solutions to the first two drawbacks of feature matching GANs, we have the following objective function of the generator:

$$\min_G \quad -\mathcal{H}(p_G) + \mathbb{E}_{x\sim p_G} \log p(x)\mathbb{I}[p(x) > \epsilon] + \|\mathbb{E}_{x\sim p_G} f(x) - \mathbb{E}_{x\sim \mathcal{U}} f(x)\|^2. \tag{4}$$

This objective is closely related to the idea of complement generator discussed in Section 3. To see that, let's first define a target complement distribution in the input space as follows

$$p^*(x) = \begin{cases} \frac{1}{Z}\frac{1}{p(x)} & \text{if } p(x) > \epsilon \text{ and } x \in \mathcal{B}_x \\ C & \text{if } p(x) \leq \epsilon \text{ and } x \in \mathcal{B}_x, \end{cases}$$

where Z is a normalizer, C is a constant, and $\mathcal{B}_x$ is the set defined by mapping $\mathcal{B}$ from the feature space to the input space. With the definition, the KL divergence (KLD) between $p_G(x)$ and $p^*(x)$ is

Published on NIPS in 2017

Good Semi-supervised Learning That Requires a Bad GAN

Zihang Dai*, Zhilin Yang*, Fan Yang, William W. Cohen, Ruslan Salakhutdinov
School of Computer Science
Carnegie Melon University
dzihang,zhiliny,fanyang1,wcohen,rsalakhu@cs.cmu.edu

5.4 Conditional Entropy

In order for the complement generator to work, according to condition (3) in Assumption 1, the discriminator needs to have strong true-fake belief on unlabeled data, i.e., $\max_{k=1}^{K} w_k^\top f(x) > 0$. However, the objective function of the discriminator in [16] does not enforce a dominant class. Instead, it only needs $\sum_{k=1}^{K} P_D(k|x) > P_D(K+1|x)$ to obtain a correct decision boundary, while the probabilities $P_D(k|x)$ for $k \leq K$ can possibly be uniformly distributed. To guarantee the strong true-fake belief in the optimal conditions, we add a conditional entropy term to the discriminator objective and it becomes,

$$\max_D \quad \mathbb{E}_{x,y\sim\mathcal{L}} \log p_D(y|x, y \leq K) + \mathbb{E}_{x\sim\mathcal{U}} \log p_D(y \leq K|x) + \\ \mathbb{E}_{x\sim p_G} \log p_D(K+1|x) + \mathbb{E}_{x\sim\mathcal{U}} \sum_{k=1}^{K} p_D(k|x) \log p_D(k|x). \tag{5}$$

By optimizing Eq. (5), the discriminator is encouraged to satisfy condition (3) in Assumption 1. Note that the same conditional entropy term has been used in other semi-supervised learning methods [19, 13] as well, but here we motivate the minimization of conditional entropy based on our theoretical analysis of GAN-based semi-supervised learning.

To train the networks, we alternatively update the generator and the discriminator to optimize Eq. (4) and Eq. (5) based on mini-batches. If an encoder is used to maximize $\mathcal{H}(p_G)$, the encoder and the generator are updated at the same time.

Good Semi-supervised Learning That Requires a Bad GAN

- **Conclusion** :In this work, we present a semi-supervised learning framework that uses generated data to boost task performance. Under this framework, we characterize the properties of various generators and theoretically prove that a complementary (i.e. bad) generator improves generalization. Empirically our proposed method improves the performance of image classification on several benchmark datasets.
- **Review:**This work extends and improves the performance of GAN based approaches to semi-supervised learning as explored in both "Improved techniques for training gans" (Salimans 2016) and "Unsupervised and semi-supervised learning with categorical generative adversarial networks" (Springenberg 2015).
- The paper introduces the notion of a complement generator which tries to sample from low-density areas of the data distribution (in feature space) and explores a variety of objective terms motivated/connected to this analysis.
- It is difficult to exactly match the motivated objective, due to various hard to estimate quantities like density and entropy, the paper uses a variety of approximations in their place.
- In addition to an illustrative case study on synthetic data, the paper has a suite of experiments on standardized semi-supervised tests including ablation studies on the various terms proposed. The overall empirical results are a significant improvement over the Feature Matching criteria proposed in "Improved techniques for training gans".
- In this paper, the authors proposed a novel semi-supervised learning algorithm based on GANs. It demonstrated that given the discriminator objective, good semi-supervised learning indeed requires a bad generator. They derive a novel formulation that substantially improves over feature matching GANs. Experiments demonstrate the state-of-the-art results on multiple benchmark datasets. The paper is interesting and well-written, which is important for the family of GANs algorithm.

PixelGAN Autoencoders

Alireza Makhzani, Brendan Frey
University of Toronto
{makhzani,frey}@psi.toronto.edu

- **Abstract** :In this paper, we describe the "PixelGAN autoencoder", a generative autoencoder in which the generative path is a convolutional autoregressive neural network on pixels (PixelCNN) that is conditioned on a latent code, and the recognition path uses a generative adversarial network (GAN) to impose a prior distribution on the latent code. We show that different priors result in different decompositions of information between the latent code and the autoregressive decoder. For example, by imposing a Gaussian distribution as the prior, we can achieve a global vs. local decomposition, or by imposing a categorical distribution as the prior, we can disentangle the style and content information of images in an unsupervised fashion. We further show how the PixelGAN autoencoder with a categorical prior can be directly used in semi-supervised settings and achieve competitive semi-supervised classification results on the MNIST. SVHN and NORB datasets.

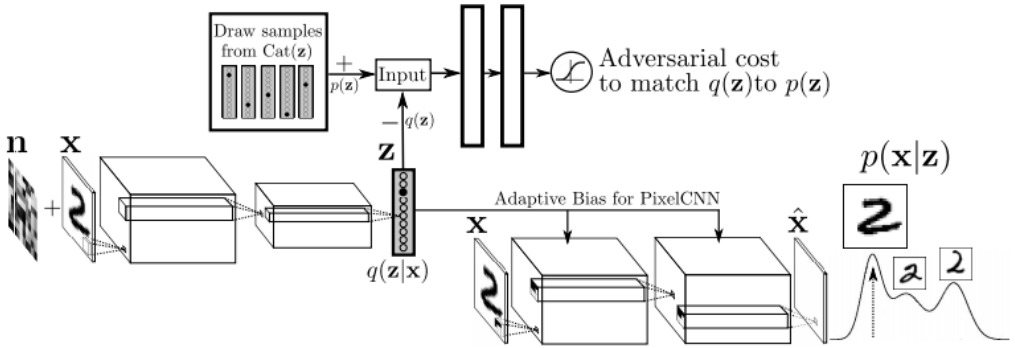


Figure 4: Architecture of the PixelGAN autoencoder with the categorical prior. $p(\mathbf{z})$ captures the class label and $p(\mathbf{x}|\mathbf{z})$ is a multi-modal distribution that captures the style distribution of a digit conditioned on the class label of that digit.

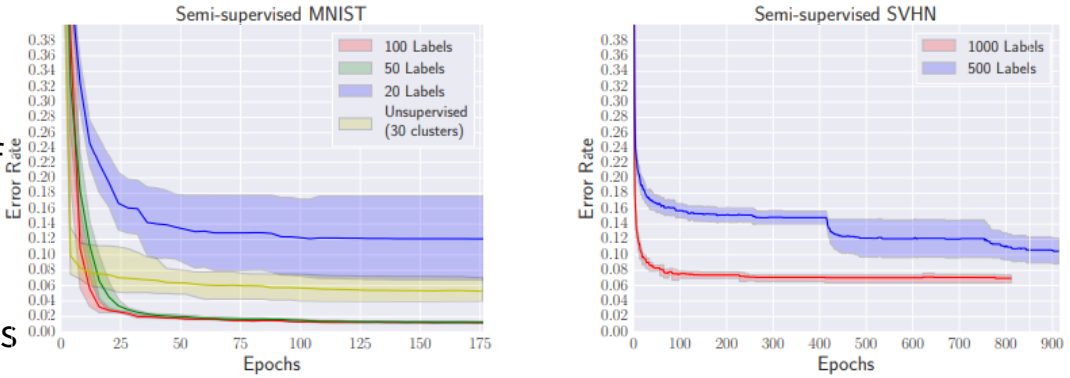


Figure 8: Semi-supervised error-rate of PixelGAN autoencoders on the MNIST and SVHN datasets

$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log p(\mathbf{x})] > -\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\Big[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[-\log p(\mathbf{x}|\mathbf{z})]\Big] - \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\Big[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))\Big]$$

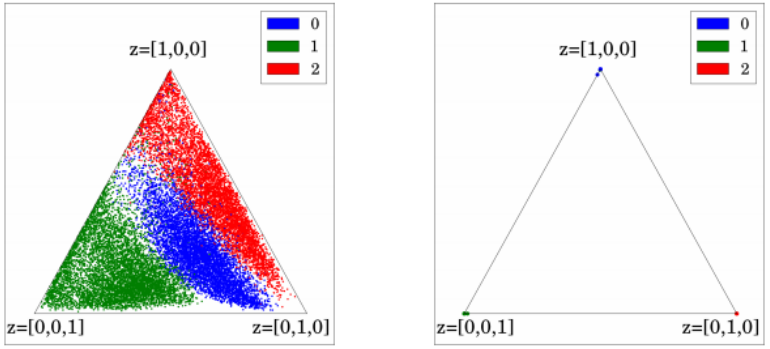
$$= -\underbrace{\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\Big[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[-\log p(\mathbf{x}|\mathbf{z})]\Big]}_{\text{reconstruction term}} - \underbrace{\mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}))}_{\text{marginal KL}} - \underbrace{\mathbb{I}(\mathbf{z};\mathbf{x})}_{\text{mutual info.}}$$

Published on NIPS in 2017

PixelGAN Autoencoders

Alireza Makhzani, Brendan Frey
University of Toronto
{makhzani,frey}@psi.toronto.edu

- **Conclusion** :In this paper, we proposed the PixelGAN autoencoder, which is a generative autoencoder that combines a generative PixelCNN with a GAN inference network that can impose arbitrary priors on the latent code. We showed that imposing different distributions as the prior enables us to learn a latent representation that captures the type of statistics that we care about, while the remaining structure of the image is captured by the PixelCNN decoder. Specifically, by imposing a Gaussian prior, we were able to disentangle the low-frequency and high-frequency statistics of the images, and by imposing a categorical prior we were able to disentangle the style and content of images and learn representations that are specifically useful for clustering and semi-supervised learning tasks. While the main focus of this paper was to demonstrate the application of PixelGAN autoencoders **in downstream tasks such as semi-supervised learning, we discussed how these architectures have many other potentials such as learning cross-domain relations between two different domains.**



(a) Without GAN Regularization (b) With GAN Regularization

Figure 5: Effect of GAN regularization (categorical prior) on the code space of PixelGAN autoencoders

	MNIST (Unsupervised)	MNIST (20 labels)	MNIST (50 labels)	MNIST (100 labels)	SVHN (500 labels)	SVHN (1000 labels)	NORB (1000 labels)
VAE [24]	-	-	-	3.33 (±0.14)	-	36.02 (±0.10)	18.79 (±0.05
VAT [25]	-	-	-	2.33	-	24.63	9.88
ADGM [26]	-	-	-	0.96 (±0.02)	-	22.86	10.06 (±0.05
SDGM [26]	-	-	-	1.32 (±0.07)	-	16.61 (±0.24)	9.40 (±0.04
Adversarial Autoencoder [6]	4.10 (±1.13)	-	-	1.90 (±0.10)	-	17.70 (±0.30)	-
Ladder Networks [27]	-	-	-	0.89 (±0.50)	-	-	-
Convolutional CatGAN [22]	4.27	-	-	1.39 (±0.28)	-	-	-
InfoGAN [16]	5.00	-	-	-	-	-	-
Feature Matching GAN [28]	-	16.77 (±4.52)	2.21 (±1.36)	0.93 (±0.06)	18.44 (±4.80)	8.11 (±1.30)	-
Temporal Ensembling [23]	-	-	-	-	7.05 (±0.30)	5.43 (±0.25)	-
PixelGAN Autoencoders	5.27 (±1.81)	12.08 (±5.50)	1.16 (±0.17)	1.08 (±0.15)	10.47 (±1.80)	6.96 (±0.55)	8.90 (±1.0)

Table 1: Semi-supervised learning and clustering error-rate on MNIST, SVHN and NORB dataset

PixelGAN Autoencoders

- **Review** :The paper proposes PixelGAN autocencoder - a generative model which is a hybrid of an adversarial autoencoder and a PixelCNN autoencoder. The authors provide a theoretical justification of the approach based on a decomposition of variational evidence lower bound (ELBO). The authors provide qualitative results with different priors on the hidden distribution, and quantitative results on semi-supervised learning on MNIST, SVHN and NORB.
-
- The paper is closely related to Adversarial autoencoders (Makhzani et al. ICLR 2016 Workshop), which, as far as I know, have only been published on arxiv and as a Workshop contribution to ICLR. Yet, the work is well known and widely cited. The novelty of the present submission significantly depends on Adversarial autoencoders being considered existing approach or not. This is a complicated situation, which, I assume, ACs and PCs are better qualified to judge about.
- The paper build and auto-encoder with pixelCNN decoder and adversarial cost on latent between uniform prior and inference distribution. With the right network design the networks separate global input information stored in the latent and local one captured by pixelCNN when trained on MNIST dataset. With categorical distribution of latents the network learns to capture very close to class information in unsupervised way. The networks perform well in semisupervised settings. The paper is yet another combination of VAE/AdvNet/PixelCNN. The paper has a nice set of experiments and discussion. The model is most closely related to VAE-pixelCNN combination with VAE loss (KL) on latents replaced by adversarial loss (even though they discuss the mathematical difference) and it would be good to run the same experiments (scaling latent loss) with that and compare.

