How good is my GAN ?

ECCV2018 勉強会

M2 中塚俊介

TL;DR

- GANの画像生成の定量的な指標が必要
  - Inception Score, Fréchet Inception Distance (FID)などが存在
  - IS, FIDはImagenetで学習されているInception Networkを使うので，
    他のデータセットへ適用して精度を保証できるものではない
- GAN-train (recall: diversity) と GAN-test (precision: quality) を提案

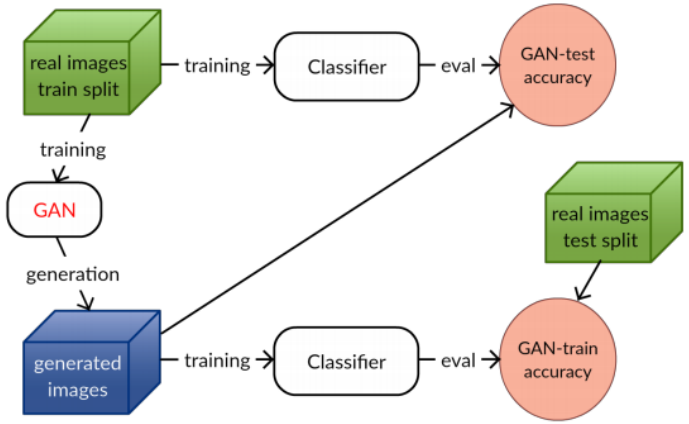
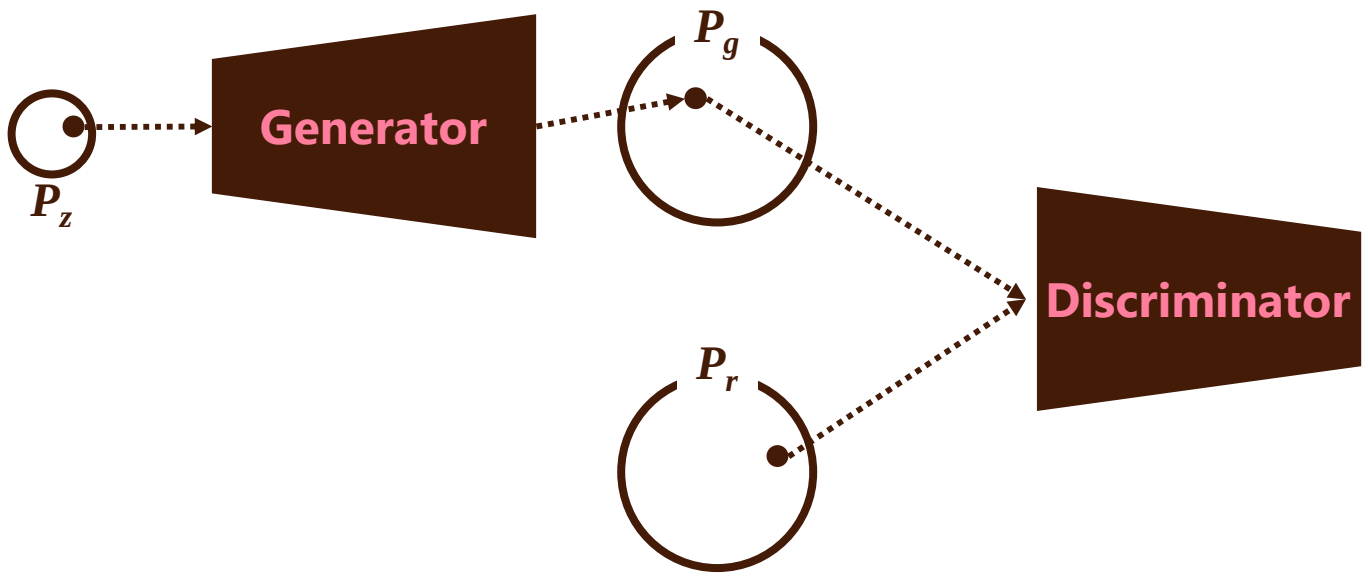


Fig. 2: Illustration of GAN-train and GAN-test. GAN-train learns a classifier on GAN generated images and measures the performance on real test images. This evaluates the diversity and realism of GAN images. GAN-test learns a classifier on real images and evaluates it on GAN images. This measures how realistic GAN images are.

