GAN survey

PRESENTED BY 薛铭龙


ICLR
2018
PART
ONE

Published on arxiv in 2017

Improved Training of Wasserstein GANs

Ishaan Gulrajani[1]*, **Faruk Ahmed**[1], **Martin Arjovsky**[2], **Vincent Dumoulin**[1], **Aaron Courville**[1,3]

*Now at Google Brain

[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

主要内容

- **Problem** : WGAN requires that the discriminator (called the critic in that work) must lie within the space of 1-Lipschitz functions, which the authors enforce through weight clipping.
- **Contribution:**

1) On toy datasets, we demonstrate how critic weight clipping can lead to undesired behavior.

2) We propose gradient penalty (WGAN-GP), which does not suffer from the same problems.

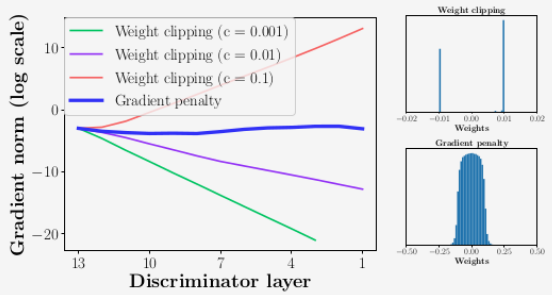
3) We demonstrate stable training of varied GAN architectures, performance improvements over weight clipping, high-quality image generation, and a character-level GAN language model without any discrete sampling.

网络结构及实验结果

WGAN目标函数：

$$\min_G \max_{D \in \mathcal{D}} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r} \left[D(\boldsymbol{x}) \right] - \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g} \left[D(\tilde{\boldsymbol{x}}) \right]$$

where $\mathcal{D}$ is the set of 1-Lipschitz functions



(b) (left) Gradient norms of deep WGAN critics during training on the Swiss Roll dataset either explode or vanish when using weight clipping, but not when using a gradient penalty. (right) Weight clipping (top) pushes weights towards two values (the extremes of the clipping range), unlike gradient penalty (bottom).

设置惩罚项使 f 满足 1 -lipschitz 条件：

$$L = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g} \left[D(\tilde{\boldsymbol{x}}) \right] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r} \left[D(\boldsymbol{x}) \right]}_{\text{Original critic loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}} \left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2 \right]}_{\text{Our gradient penalty}}. \tag{3}$$

On the regularization of Wasserstein GANs

Henning Petzka*
Fraunhofer Institute IAIS,
Sankt Augustin, Germany
henning.petzka@gmail.com

Asja Fischer* & Denis Lukovnikov
Department of Computer Science,
University of Bonn, Germany
asja.fischer@gmail.com
lukovnik@cs.uni-bonn.de

主要内容

- **Problem** : Weight clipping is a clearly terrible way to enforce a Lipschitz constraint. If the clipping parameter is large, then it can take a long time for any weights to reach their limit, thereby making it harder to train the critic till optimality. If the clipping is small, this can easily lead to vanishing gradients when the number of layers is big, or batch normalization is not used (such as in RNNs).
- **Method** : We present theoretical arguments why using a weaker regularization term enforcing the Lipschitz constraint is preferable. For stable training of Wasserstein GANs, we propose to use the following penalty term to enforce the Lipschitz constraint that appears in the objective function:

$$\mathbb{E}_{\hat{x}\sim\tau}[(\max\{0, ||\nabla f(\hat{x})|| - 1\})^2]$$

- **Result**: We presented theoretical and empirical evidence that this gradient penalty performs better than the previously considered approaches of clipping weights and of applying the stronger gradient penalty given by $\mathbb{E}_{\hat{x}\sim\tau}[(||\nabla f(\hat{x})||_2 - 1)^2]$ In addition to more stable learning behavior, the proposed regularization term leads to lower sensitivity to the value of the penalty weight λ

实验结果

$$\mathbb{E}_{y\sim\nu}[f(y)] - \mathbb{E}_{x\sim\mu}[f(x)] + \lambda\mathbb{E}_{\hat{x}\sim\tau}[(||\nabla f(\hat{x})||_2 - 1)^2] \ ,$$

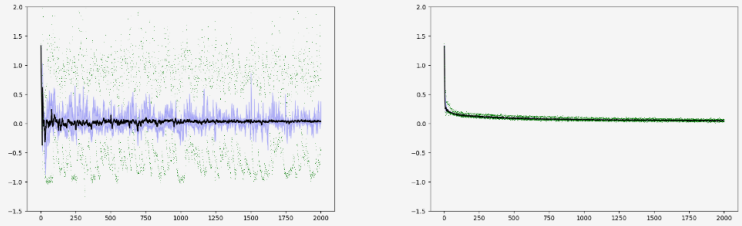


Figure 4: Evolution of the negative of WGAN critic's loss (without the regularization term) for $\lambda = 5$. Median results over the 20 runs (blue area indicates quantiles, green dots outliers). Left: For the GP-penalty. Right: For the LP-penalty.

PENALTY WEIGHT	WGAN-GP	WGAN-LP
0.1	7.781 (± 0.104)	8.017(± 0.075)
5	7.817 (± 0.095)	7.859 (± 0.085)
10	7.840 (± 0.066)	7.989 (± 0.119)
100	7.548 (± 0.102)	7.815 (± 0.038)
200	7.472 (± 0.070)	7.721 (± 0.105)

Table 1: Inception Score on CIFAR-10. Reported are the maximal mean values reached during training. Means are computed over 10 image sets, variances given in parenthesis.

Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect

Xiang Wei[1,2*], **Boqing Gong**[3*], **Zixia Liu**[1], **Wei Lu**[2], **Liqiang Wang**[1]
[1]Department of Computer Science, University of Central Florida, Orlando, FL, USA 32816
[2]School of Software Engineering, Beijing Jiaotong University, Beijing, China 100044
[3]Tencent AI Lab, Bellevue, WA, USA 98004
yqweixiang@knights.ucf.edu, boqinggo@outlook.com
zixia@knights.ucf.edu, luwei@bjtu.edu.cn, lwang@cs.ucf.edu

主要内容

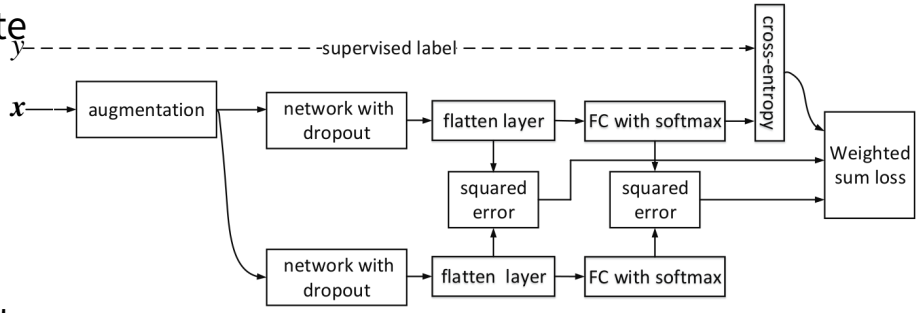
- **Problem** : Unlike the weight clipping, however, by no means one can penalize everywhere using this term through a finite number of training iterations. As a result, the gradient penalty term b , leaving significant parts of the support domain not GP takes effect only upon the sampled points x examined at all
- **Method**:Moreover, instead of focusing on one particular data point at a time, we devise a regularization over a pair of data points drawn near the manifold following the most basic definition of the 1-Lipschitz continuity.