Published on NIPS in 2017

Semi-Supervised Learning for Optical Flow with Generative Adversarial Networks

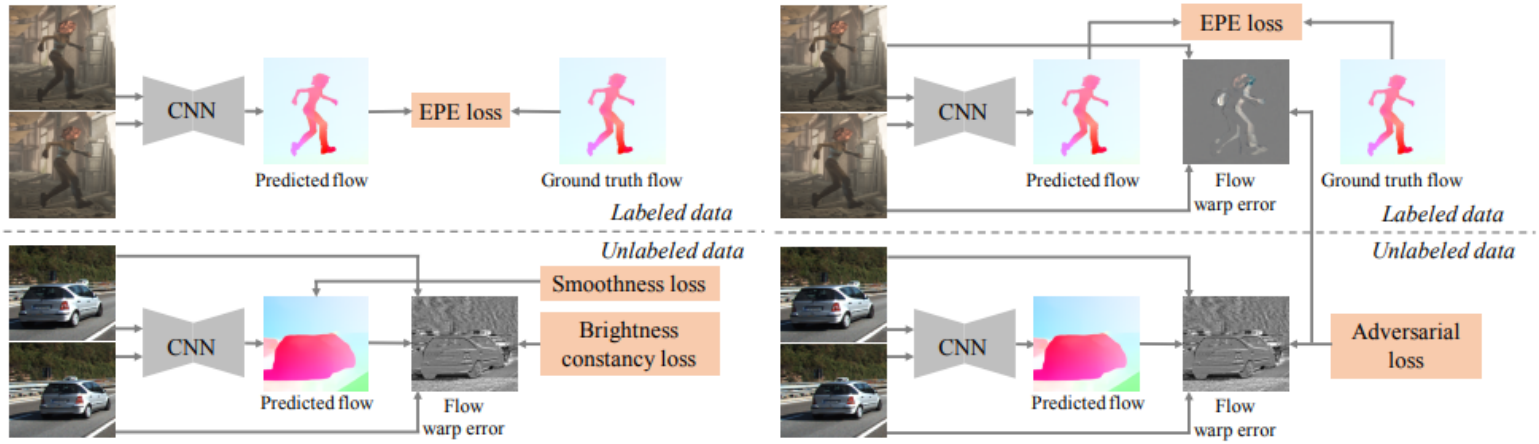
Wei-Sheng Lai[1] **Jia-Bin Huang**[2] **Ming-Hsuan Yang**[1,3]
[1]University of California, Merced [2]Virginia Tech [3]Nvidia Research
[1]{wlai24|mhyang}@ucmerced.edu [2]jbhuang@vt.edu

- **Abstract** :Convolutional neural networks (CNNs) have recently been applied to the optical flow estimation problem. As training the CNNs requires sufficiently large amounts of labeled data, existing approaches resort to synthetic, unrealistic datasets. On the other hand, unsupervised methods are capable of leveraging real-world videos for training where the ground truth flow fields are not available. These methods, however, rely on the fundamental assumptions of brightness constancy and spatial smoothness priors that do not hold near motion boundaries. In this paper, we propose to exploit unlabeled videos for semi-supervised learning of optical flow with a Generative Adversarial Network. Our key insight is that the adversarial loss can capture the structural patterns of flow warp errors without making explicit assumptions. Extensive experiments on benchmark datasets demonstrate that the proposed semi-supervised algorithm performs favorably against purely supervised and baseline semi-supervised learning schemes.

In this work, we propose a generative adversarial network for learning optical flow in a semi-supervised manner. We use a discriminative network and an adversarial loss to learn the structural patterns of the flow warp error without making assumptions on brightness constancy and spatial smoothness. The adversarial loss serves as guidance for estimating optical flow from both labeled and unlabeled datasets. Extensive evaluations on benchmark datasets validate the effect of the adversarial loss and demonstrate that the proposed method performs favorably against the purely supervised and the straightforward semi-supervised learning approaches for learning optical flow

Semi-Supervised Learning for Optical Flow with Generative Adversarial Networks

Wei-Sheng Lai[1] **Jia-Bin Huang**[2] **Ming-Hsuan Yang**[1,3]
[1]University of California, Merced [2]Virginia Tech [3]Nvidia Research
[1]{wlai24|mhyang}@ucmerced.edu [2]jbhuang@vt.edu



(a) Baseline semi-supervised learning (b) The proposed semi-supervised learning

Figure 1: **Semi-supervised learning for optical flow estimation.** (a) A baseline semi-supervised algorithm utilizes the assumptions of brightness constancy and spatial smoothness to train CNN from unlabeled data (e.g., [1, 40]). (b) We train a generative adversarial network to capture the structure patterns in flow warp error images without making any prior assumptions.

Updating discriminator D.

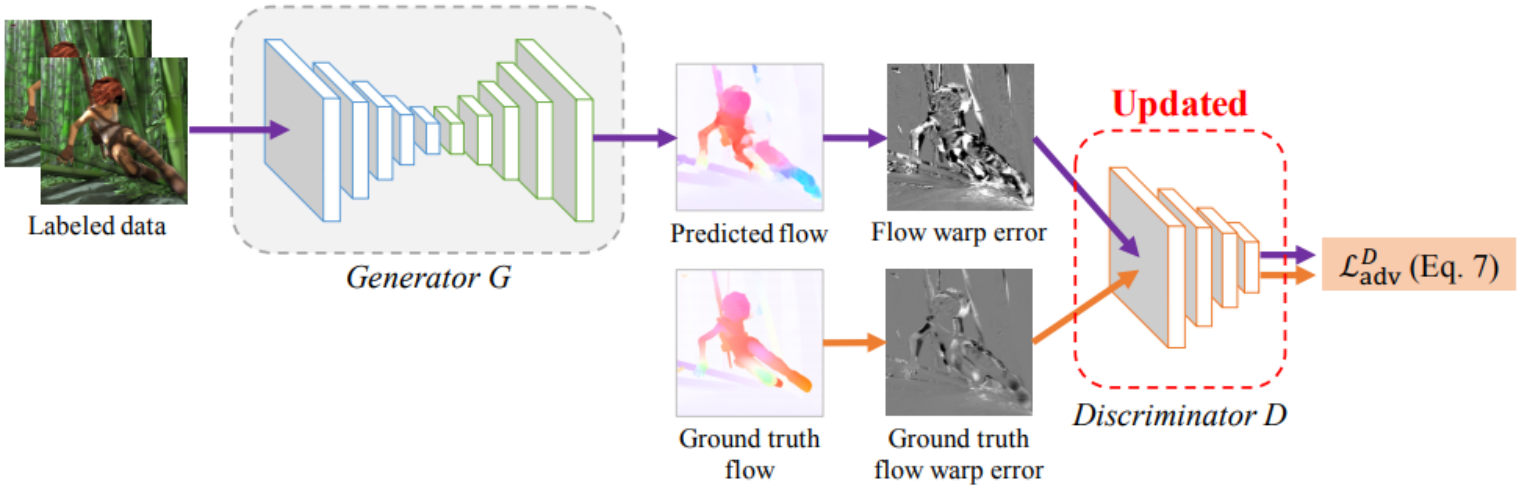
$$\begin{aligned}\mathcal{L}_{\text{adv}}^{D}(y,\hat{y}) &= \mathcal{L}_{\text{BCE}}(D(\hat{y}),1) + \mathcal{L}_{\text{BCE}}(D(y),0)\\ &= -\log D(\hat{y}) - \log(1-D(y)).\end{aligned}$$

Updating generator G.

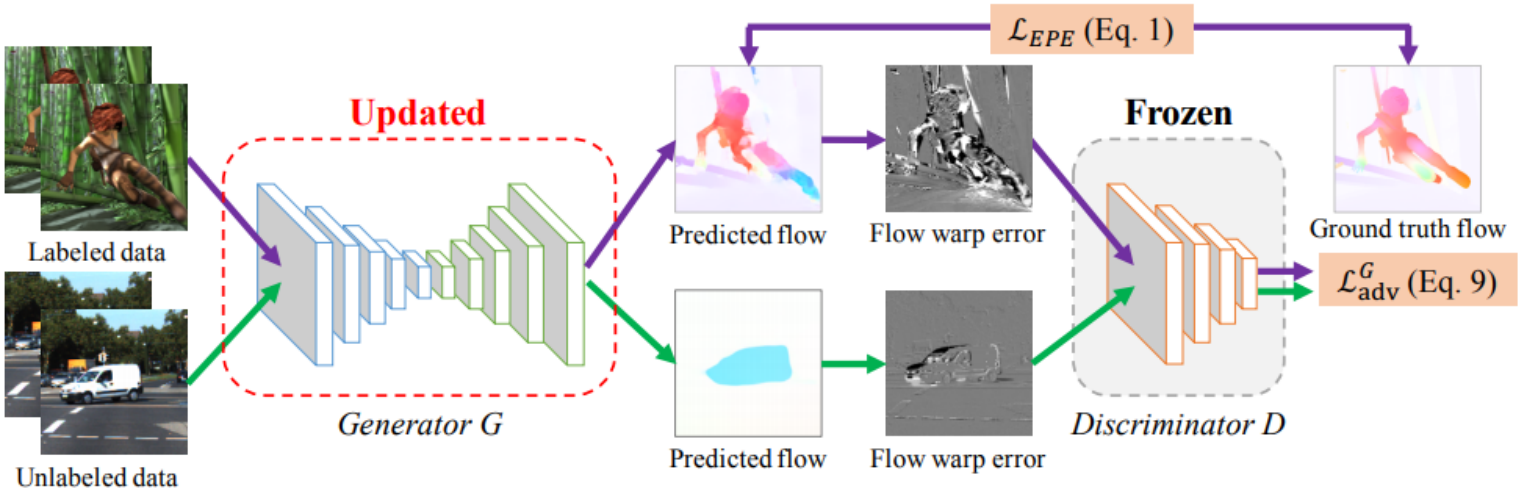
$$\mathcal{L}_{\text{adv}}^{G}(y) = \mathcal{L}_{\text{BCE}}(D(y),1) = -\log(D(y)).$$

Semi-Supervised Learning for Optical Flow with Generative Adversarial Networks

Wei-Sheng Lai[1] **Jia-Bin Huang**[2] **Ming-Hsuan Yang**[1,3]
[1]University of California, Merced [2]Virginia Tech [3]Nvidia Research
[1]{wlai24|mhyang}@ucmerced.edu [2]jbhuang@vt.edu



(a) Update discriminator D using labeled data



(b) Update generator G using both labeled and unlabeled data

Figure 3: **Adversarial training procedure.** Training a generative adversarial network involves the alternative optimization of the discriminator D and generator G.

Semi-Supervised Learning for Optical Flow with Generative Adversarial Networks

Wei-Sheng Lai[1] **Jia-Bin Huang**[2] **Ming-Hsuan Yang**[1,3]
[1]University of California, Merced [2]Virginia Tech [3]Nvidia Research
[1]{wlai24|mhyang}@ucmerced.edu [2]jbhuang@vt.edu

- **Conclusion** :In this work, we propose a generative adversarial network for learning optical flow in a semi-supervised manner. We use a discriminative network and an adversarial loss to learn the structural patterns of the flow warp error without making assumptions on brightness constancy and spatial smoothness. The adversarial loss serves as guidance for estimating optical flow from both labeled and unlabeled datasets. Extensive evaluations on benchmark datasets validate the effect of the adversarial loss and demonstrate that the proposed method performs favorably against the purely supervised and the straightforward semi-supervised learning approaches for learning optical flow.
- **Review:**The paper presents a semi-supervised approach to learning optical flow using a generative adversarial network (GAN) on flow warp errors. Rather than using a handcrafted loss (e.g., deviation of brightness constancy + deviation from smoothness) the paper explores the use of a GAN applied to flow warp errors.
- The paper proposes a semi-supervised learning scheme with GANs for optical flow. The proposed approach is reasonable, and some tiny improvements (Table 2 and 3) are achieved over fully supervised or baseline semi-supervised (similar) architectures. However, this is the case when the proposed approach uses all the train data of the supervised architectures and extra train data.
- This paper discusses the training of an optical flow estimation model, and proposes two things: 1) train the model in a semi-supervised way by using a supervised loss on labeled data, and an unsupervised loss on unlabeled data.2) use the labeled data to also learn the unsupervised loss using a discriminator (in a GAN setting).

Published on NIPS in 2017

Stabilizing Training of Generative Adversarial Networks through Regularization

Kevin Roth
Department of Computer Science
ETH Zürich
kevin.roth@inf.ethz.ch

Aurelien Lucchi
Department of Computer Science
ETH Zürich
aurelien.lucchi@inf.ethz.ch

- **Abstract** :Deep generative models based on Generative Adversarial Networks (GANs) have demonstrated impressive sample quality but in order to work they require a careful choice of architecture, parameter initialization, and selection of hyper-parameters. This fragility is in part due to a dimensional **mismatch or non-overlapping support between the model distribution and the data distribution**, causing their density ratio and the associated f-divergence to be undefined. We overcome this fundamental limitation and propose a **new regularization** approach with low computational cost that yields a stable GAN training procedure. We demonstrate the effectiveness of this regularizer accross several architectures trained on common benchmark image generation tasks. Our regularization turns GAN models into reliable building blocks for deep learning.

Objective function:

$$F(\mathbb{P}, \mathbb{Q}; \phi) = \mathbf{E}_{\mathbb{P}}\left[g(\phi(\mathbf{x}))\right] + \mathbf{E}_{\mathbb{Q}}\left[g(-\phi(\mathbf{x}))\right],$$

Contribution:

- **We systematically derive a novel, efficiently computable regularization method for f-GAN.**
- **We show how this addresses the dimensional misspecification challenge.**
- **We empirically demonstrate stable GAN training across a broad set of models.**

Regularized f-GAN

$$F_\gamma(\mathbb{P}, \mathbb{Q}; \psi) = \mathbf{E}_{\mathbb{P}}\left[\psi\right] - \mathbf{E}_{\mathbb{Q}}\left[f^c \circ \psi\right] - \frac{\gamma}{2}\Omega_f(\mathbb{Q}; \psi)$$
$$\Omega_f(\mathbb{Q}; \psi) := \mathbf{E}_{\mathbb{Q}}\left[(f^{c\prime\prime} \circ \psi)\,\|\nabla\psi\|^2\right] \quad (19)$$

Regularized Jensen-Shannon GAN

$$F_\gamma(\mathbb{P}, \mathbb{Q}; \varphi) = \mathbf{E}_{\mathbb{P}}\left[\ln(\varphi)\right] + \mathbf{E}_{\mathbb{Q}}\left[\ln(1-\varphi)\right] - \frac{\gamma}{2}\Omega_{JS}(\mathbb{P}, \mathbb{Q}; \varphi)$$
$$\Omega_{JS}(\mathbb{P}, \mathbb{Q}; \varphi) := \mathbf{E}_{\mathbb{P}}\left[(1-\varphi(\mathbf{x}))^2||\nabla\phi(\mathbf{x})||^2\right] + \mathbf{E}_{\mathbb{Q}}\left[\varphi(\mathbf{x})^2||\nabla\phi(\mathbf{x})||^2\right] \quad (20)$$

Published on NIPS in 2017

Stabilizing Training of Generative Adversarial Networks through Regularization

Kevin Roth
Department of Computer Science
ETH Zürich
kevin.roth@inf.ethz.ch

Aurelien Lucchi
Department of Computer Science
ETH Zürich
aurelien.lucchi@inf.ethz.ch

- **Conclusion** :We introduced a regularization scheme to train deep generative models based on generative adversarial networks (GANs). While dimensional misspecifications or non-overlapping support between the data and model distributions can cause severe failure modes for GANs, we showed that this can be addressed by adding a penalty on the weighted gradient-norm of the discriminator. Our main result is a simple yet effective modification of the standard training algorithm for GANs, turning them into reliable building blocks for deep learning that can essentially be trained indefinitely without collapse. Our experiments demonstrate that our regularizer improves stability, prevents GANs from overfitting and therefore leads to better generalization properties (cf cross-testing protocol). Further research on the optimization of GANs as well as their convergence and generalization can readily be built upon our theoretical results.