GANのおさらい

- Discriminator
  - 入力が $x \sim \boldsymbol{P}_r$ もしくは，$x' \sim \boldsymbol{P}_g$ であるかを判定する
  - 実質的には，データ分布 $\boldsymbol{P}_r$ と生成分布 $\boldsymbol{P}_g$ がどれだけ離れているかを測る
- Generator
  - Discriminatorの提示するどれだけ離れているかという指標を最小化する
  - 最終的には，本物の画像 $(x \sim \boldsymbol{P}_r)$ に近い画像 $(x' \sim \boldsymbol{P}_g)$ を出力する



GANのおさらい

- P_r とP_g がどれだけ離れているかを表す指標は様々
  - JSD
  - Energy
  - Wasserstein Distance
  - ... and more !!
- GANのテクニックも様々
  - Unrolled
  - Gradient Penalty
  - Spectral Normalization
  - Progressive Growing
  - ... and more !!

何をモチベーションに色々な研究がされているのか？

GANに求められてるもの

- 多様性のある生成
  - データ分布を完全にカバーできるわけではない
  - Mode Collapse （同じような出力ばかりする）
- 質の高い生成
  - 人間がみると生成と本物の差はわかってしまう
- 安定的な学習
  - 学習が破綻しやすい
  - 2つのネットワークのバランス調整が必要

生成に関する指標が必要
⇒ Inception Score, FID

Inception Score

- Imagenet で学習済みのInception Networkを使った評価指標
- Inception Network が識別しやすい（質が高い），
  識別されるラベルのバリエーションが大きい（多様性）ほど，大きくなる指標

$$IS = \exp\left[E_{x \sim P_g} \left[D_{KL} \left(p(y \mid x) \| p(y) \right) \right] \right]$$

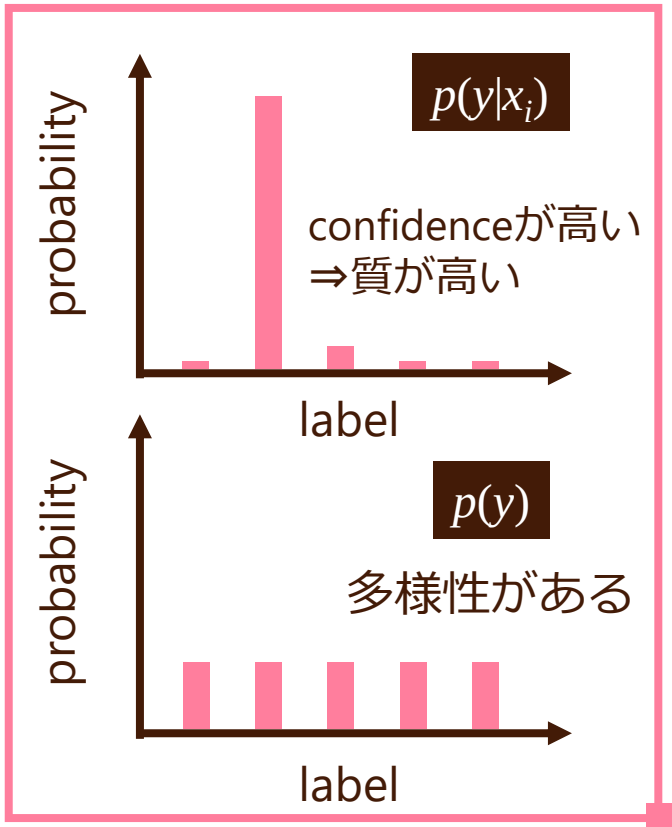
$p(y \mid x_i) = F(x_i)$

ネットワークの出力

$p(y) = E_{x_i \sim X} \left[p(y \mid x_i) \right]$

ネットワークの出力の周辺確率

P_rの情報は一切考慮に入っていない



理想的な状態

Fréchet Inception Distance (FID)