Immediately, we can add the following soft consistency term (CT) to the value function of WGAN in order to penalize the violations to the inequality in eq. (3),

$$CT|_{\boldsymbol{x}_1,\boldsymbol{x}_2} = \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\max\left(0, \frac{d(D(\boldsymbol{x}_1), D(\boldsymbol{x}_2))}{d(\boldsymbol{x}_1, \boldsymbol{x}_2)} - M'\right)\right] \quad (4)$$

- Following (Laine & Aila, 2016), we add our consistency regularization $CT|_{\boldsymbol{x}',\boldsymbol{x}''}$ to the objective function of the semi-supervised training, in order to take advantage of the effect of temporal ensembling.

网络结构及实验结果



(a) GP-WGAN (b) Ours

Method	Test error (%)
Ladder (Rasmus et al., 2015)	1.06±0.37
VAT (Miyato et al., 2017)	1.36
CatGAN (Springenberg, 2015)	1.39±0.28
Improved GAN (Salimans et al., 2016)	0.93 ± 0.065
Triple GAN (Li et al., 2017a)	0.91 ± 0.58
Our CT-GAN	**0.89 ± 0.13**

Spectral Normalization for Generative Adversarial Networks

Youngjin Kim, Minjung Kim & Gunhee Kim
Department of Computer Science and Engineering, Seoul National University, Seoul, Korea
{youngjin.kim,minjung.kim,gunhee.kim}@vision.snu.ac.kr

主要内容

网络结构及实验结果

- **Problem** :One of the challenges in the study of generative adversarial networks is the instability of its training.
- **Method**:In this paper, we propose a novel weight normalization technique called spectral normalization to stabilize the training of the discriminator.
- **Result**:We tested the efficacy of spectral normalization onCIFAR10, STL-10, and ILSVRC2012 dataset, and we experimentally confirmed that spectrally normalized GANs (SN-GANs) is capable of generating images of better or equal quality relative to the previous training stabilization techniques.

Method	Inception score		FID	
	CIFAR-10	STL-10	CIFAR-10	STL-10
Real data	11.24±.12	26.08±.26	7.8	7.9
-Standard CNN-				
Weight clipping	6.41±.11	7.57±.10	42.6	64.2
GAN-GP	6.93±.08		37.7	
WGAN-GP	6.68±.06	8.42±.13	40.2	55.1
Batch Norm.	6.27±.10		56.3	
Layer Norm.	7.19±.12	7.61±.12	33.9	75.6
Weight Norm.	6.84±.07	7.16±.10	34.7	73.4
Orthonormal	7.40±.12	8.56±.07	29.0	46.7
(ours) SN-GANs	7.42±.08	8.28±.09	29.3	53.1
Orthonormal (2x updates)		8.67±.08		44.2
(ours) SN-GANs (2x updates)		8.69±.09		47.5
(ours) SN-GANs, Eq.(17)	7.58±.12		25.5	
(ours) SN-GANs, Eq.(17) (2x updates)		8.79±.14		43.2
-ResNet-[5]				
Orthonormal, Eq.(17)	7.92±.04	8.72±.06	23.8±.58	42.4±.99
(ours) SN-GANs, Eq.(17)	**8.22**±.05	**9.10**±.04	**21.7**±.21	**40.1**±.50
DCGAN†	6.64±.14	7.84±.07		
LR-GANs‡	7.17±.07			
Warde-Farley et al.*	7.72±.13	8.51±.13		
WGAN-GP (ResNet)††	7.86±.08			

MGAN: Training generative adversarial nets with multiple generators

Quan Hoang
University of Massachusetts-Amherst
Amherst, MA, USA
qhoang@umass.edu

Tu Dinh Nguyen, Trung Le, Dinh Phung
PRaDA Centre, Deakin University
Geelong, Australia
{tu.nguyen,trung.l,dinh.phung}
@deakin.edu.au

主要内容

- **Problem** : We propose in this paper a new approach to train the Generative Adversarial Nets (GANs) with a mixture of generators to overcome the **mode collapsing** problem.
- **Method**:The main intuition is to employ multiple generators, instead of using a single one as in the original GAN. Our idea is to approximate data distribution using a mixture of multiple distributions wherein each distribution captures a subset of data modes separately from those of others. To achieve this goal, we propose a minimax game of one discriminator, one classifier and many generators to formulate an optimization problem that minimizes the JSD between P_{data} and P_{model}
- **Result**:

1) achieving state-of-the-art Inception scores;
2) generating diverse and appealing recognizable objects at different resolutions;
3) specializing in capturing different types of objects by the generators

网络结构及实验结果

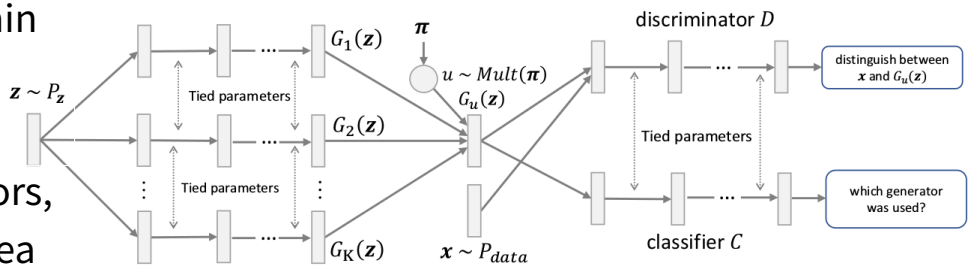


Figure 1: MGAN's architecture with K generators, a binary discriminator, a multi-class classifier.

Fréchet Inception Distance results. One disadvantage of the Inception score is that it does not compare the statistics of real world samples and those of synthetic examples. Therefore, we further evaluate MGAN using the *Fréchet Inception Distance* (FID) proposed in (Heusel et al., 2017). Let p and q be the distributions of the representations obtained by projecting real and synthetic samples to the last hidden layer in Inception model (Szegedy et al., 2015). Assuming that p and q are both multivariate Gaussian distributions, FID measures the *Fréchet distance* (Dowson & Landau, 1982) which is also the 2-Wasserstein distance, between the two distributions. Tab. 2 compares the FIDs obtained by MGAN with baselines collected in (Heusel et al., 2017). It is noteworthy that lower FID is better, and that WGAN-GP and WGAN-GP + TTUR uses the ResNet architecture while MGAN employs the DCGAN architecture. In terms of FID, MGAN is roughly 28% better than DCGAN and DCGAN + TTUR, 9% better than WGAN-GP and 8% weaker than WGAN-GP + TTUR. This result further proves that MGAN helps address the mode collapsing problem.

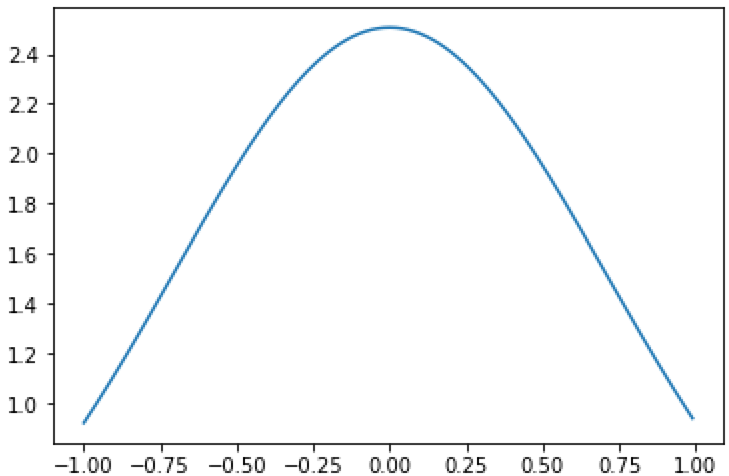
Table 2: FIDs (lower is better) on CIFAR-10.