		True condition	
		Positive	Negative
Predicted	Positive	0.9688	0.0002
	Negative	0.0312	0.9998

Cross-testing: FP: 0.0

		True condition	
		Positive	Negative
Predicted	Positive	1.0	0.0013
	Negative	0.0	0.9987

Cross-testing: FP: 1.0

For both models, the discriminator is able to recognize his own generator’ s samples (low FP in the confusion matrix). The regularized GAN also manages to perfectly classify the unregularized GAN’ s samples as fake (cross-testing FP 0.0) whereas the unregularized GAN classifies the samples of the regularized GAN as real (cross-testing FP 1.0). In other words, the regularized model is able to fool the unregularized one, whereas the regularized variant cannot be fooled.

Stabilizing Training of Generative Adversarial Networks through Regularization

Wei-Sheng Lai[1] **Jia-Bin Huang**[2] **Ming-Hsuan Yang**[1,3]
[1]University of California, Merced [2]Virginia Tech [3]Nvidia Research
[1]{wlai24|mhyang}@ucmerced.edu [2]jbhuang@vt.edu

- **Review** :In this paper, the authors have introduced a regularization scheme to train generative models with GANs. They added a penalty on the weighted gradient norm to address the severe failure modes caused by dimensional misspecification between the true and model distributions. In general, this paper is well written and easy to follow.
- (1) The implementation details of the proposed algorithm are missing. The authors should discuss the implementation details in the main paper. Also, please discuss the trick while training the proposed model. GAN is really hard to train and also difficult to reproduce the results reported by the authors. It will be easier if the authors can include the training tricks in their paper.
- (2) The training time is not discussed in this paper. There is always a balance between the performance gain and the training time.
- (3) The experimental part is the weakest part of this paper. The authors may want to include more exciting experimental results in the main paper.

This paper proposed to stabilize the training of GAN using proposed gradient-norm regularizer.
This regularization is designed for conventional GAN, or more general f-GAN proposed last year. The idea is interesting but the justification is a little bit coarse.

- In order to solve the problem of requiring careful choice of architecture and parameters in Generative Adversarial Networks (GANs) based deep generative models. The author proposed a new regularization approach which can sovle the problem cause by the dimensional mismatch between the model distribution and the true distribution. By analysis noise covolution, conbobived discriminants and the efficient gradient norm-based regularization the author propose a Gradient-Norm Regularizer for f-GAN. The experiment on MNIST and CIFAR-10 also should that the proposed regularizer is useful. This work is very interesting and enlightening which can help people to build more stable GAN.

Published on NIPS in 2017

Temporal Coherency based Criteria for Predicting Video Frames using Deep Multi-stage GenerativeAdversarial Networks

Prateep Bhattacharjee[1], **Sukhendu Das**[2]
Visualization and Perception Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai, India
prateepb@cse.iitm.ac.in, [2]sdas@iitm.ac.in

- **Abstract** :Predicting the future from a sequence of video frames has been recently a sought after yet challenging task in the field of computer vision and machine learning. Although there have been efforts for tracking using motion trajectories and flow features, the complex problem of generating unseen frames has not been studied extensively. In this paper, we deal with this problem using convolutional models within a multi-stage Generative Adversarial Networks (GAN) framework. The proposed method uses two stages of GANs to generate crisp and clear set of future frames. Although GANs have been used in the past for predicting the future, none of the works consider the relation between subsequent frames in the temporal dimension. Our main contribution lies in formulating two objective functions based on the Normalized Cross Correlation (NCC) and the Pairwise Contrastive Divergence (PCD) for solving this problem. This method, coupled with the traditional L1 loss, has been experimented with three real-world video datasets viz. Sports-1M, UCF-101 and the KITTI.

  Performance analysis reveals superior results over the recent state-of-the-art methods.

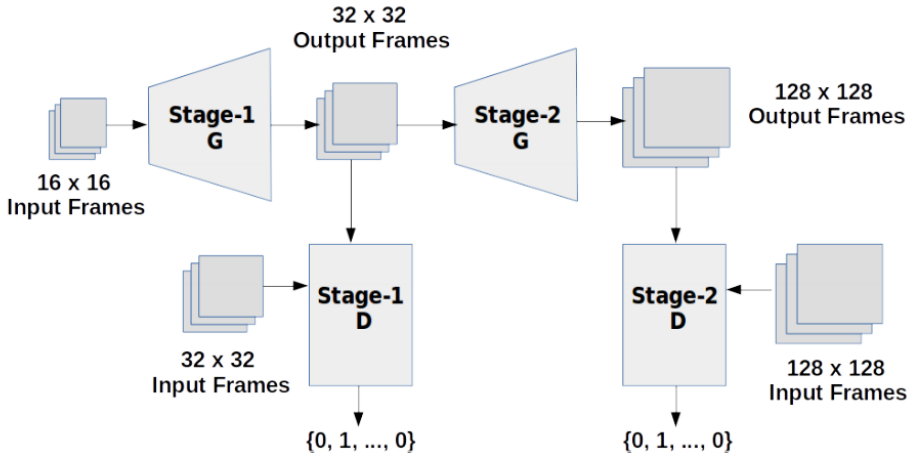


Figure 1: The proposed multi-stage GAN framework. The stage-1 generator network produces a low-resolution version of predicted frames which are then fed to the stage-2 generator. Discriminators at both the stages predict 0 or 1 for each predicted frame to denote its origin: synthetic or original.

6 Combined Loss

Finally, we combine the objective functions given in eqns. 5 - 8 along with the general $L1$-loss with different weights as:

$$\mathcal{L}_{Combined} = \lambda_{adv}\mathcal{L}^{G}_{adv}(X) + \lambda_{L1}\mathcal{L}_{L1}(X,Y) + \lambda_{NCCL}\mathcal{L}_{NCCL}(Y,\hat{Y}) + \lambda_{PCDL}\mathcal{L}_{PCDL}(\hat{Y},\vec{p}) + \lambda_{PCDL}\mathcal{L}_{3-PCDL}(\hat{Y},\vec{p}) \quad (10)$$

All the weights *viz.* $\lambda_{L1}, \lambda_{NCCL}, \lambda_{PCDL}$ and λ_{3-PCDL} have been set as 0.25, while λ_{adv} equals 0.01. This overall loss is minimized during the training stage of the multi-stage GAN using Adam optimizer [11].

We also evaluate our models by incorporating another loss function described in section A of the supplementary document, the Smoothed Normalized Cross-Correlation Loss (SNCCL). The weight for SNCCL, λ_{SNCCL} equals 0.33 while λ_{3-PCDL} and λ_{PCDL} is kept at 0.16.

Temporal Coherency based Criteria for Predicting Video Frames using Deep Multi-stage GenerativeAdversarial Networks

- **Conclusion** :In this paper, we modified the Generative Adversarial Networks (GAN) framework with the use of unpooling operations and introduced two objective functions based on the normalized crosscorrelation (NCCL) and the contrastive divergence estimate (PCDL), to design an efficient algorithm for video frame(s) prediction. Studies show significant improvement of the proposed methods over the recent published works. Our proposed objective functions can be used with more complex networks involving 3D convolutions and recurrent neural networks. In the future, we aim to learn weights for the cross-correlation such that it focuses adaptively on areas involving varying amount of motion.
- **Review:**This method provides 2 contributions for next frame prediction from video sequences. The first is the introduction of a normalized cross correlation loss, which provide a better similarity score to judge if the predicted frame is close to the true future.
- The second is the pairwise contrastive divergence loss, based on the idea of similarity of the image features. Results are presented on the UCF101 and Kitti datasets, and a numerical comparison using image similarity metrics (PSNR, SSIM) with Mathieu et al ICLR16 is performed.

1. It is addressing an interesting and relevant problem, that is going to be of wide interest, especially in the computer vision community.
2. It is relatively novel. The new things are comming through the usage of the contrastive loss and the cross-correlation loss as well as the multi-stage GAN idea.
3. Relatively well written and easy to read.
4. Positive experimental evaluation. The results on the UCF dataset suggest that the combined method is better than the baselines, and the different components do contribute to a better overall performance.

Published on NIPS in 2017

Multi-Modal Imitation Learning from Unstructured Demonstrations using Generative Adversarial Nets

Karol Hausman[*†], **Yevgen Chebotar**[*†‡], **Stefan Schaal**[†‡], **Gaurav Sukhatme**[†], **Joseph J. Lim**[†]

[†]University of Southern California, Los Angeles, CA, USA
[‡]Max-Planck-Institute for Intelligent Systems, Tübingen, Germany
{hausman, ychebota, sschaal, gaurav, limjj}@usc.edu

- **Abstract** :Imitation learning has traditionally been applied to learn a single task from demonstrations thereof. The requirement of structured and isolated demonstrations limits the scalability of imitation learning approaches as they are difficult to apply to real-world scenarios, where robots have to be able to execute a multitude of tasks. In this paper, we propose a multi-modal imitation learning framework that is able to segment and imitate skills from unlabelled and unstructured demonstrations by learning skill segmentation and imitation learning jointly. The extensive simulation results indicate that our method can efficiently separate the demonstrations into individual skills and learn to imitate them using a single multi-modal policy.

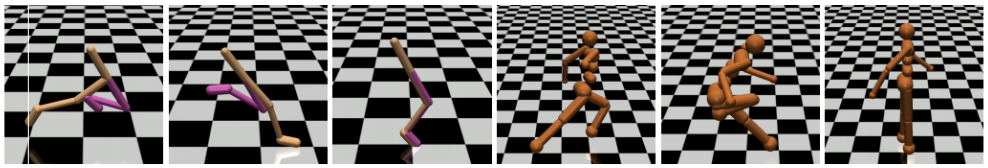


Figure 1: *Left:* Walker-2D running forwards, running backwards, jumping. *Right:* Humanoid running forwards, running backwards, balancing.

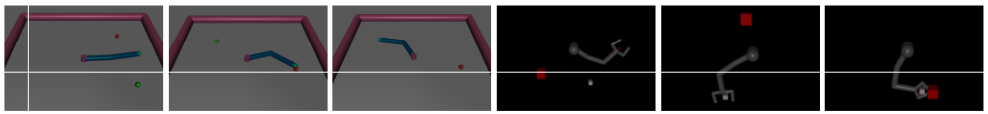


Figure 2: *Left:* Reacher with 2 targets: random initial state, reaching one target, reaching another target. *Right:* Gripper-pusher: random initial state, grasping policy, pushing (when grasped) policy.

$$\max_{\theta} \min_{w} \mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta}[\log(D_w(s,a))] + \mathbb{E}_{(s,a)\sim\pi_E}[1-\log(D_w(s,a))] \quad (8)$$
$$+(\lambda_H-\lambda_I)H(\pi_\theta(a|s)) + \lambda_I\mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta}\log(p(i|s,a)) + \lambda_I H(i),$$

where $H(i)$ is a constant that does not influence the optimization. This results in the same optimization objective as for the single expert policy (see Eq. (2)) with an additional term $\lambda_I\mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta}\log(p(i|s,a))$ responsible for rewarding state-action pairs that make the latent intention inference easier. We refer to this cost as the latent intention cost and represent $p(i|s,a)$ with a neural network. The final reward function for the generator is:

$$\mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta}[\log(D_w(s,a))] + \lambda_I\mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta}\log(p(i|s,a)) + \lambda_{H'}H(\pi_\theta(a|s)). \quad (9)$$

Multi-Modal Imitation Learning from Unstructured Demonstrations using Generative Adversarial Nets

- **Conclusion** :We present a novel imitation learning method that learns a multi-modal stochastic policy, which is able to imitate a number of automatically segmented tasks using a set of unstructured and unlabeled demonstrations. The presented approach learns the notion of intention and is able to perform different tasks based on the policy intention input. We evaluated our method on a set of simulation scenarios where we show that it is able to segment the demonstrations into different tasks and to learn a multi-modal policy that imitates all of the segmented skills. We also compared our method to a baseline approach that performs imitation learning without explicitly separating the tasks. In the future work, we plan to focus on autonomous discovery of the number of tasks in the given pool of demonstrations as well as evaluating this method on real robots. We also plan to learn an additional hierarchical policy over the discovered intentions as an extension of this work.