- Imagenet で学習済みのInception Networkを使った評価指標
- Inception Networkの特徴ベクトルの平均と分散共分散行列を $\boldsymbol{P}_r$ と $\boldsymbol{P}_g$ のそれぞれから算出（2つのガウス分布）
- 小さいほど，良い指標
- 現在のスタンダード

$$d^2 = \left|\mu_r - \mu_g\right|^2 + tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right)$$

Sliced Wasserstein Distance

- PGGANで使われてる指標
- Laplacian Pyramid の各レベルから，7×7のパッチを切り出す
  論文内では，16,384枚の画像から各レベルで128枚のパッチ切り出して，
  2^{21}個のdescriptorを作成
- それぞれを1次元に変換して，SWDを測る

$$y' = top\_k(y \cdot \theta)$$

$$x' = top\_k(x \cdot \theta)$$

$$\theta \sim N(\mathbf{o}, \mathbf{I}), \theta \in R^N$$

$$SWD = |x' - y'|$$

- Wasserstein Distanceは
  高次元ベクトルに対して，計算が大変
- 1次元に関しては，closed formが存在する
- だから，slice して1次元ごとにしよう

それぞれの弱点

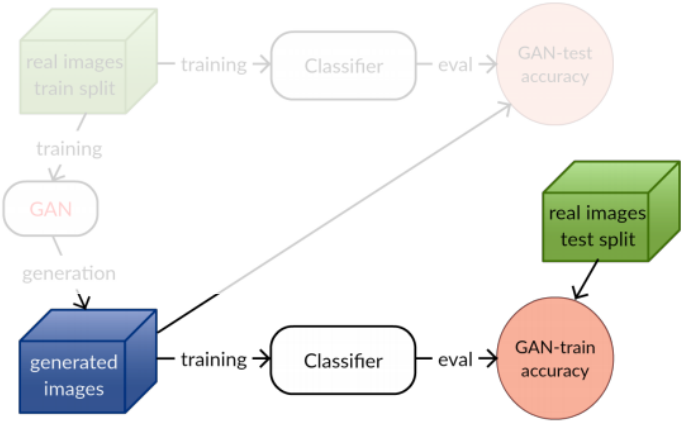
- Inception Score
  - Imagenet で学習済みのInception Networkに依存している
  - P_rの情報は一切考慮に入っていない
- FID
  - Imagenet で学習済みのInception Networkに依存している
  - 結局のところ，ガウス分布を仮定している
- Sliced Wasserstein Distance
  - 解像度が低い画像に関しては適用できない

前提（手法に入る前に）

- この論文で扱うのは，ConditionalなGAN
- GANにおいて大事なことは
  - realistic
  - recognizable as coming from a given class
- Target distributionを完全につかんだGAN
  - 生成する分布 S_g はデータセットの分布 S_t とは区別できない
  - S_g と S_t の容量が同じ and MNISTのような単純なデータセットなら同じvalidation accuracy を持つはず

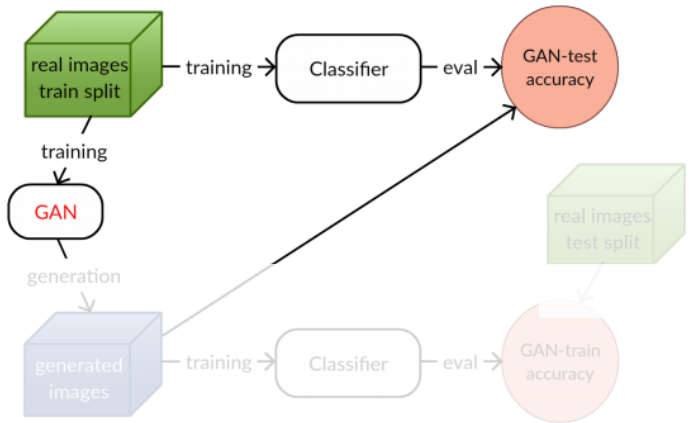
GAN-train

- S_gで訓練された分類器を S_t を使って評価する
- GANが完璧でなければ
  - Mode Collapseして多様性がない
  - 識別に有用な特徴をとるのに十分な品質でない
  - Conditionalな生成がうまくできていない
- train acc と valid acc が近いなら,
  S_g は S_t と同じくらいのQuality かつ Diverse である



