TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2019

Generative Adversarial Networks (GANs)

Generative Adversarial Nets

Goodfellow et al., June 2014

Let $\mathrm{Pop} \uplus P_\Phi$ be the distribution defined flipping an unbiased coin and, if heads, returning $(1, y)$ with $y \sim \mathrm{Pop}$ and, if tails, returning $(-1, y)$ with $y \sim P_\Phi$.

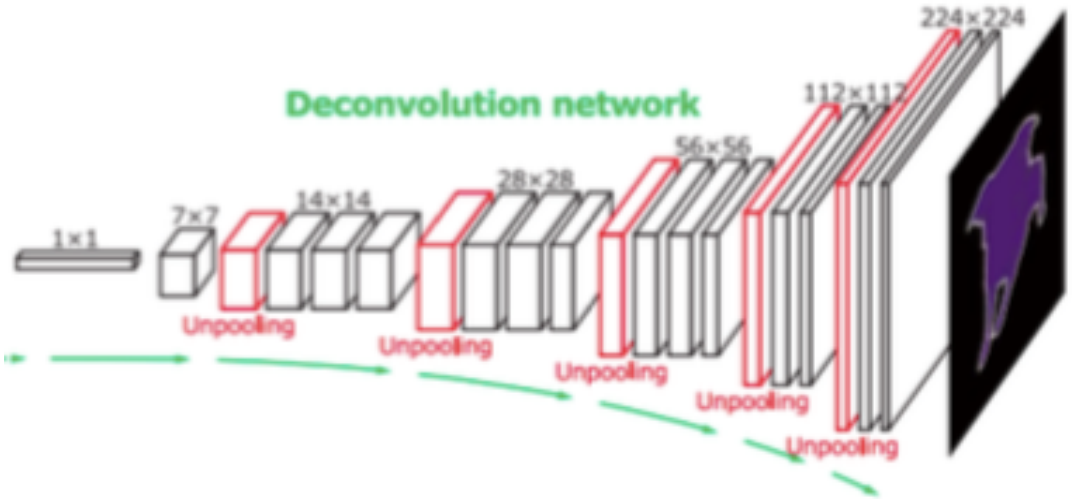
$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \min_{\Psi} \; E_{(i,y)\sim(\mathrm{Pop}\uplus P_\Phi)} - \ln P_\Psi(i|y)$$