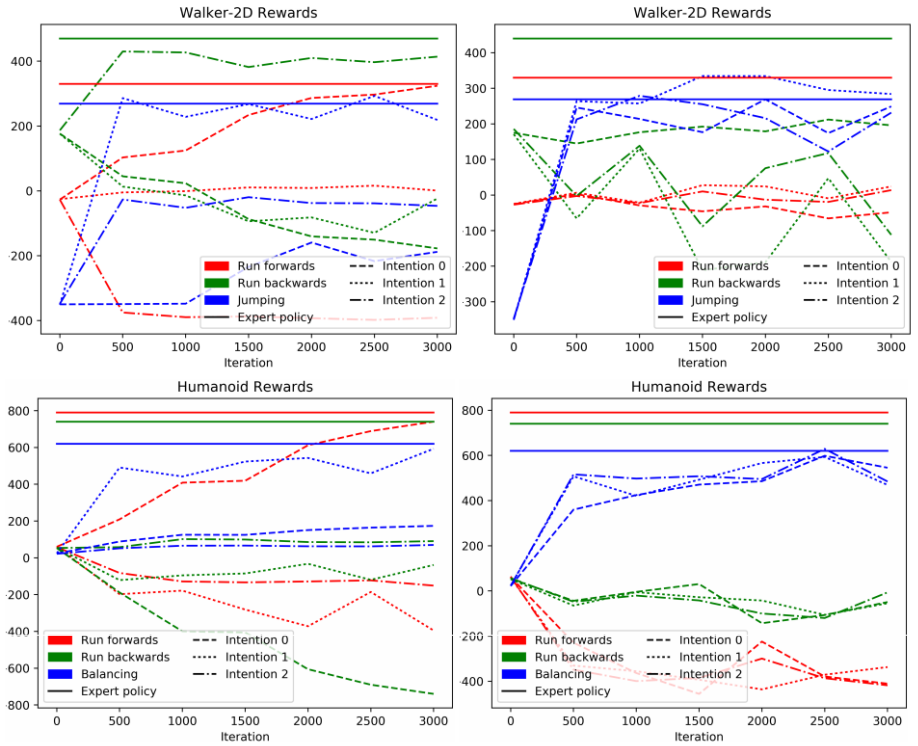


Figure 5: *Top:* Rewards of Walker-2D policies for different intention values over the training iterations with (left) and without (right) the latent intention cost. *Bottom:* Rewards of Humanoid policies for different intention values over the training iterations with (left) and without (right) the latent intention cost.

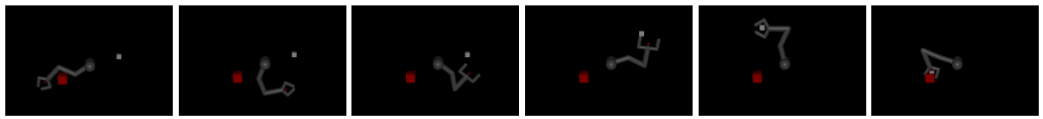


Figure 6: Time-lapse of the learned Gripper-pusher policy. The intention variable is changed manually in the fifth screenshot, once the grasping policy has grasped the block.

Multi-Modal Imitation Learning from Unstructured Demonstrations using Generative Adversarial Nets

- **Review** :The paper describes a new learning model able to discover 'intentions' from expert policies by using an imitation learning framework. The idea is mainly based on the GAIL model which aims at learning by imitation a policy using a GAN approach. The main difference in the article is that the learned policy is, in fact, a mixture of sub-policies, each sub-policy aiming at automatically matching a particular intention in the expert behavior. The GAIL algorithm is thus derived with this mixture, resulting in an effective learning technique. Another approach is also proposed where the intention will be captured through a latent vector by derivating the InfoGAN algorithm for this particular case. Experiments are made on 4 different settings and show that the model is able to discover the underlying intentions contained in the demonstration trajectories.
- This paper proposes to learn stills from demonstrations without any a priori knowledge. Using GAN, it generates trajectories from a mixture of policies, and imitate the demonstration. This idea is very simple and easy to be understood. I accept that the idea could works.
- This paper considers a multi-task imitation learning problem where the agent should learn to imitate multiple expert policies without having access to the identity of the tasks. The proposed method is based on GAIL (generative adversarial imitation learning) with an additional objective that encourages some of the latent variables (called intention variables) to be easily inferred from the generated trajectories. The results show that the intention variables can captures different modes of the expert behaviors on several Mujoco tasks.

Published on NIPS in 2017

Approximation and Convergence Properties of Generative Adversarial Learning

Shuang Liu
University of California, San Diego
shuangliu@ucsd.edu

Olivier Bousquet
Google Brain
obousquet@google.com

Kamalika Chaudhuri
University of California, San Diego
kamalika@cs.ucsd.edu

- **Abstract** :Generative adversarial networks (GAN) approximate a target data distribution by jointly optimizing an objective function through a "two-player game" between a generator and a discriminator. Despite their empirical success, however, two very basic questions on how well they can approximate the target distribution remain unanswered. First, it is not known how restricting the discriminator family affects the approximation quality. Second, while a number of different objective functions have been proposed, we do not understand when convergence to the global minima of the objective function leads to convergence to the target distribution under various notions of distributional convergence. In this paper, we address these questions in a broad and unified setting by defining a notion of adversarial divergences that includes a number of recently proposed objective functions. We show that if the objective function is an adversarial divergence with some additional conditions, then using a restricted discriminator family has a moment-matching effect. Additionally, we show that for objective functions that are strict adversarial divergences, convergence in the objective function implies weak convergence, thus generalizing previous results.

(a) GAN [7].
$$\mathcal{F} = \{x, y \mapsto \log(u(x)) + \log(1 - u(y)) : u \in \mathcal{V}\}$$
$$\mathcal{V} = (0,1)^X \cap C_b(X).$$

(b) f-GAN [10]. Let $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a convex lower semi-continuous function. Assume $f^*(x) \geq x$ for any $x \in \mathbb{R}$, f^* is continuously differentiable on $\mathrm{int}(\mathrm{dom}\, f^*)$, and there exists $x_0 \in \mathrm{int}(\mathrm{dom}\, f^*)$ such that $f^*(x_0) = x_0$.
$$\mathcal{F} = \{x, y \mapsto v(x) - f^*(v(y)) : v \in \mathcal{V}\},$$
$$\mathcal{V} = (\mathrm{dom}\, f^*)^X \cap C_b(X).$$

(c) MMD-GAN [5, 9]. Let $k : X^2 \to \mathbb{R}$ be a universal reproducing kernel. Let $\mathcal{M}$ be the set of signed measures on X.
$$\mathcal{F} = \{x, y \mapsto v(x) - v(y) : v \in \mathcal{V}\},$$
$$\mathcal{V} = \left\{x \mapsto \mathbb{E}_\mu\left[k(x, \cdot)\right] : \mu \in \mathcal{M},\ \mathbb{E}_{\mu^2}[k] \leq 1\right\}.$$

(d) Wasserstein-GAN (WGAN) [2]. Assume X is a metric space.
$$\mathcal{F} = \{x, y \mapsto v(x) - v(y) : v \in \mathcal{V}\},$$
$$\mathcal{V} = \left\{v \in C_b(X) : \|v\|_{\mathrm{Lip}} \leq K\right\},$$
where K is a positive constant, $\|\cdot\|_{\mathrm{Lip}}$ denotes the Lipschitz constant.

(e) WGAN-GP (Improved WGAN) [8]. Assume X is a convex subset of a Euclidean space.
$$\mathcal{F} = \{x, y \mapsto v(x) - v(y) - \eta\mathbb{E}_{t\sim U}\left[(\|\nabla v(tx + (1-t)y)\|_2 - 1)^p\right] : v \in \mathcal{V}\},$$
$$\mathcal{V} = C^1(X),$$
where U is the uniform distribution on $[0, 1]$, η is a positive constant, $p \in (1, \infty)$.

Approximation and Convergence Properties of Generative Adversarial Learning

- **Conclusion** :In conclusion, our results provide insights on the cost or loss functions that should be used in GANs. The choice of cost function plays a very important role in this case – more so, for example, than data domains or network architectures. For example, most works still use the DCGAN architecture, while changing the cost functions to achieve different levels of performance, and which cost function is better is still a matter of debate. In particular we provide a framework for studying man different GAN criteria in a way that makes them mor directly comparable, and under this framework, we study both approximation and convergence properties of various loss functions.

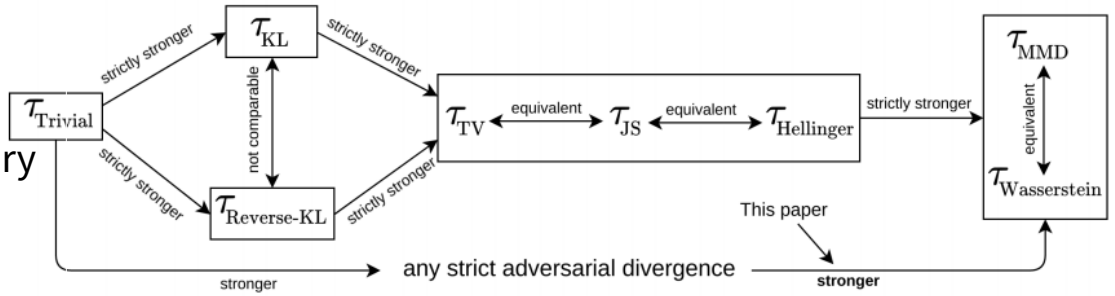


Figure 1: Structure of the class of strict adversarial divergences

(a) GAN. Note that for any $x \in (0,1)$, $\log\left(1/(x(1-x))\right) \geq \log(4)$. Let $u_{\theta^\mu_\nu} = \frac{1}{2}$,

$$f_\theta(x,y) = \log(u_\theta(x)) + \log(1-u_\theta(y)) = \underbrace{\log(u_\theta(x)) - \log(u_\theta(y))}_{m_\theta(x,y)\left(\text{note } \mathbb{E}_{\mu\otimes\nu}\left[m_{\theta^\mu_\nu}\right]=0\right)} - \underbrace{\log\left(1/\left(u_\theta(y)(1-u_\theta(y))\right)\right)}_{r_\theta(x,y)\left(\text{note } r_\theta(x,y)\geq r_{\theta^\mu_\nu}(x,y)=\log(4)\right)}.$$

(b) f-GAN. Recall that $f^*(x) - x \geq 0$ for any $x \in \mathbb{R}$ and $f^*(x_0) = x_0$. Let $v_{\theta^\mu_\nu} = \mathbf{x_0}$,

$$f_\theta(x,y) = v_\theta(x) - f^*(v_\theta(y)) = \underbrace{v_\theta(x) - v_\theta(y)}_{m_\theta(x,y)\left(\text{note } \mathbb{E}_{\mu\otimes\nu}\left[m_{\theta^\mu_\nu}\right]=0\right)} - \underbrace{\left(f^*(v_\theta(y)) - v_\theta(y)\right)}_{r_\theta(x,y)\left(\text{note } r_\theta(x,y)\geq r_{\theta^\mu_\nu}(x,y)=0\right)}.$$

(c, d) MMD-GAN or Wasserstein-GAN. Let $v_{\theta^\mu_\nu} = \mathbf{0}$,

$$f_\theta(x,y) = \underbrace{v_\theta(x) - v_\theta(y)}_{m_\theta(x,y)\left(\text{note } \mathbb{E}_{\mu\otimes\nu}\left[m_{\theta^\mu_\nu}\right]=0\right)} - \underbrace{0}_{r_\theta(x,y)\left(\text{note } r_\theta(x,y)=r_{\theta^\mu_\nu}(x,y)=0\right)}.$$

(e) WGAN-GP. Note that the function $x \mapsto x^p$ is nonnegative on $\mathbb{R}$. Let

$$v_{\theta^\mu_\nu} = \begin{cases} (x_1, x_2, \cdots, x_n) \mapsto \frac{\sum_{i=1}^n x_i}{\sqrt{n}}, & \text{if } \mathbb{E}_\mu[\sum_{i=1}^n x_i] \geq \mathbb{E}_\nu[\sum_{i=1}^n x_i], \\ (x_1, x_2, \cdots, x_n) \mapsto -\frac{\sum_{i=1}^n x_i}{\sqrt{n}}, & \text{otherwise,} \end{cases}$$

$$f_\theta(x,y) = \underbrace{v_\theta(x) - v_\theta(y)}_{m_\theta(x,y)\left(\text{note } \mathbb{E}_{\mu\otimes\nu}\left[m_{\theta^\mu_\nu}\right]\geq 0\right)} \underbrace{- \eta\mathbb{E}_{t\sim U}\left[\left(\|\nabla v(tx+(1-t)y)\|_2 - 1\right)^2\right]}_{r_\theta(x,y)\left(\text{note } r_\theta(x,y)\geq r_{\theta^\mu_\nu}(x,y)=0\right)}.$$

Published on NIPS in 2017

Triple Generative Adversarial Nets

Chongxuan Li, Kun Xu, Jun Zhu*, Bo Zhang
Dept. of Comp. Sci. & Tech., TNList Lab, State Key Lab of Intell. Tech. & Sys.,
Center for Bio-Inspired Computing Research, Tsinghua University, Beijing, 100084, China
{licx14, xu-k16}@mails.tsinghua.edu.cn, {dcszj, dcszb}@mail.tsinghua.edu.cn

- **Review** :Generative Adversarial Nets (GANs) have shown promise in image generation and semi-supervised learning (SSL). However, existing GANs in SSL have two problems: (1) the generator and the discriminator (i.e. the classifier) may not be optimal at the same time; and (2) the generator cannot control the semantics of the generated samples. The problems essentially arise from the two-player formulation, where a single discriminator shares incompatible roles of identifying fake samples and predicting labels and it only estimates the data without considering the labels. To address the problems, we present triple generative adversarial net (Triple-GAN), which consists of three players—a generator, a discriminator and a classifier. The generator and the classifier characterize the conditional distributions between images and labels, and the discriminator solely focuses on identifying fake image-label pairs.

- We design compatible utilities to ensure that the distributions characterized by the classifier and the generator both converge to the data distribution. Our results on various datasets demonstrate that Triple-GAN as a unified model can simultaneously (1) achieve the state-of-the-art classification results among deep generative models, and (2) disentangle the classes and styles of the input and transfer smoothly in the data space via interpolation in the latent space class-conditionally.

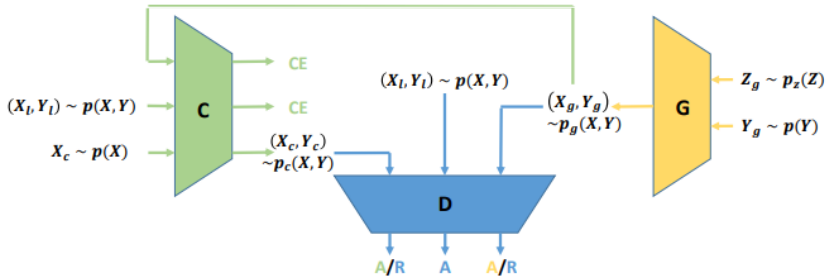


Figure 1: An illustration of Triple-GAN (best view in color). The utilities of D, C and G are colored in blue, green and yellow respectively, with "R" denoting rejection, "A" denoting acceptance and "CE" denoting the cross entropy loss for supervised learning. "A"s and "R"s are the adversarial losses and "CE"s are unbiased regularizations that ensure the consistency between p_g, p_c and p, which are the distributions defined by the generator, classifier and true data generating process, respectively.