GAN-test

- S_tで訓練された分類器を S_g を使って評価する
  - GANがうまく学習できていれば，簡単なタスクのはず
  - S_t と S_gは同じ分布になっているはずだから
- 理想は「 train acc と valid acc が近い」こと
  - valid acc が高いなら，GANがOverfitting してる and S_t を単に記憶しているだけ
  - valid acc が低いなら，GANが S_t を捉えきれていない and Low Quality
- この指標は，分布の多様性を測ることは不可能
  - GANが1つのサンプルだけを記憶すれば，
    GAN-test は高くなってしまうから．．．



Datasets

- MNIST
  - Shape : (28, 28, 1)
  - Train : 60k
  - Test : 10k
- CIFAR10 and CIFAR100
  - Shape : (32, 32, 1)
  - Train : 50k
  - Test : 10k
- ImageNet (1000 classes)
  - Shape : (128, 128, 3) or (64, 64, 3)
  - Train : 1.3M
  - Test : 50k

Models

- PixelCNN++
  - Autoregressionモデル
- WGAN
  - Wasserstein Distance を metricsとしたGAN
  - with Gradient Penalty
  - 現在のStandard Model
- SNGAN
  - Spectral Normalization を適用したGAN
  - 2017年のConditional GANの最高のモデル
- DCGAN
  - Base Model

Experiment (MNIST)

- 4層のConvnetで99.3% のtest acc が出た
  - SNGANの GAN-train: 99.0%
  - SNGANの GAN-test : 99.2%
- S_g は S_t と同じくらいのQuality かつ Diversity である

Experiment (CIFAR10)

model	IS	FID-5K	FID	GAN-train	GAN-test	SWD 16	SWD 32
real images	11.33	9.4	2.1	92.8	-	2.8	2.0
SNGAN パラメータ数	8.43	18.8	11.8	82.2	87.3	3.9	24.4
WGAN-GP (10M)	8.21	21.5	14.1	79.5	85.0	3.8	6.2
WGAN-GP (2.5M)	8.29	22.1	15.0	76.1	80.7	3.4	6.9
DCGAN	6.69	42.5	35.6	65.0	58.2	6.5	24.7
PixelCNN++	5.36	121.3	119.5	34.0	47.1	14.9	56.6

Table 1: CIFAR10 experiments. IS: higher is better. FID and SWD: lower is better. SWD values here are multiplied by 10^3 for better readability. GAN-train and GAN-test are accuracies given as percentage (higher is better).

WDを最小化しているので小さくなる
⇒他のModelと比較するのは Fairじゃない？

S_g (SNGAN)　　S_t

Classifierの特徴ベクトルが
近いサンプルを持ってきた

SNGANは，訓練データセットに近い，
同じクラスの画像を見つけることができる

Experiment (CIFAR10)

GAN-trainとGAN-testの関係を強調するために
Subsampling と Corruptingしてみた

- Subsampling（データを少なくしてみる）
  - GAN-test ：鈍感
  - GAN-train : 敏感
- Corrupting （1%～20%のごま塩ノイズ）
  - GAN-test ：ほとんど影響なし
  - GAN-train : 82% ⇒ 15%

Experiment (CIFAR10)

「CIFAR10において，FIDを使うのは不十分ではないか？」

- CIFAR10にガウシアンノイズ（σ=5）を付与すると，FIDは27.1に
  - その中から，5k random samplingしても 29.6
  - この微差は，Diversity によるものか Quality によるものか判別が難しい
- 提案手法では
  - GAN-test ：95%⇒95%
  - GAN-train : 91%⇒80%
  - つまり， Diversityが失われていたことがわかる！

CIFAR100

差が顕著にわかる

model	IS	FID-5K	FID	GAN-train	GAN-test	SWD 16	SWD 32
real images	14.9	10.8	2.4	69.4	-	2.7	2.0
SNGAN	9.30	23.8	15.6	45.0	59.4	4.0	15.6
WGAN-GP (10M)	9.10	23.5	15.6	26.7	40.4	6.0	9.1
WGAN-GP (2.5M)	8.22	28.8	20.6	5.4	4.3	3.7	7.7
DCGAN	6.20	49.7	41.8	3.5	2.4	9.9	20.8
PixelCNN++	6.27	143.4	141.9	4.8	27.5	8.5	25.9