Generator / Discriminator

Assuming Universality: $P_{\Phi^*} = \mathrm{Pop}$

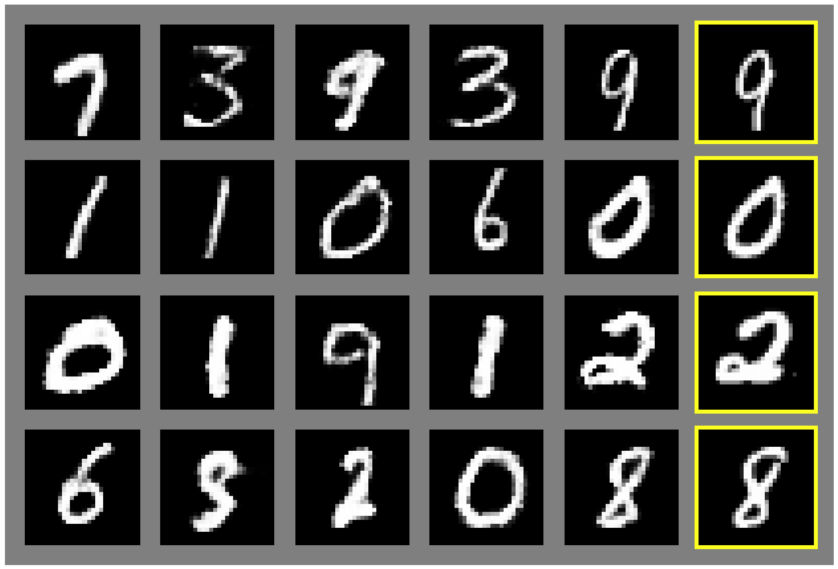
Sampling from the Generator

$z \sim \mathcal{N}(0, I)$ $\qquad\qquad$ $y \sim P_\Phi$



Generative Adversarial Nets

Goodfellow et al., June 2014

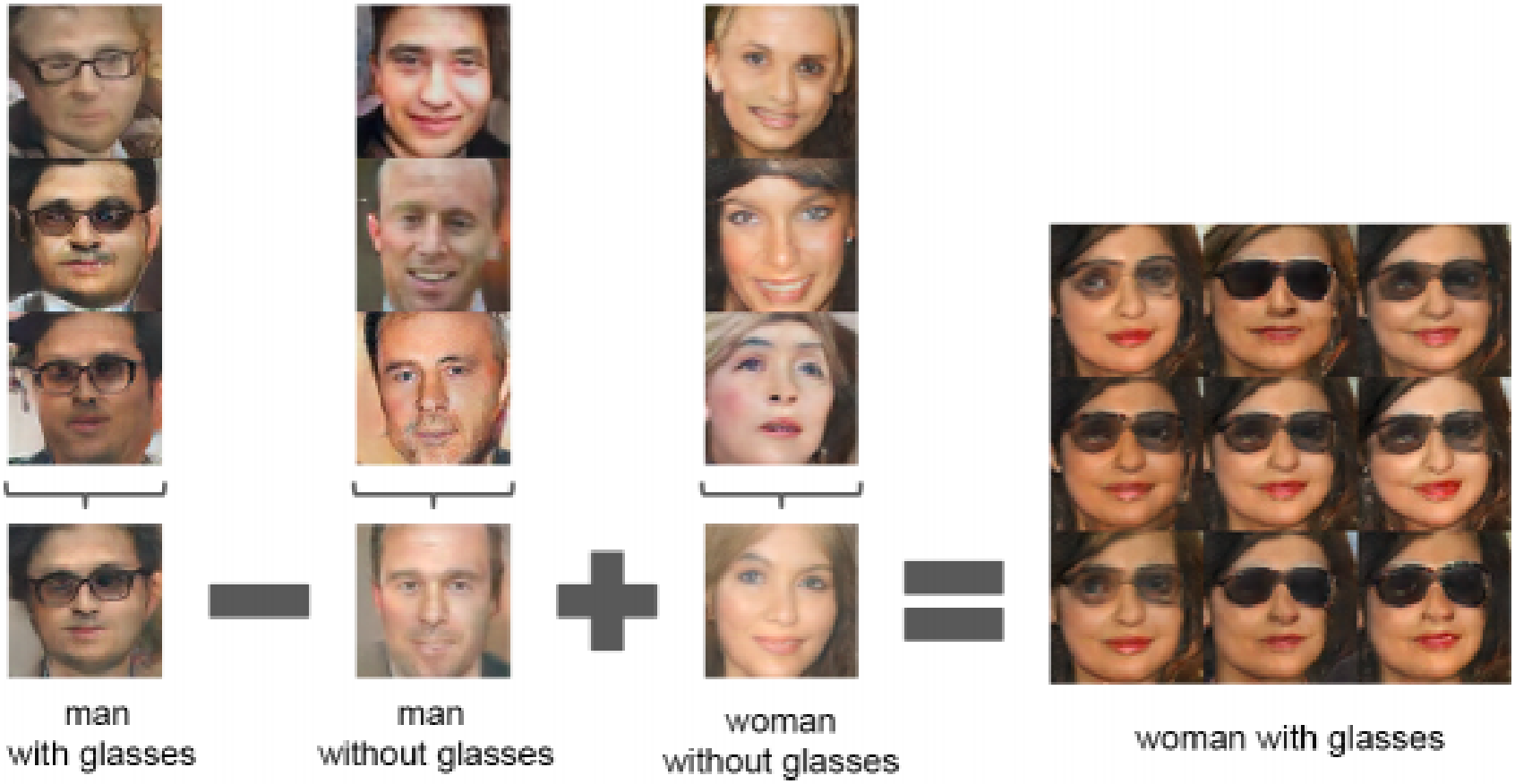


Unsupervised Representation Learning ... (DC GANS)

Radford et al., Nov. 2015

Unsupervised Representation Learning ... (DC GANS)