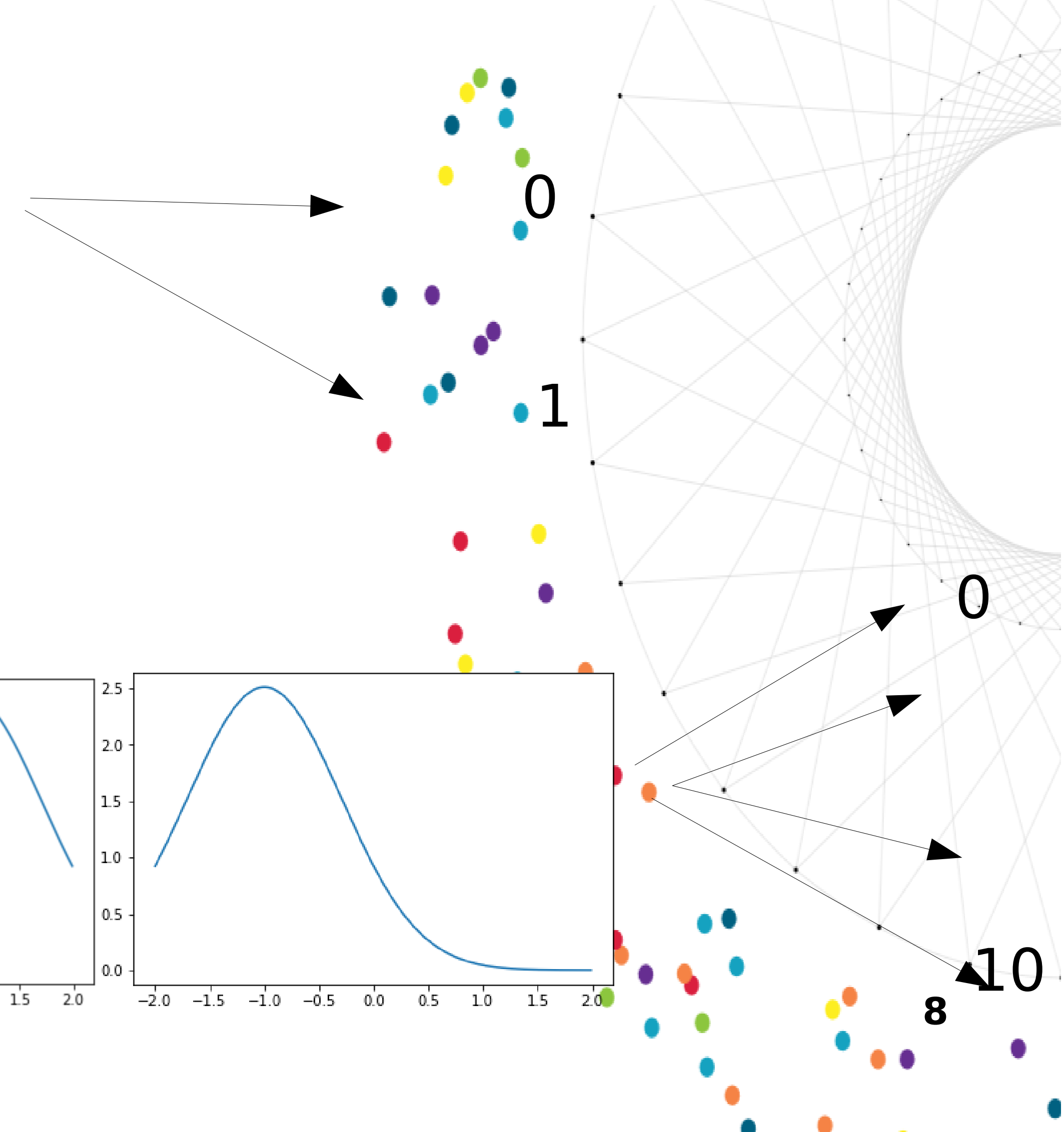
Model	FID
DCGAN (Radford et al., 2015)	37.7
DCGAN + TTUR (Heusel et al., 2017)	36.9
WGAN-GP (Gulrajani et al., 2017)	29.3
WGAN-GP + TTUR (Heusel et al., 2017)	24.8
MGAN	**26.7**

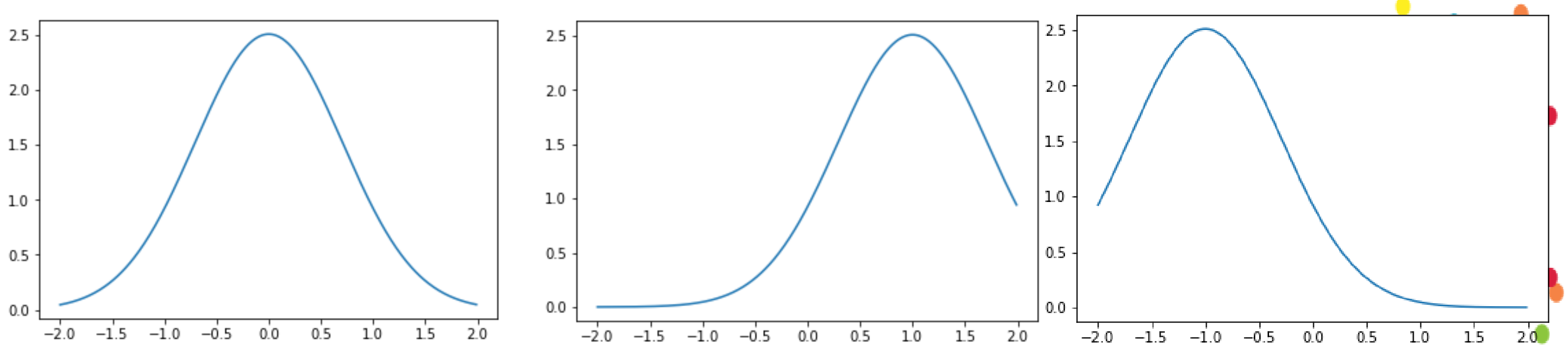
MGAN: Training generative adversarial nets with multiple generators

Quan Hoang
University of Massachusetts-Amherst
Amherst, MA, USA
qhoang@umass.edu

Tu Dinh Nguyen, Trung Le, Dinh Phung
PRaDA Centre, Deakin University
Geelong, Australia
{tu.nguyen,trung.l,dinh.phung}
@deakin.edu.au







CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training

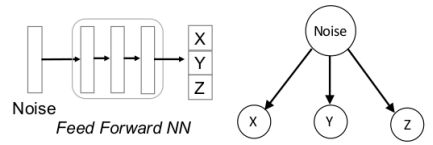
Murat Kocaoglu*, Christopher Snyder*, Alexandros G. Dimakis, Sriram Vishwanath

Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX, USA
mkocaoglu@utexas.edu, 22csnyder@gmail.com,
dimakis@austin.utexas.edu, sriram@austin.utexas.edu

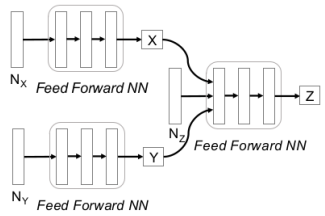
主要内容

网络结构及实验结果

- **Problem** : An extension of GANs is to enable sampling from the class conditional data distributions by feeding class labels to the generator alongside the noise vectors. However, these architectures do not capture the dependence between the labels. **As far as we are aware of, in all of these works, the class labels are chosen independently from one another.**
- **Method**:We proposed a novel generative model with label inputs. In addition to being able to create samples conditioned on labels, our generative model can also sample from the interventional distributions.
- **Result**:We show that the proposed architectures can be used to sample from observational and interventional image distributions, even for interventions which do not naturally occur in the dataset.