High Quality だけど Poor Diversity

Random Forestで試しても順位変わらず

CIFAR100

- 官能的な検証も行った
  - 5人の被験者に，特定のクラスに対して生成された2枚のサンプルから“どっちがRealistic か？”を聞いてみる
  - “SNGAN vs DCGAN“ “SNGAN vs WGAN(2.5M)” “SNGAN vs WGAN(10M)”
  - ↑の3回の判定を1人に100回ずつやってもらった
- SNGAN vs DCGAN
  - 368 : 132
- SNGAN vs WGAN(2.5M)
  - 274 : 226
- SNGAN vs WGAN(10M)
  - 230 : 270

ImageNet

res	model	IS	FID-5K	FID	GAN-train top-1	GAN-train top-5	GAN-test top-1	GAN-test top-5
64px	real images	63.8	15.6	2.9	55.0	78.8	-	-
	SNGAN	12.3	44.5	34.4	3	8.4	12.9	28.9
	WGAN-GP	11.3	46.7	35.8	0.1	0.7	0.1	0.5
128px	real images	203.2	17.4	3.0	59.1	81.9	-	-
	SNGAN*	35.3	44.9	33.2	9.3	21.9	39.5	63.4
	WGAN-GP	11.6	91.6	79.5	0.1	0.5	0.1	0.5

かなり精度が上がった

変化がわからない

データセットの大きさとDiversityの関係

- 意外と全体的にサチるのは早い
- SNGANが一番Diversity がある

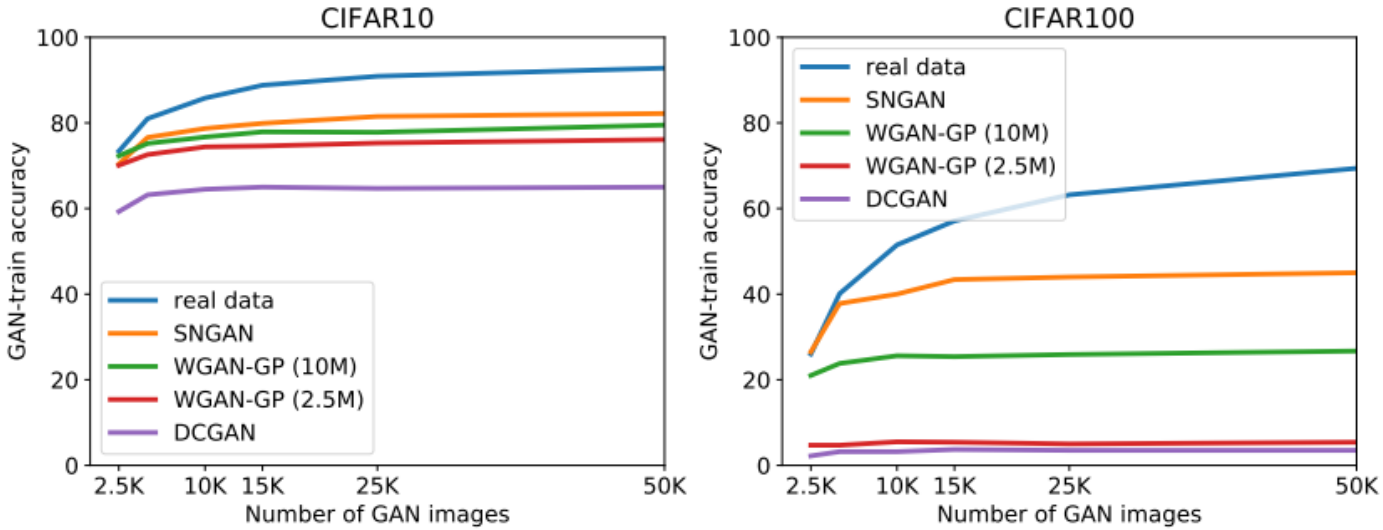
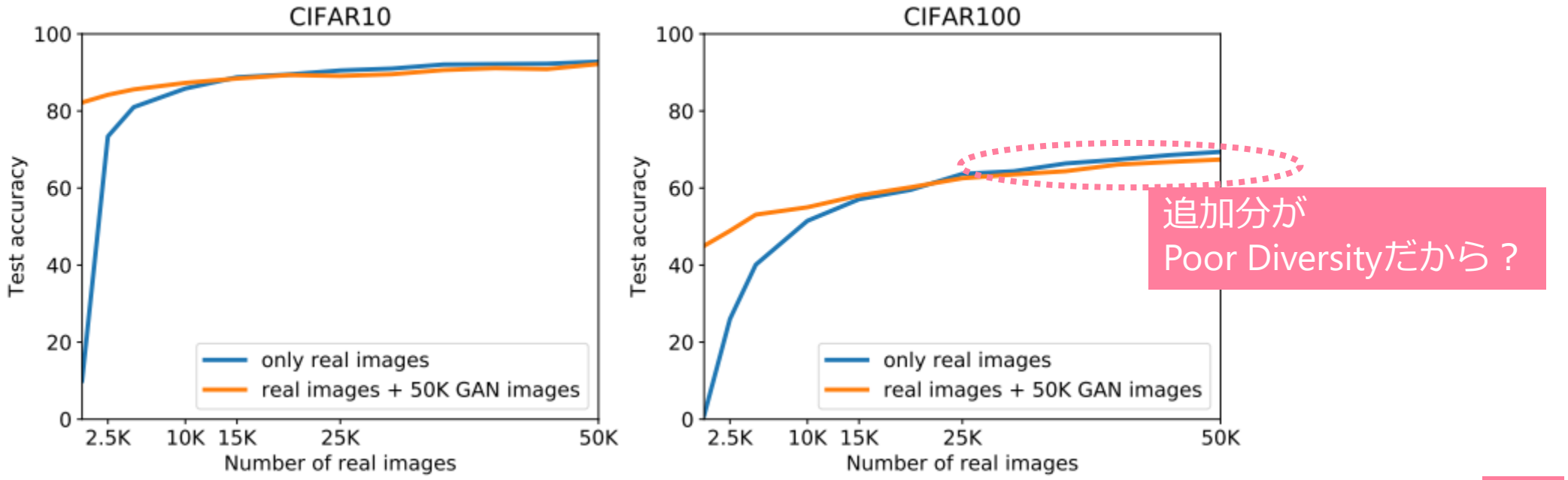