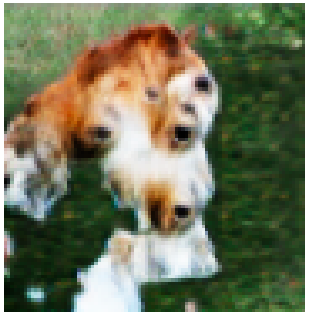
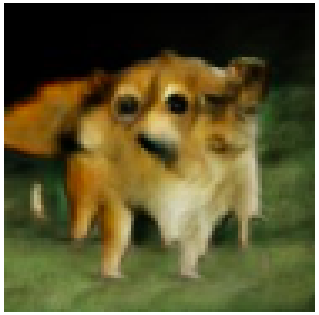
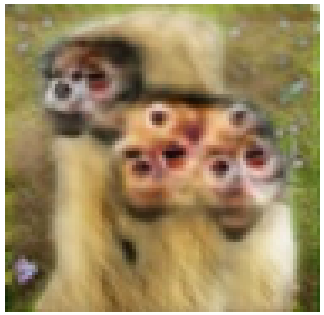
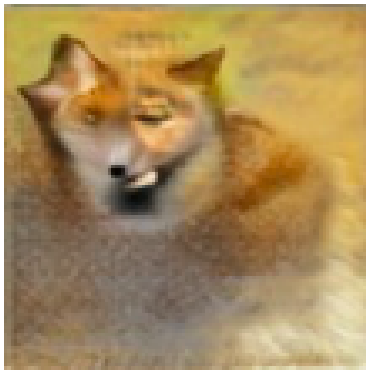
Radford et al., Nov. 2015

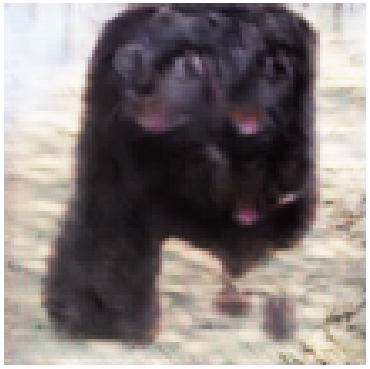


Interpolated Faces

[Ayan Chakrabarti]

Early GANs on ImageNet



Conditional GANs

All distribution modeling methods apply to conditional distributions.

$$\Phi^* = \operatorname*{argmax}_{\Phi} \min_{\Psi} E_{i,x,y\sim(\mathrm{Pop} \uplus \mathrm{Pop}(x)P_\Phi(y|x))} - \ln P_\Psi(i|x,y)$$

If x is never repeated we can replace $P_\Phi(y|x)$ by a deterministic function $\hat{y}_\Phi(x)$.

$$\Phi^* = \operatorname*{argmax}_{\Phi} \min_{\Psi} E_{i,x,y\sim(\mathrm{Pop} \uplus (\mathrm{Pop}(x);\hat{y}(x)))} - \ln P_\Psi(i|x,y)$$

Discrimination Loss can Augment Distortion Loss

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\ E_{x,y\sim \mathrm{Pop}}\ \mathrm{Dist}(y, \hat{y}_\Phi(x))$$

This is the fundamental equation if $\mathrm{Dist}(y, \hat{y}_\Phi(x)) = -\ln P_\Phi(y|x)$.

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\ \mathcal{L}_{\mathrm{Dist}}(\Phi) + \lambda \mathcal{L}_{\mathrm{Disc}}(\Phi)$$

$$\mathcal{L}_{\mathrm{Dist}}(\Phi) = E_{x,y\sim \mathrm{Pop}}\ \mathrm{Dist}(y, \hat{y}_\Phi(x))$$

$$\mathcal{L}_{\mathrm{Disc}}(\Phi) = \max_{\Psi}\ E_{i,x,y\sim(\mathrm{Pop} \uplus (\mathrm{Pop}(x);\hat{y}_\Phi(x)))}\ \ln P_\Psi(i|x,y)$$

Image-to-Image Translation (Pix2Pix)

Isola et al., Nov. 2016

We assume a corpus of "image translation pairs" such as images paired with semantic segmentations.