(a) Naive feedforward generator architecture and the causal graph it represents.



(b) Generator neural network architecture that represent the causal graph $X \to Z \leftarrow Y$.

Figure 2: (a) The causal graph implied by the naive feedforward generator architecture. (b) A neural network implementation of the causal graph $X \to Z \leftarrow Y$: Each feed forward neural net captures the function f in the structural equation model $V = f(Pa_V, E)$.

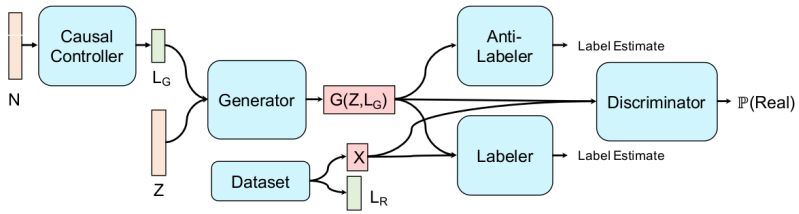


Figure 3: CausalGAN architecture: Causal controller is a pretrained causal implicit generative model for the image labels. Labeler is trained on the real data, Anti-Labeler is trained on generated data. Generator minimizes Labeler loss and maximizes Anti-Labeler loss.

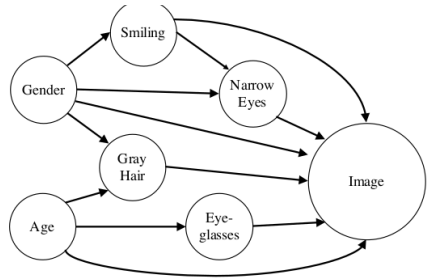


Figure 3: A plausible causal model for image generation.

cGANs with Projection Discriminator

Takeru Miyato[1], **Masanori Koyama**[2]
miyato@preferred.jp
koyama.masanori@gmail.com
[1]Preferred Networks, Inc. [2]Ritsumeikan University

主要内容

- **Problem** : Conditional GANs (cGANs) are a type of GANs that use conditional information. Unlike in standard GANs, the discriminator of cGANs discriminates between the generator distribution and the target distribution on the set of the pairs of generated samples x and its intended conditional variable y. To the authors' knowledge, most frameworks of discriminators in cGANs at the time of writing feeds the pair the conditional information y into the discriminator by naively concatenating (embedded) y to the input or to the feature vector at some middle layer. **Mode collapse**
- **Method**:We propose a novel, projection based way to incorporate the conditional information into the discriminator of GANs that respects the role of the conditional information in the underlining probabilistic model.

$$\mathcal{L}(D) = -E_{q(\boldsymbol{y})}\left[E_{q(\boldsymbol{x}|\boldsymbol{y})}\left[\log(D(\boldsymbol{x},\boldsymbol{y}))\right]\right] - E_{p(\boldsymbol{y})}\left[E_{p(\boldsymbol{x}|\boldsymbol{y})}\left[\log(1-D(\boldsymbol{x},\boldsymbol{y}))\right]\right], \quad (1)$$

$$f^*(\boldsymbol{x},\boldsymbol{y}) = \log\frac{q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})} = \log\frac{q(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{x})} + \log\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} := r(\boldsymbol{y}|\boldsymbol{x}) + r(\boldsymbol{x}). \quad (2)$$

网络结构及实验结果

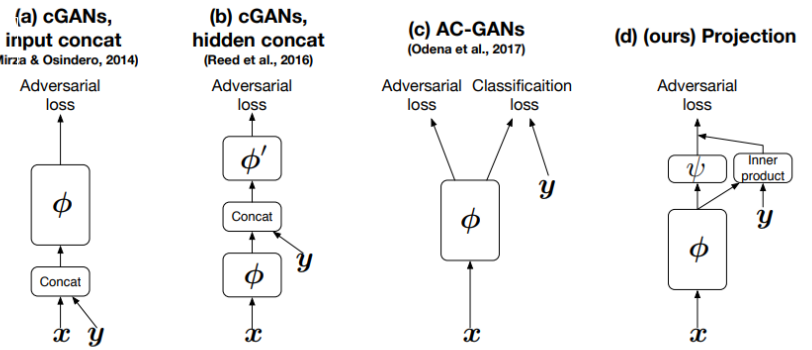


Figure 1: Discriminator models for conditional GANs

Table 1: Inception score and intra FIDs on ImageNet.

Method	Inception Score	Intra FID
AC-GANs	28.5±.20	260.0
concat	21.1±.35	141.2
projection	**29.7**±.61	**103.1**
*projection (850K iteration)	**36.8**±.44	**92.4**

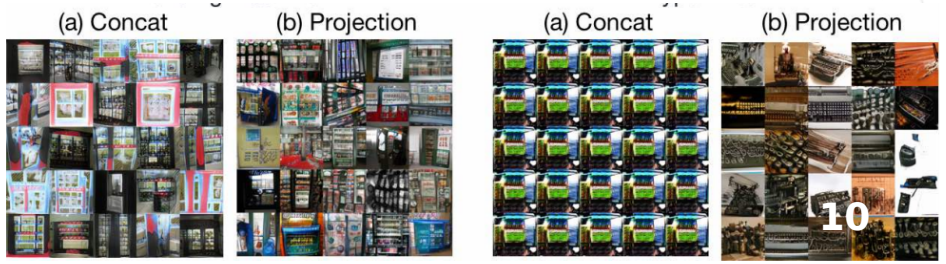


Figure 6: Collapsed images on the *concat* model.

Activation Maximization Generative Adversarial Nets

Zhiming Zhou, Han Cai
Shanghai Jiao Tong University
heyohai,hcai@apex.sjtu.edu.cn

Shu Rong
Yitu Tech
shu.rong@yitu-inc.com

Yuxuan Song, Kan Ren
Shanghai Jiao Tong University
songyuxuan,kren@apex.sjtu.edu.cn

Jun Wang
University College London
j.wang@cs.ucl.ac.uk

Weinan Zhang, Yu Yong
Shanghai Jiao Tong University
wnzhang@sjtu.edu.cn, yyu@apex.sjtu.edu.cn

主要内容

网络结构及实验结果

- **Problem** :By taking the class labels into account, these GAN models show improved generation quality and stability. However, the mechanisms behind them have not been fully explored (Goodfellow, 2016).

1) CatGAN (Springenberg,2015) builds the discriminator as a multi-class classifier

2) LabelGAN (Salimans et al., 2016) extends the discriminator with one extra class for the generated samples

3) AC-GAN (Odena et al., 2016) jointly trains the real-fake discriminator and an auxiliary classifier for the specific real classes.