Fig. 4: The effect of varying the size of the generated image set on GAN-train accuracy. For comparison, we also show the result (in blue) of varying the size of the real image training dataset. (Best viewed in pdf.)

Data Augmentation

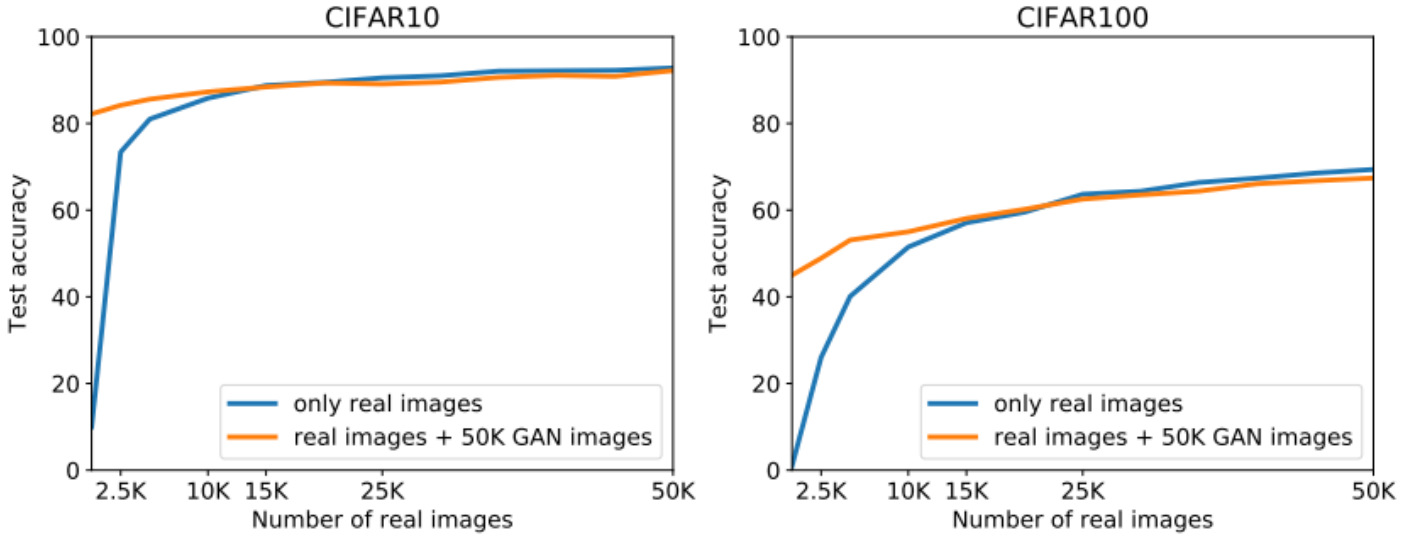
- すべてのデータで学習されたSNGANで生成されたデータを50k追加してみた
- Performanceの向上を確認
  - でも，これ“すべてのデータで学習されたSNGAN”を使っているため，
    ニーズに則した実験ではないのでは…?