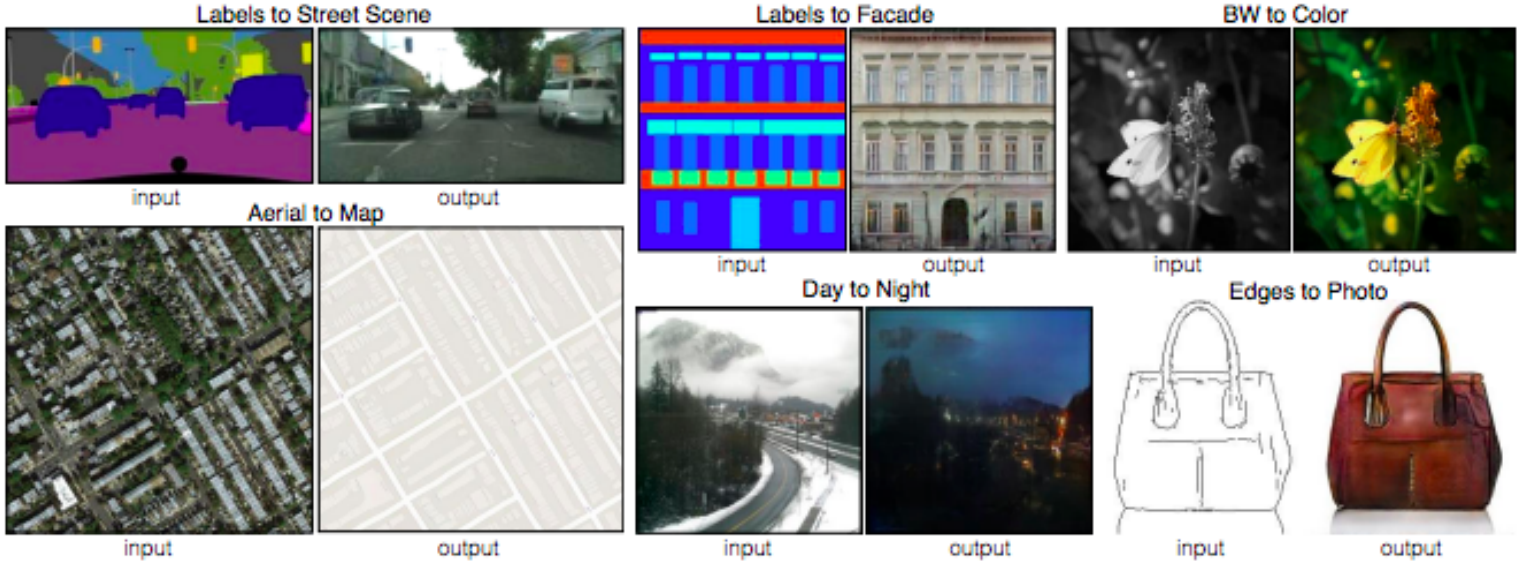
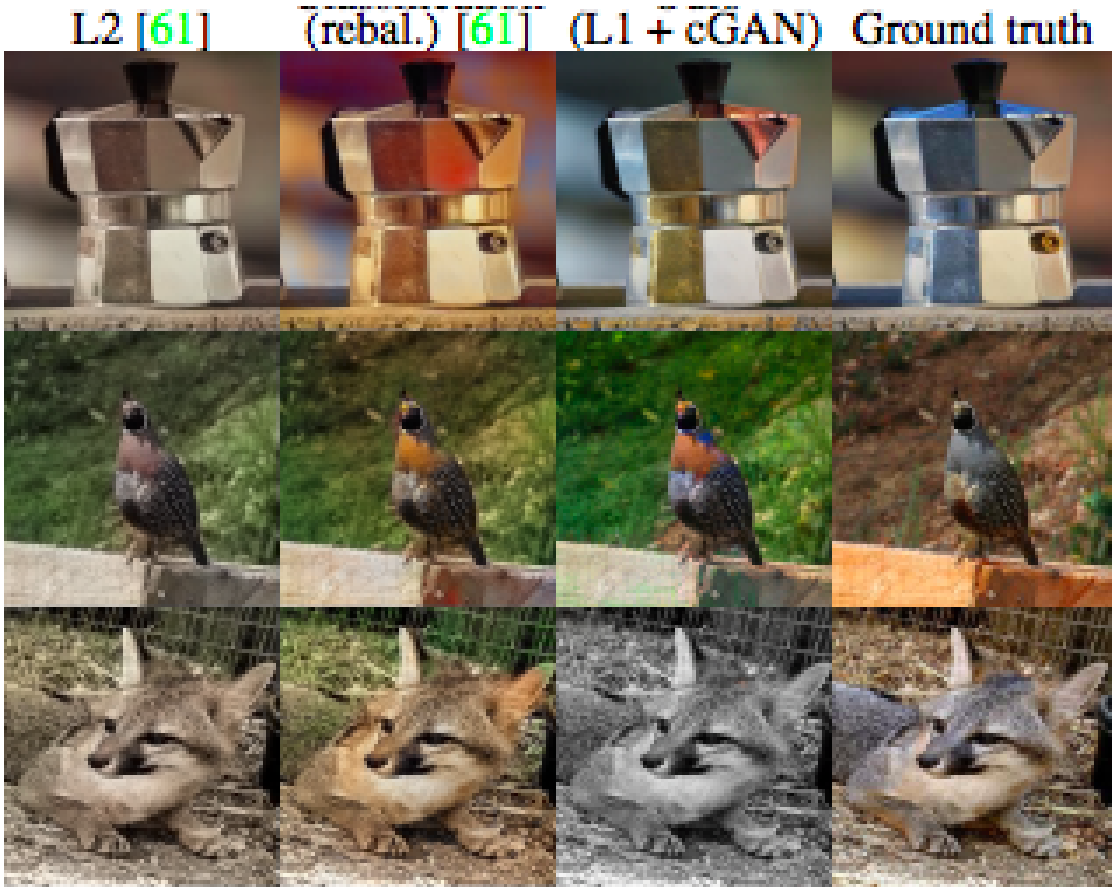