Triple Generative Adversarial Nets

- **Conclusion** :We present triple generative adversarial networks (Triple-GAN), a unified game-theoretical framework with three players—a generator, a discriminator and a classifier, to do semi-supervised learning with compatible utilities. With such utilities, Triple-GAN addresses two main problems of existing methods [26, 25]. Specifically, Triple-GAN ensures that both the classifier and the generator can achieve their own optima respectively in the perspective of game theory and enable the generator to sample data in a specific class. Our empirical results on MNIST, SVHN and CIFAR10 datasets demonstrate that as a unified model, Triple-GAN can simultaneously achieve the state-of-the-art classification results among deep generative models and disentangle styles and classes and transfer smoothly on the data level via interpolation in the latent space.

Algorithm 1 Minibatch stochastic gradient descent training of Triple-GAN in SSL.

for number of training iterations **do**

- Sample a batch of pairs $(x_g, y_g) \sim p_g(x, y)$ of size m_g, a batch of pairs $(x_c, y_c) \sim p_c(x, y)$ of size m_c and a batch of labeled data $(x_d, y_d) \sim p(x, y)$ of size m_d.
- Update D by ascending along its stochastic gradient:

$$\nabla_{\theta_d}\left[\frac{1}{m_d}\Big(\sum_{(x_d,y_d)}\log D(x_d,y_d)\Big)+\frac{\alpha}{m_c}\sum_{(x_c,y_c)}\log(1-D(x_c,y_c))+\frac{1-\alpha}{m_g}\sum_{(x_g,y_g)}\log(1-D(x_g,y_g))\right]$$

- Compute the unbiased estimators $\tilde{\mathcal{R}}_{\mathcal{L}}$ and $\tilde{\mathcal{R}}_{\mathcal{P}}$ of $\mathcal{R}_{\mathcal{L}}$ and $\mathcal{R}_{\mathcal{P}}$ respectively.
- Update C by descending along its stochastic gradient:

$$\nabla_{\theta_c}\left[\frac{\alpha}{m_c}\sum_{(x_c,y_c)} p_c(y_c|x_c)\log(1-D(x_c,y_c))+\tilde{\mathcal{R}}_{\mathcal{L}}+\alpha_{\mathcal{P}}\tilde{\mathcal{R}}_{\mathcal{P}}\right].$$

- Update G by descending along its stochastic gradient:

$$\nabla_{\theta_g}\left[\frac{1-\alpha}{m_g}\sum_{(x_g,y_g)}\log(1-D(x_g,y_g))\right].$$

end for

Table 1: Error rates (%) on partially labeled MNIST, SHVN and CIFAR10 datasets, averaged by 10 runs. The results with † are trained with more than 500,000 extra unlabeled data on SVHN.

Algorithm	MNIST $n = 100$	SVHN $n = 1000$	CIFAR10 $n = 4000$
M1+M2 [11]	3.33 (±0.14)	36.02 (±0.10)	
VAT [18]	2.33		24.63
Ladder [23]	1.06 (±0.37)		20.40 (±0.47)
Conv-Ladder [23]	**0.89** (±0.50)		
ADGM [17]	0.96 (±0.02)	22.86 †	
SDGM [17]	1.32 (±0.07)	16.61(±0.24)†	
MMCVA [15]	1.24 (±0.54)	**4.95** (±0.18) †	
CatGAN [26]	1.39 (±0.28)		19.58 (±0.58)
Improved-GAN [25]	0.93 (±0.07)	8.11 (±1.3)	18.63 (±2.32)
ALI [5]		7.3	18.3
Triple-GAN (**ours**)	**0.91** (±0.58)	**5.77**(±0.17)	**16.99** (±0.36)

Table 2: Error rates (%) on MNIST with different number of labels, averaged by 10 runs.

Algorithm	$n = 20$	$n = 50$	$n = 200$
Improved-GAN [25]	16.77 (±4.52)	2.21 (±1.36)	0.90 (±0.04)
Triple-GAN (**ours**)	**4.81** (±4.95)	**1.56** (±0.72)	**0.67** (±0.16)

Triple Generative Adversarial Nets

- **Review** :In this paper, the authors propose a new formulation of adversarial networks for image generation, that incorporates three networks instead of the usual generator G and discriminator D. In addition, they include a classifier C, which cooperates with G to learn a compatible joint distribution (X,Y) over images and labels. The authors show how this formulation overcomes pitfalls of previous class-conditional GANs; namely that class-conditional generator and discriminator networks have competing objectives that may prevent them from learning the true distribution and preventing G from accurately generating class-conditional samples.
- The paper presents a GAN-like architecture called Triple-GAN that, given partially labeled data, is designed to achieve simultaneously the following two goals: (1) Get a good generator that generates realistically-looking samples conditioned on class labels; (2) Get a good classifier, with smallest possible prediction error.

The paper shows that other similar GAN-based approaches always implicitly privileged either (1) or (2), and or needed much more labeled data to train the classifier. By separating the discrimination task between true and fake data from the classification task, the paper outperforms the state-of-the-art, both in (1) and (2). In particular, the classifier achieves high accuracy with only very few labeled dataset, while the generator produces state-of-the-art images, even when conditioned on y labels.

- This paper proposes a three-player adversarial game to overcome the fact that a discriminator in a semi-supervised setting has two incompatible roles, namely to classify and separate real data from fake data.
- The paper is well-written and the authors display a good knowledge of the GAN literature. I think it proposes a solution to a relevant problem, and the empirical evidence presented to back up the claims being made is convincing.

Published on NIPS in 2017

Structured Generative Adversarial Networks

Chongxuan Li, Kun Xu, Jun Zhu, Bo Zhang
Dept. of Comp. Sci. & Tech., TNList Lab, State Key Lab of Intell. Tech. & Sys.,
Center for Bio-Inspired Computing Research, Tsinghua University, Beijing, 100084, China
{licx14, xu-k16}@mails.tsinghua.edu.cn, {dcszj, dcszb}@mail.tsinghua.edu.cn

- **Abstract** :We study the problem of conditional generative modeling based on designated semantics or structures. Existing models that build conditional generators either require massive labeled instances as supervision or are unable to accurately control the semantics of generated samples. We propose structured generative adversarial networks (SGANs) for semi-supervised conditional generative modeling. SGAN assumes the data x is generated conditioned on two independent latent variables: y that encodes the designated semantics, and z that contains other factors of variation. To ensure disentangled semantics in y and z, SGAN builds two collaborative games in the hidden space to minimize the reconstruction error of y and z, respectively. Training SGAN also involves solving two adversarial games that have their equilibrium concentrating at the true joint data distributions p(x, z) and p(x, y), avoiding distributing the probability mass diffusely over data space that MLE-based methods may suffer. We assess SGAN by evaluating its trained networks, and its performance on downstream tasks. We show that SGAN delivers a highly controllable generator, and disentangled representations; it also establishes start-of-the-art results across multiple datasets when applied for semi-supervised image classification (1.27%, 5.73%, 17.26% error rates on MNIST, SVHN and CIFAR-10 using 50, 1000 and 4000 labels, respectively). Benefiting from the separate modeling of y and z, SGAN can generate images with high visual quality and strictly following the designated semantic, and can be extended to a wide spectrum of applications, such as style transfer.

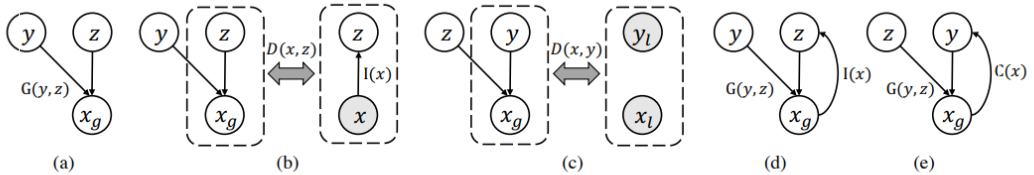


Figure 1: An overview of the SGAN model: (a) the generator $p_g(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})$; (b) the adversarial game $\mathcal{L}_{\boldsymbol{xz}}$; (c) the adversarial game $\mathcal{L}_{\boldsymbol{xy}}$; (d) the collaborative game $\mathcal{R}_{\boldsymbol{z}}$; (e) the collaborative game $\mathcal{R}_{\boldsymbol{y}}$.

Structured Generative Adversarial Networks

- **Conclusion** :We have presented SGAN for semi-supervised conditional generative modeling, which learns from a small set of labeled instances to disentangle the semantics of our interest from other elements in the latent space. We show that SGAN has improved disentanglability and controllability compared to baseline frameworks. SGAN’s design is beneficial to a lot of downstream applications: it establishes new state-of-the-art results on semi-supervised classification, and outperforms strong baseline in terms of the visual quality and inception score on controllable image generation.
- **Review:**This paper proposes a novel GAN structure for semi-supervised learning, a setting in which there exist a small dataset with class labels along with a larger unlabeled dataset. The main idea of this paper is to disentangle the labels (y) from the hidden states (z) using two GAN problems that represent p(x,y) and p(x,z). The generator is shared between both GAN problems, but each problem is trained simultaneously using ALI[4].

There are two adversarial games defined for training the joints p(x, y) and p(x, z). Two "collaborative games" are also defined in order to better disentangle y from z and enforce structure on y.

- This paper proposes the SGAN model which can learn an inference network for GAN architectures. There are two sets of latent variables, y for the label information and z for all other variations in the image (style). The generator p_g conditions on y and z and generate an image x. The adversarial cost L_xz, uses an ALI-like framework to infer z, and the adversarial cost L_xy uses another discriminator to train a standard conditional GAN on the few supervised labeled data by concatenating the labels to both the input of the generator and the input of the discriminator. The SGAN network also uses R_y and R_z to auto-encode both y and z latent variables. R_y has an additional term that minimizes the cross-entropy cost of the inference network of y on the labeled data.

Structured Generative Adversarial Networks

- **Conclusion** :We present triple generative adversarial networks (Triple-GAN), a unified game-theoretical framework with three players—a generator, a discriminator and a classifier, to do semi-supervised learning with compatible utilities. With such utilities, Triple-GAN addresses two main problems of existing methods [26, 25]. Specifically, Triple-GAN ensures that both the classifier and the generator can achieve their own optima respectively in the perspective of game theory and enable the generator to sample data in a specific class. Our empirical results on MNIST, SVHN and CIFAR10 datasets demonstrate that as a unified model, Triple-GAN can simultaneously achieve the state-of-the-art classification results among deep generative models and disentangle styles and classes and transfer smoothly on the data level via interpolation in the latent space.

Algorithm 1 Minibatch stochastic gradient descent training of Triple-GAN in SSL.

for number of training iterations **do**

- Sample a batch of pairs $(x_g, y_g) \sim p_g(x, y)$ of size m_g, a batch of pairs $(x_c, y_c) \sim p_c(x, y)$ of size m_c and a batch of labeled data $(x_d, y_d) \sim p(x, y)$ of size m_d.
- Update D by ascending along its stochastic gradient:

$$\nabla_{\theta_d}\left[\frac{1}{m_d}\Big(\sum_{(x_d,y_d)}\log D(x_d,y_d)\Big)+\frac{\alpha}{m_c}\sum_{(x_c,y_c)}\log(1-D(x_c,y_c))+\frac{1-\alpha}{m_g}\sum_{(x_g,y_g)}\log(1-D(x_g,y_g))\right]$$

- Compute the unbiased estimators $\tilde{\mathcal{R}}_{\mathcal{L}}$ and $\tilde{\mathcal{R}}_{\mathcal{P}}$ of $\mathcal{R}_{\mathcal{L}}$ and $\mathcal{R}_{\mathcal{P}}$ respectively.
- Update C by descending along its stochastic gradient:

$$\nabla_{\theta_c}\left[\frac{\alpha}{m_c}\sum_{(x_c,y_c)} p_c(y_c|x_c)\log(1-D(x_c,y_c))+\tilde{\mathcal{R}}_{\mathcal{L}}+\alpha_{\mathcal{P}}\tilde{\mathcal{R}}_{\mathcal{P}}\right].$$

- Update G by descending along its stochastic gradient:

$$\nabla_{\theta_g}\left[\frac{1-\alpha}{m_g}\sum_{(x_g,y_g)}\log(1-D(x_g,y_g))\right].$$

end for

Table 1: Error rates (%) on partially labeled MNIST, SHVN and CIFAR10 datasets, averaged by 10 runs. The results with † are trained with more than 500,000 extra unlabeled data on SVHN.

Algorithm	MNIST $n = 100$	SVHN $n = 1000$	CIFAR10 $n = 4000$
M1+M2 [11]	3.33 (±0.14)	36.02 (±0.10)	
VAT [18]	2.33		24.63
Ladder [23]	1.06 (±0.37)		20.40 (±0.47)
Conv-Ladder [23]	**0.89** (±0.50)		
ADGM [17]	0.96 (±0.02)	22.86 †	
SDGM [17]	1.32 (±0.07)	16.61(±0.24)†	
MMCVA [15]	1.24 (±0.54)	**4.95** (±0.18) †	
CatGAN [26]	1.39 (±0.28)		19.58 (±0.58)
Improved-GAN [25]	0.93 (±0.07)	8.11 (±1.3)	18.63 (±2.32)
ALI [5]		7.3	18.3
Triple-GAN (**ours**)	**0.91** (±0.58)	**5.77**(±0.17)	**16.99** (±0.36)

Table 2: Error rates (%) on MNIST with different number of labels, averaged by 10 runs.