Data Augmentation

- SNGANに与えるデータも制限してみた
  - 精度が落ちる？
  - Poor Diversityだから？

Num real images	real C10	real+GAN C10	real C100	real+GAN C100
2.5k	73.4	67.0	25.6	23.9
5k	80.9	77.9	40.0	33.5
10k	85.8	83.5	51.5	45.5



まとめ

- GAN-train (recall: diversity) と GAN-test (precision: quality) を提案
  - GAN-train と GAN-test は従来手法に比べて有用
- 実験・評価・考察を頑張りました
- GANでのAugmentationは，Diversity を失いやすい
  - 小さいDataset からDiversity のあるGANを学習しないといけない ~~（それは夢）~~

- 感想
  - Augmentationしたら精度が上がるという論文の是非が気になる
    もちろん，Datasetの違い（枚数，種類，複雑さ）もあるが．．．
  - クラス情報なしの評価指標も必要では？

Questions ?

Distance, Divergence （過去スライドより抜粋）

分布間の距離をどう定義するかでGANのObjectiveは変わる

- JSD(Jensen Shannon Divergence)　→ Standard GAN

$$KLD\left[p_r \,\|\, p_g\right] = E_{x\sim p_r}\left[\log \frac{p_r(x)}{p_g(x)}\right]$$

$$m = \frac{p_r(x) + p_g(x)}{2}$$

$$JSD\left[p_r \,\|\, p_g\right] = \frac{1}{2}\left(KLD\left[p_r \,\|\, m\right] + KLD\left[p_g \,\|\, m\right]\right)$$

- EMD(Earth Mover's Distance)　→ WGAN

$$W\left(p_r, p_g\right) = \inf_{\gamma \subseteq \Pi\left(p_r, p_g\right)} E_{(x,y)\sim\gamma}\left\|x - y\right\|$$

JSD vs EMD（過去スライドより抜粋）

- Standard GAN
  - 不連続なp_g を生成していたのは、JSD自体が不連続だから
  - また、勾配消失問題もJSDのせい

- WGAN
  - Gがθにおいて連続なら、EMDも連続
  - DがK-Lipschitz連続性(関数の傾きがある定数Kで抑えられるような一様連続性)を保つなら、EMDはどこでも連続かつ微分可能