Image-to-Image Translation (Pix2Pix)

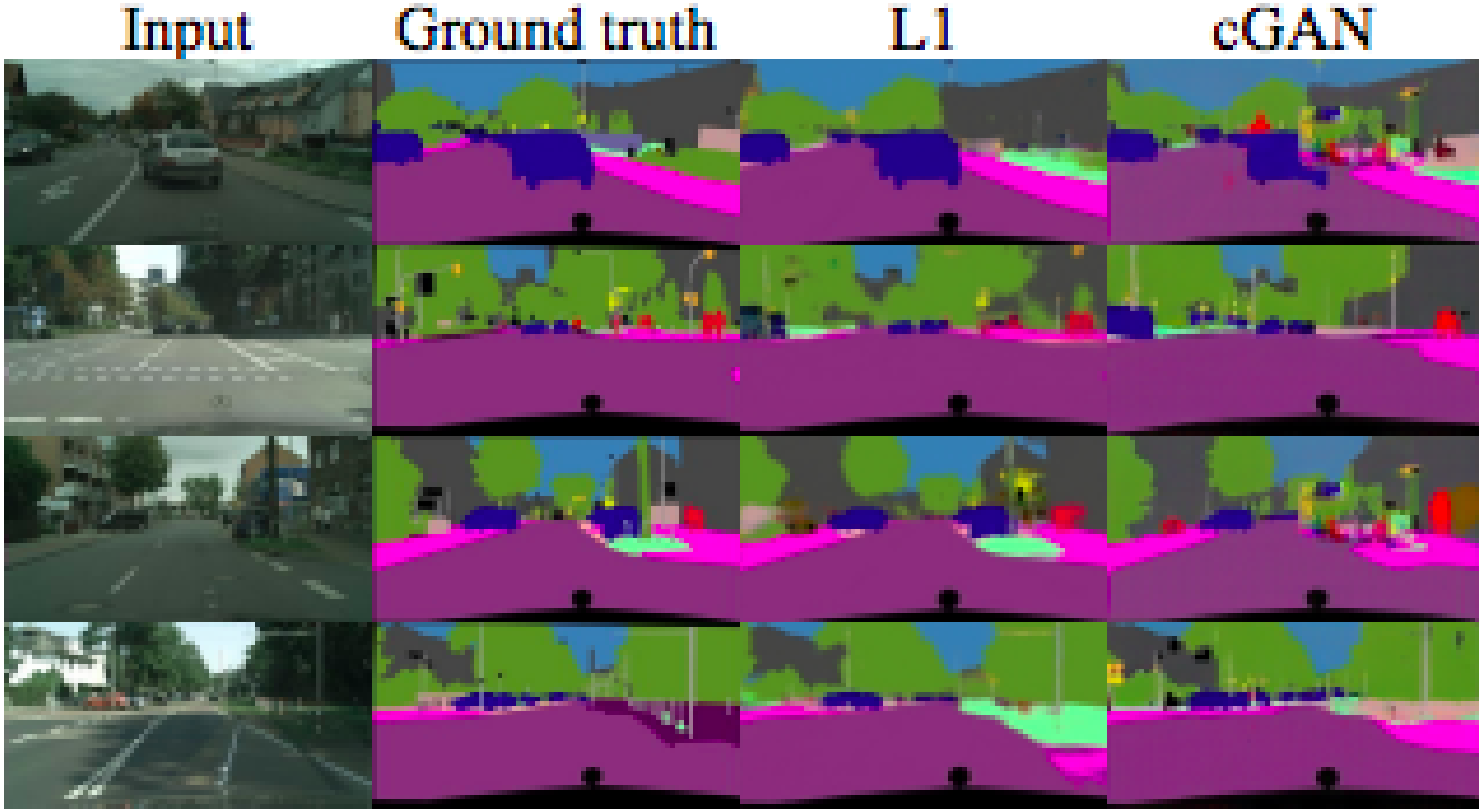
Isola et al., Nov. 2016

Arial Photo to Map and Back

Colorization



Semantic Segmentation