Algorithm	$n = 20$	$n = 50$	$n = 200$
Improved-GAN [25]	16.77 (±4.52)	2.21 (±1.36)	0.90 (±0.04)
Triple-GAN (**ours**)	**4.81** (±4.95)	**1.56** (±0.72)	**0.67** (±0.16)

Published on NIPS in 2017

Dual Discriminator Generative Adversarial Nets

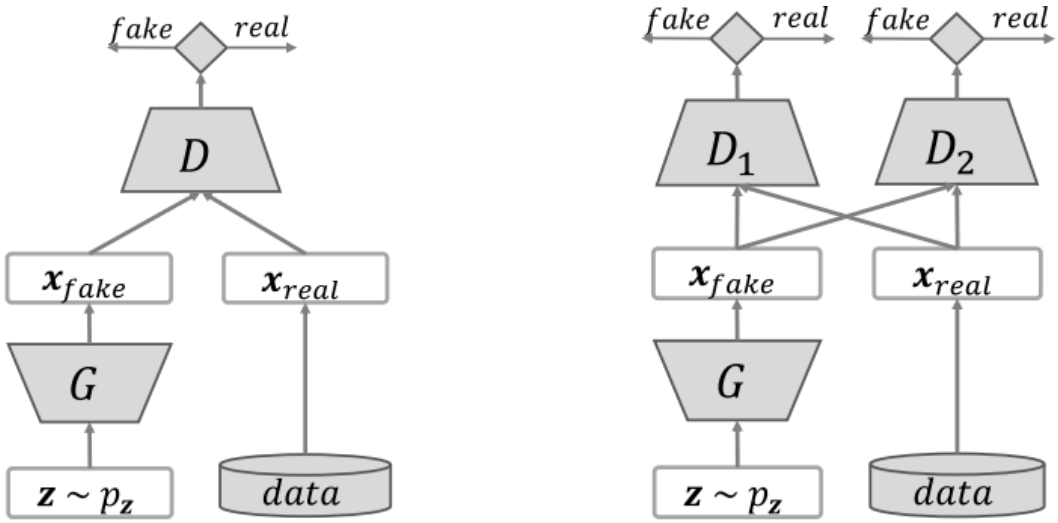
Tu Dinh Nguyen, Trung Le, Hung Vu, Dinh Phung
Deakin University, Geelong, Australia
Centre for Pattern Recognition and Data Analytics
{tu.nguyen, trung.l, hungv, dinh.phung}@deakin.edu.au

- **Abstract** :We propose in this paper a novel approach to tackle the problem of mode collapse encountered in generative adversarial network (GAN). Our idea is intuitive but proven to be very effective, especially in addressing some key limitations of GAN. In essence, it combines the Kullback-Leibler (KL) and reverse KL divergences into a unified objective function, thus it exploits the complementary statistical properties from these divergences to effectively diversify the estimated density in capturing multi-modes. We term our method dual discriminator generative adversarial nets (D2GAN) which, unlike GAN, has two discriminators; and together with a generator, it also has the analogy of a minimax game, wherein a discriminator rewards high scores for samples from data distribution whilst another discriminator, conversely, favoring data from the generator, and the generator produces data to fool both two discriminators.
- We develop theoretical analysis to show that, given the maximal discriminators, optimizing the generator of D2GAN reduces to minimizing both KL and reverse KL divergences between data distribution and the distribution induced from the data generated by the generator, hence effectively avoiding the mode collapsing problem. We conduct extensive experiments on synthetic and real-world large-scale datasets (MNIST, CIFAR-10, STL-10, ImageNet), where we have made our best effort to compare our D2GAN with the latest state-of-the-art GAN’ s variants in comprehensive qualitative and quantitative evaluations. The experimental results demonstrate the competitive and superior performance of our approach in generating good quality and diverse samples over baselines, and the capability of our method to scale up to ImageNet database.

Published on NIPS in 2017

Dual Discriminator Generative Adversarial Nets

Tu Dinh Nguyen, Trung Le, Hung Vu, Dinh Phung
Deakin University, Geelong, Australia
Centre for Pattern Recognition and Data Analytics
{tu.nguyen, trung.l, hungv, dinh.phung}@deakin.edu.au



(a) GAN. (b) D2GAN.

Figure 1: An illustration of the standard GAN and our proposed D2GAN.

Table 2: Inception scores on CIFAR-10.

Model	Score
Real data	11.24±0.16
WGAN [2]	3.82±0.06
MIX+WGAN [3]	4.04±0.07
Improved-GAN [27]	4.36±0.04
ALI [8]	5.34±0.05
BEGAN [4]	5.62
MAGAN [30]	5.67
DCGAN [24]	6.40±0.05
DFM [31]	7.72±0.13
D2GAN	**7.15±0.07**

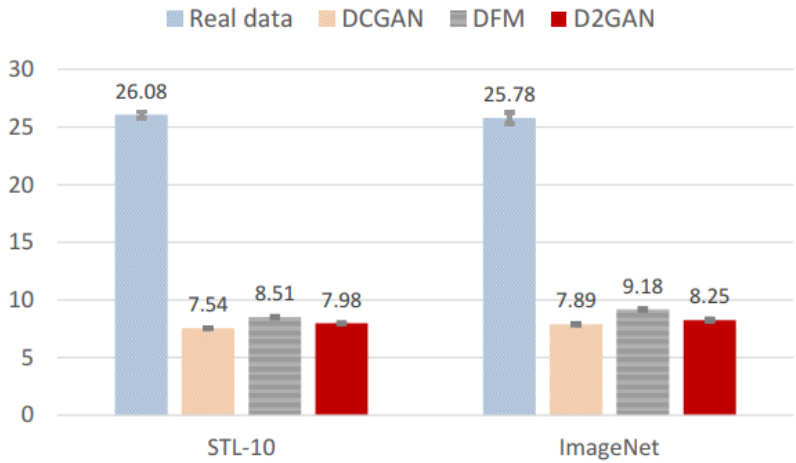


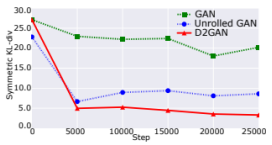
Figure 4: Inception scores on STL-10 and ImageNet.

Dual Discriminator Generative Adversarial Nets

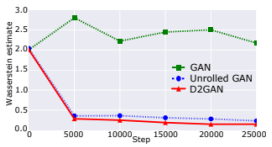
Tu Dinh Nguyen, Trung Le, Hung Vu, Dinh Phung
Deakin University, Geelong, Australia
Centre for Pattern Recognition and Data Analytics
{tu.nguyen, trung.l, hungv, dinh.phung}@deakin.edu.au

- **Conclusion** :To summarize, we have introduced a novel approach to combine Kullback-Leibler (KL) and reverse KL divergences in a unified objective function of the density estimation problem. Our idea is to exploit the complementary statistical properties of two divergences to improve both the quality and diversity of samples generated from the estimator. To that end, we propose a novel framework based on generative adversarial nets (GANs), which formulates a minimax game of three players: two discriminators and one generator, thus termed dual discriminator GAN (D2GAN). Given two discriminators fixed, the learning of generator moves towards optimizing both KL and reverse KL divergences simultaneously, and thus can help avoid mode collapse, a notorious drawback of GANs.

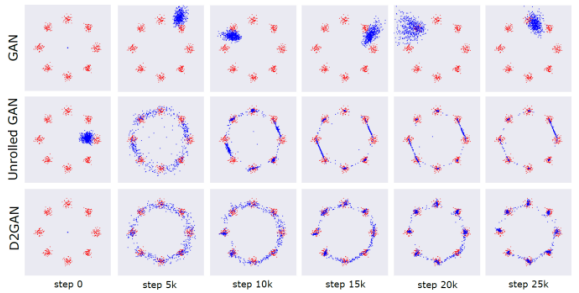
We have established extensive experiments to demonstrate the effectiveness and scalability of our proposed approach using synthetic and large-scale real-world datasets. Compared with the latest state-of-the-art baselines, our model is more scalable, can be trained on the large-scale ImageNet dataset, and obtains Inception scores lower than those of the combination of denoising autoencoder and GAN (DFM), but significantly higher than the others. Finally, we note that our method is orthogonal and could integrate techniques in those baselines such as semi-supervised learning [27], conditional architectures [21, 7, 25] and autoencoder [5, 31].



(a) Symmetric KL divergence.



(b) Wasserstein distance.



(c) Evolution of data (in blue) generated from GAN (top row), UnrolledGAN (middle row) and our D2GAN (bottom row) on 2D data of 8 Gaussians. Data sampled from the true mixture are red.

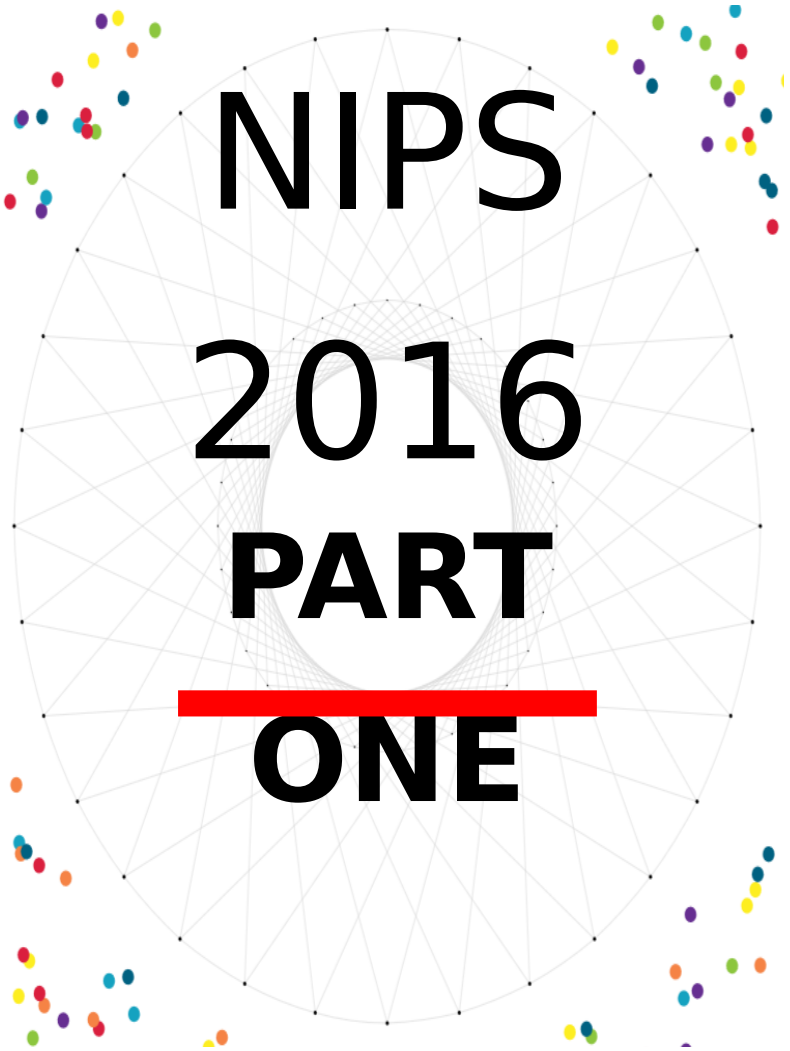
Figure 2: The comparison of standard GAN, UnrolledGAN and our D2GAN on 2D synthetic dataset.

Published on NIPS in 2017

Dual Discriminator Generative Adversarial Nets

Tu Dinh Nguyen, Trung Le, Hung Vu, Dinh Phung
Deakin University, Geelong, Australia
Centre for Pattern Recognition and Data Analytics
{tu.nguyen, trung.l, hungv, dinh.phung}@deakin.edu.au

- **Review** :The paper proposes to train GANs by minimizing lower bounds of KL and reverse KL. The KL and the reverse KL costs are weighted by two hyperparameters. The KL and reverse KL estimators were previously mentioned in f-GAN. It seems that this paper rediscovers the estimators. The theoretical analysis is consistent with the f-GAN properties.
- This paper presents a variant of generative adversarial networks (GANs) that utilizes two discriminators, one tries to assign high scores for data, and the other tries to assign high scores for the samples, both discriminating data from samples, and the generator tries to fool both discriminators. It has been shown in section 3 that the proposed approach effectively optimizes the sum of KL and reverse KL between generator distribution and data distribution in the idealized non-parametric setup, therefore encouraging more mode coverage than other GAN variants.
- The paper proposes dual discriminator GAN (D2GAN), which uses two discriminators and a different loss function for training the discriminators. The approach is clearly explained and I enjoyed reading the paper. The experiments on MoG, MNIST-1K, CIFAR-10, STL-10, ImageNet support the main claims. Overall, I think this is a good paper and I vote for acceptance.
- It is good to know that the minimization of the symmetric KL works well.The experiments show good results on 2D toy data, MNIST-1K and natural images.


NIPS
2016
PART
ONE

Published on NIPS in 2016

f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Sebastian Nowozin, Botond Cseke, Ryota Tomioka
Machine Intelligence and Perception Group
Microsoft Research
{Sebastian.Nowozin, Botond.Cseke, ryoto}@microsoft.com

- **Abstract** :Generative neural samplers are probabilistic models that implement sampling using feed forward neural networks: they take a random input vector and produce a sample from a probability distribution defined by the network weights. These models are expressive and allow efficient computation of samples and derivatives, but cannot be used for computing likelihoods or for marginalization. The generative adversarial training method allows to train such models through the use of an auxiliary discriminative neural network. We show that the generative-adversarial approach is a special case of an existing more general variational divergence estimation approach. We show that any f-divergence can be used for training generative neural samplers. We discuss the benefits of various choices of divergence functions on training complexity and the quality of the obtained generative models.

- **Discussion**:Generative neural samplers offer a powerful way to represent complex distributions without limiting factorizing assumptions. However, while the purely generative neural samplers as used in this paper are interesting their use is limited because after training they cannot be conditioned on observed data and thus are unable to provide inferences. We believe that in the future the true benefits of neural samplers for representing uncertainty will be found in discriminative models and our presented methods extend readily to this case by providing additional inputs to both the generator and variational function as in the conditional GAN model [8]. We hope that the practical difficulties of training with saddle point objectives are not an underlying feature of the model but instead can be overcome with novel optimization algorithms. Further investigations, such as [30], are needed to investigate and hopefully overcome these difficulties.

Acknowledgements. We thank Ferenc Huszar for discussions on the generative-adversarial approach.

f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

More concretely, we make the following contributions over the state-of-the-art:

- We derive the GAN training objectives for all f-divergences and provide as example additional divergence functions, including the Kullback-Leibler and Pearson divergences.
- We simplify the saddle-point optimization procedure of Goodfellow et al. [10] and provide a theoretical justification.
- We provide experimental insight into which divergence function is suitable for estimating generative neural samplers for natural images.

Name	$D_f(P\|Q)$	Generator $f(u)$	$T^*(x)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$	$u \log u$	$1 + \log \frac{p(x)}{q(x)}$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$	$-\log u$	$-\frac{q(x)}{p(x)}$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \, \mathrm{d}x$	$(u-1)^2$	$2(\frac{p(x)}{q(x)} - 1)$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \, \mathrm{d}x$	$(\sqrt{u}-1)^2$	$(\sqrt{\frac{p(x)}{q(x)}} - 1) \cdot \sqrt{\frac{q(x)}{p(x)}}$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x$	$-(u+1) \log \frac{1+u}{2} + u \log u$	$\log \frac{2p(x)}{p(x)+q(x)}$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x - \log(4)$	$u \log u - (u+1) \log(u+1)$	$\log \frac{p(x)}{p(x)+q(x)}$

Table 1: List of f-divergences $D_f(P\|Q)$ together with generator functions. Part of the list of divergences and their generators is based on [26]. For all divergences we have $f : \mathrm{dom}_f \to \mathbb{R} \cup \{+\infty\}$, where f is convex and lower-semicontinuous. Also we have $f(1) = 0$ which ensures that $D_f(P\|P) = 0$ for any distribution P. As shown by [10] GAN is related to the Jensen-Shannon divergence through $D_{\mathrm{GAN}} = 2D_{\mathrm{JS}} - \log(4)$.