Gradient Penalty（過去スライドより抜粋）

Optimal Critic D^*は

$$\hat{x} = (1-\varepsilon)G(z) + \varepsilon x$$

という直線上のどこにおいても、以下の勾配を持つ

$$\nabla_{\hat{x}} D^*(\hat{x}) = \frac{x - \hat{x}}{\|x - \hat{x}\|}$$

正規化されているのでnormは1

つまり、勾配のnormは1

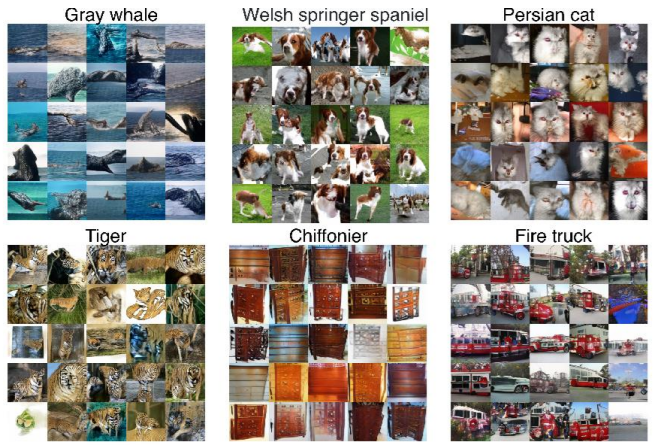
だから、GP項はnorm=1からどれだけ離れているかを表している

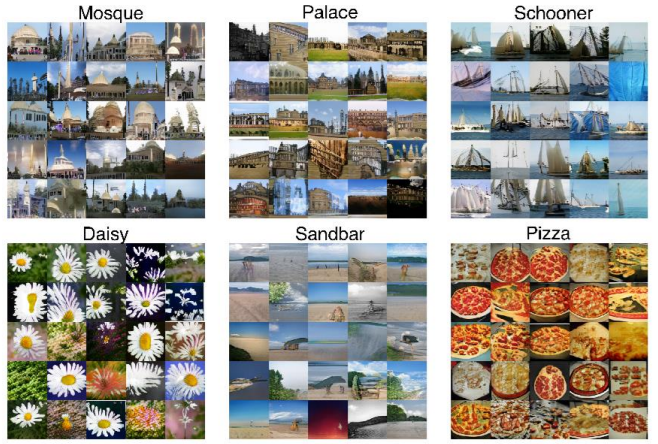
これで、1-Lipschitz連続性が担保できる！

$$\left[\left(\left\| \nabla_{\hat{x}} D(\hat{x}) \right\|_2 - 1 \right)^2 \right]$$

Spectral Normalization for Generative Adversarial Networks

- GANにSpectral Normalization を適用した
- Contributions
  - GANの安定性の向上
  - ImageNet等の多クラスデータセットに対して，生成を可能にした
  - 簡単な実装，計算
- PFNの論文
- GoodfellowがGANで読むべき論文10選に選出





Spectral Normalization for Generative Adversarial Networks

■ 前提

- GANにおいて，Discriminatorの学習が重要
- Discriminatorが，Lipschitz連続性を保つことが安定に繋がる
- NNの演算は，$a(\boldsymbol{W}\boldsymbol{x})$の繰り返し

$$y = a_L(\boldsymbol{W_L}(a_{L-1}(\boldsymbol{W_{L-1}}(a_{L-2}(\cdots a_0(\boldsymbol{W_0}x))))))$$

Spectral Normalization

- 入力$\boldsymbol{h}$ のL層目から，$L+1$層目への写像g のLipschitz norm は，

$$\sup_{h} \sigma(\nabla g(h))$$

- 上式のσ は，行列$\boldsymbol{A}$ のSpectral Normである

$$\sigma(A) = \max_{h \neq 0} \frac{\|Ah\|_2}{\|h\|_2} = \max_{\|h\|_2 \leq 1} \|Ah\|_2$$

- これは，行列$\boldsymbol{A}$ の最大特異値に等しい
- 各レイヤーの演算は，線形であるため

$$\sup_{h} \sigma(\nabla g(h)) = \sup_{h} \sigma(W) = \sigma(W)$$

- つまり，重みW の最大特異値がLipschitz Normである

Spectral Normalization

- NNの演算は，$a(\boldsymbol{W}\boldsymbol{x})$の繰り返し
  - 各レイヤーのLipschitz NormがK以下ならば，NN全体のLipschitz NormもK以下
  - $K = 1$ ならば...

$$W_{SN} = \frac{W}{\sigma(W)}$$

- 計算時は，更新ごとに全重みを特異値分解する必要がある
  - 計算コスト大
  - Power Iteration（近似）で解決