- **Method**:With class aware gradient and cross-entropy decomposition, we reveal how classlabels and associated losses influence GAN’ s training. Based on that, we propose Activation Maximization Generative Adversarial Networks (AM-GAN) as an advanced solution.
- **Result**:where AM-GAN outperforms other strong baselines and achieves state-of-the-art Inception Score (8.91)on CIFAR-10.Inception Score mainly tracks the diversity of the generator

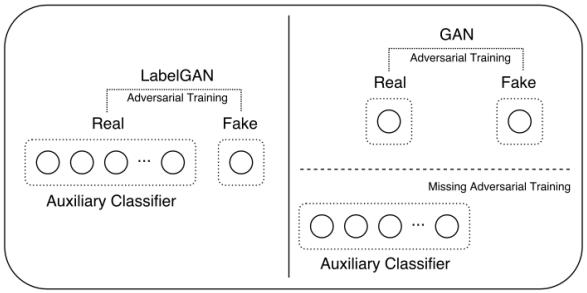


Figure 2: AM-GAN (left) v.s. AC-GAN* (right). AM-GAN can be viewed as a combination of LabelGAN and auxiliary classifier, while AC-GAN* is a combination of vanilla GAN and auxiliary classifier. AM-GAN can naturally conduct adversarial training among all the classes, while in AC-GAN*, adversarial training is only conducted at the real-fake level and missing in the auxiliary classifier.

Model	Score ± Std.
DFM (Warde-Farley & Bengio, 2017)	7.72 ± 0.13
Improved GAN (Salimans et al., 2016)	8.09 ± 0.07
AC-GAN (Odena et al., 2016)	8.25 ± 0.07
WGAN-GP + AC (Gulrajani et al., 2017)	8.42 ± 0.10
SGAN (Huang et al., 2016b)	8.59 ± 0.12
AM-GAN (our work)	**8.91** ± **0.11**
Splitting GAN (Guillermo et al., 2017)	8.87 ± 0.09
Real data	11.24 ± 0.12

Table 3: Inception Score comparison on CIFAR-10. Splitting GAN uses the class splitting technique to enhance the class label information, which is orthogonal to AM-GAN.

Improving GAN Training via Binarized Representation Entropy (BRE) Regularization

Yanshuai Cao, Gavin Weiguang Ding, Kry Yik-Chau Lui, Ruitong Huang
Borealis AI
Canada

主要内容

- **Problem** : The capacity of D plays an essential role in giving G sufficient learning guidances to model the complex real data distribution. With insufficient capacity, D could fail to distinguish real and generated data distributions even when their Jensen-Shannon divergence or Wasserstein distance is not small. In this work, we demonstrate that even with sufficient maximum capacity, D might not allocate its capacity in a desirable way that facilitates convergence to a good equilibrium.
- **Method** : We then propose a novel regularizer to guide D to have a better model capacity allocation. Our regularizer is constructed to encourage D’ s hidden binary activation patterns to have high joint entropy, based on a connection between the model capacity of a rectifier net and its internal binary activation patterns.
- **Result** : Our experiments show that such high entropy representation leads to faster convergences, improved sample quality, as well as lower errors in semi-supervised learning.

网络结构

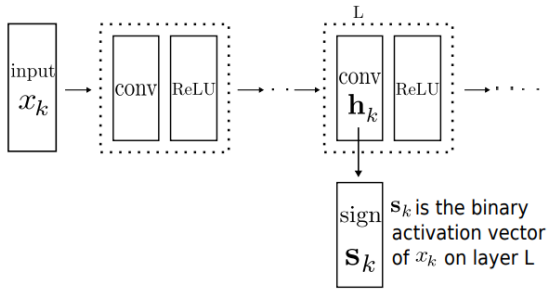


Figure 2: Activation vector $\mathbf{s}_k$ of a sample x_k on a layer L immediately before nonlinearity.

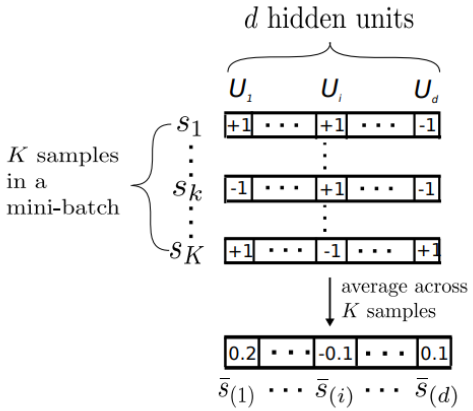


Figure 3: Notations for defining R_{BRE}.

Quantitatively Evaluating GANs With Divergences Proposed for Training

Daniel Jiwoong Im[1,2], **He Ma**[3,4], **Graham Taylor**[3,4], **& Kristin Branson**[1]
[1]Janelia Research Campus, HHMI, [2]AIFounded Inc. [3]University of Guelph, [4]Vector Institute
{imd, bransonk}@janelia.hhmi.org
{hma02, gtaylor}@uoguelph.ca

主要内容

- **Problem** : Generative adversarial networks (GANs) have been extremely effective in approximating complex distributions of high-dimensional, input data samples, and substantial progres has been made in understanding and improving GAN performance in terms of both theory and application. **However, we currently lack quantitative methods for model assessment.**
- **Method** : In this paper, we proposed to use four well-known distance functions as an evaluation metrics, and empirically investigated the DCGAN, W-DCGAN, and LS-DCGAN families under these metrics. Previously, these models were compared based on **visual assessment** of sample quality and difficulty of training. In our experiments, we showed that there are performance differences in terms of average experiments, but that some are not statistically significant. Moreover, we thoroughly analyzed the performance of GANs under different hyper-parameter settings.
- **Result** :

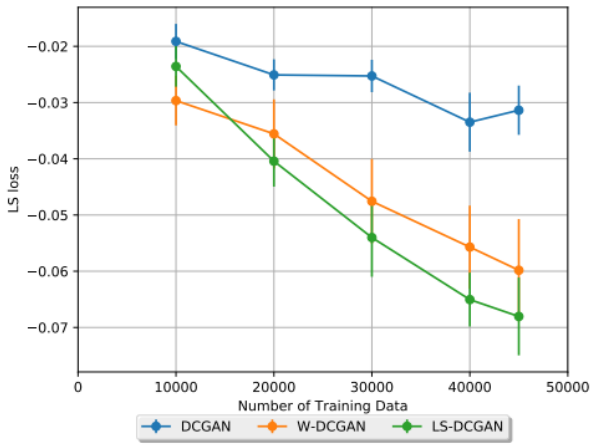
1) The larger the GAN architecture, the better the results;
2) Having a generator network larger than the discriminator network does not yield good results;
3) the best ratio between discriminator and generator updates depend on the data set;
4) the W-DCGAN and LS-DCGAN performance increases much faster than DCGAN as the number of training examples grows

损失函数及实验结果

Metric	μ	v	Function Class
GAN (GC)	$\log f$	$-\log(1-f)$	$\mathcal{X} \to \mathbb{R}_+, \exists M \in \mathbb{R} : \lvert f(x)\rvert \leq M$
Least-Squares GAN (LS)	$-(f-b)^2$	$(f-a)^2$	$\mathcal{X} \to \mathbb{R}, \exists M \in \mathbb{R} : \lvert f(x)\rvert \leq M$
MMD	f	f	$f : \lVert f\rVert_{\mathcal{H}} \leq 1$
Wasserstein (IW)	f	f	$f : \lVert f\rVert_L \leq 1$

Algorithm 1 Compute the divergence/distance.

procedure DIVERGENCECOMPUTATION(Dataset $\{X_{tr}, X_{te}\}$, generator G_θ, learning rate η, evaluation criterion $J(\varphi, X, Y)$)
  Initialize critic network parameter φ.
  for $i = 1 \cdots N$ **do**
    Sample data points from X, $\{x_m\} \sim X_{tr}$.
    Sample points from generative model, $\{s_m\} \sim G_\theta$.
    $\varphi \leftarrow \varphi + \eta \nabla_\varphi J(\{x_m\}, \{s_m\}; \varphi)$.
  Sample points from generative model, $\{s_m\} \sim G_\theta$.
  return $J(\varphi, X_{te}, \{s_m\})$.