Published on NIPS in 2016

f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) \sup_{t \in \mathrm{dom}_{f^*}} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} \mathrm{d}x$$

$$\geq \sup_{T \in \mathcal{T}} \left(\int_{\mathcal{X}} p(x)\, T(x)\, \mathrm{d}x - \int_{\mathcal{X}} q(x)\, f^*(T(x))\, \mathrm{d}x \right)$$

$$= \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{x \sim P}\left[T(x)\right] - \mathbb{E}_{x \sim Q}\left[f^*(T(x))\right] \right),$$

Algorithm 1 Single-Step Gradient Method

1: **function** SINGLESTEPGRADIENTITERATION($P, \theta^t, \omega^t, B, \eta$)
2: Sample $X_P = \{x_1, \ldots, x_B\}$ and $X_Q = \{x'_1, \ldots, x'_B\}$, from P and Q_{θ^t}, respectively.
3: Update: $\omega^{t+1} = \omega^t + \eta \nabla_\omega F(\theta^t, \omega^t)$.
4: Update: $\theta^{t+1} = \theta^t - \eta \nabla_\theta F(\theta^t, \omega^t)$.
5: **end function**

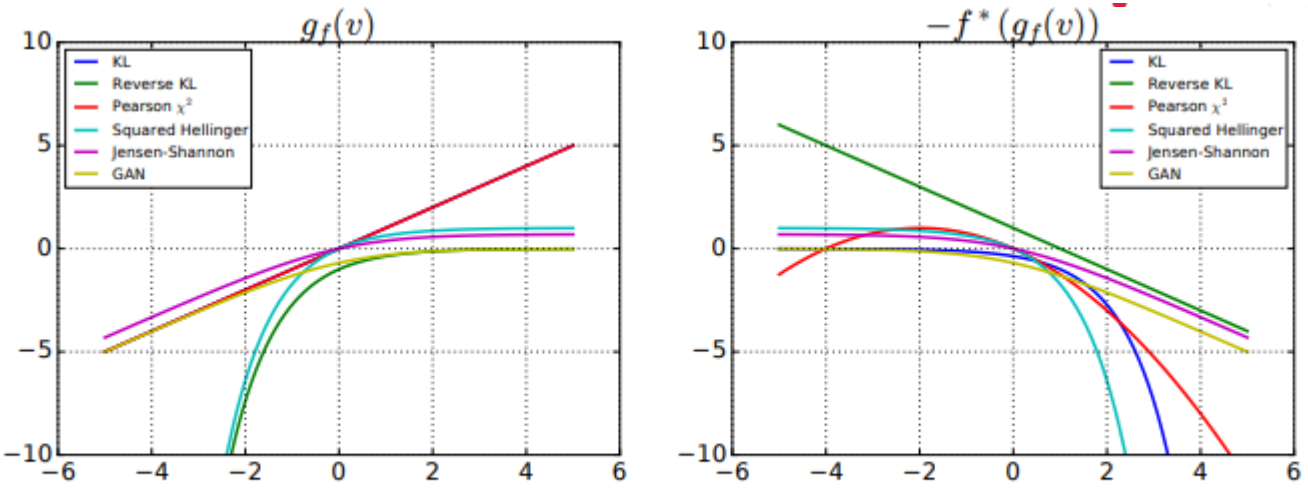


Figure 1: The two terms in the saddle objective (7) are plotted as a function of the variational function $V_\omega(x)$.

f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Name	Output activation g_f	dom_{f^*}	Conjugate $f^*(t)$	$f'(1)$
Kullback-Leibler (KL)	v	$\mathbb{R}$	$\exp(t-1)$	1
Reverse KL	$-\exp(-v)$	$\mathbb{R}_-$	$-1-\log(-t)$	-1
Pearson χ^2	v	$\mathbb{R}$	$\frac{1}{4}t^2+t$	0
Squared Hellinger	$1-\exp(-v)$	$t<1$	$\frac{t}{1-t}$	0
Jensen-Shannon	$\log(2)-\log(1+\exp(-v))$	$t<\log(2)$	$-\log(2-\exp(t))$	0
GAN	$-\log(1+\exp(-v))$	$\mathbb{R}_-$	$-\log(1-\exp(t))$	$-\log(2)$

Table 2: Recommended final layer activation functions and critical variational function level defined by $f'(1)$. The critical value $f'(1)$ can be interpreted as a classification threshold applied to $T(x)$ to distinguish between true and generated samples.

- **Review**:Nicely written and readable paper. The construction of the VDM approach is elegant and is an interesting extension of Nguyen et al. from estimating a divergence to estimating model parameters. The results also show that there are no issues optimizing the other f-divergences and it provides a reason why Goodfellow et al.'s modified objective works better. It would also be interesting to know if the other f-divergences or Alg. 1 make GAN training more stable. The experiments on MNIST are unconvincing as they stand. I also wonder if the KDE estimator approach will be biased towards a particular divergence.

Published on NIPS in 2016

f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

- **Review** :This paper introduces Variational Divergence Minimization (VDM), a novel training criteria for generative models based on f-divergence minimization. Starting from the idea of learning a model Q by minimizing the f-divergence between the empirical distribution P and model Q, the authors derive a mini-max objective function which generalizes the objective optimized by Generative Adversarial Networks (GAN). The GAN objective is recovered for a particular choice of generator function f(u), linked to the Jensen-Shannon divergence. Interestingly, other choices of f(u) can lead to GAN-like objectives which minimize KL-divergence (in either direction) or other f-divergences. Experiments confirm the efficacy of the generalized GAN objective, with visualizations of samples and coarse likelihood estimates (via non-parametric density estimate) on a synthetic toy dataset, MNIST and LSUN datasets.
- This paper shows how the GAN framework can be extended to train the model with many different divergences, rather than just the Jensen-Shannon Divergence.
- The author generalizes the generative adversarial network objective to a rich family of divergences and explore their different behaviors. This includes kullback-leibler and reverse kullback leibler divergences. They also provide a simplified version of the optimization algorithm. Experiments are run on a small synthetic dataset for in depth analysis and on MNIST and LSUN for real world results.

Published on NIPS in 2016

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Xi Chen[†‡], **Yan Duan**[†‡], **Rein Houthooft**[†‡], **John Schulman**[†‡], **Ilya Sutskever**[‡], **Pieter Abbeel**[†‡]
† UC Berkeley, Department of Electrical Engineering and Computer Sciences
‡ OpenAI

- **Abstract** :This paper describes InfoGAN, an information-theoretic extension to the Generative Adversarial Network that is able to learn disentangled representations in a completely unsupervised manner. InfoGAN is a generative adversarial network that also maximizes the mutual information between a small subset of the latent variables and the observation. We derive a lower bound of the mutual information objective that can be optimized efficiently. Specifically, InfoGAN successfully disentangles writing styles from digit shapes on the MNIST dataset, pose from lighting of 3D rendered images, and background digits from the central digit on the SVHN dataset. It also discovers visual concepts that include hair styles, presence/absence of eyeglasses, and emotions on the CelebA face dataset. Experiments show that InfoGAN learns interpretable representations that are competitive with representations learned by existing supervised methods.

- **Conclusion**:This paper introduces a representation learning algorithm called Information Maximizing Generative Adversarial Networks (InfoGAN). In contrast to previous approaches, which require supervision, InfoGAN is completely unsupervised and learns interpretable and disentangled representations on challenging datasets. In addition, InfoGAN adds only negligible computation cost on top of GAN and is easy to train. The core idea of using mutual information to induce representation can be applied to other methods like VAE [3], which is a promising area of future work. Other possible extensions to this work include: learning hierarchical latent representations, improving semi-supervised learning with better codes [31], and using InfoGAN as a high-dimensional data discovery tool.

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

$$
\begin{aligned}
I(c; G(z,c)) &= H(c) - H(c|G(z,c)) \\
&= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\
&= \mathbb{E}_{x\sim G(z,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot|x) \parallel Q(\cdot|x))}_{\geq 0} + \mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\
&\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c)
\end{aligned}
$$

Hence, InfoGAN is defined as the following minimax game with a variational regularization of mutual information and a hyperparameter λ:

$$
\min_{G,Q} \max_{D} V_{\mathrm{InfoGAN}}(D,G,Q) = V(D,G) - \lambda L_I(G,Q) \tag{6}
$$

7.1 Mutual Information Maximization

To evaluate whether the mutual information between latent codes c and generated images $G(z,c)$ can be maximized efficiently with proposed method, we train InfoGAN on MNIST dataset with a uniform categorical distribution on latent codes $c \sim \mathrm{Cat}(K=10, p=0.1)$. In Fig 1, the lower bound $L_I(G,Q)$ is quickly maximized to $H(c) \approx 2.30$, which means the bound (4) is tight and maximal mutual information is achieved.

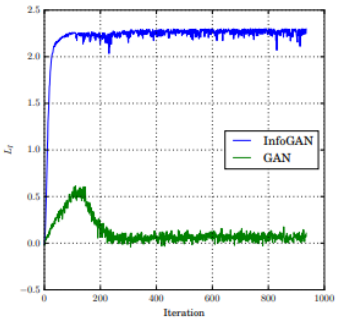


Figure 1: Lower bound L_I over training iterations

As a baseline, we also train a regular GAN with an auxiliary distribution Q when the generator is not explicitly encouraged to maximize the mutual information with the latent codes. Since we use expressive neural network to parametrize Q, we can assume that Q reasonably approximates the true posterior $P(c|x)$ and hence there is little mutual information between latent codes and generated images in regular GAN. We note that with a different neural network architecture, there might be a higher mutual information between latent codes and generated images even though we have not observed such case in our experiments. This comparison is meant to demonstrate that in a regular GAN, there is no guarantee that the generator will make use of the latent codes.

Published on NIPS in 2016

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

- **Review** :The paper presents an extension of the standard GAN framework, that allows to learn a set of disentangled interpretable codes in an unsupervised manner. The approach is motivated from the information-theoretic point of view and is based on minimizing the mutual information between the latent codes and the generated images. The experiments show the results on several types of image data: handwritten digits, house numbers, faces, chairs.
-
- This paper presents an extension of GAN to learn interpretable representations (codes to generate images). The main observation is that the original GAN tends to ignore the additional condition (code) to generate images P(x|z) = P(x|z,c) where c is the code. To address this issue, this paper proposes to maximize the mutual information between the code c and the generated images G(z,c), and derive a variational lower bound for learning with a proposal posterior distribution Q(c|x). The final learning objective is a combination of minmax game loss of GAN and log likelihood of the proposal code distribution. In network implementation, this new code likelihood term is simply added at the end of discriminator with additional linear layers. Experiments are carried out on MNIST, renderings of faces and chairs, SVHN and CelebA with codes being categorical or uniform.
-
- The paper presents InfoGAN that adds information maximization term, which maximizes the information between the latent codes and the generated samples guided by the latent codes, to GAN objective function. By doing so, the latent codes can disentangle latent factors of variation in the training data without explicit supervision on those factors. To train with the mutual information maximization objective, the paper proposes the variational method that lower bounds the mutual information and the objective function becomes tractable. In experiments, the paper demonstrates the effectiveness of the model in disentangling interpretable visual factors using a few latent codes, such as azimuth, lighting, elevation in 3D synthesized face images.

Published on NIPS in 2016

Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling

Jiajun Wu* MIT CSAIL
Chengkai Zhang* MIT CSAIL
Tianfan Xue MIT CSAIL
William T. Freeman MIT CSAIL, Google Research
Joshua B. Tenenbaum MIT CSAIL

- **Abstract** :We study the problem of 3D object generation. We propose a novel framework, namely 3D Generative Adversarial Network (3D-GAN), which generates 3D objects from a probabilistic space by leveraging recent advances in volumetric convolutional networks and generative adversarial nets. The benefits of our model are three-fold: first, the use of an adversarial criterion, instead of traditional heuristic criteria, enables the generator to capture object structure implicitly and to synthesize high-quality 3D objects; second, the generator establishes a mapping from a low-dimensional probabilistic space to the space of 3D objects, so that we can sample objects without a reference image or CAD models, and explore the 3D object manifold; third, the adversarial discriminator provides a powerful 3D shape descriptor which, learned without supervision, has wide applications in 3D object recognition. Experiments demonstrate that our method generates high-quality 3D objects, and our unsupervisedly learned features achieve impressive performance on 3D object recognition, comparable with those of supervised learning methods.
- **Conclusion**:In this paper, we proposed 3D-GAN for 3D object generation, as well as 3D-VAE-GAN for learning an image to 3D model mapping. We demonstrated that our models are able to generate novel objects and to reconstruct 3D objects from images. We showed that the discriminator in GAN, learned without supervision, can be used as an informative feature representation for 3D objects, achieving impressive performance on shape classification. We also explored the latent space of object vectors, and presented results on object interpolation, shape arithmetic, and neuron visualization.

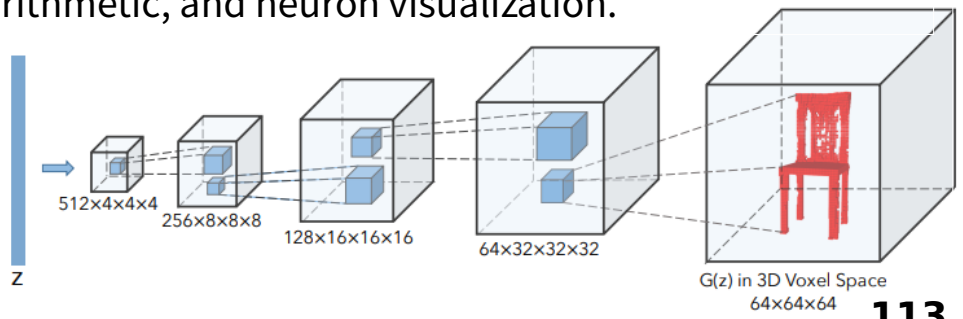


Figure 1: The generator in 3D-GAN. The discriminator mostly mirrors the generator.

developed a recurrent adversarial network for image generation. While previous approaches focus on modeling 2D images, we discuss the use of an adversarial component in modeling 3D objects.

Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling

Supervision	Pretraining	Method	Classification (Accuracy)	
			ModelNet40	ModelNet10
Category labels	ImageNet	MVCNN [Su et al., 2015a]	90.1%	-
		MVCNN-MultiRes [Qi et al., 2016]	**91.4%**	-
	None	3D ShapeNets [Wu et al., 2015]	77.3%	83.5%
		DeepPano [Shi et al., 2015]	77.6%	85.5%
		VoxNet [Maturana and Scherer, 2015]	83.0%	92.0%
		ORION [Sedaghat et al., 2016]	-	**93.8%**
Unsupervised	-	SPH [Kazhdan et al., 2003]	68.2%	79.8%
		LFD [Chen et al., 2003]	75.5%	79.9%
		T-L Network [Girdhar et al., 2016]	74.4%	-
		VConv-DAE [Sharma et al., 2016]	75.5%	80.5%
		3D-GAN (ours)	**83.3%**	**91.0%**

Table 1: Classification results on the ModelNet dataset. Our 3D-GAN outperforms other unsupervised learning methods by a large margin, and is comparable to some recent supervised learning frameworks.

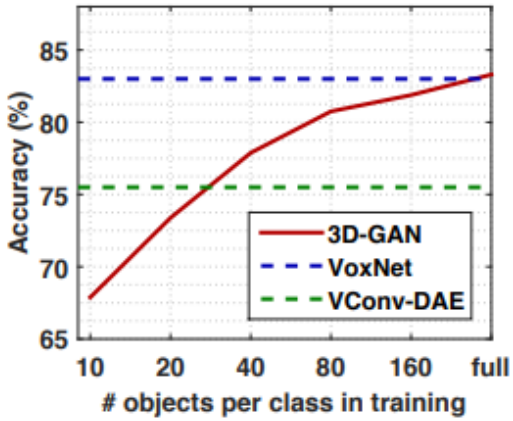


Figure 4: ModelNet40 classification with limited training data

Figure 5: The effects of individual dimensions of the object vector

Figure 6: Intra/inter-class interpolation between object vectors

Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling

- **Review** :The paper presents an extension of the standard GAN framework, that allows to learn a set of disentangled interpretable codes in an unsupervised manner. The approach is motivated from the information-theoretic point of view and is based on minimizing the mutual information between the latent codes and the generated images. The experiments show the results on several types of image data: handwritten digits, house numbers, faces, chairs.
- The paper proposes a generative adversarial networks (GAN) framework for 3D object generation, which is called the Volumetric Adversarial Networks (VAN). In order to adapt the GAN framework for the 3D object generation task, 3D volumetric convolutional architecture is used for the the generator and discriminator. The paper also combines the VAN network with a variational autoencoder for the purpose of synthesizing 3D shape from 2D query image, which is termed as VAE-VAN. A set of experiment results were presented for evaluating the proposed method, which include 1) visual comparison of the generated 3D shape with a prior work, 2) shape classification performance using the unsupervised feature extracted from the discriminator network with several prior works 3) visualization of the generated 3D objects from color images using the VAE-VAN framework, and 4) quantitative comparison of the generated 3D objects from color images using the VAE-VAN framework with several works. The paper also shows the shape arithmetic operation enabled by the VAN and visualizes the neurons in the discriminator.
- The authors propose a latent representation of voxelized volumes suitable both for generative and discriminative purposes. This representation is also linked to cropped 2D images. The system is based on Generative Adversarial Nets (GANs, Goodfellow et al. 2014), where the generative net is replaced by an VAE encoder, as in Larsen et al. 2016. The present manuscript focuses on 3D volume data, as opposed to 2D images from those two publications. The method shows qualitative and quantitative experiments related to 3D object generation, 3D object classification and 3D reconstruction from RGB images, with good results compared to the state of the art.

Published on NIPS in 2016

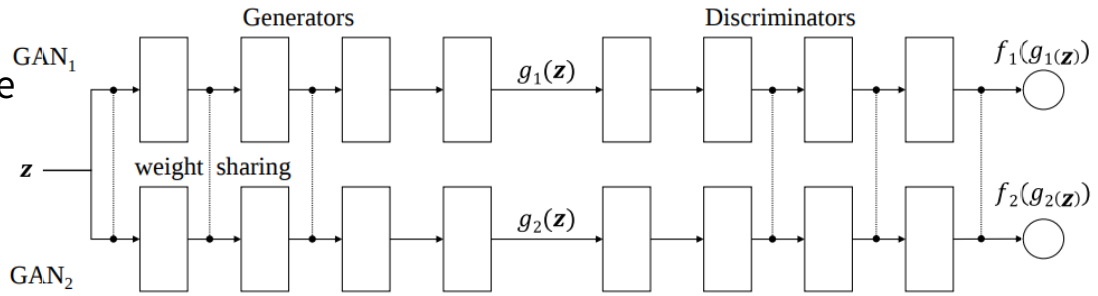
Coupled Generative Adversarial Network

Ming-Yu Liu
Mitsubishi Electric Research Labs (MERL),
mliu@merl.com

Oncel Tuzel
Mitsubishi Electric Research Labs (MERL),
oncel@merl.com

- **Abstract** :We propose coupled generative adversarial network (CoGAN) for learning a joint distribution of multi-domain images. In contrast to the existing approaches, which require tuples of corresponding images in different domains in the training set, CoGAN can learn a joint distribution without any tuple of corresponding images. It can learn a joint distribution with just samples drawn from the marginal distributions. This is achieved by enforcing a weight-sharing constraint that limits the network capacity and favors a joint distribution solution over a product of marginal distributions one. We apply CoGAN to several joint distribution learning tasks, including learning a joint distribution of color and depth images, and learning a joint distribution of face images with different attributes. For each task it successfully learns the joint distribution without any tuple of corresponding images. We also demonstrate its applications to domain adaptation and image transformation.

Conclusion:We presented the CoGAN framework for learning a joint distribution of multi-domain images. We showed that via enforcing a simple weight-sharing constraint to the layers that are responsible for decoding abstract semantics, the CoGAN learned the joint distribution of images by just using samples drawn separately from the marginal distributions. In addition to convincing image generation results on faces and RGBD images, we also showed promising results of the CoGAN framework for the image transformation and unsupervised domain adaptation tasks.



Coupled Generative Adversarial Network

Learning: The CoGAN framework corresponds to a constrained minimax game given by

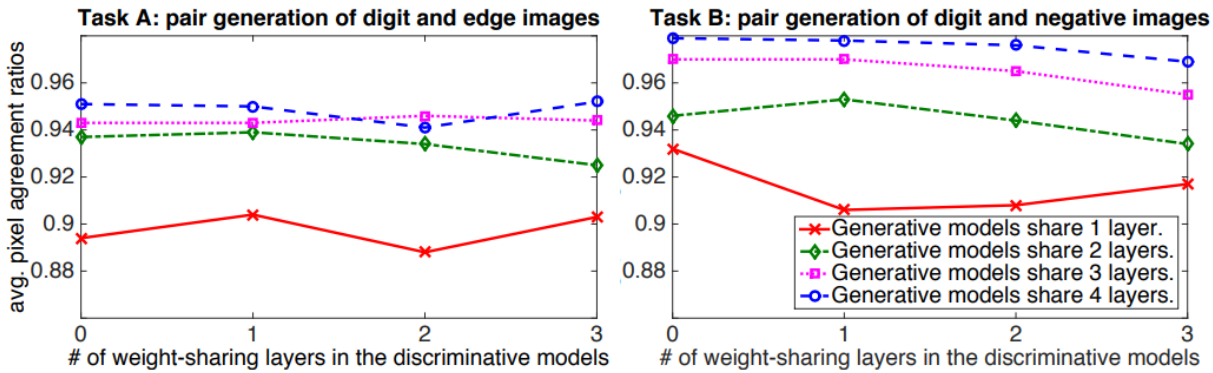
$$\max_{g_1,g_2} \min_{f_1,f_2} V(f_1, f_2, g_1, g_2), \quad \text{subject to} \quad \boldsymbol{\theta}_{g_1^{(i)}} = \boldsymbol{\theta}_{g_2^{(i)}}, \quad \text{for } i = 1, 2, ..., k \tag{2}$$

$$\boldsymbol{\theta}_{f_1^{(n_1-j)}} = \boldsymbol{\theta}_{f_2^{(n_2-j)}}, \text{ for } j = 0, 1, ..., l-1$$

where the value function V is given by

$$\begin{aligned} V(f_1, f_2, g_1, g_2) = & E_{\mathbf{x}_1 \sim p_{\mathbf{X}_1}}[-\log f_1(\mathbf{x}_1)] + E_{\mathbf{z} \sim p_{\mathbf{Z}}}[-\log(1 - f_1(g_1(\mathbf{z})))] \\ & + E_{\mathbf{x}_2 \sim p_{\mathbf{X}_2}}[-\log f_2(\mathbf{x}_2)] + E_{\mathbf{z} \sim p_{\mathbf{Z}}}[-\log(1 - f_2(g_2(\mathbf{z})))]. \end{aligned} \tag{3}$$

In the game, there are two teams and each team has two players. The generative models form a team and work together for synthesizing a pair of images in two different domains for confusing the discriminative models. The discriminative models try to differentiate images drawn from the training data distribution in the respective domains from those drawn from the respective generative models. The collaboration between the players in the same team is established from the weight-sharing constraint. Similar to GAN, CoGAN can be trained by back propagation with the alternating gradient update steps. The details of the learning algorithm are given in the supplementary materials.



Published on NIPS in 2016

Coupled Generative Adversarial Network

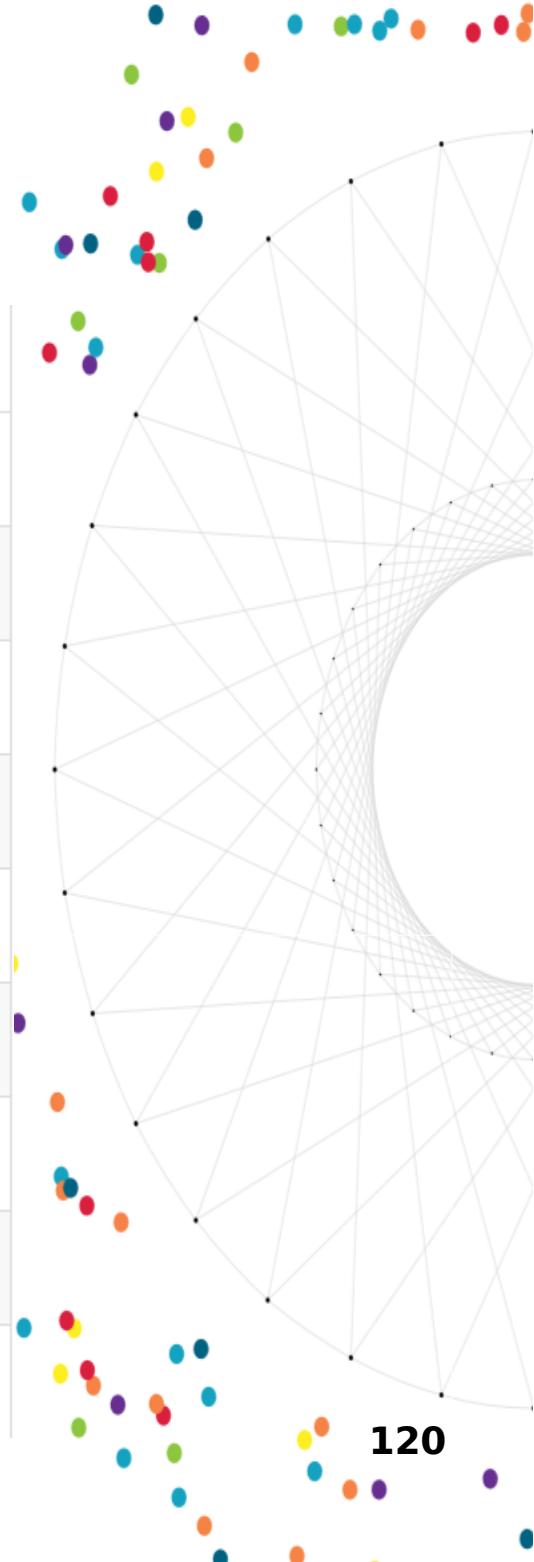
- **Review** :The paper proposes a method for generating pairs of corresponding images in two different domains. The method does not rely on existence of paired images in the training set. The idea is to train a pair of GANs that share several layers of weights.
-
- The paper proposes a method for learning generative models of pairs of corresponding images belonging to two different domains (e.g. the RGB image of a scene and the corresponding depth image). The method is based on two adversarial networks with partially shared weights. The generative networks share the weights that map the noise to an intermediate code but have separate weights that map from the intermediate code to each image type. This induces the model to generate pairs of corresponding images even when it is never trained with corresponding image pairs. The authors evaluate the proposed method with several image datasets, and also provide a demonstration of how it can be applied to domain adaptation and cross-domain transformation problems.
- The paper presents the coupled GAN (CoGAN) for learning to generate pair of images with different attributes or from different domains without knowing per-example correspondences. The key idea is to train a coupled GAN model jointly while sharing weights of the higher layers of both generative and discriminative models of GAN. In experiments, the paper demonstrates effectiveness of CoGAN at generating images at different domains and with different attributes. As an application of CoGAN, the paper proposes unsupervised domain adaptation as well as cross-domain image transformation, which shows promising results.
- This paper proposed a simple yet effective way to learn common semantics from data that have different low-level feature statistics. Fo example, for digits and their intensity inverted versions, the semantics are exactly the same while the pixel values are highly different. The paper achieves this by having two GANs with the high-level layers, which loosely correspond to semantics rather than feature encoding/decoding, shared between the two GANs. Training are then carried out in a purely unsupervised way without the need of explicit pairs of samples.

Summary

PART Three

Summary

Solved Problem	nums of paper
Unstable(Generator hard to train)	20
Pull in AutoEncoder(Semi/Unsupervised)	12
Mode Collapse	7
GAN application	6
Nash Equilibrium	3
Leverage Auxiliary Information	3
Computational efficiency	2
Lipschitz Continuous	1
Lable GAN	1




Thanks for watching
PRESENTED BY 薛铭龙