Published as a conference paper at ICML 2017

Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models

Pouya Samangouei*, **Maya Kabkab***, **and Rama Chellappa**
Department of Electrical and Computer Engineering
University of Maryland Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
{pouya, mayak, rama}@umiacs.umd.edu

主要内容 网络结构及实验结果

- **Problem** :Despite their outstanding performance on several machine learning tasks, **deep neural networks have been shown to be susceptible to adversarial attacks**. These attacks come in the form of adversarial examples: carefully crafted perturbations added to a legitimate input sample.
- **Method**:We propose Defense-GAN, a new framework leveraging the expressive capability of generative models to defend deep neural networks **against black-box and white-box adversarial attacks**. Defense-GAN is trained to model the distribution of unperturbed images. At inference time, it finds a close output to a given image which **does not contain the adversarial changes**. This output is then fed to the classifier. Our proposed method can be used with **any classification model** and **does not modify the classifier structure** or **training procedure**. It can also be used as a defense against any attack as it does not assume knowledge of the process for generating the adversarial examples.
- **Result**:We empirically show that Defense-GAN is consistently effective against different attack methods and improves on existing defense strategies.

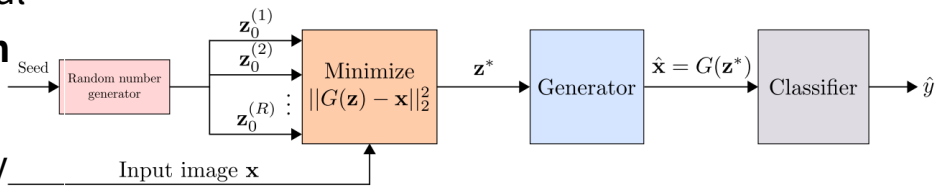


Figure 1: Overview of the Defense-GAN algorithm.

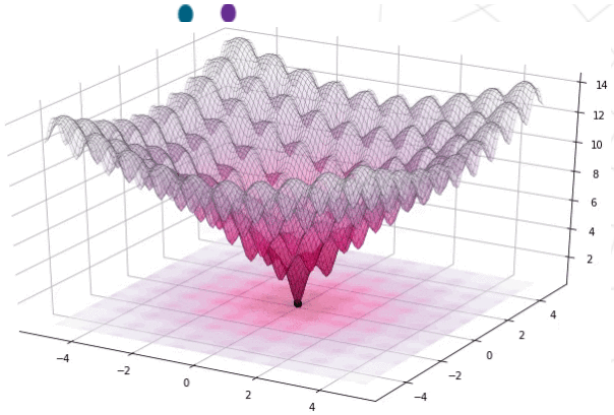


Table 1: Classification accuracies of different classifier and substitute model combinations using various defense strategies on the MNIST dataset, under FGSM black-box attacks with $\epsilon = 0.3$. Defense-GAN has $L = 200$ and $R = 10$.

Classifier/ Substitute	No Attack	No Defense	**Defense-GAN-Rec**	**Defense-GAN-Orig**	MagNet	Adv. Tr. $\epsilon = 0.3$	Adv. Tr. $\epsilon = 0.15$
A/B	0.9970	0.6343	0.9312	0.9282	0.6937	**0.9654**	0.6223
A/E	0.9970	0.5432	0.9139	0.9221	0.6710	**0.9668**	0.9327
B/B	0.9618	0.2816	0.9057	**0.9105**	0.5687	0.2092	0.3441
B/E	0.9618	0.2128	0.8841	**0.8892**	0.4627	0.1120	0.3354
C/B	0.9959	0.6648	0.9357	0.9322	0.7571	**0.9834**	0.9208
C/E	0.9959	0.8050	0.9223	0.9182	0.6760	**0.9843**	0.9175
D/B	0.9920	0.4641	0.9272	**0.9323**	0.6817	0.7667	0.8514
D/E	0.9920	0.3931	**0.9164**	0.9155	0.6073	0.7676	0.7129

Coulomb GANs: Provably Optimal Nash Equilibria via Potential Fields

Thomas Unterthiner[1] **Bernhard Nessler**[1] **Calvin Seward**[1,2] **Günter Klambauer**[1]

Martin Heusel[1] **Hubert Ramsauer**[1] **Sepp Hochreiter**[1]

[1]LIT AI Lab & Institute of Bioinformatics, Johannes Kepler University Linz, Austria
[2]Zalando Research, Mühlenstraße 25, 10243 Berlin, Germany
{unterthiner,nessler,seward,klambauer,mhe,ramsauer,hochreit}@bioinf.jku.at

原理性很强

主要内容

网络结构及实验结果

- **Problem** : 1) GANs suffer from “**mode collapsing**”, where the model generates samples only in certain regions which are called modes. 2) GANs cannot assure that the **density of training samples** is correctly modeled by the generator. 3) the discriminator of GANs may forget previous modeling errors of the generator which then may reappear, a property that leads to oscillatory behavior instead of convergence
- **Method:** we introduce the Coulomb GAN, which has only one **Nash equilibrium**. We are later going to show that this Nash equilibrium is optimal, i.e., the model distribution matches the target distribution. We propose Coulomb GANs to avoid the GAN shortcoming (1) to (3) by using a potential field created by point charges analogously to the electric field in physics. Its only solution is optimal. We will then show how learning the discriminator and generator works in a Coulomb GAN and discuss the assumptions needed for our optimality proof. Coulomb GAN does indeed work well in practice and that the samples it produces have very large variability and appear to capture the original distribution very well.

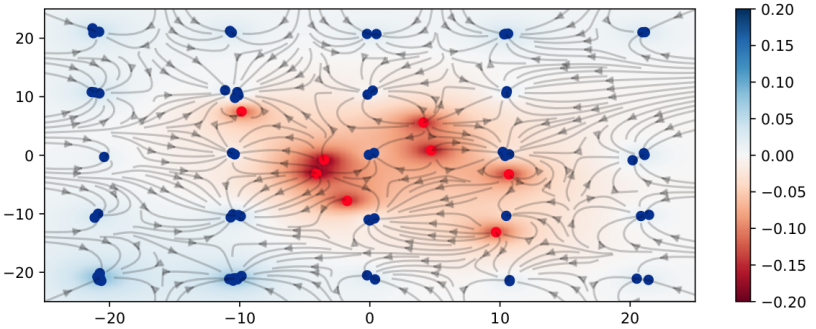
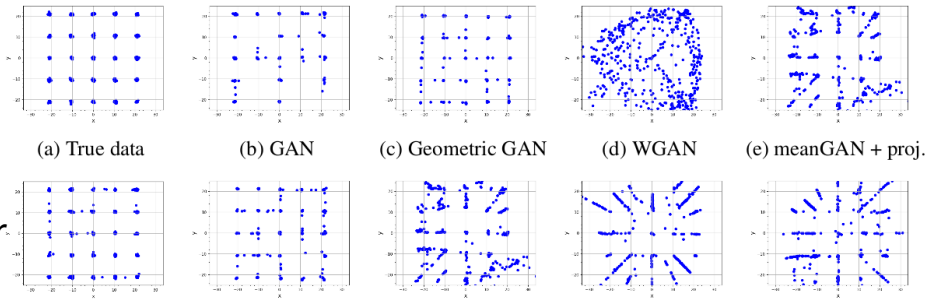


Figure 1: The vector field of a Coulomb GAN. The basic idea behind the Coulomb GAN: true samples (blue) and generated samples (red) create a potential field (scalar field). Blue samples act as sinks that attract the red samples, which repel each other. The superimposed vector field shows the forces acting on the generator samples to equalize potential differences, and the background color shows the potential at each position. Best viewed in color.



AmbientGAN: Generative models from lossy measurements

Ashish Bora
Department of Computer Science
University of Texas at Austin
ashish.bora@utexas.edu

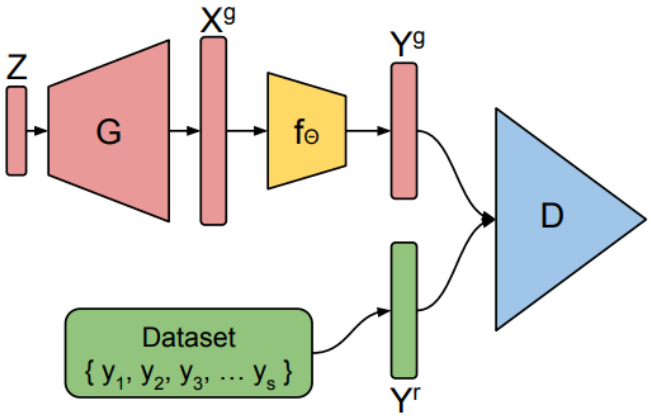
Eric Price
Department of Computer Science
University of Texas at Austin
ecprice@cs.utexas.edu

Alexandros G. Dimakis
Department of Electrical and Computer Engineering
University of Texas at Austin
dimakis@austin.utexas.edu

主要内容

- **Problem** : GAN model requires access to a large number of fully-observed samples from the desired distribution. Unfortunately, obtaining multiple high-resolution samples can be expensive or impractical for some applications. For example, many sensing and tomography problems (e.g. MRI, CT Scan) require a large number of projections for good reconstruction.
- **Method**: rather than distinguish a real image from a generated image as in a traditional GAN, our discriminator must distinguish a real measurement from a simulated measurement of a generated imageGenerative models are powerful tools, but constructing a generative model requires a large, high quality dataset of the distribution of interest. We show how to relax this requirement, by learning a distribution from a dataset that only contains incomplete, noisy measurements of the distribution.

网络结构及实验结果



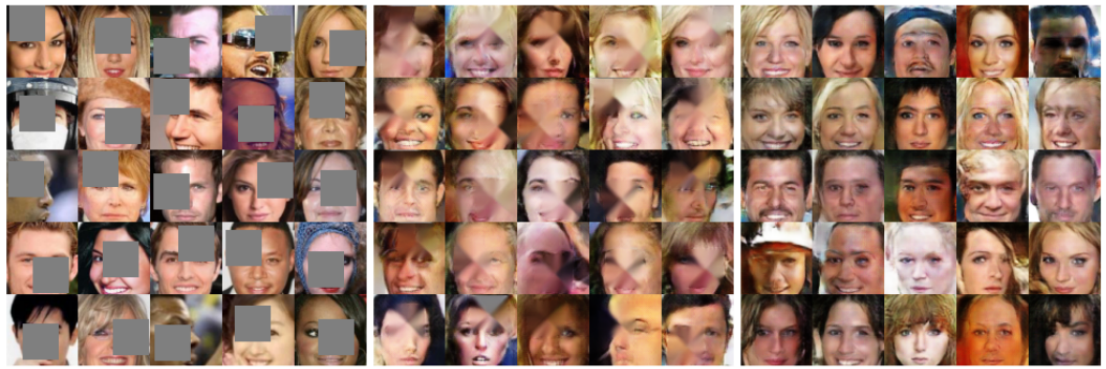


Figure 2: (Left) Samples of lossy measurements used for training. Samples produced by (middle) a baseline that trains from inpainted images, and (right) our model.

Sobolev GAN

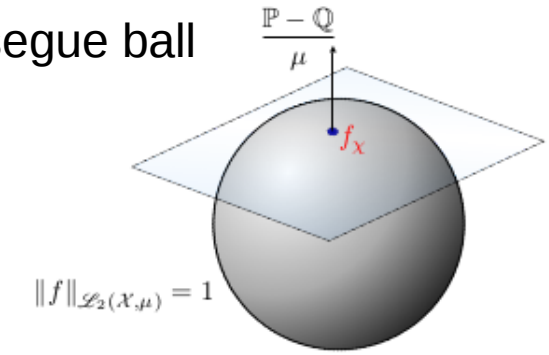
Youssef Mroueh[†], **Chun-Liang Li**[○,⋆], **Tom Sercu**[†,⋆], **Anant Raj**[◇,⋆] **& Yu Cheng**[†]
† IBM Research AI
○ Carnegie Mellon University
◇ Max Planck Institute for Intelligent Systems
⋆ denotes Equal Contribution
{mroueh,chengyu}@us.ibm.com, chunlial@cs.cmu.edu,
tom.sercu1@ibm.com,anant.raj@tuebingen.mpg.de

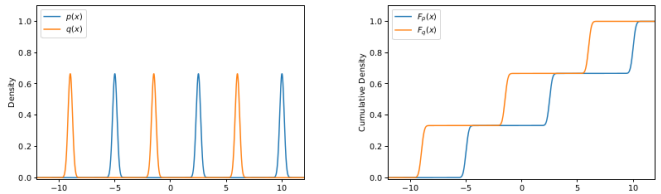
主要内容

- **Problem** :When moving to neural generators of discrete sequences generative adversarial networks theory and practice are still not very well understood.
- **Method**:We propose a new Integral Probability Metric (IPM) between distributions: the Sobolev IPM. The Sobolev IPM compares the mean discrepancy of two distributions for functions (critic) restricted to a Sobolev ball defined with respect to a dominant measure μ. We show that the Sobolev IPM compares two distributionsin high dimensions based on weighted conditional Cumulative Distribution Functions (CDF) of each coordinate on a leave one out basis.
- **Result**:we show that a variant of Sobolev GAN achieves competitive results in semi-supervised learning on CIFAR-10, thanks to the smoothness enforced on the critic by Sobolev GAN which relates to Laplacian regularization.

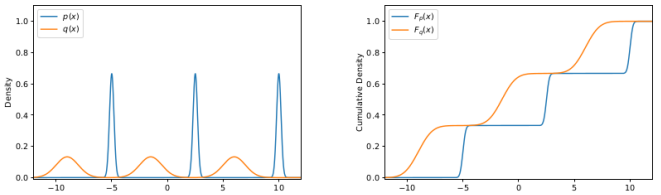
网络结构及实验结果

Lebesgue ball





(a) Smoothed discrete densities: PDF versus CDF of smoothed discrete densities with non overlapping supports.



(b) Smoothed Discrete and Continuous densities: PDF versus CDF of a smoothed discrete density and a continuous density with non overlapping supports.

Training Generative Adversarial Networks via Primal-Dual Subgradient Methods: A Lagrangian Perspective on GAN

†Xu Chen , ‡Jiang Wang, †Hao Ge *
† Department of EECS, Northwestern University, Evanston, IL, USA
‡ Google Inc.
{chenx,haoge2013}@u.northwestern.edu
wangjiangb@gmail.com

主要内容

- **Problem** : “For GANs, there is no theoretical prediction as to whether simultaneous gradient descent should converge or not. One notable inconvergence issue with GAN training is referred to as **mode collapse**, where the generator characterizes only a few modes of the true data distribution
- **Method**: We relate the minimax game of generative adversarial networks (GANs) to finding the saddle points of the **Lagrangian function** for a convex optimization problem,where the discriminator outputs and the distribution of generator outputs play the roles of primal variables and dual variables, respectively. The modified objective function forces the distribution of generator outputs to be updated along the direction according to the primal-dual subgradient methods.
- **Result**:A toy example shows that the proposed method is able to resolve mode collapse, which in this case cannot be avoided by the standard GAN or Wasserstein GAN.

实验结果

In this paper, we do not aim at achieving superior performance over other GANs, but rather provide a new perspective of understanding GANs, and propose an improved training technique that can be applied on top of existing GANs

Method	Score
Real data	11.24 ± 0.16
WGAN (Arjovsky et al., 2017)	3.82 ± 0.06
MIX + WGAN (Arora et al., 2017)	4.04 ± 0.07
Improved-GAN (Salimans et al., 2016)	4.36 ± 0.04
ALI (Dumoulin et al., 2016)	5.34 ± 0.05
DCGAN (Radford et al., 2015)	6.40 ± 0.05
Proposed method	4.53 ± 0.04

Table 2: Inception scores on CIFAR-10 dataset.

Inception Score mainly tracks the diversity of generator, while there is no reliable evidence that it can measure the true sample quality.

Improved Techniques for Training GANs

Tim Salimans tim@openai.com
Ian Goodfellow ian@openai.com
Wojciech Zaremba woj@openai.com
Vicki Cheung vicki@openai.com
Alec Radford alec@openai.com
Xi Chen peter@openai.com

主要内容

- **Problem** :GANs are typically trained using gradient descent techniques that are designed to find a low value of a cost function, rather than to find the Nash equilibrium of a game. When used to seek for a Nash equilibrium, these algorithms may fail to converge Unfortunately, a modification to θ (D) that reduces J (D) can increase J (G) , and a modification to θ (G) that reduces J (G) can increase J (D) . Gradient descent thus fails to converge for many games. **Semi-Supervised**

- **Method**:We present a variety of new architectural features and training procedures that we apply to the generative adversarial networks (GANs) framework.

1) Feature matching
2) Minibatch discrimination
3) Historical averaging
4) One-sided label smoothing
5) Virtual batch normalization

- **Result** :our model generates MNIST samples that humans cannot distinguish from real data, and CIFAR-10 samples that yield a human error rate of 21.3%.

Published as a conference paper at ICLR 2018

Semantically Decomposing the Latent Spaces of Generative Adversarial Networks

Chris Donahue
Department of Music
University of California, San Diego
cdonahue@ucsd.edu

Zachary C. Lipton
Carnegie Mellon University
Amazon AI
zlipton@cmu.edu

Akshay Balsubramani
Department of Genetics
Stanford University
abalsubr@stanford.edu

Julian McAuley
Department of Computer Science
University of California, San Diego
jmcauley@eng.ucsd.edu

主要内容

网络结构及实验结果

- **Problem** :While we may know the identity of the subject in each photograph, we may not know the contingent aspects of the observation (such as lighting, pose and background). This kind of data is ubiquitous; given a set of commonalities, we might want to incorporate this structure into our latent representations. While GANs are popular, owing to their ability to generate high-fidelity images, they do not, in their original form, explicitly disentangle the latent factors according to known commonalities.
- **Method**:We propose a new algorithm for training generative adversarial networks that jointlylearns latent codes for both identities (e.g. individual humans) and observations(e.g. specific photographs). By fixing the identity portion of the latent codes, we can generate diverse images of the same subject, and by fixing the observation portion, we can traverse the manifold of subjects while maintaining contingent aspects such as lighting and pose.
- **Result**:Experiments with human judges and an off-the-shelf face verification system demonstrate our algorithm’ s ability to generate convincing, identity-matched photographs

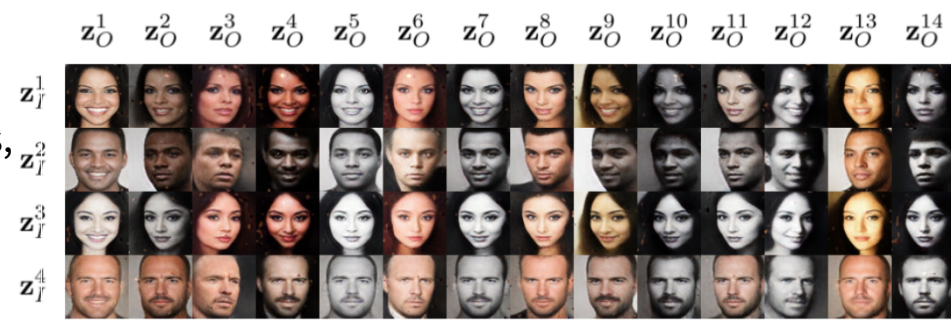
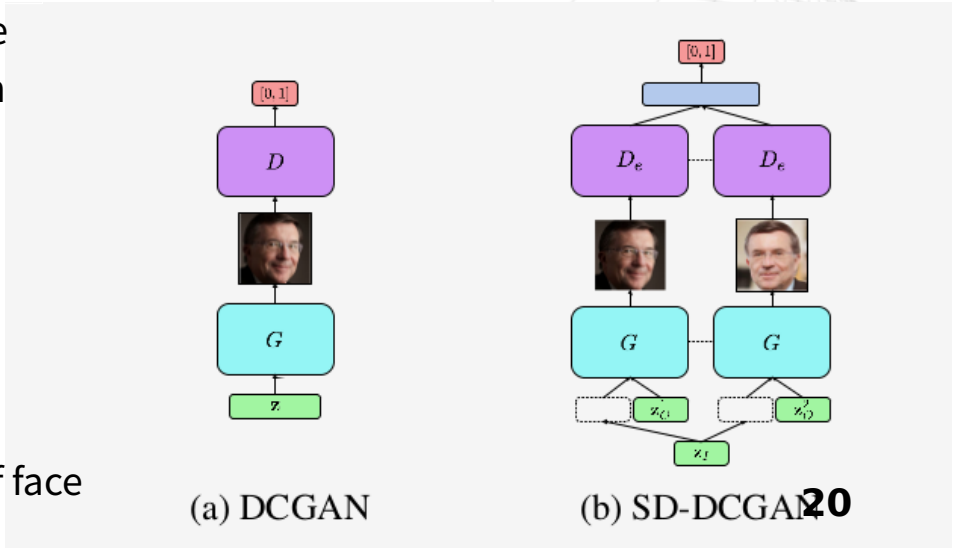


Figure 1: Generated samples from SD-BEGAN. Each of the four rows has the same identity code $\mathbf{z}_I$ and each of the fourteen columns has the same observation code $\mathbf{z}_O$.



(a) DCGAN (b) SD-DCGAN

THANK YOU FOR WATCHING

PRESENTED BY OfficePLUS