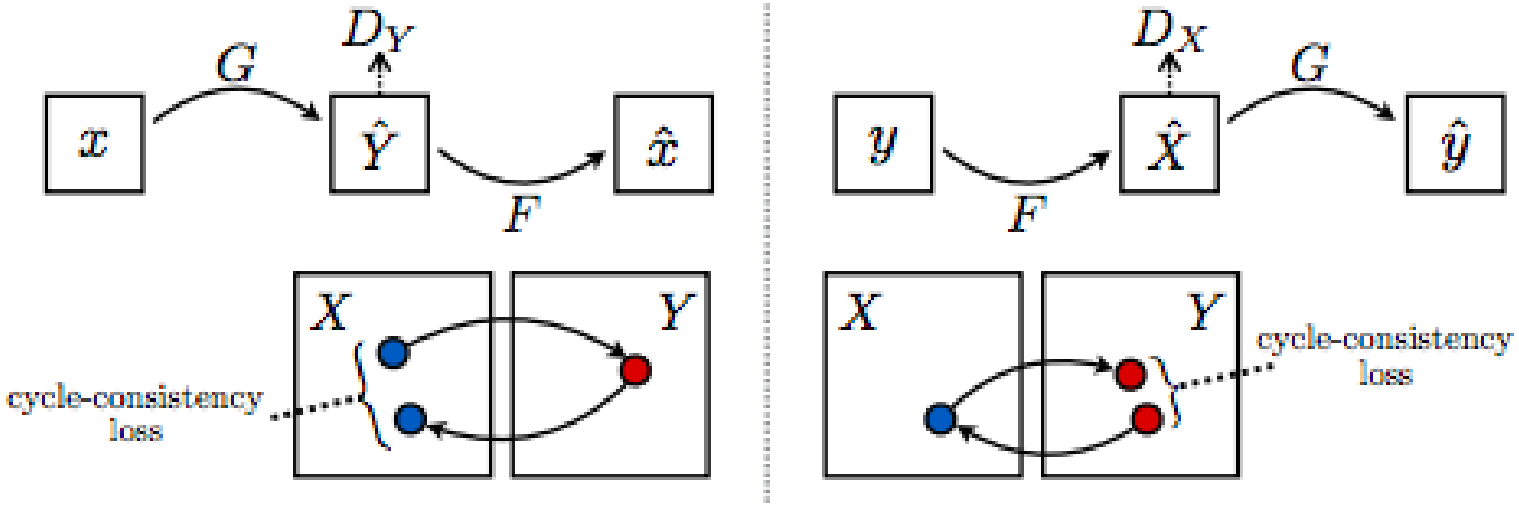


Unpaired Image-to-Image Translation (Cycle GANs)

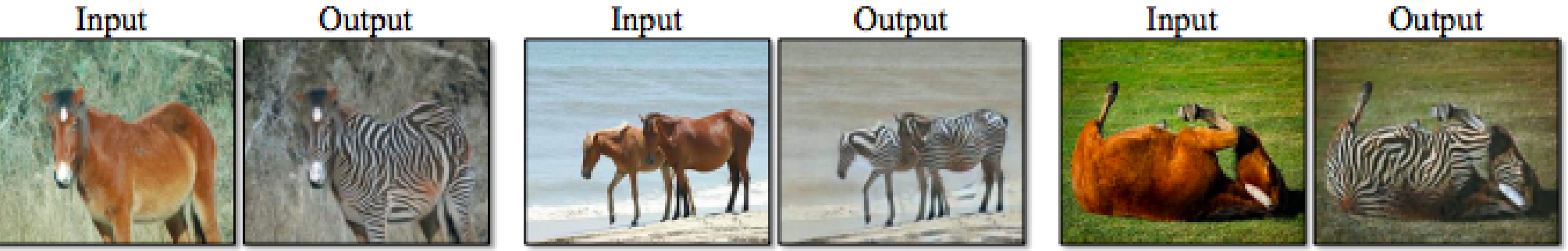
Zhu et al., March 2017

We have two corpora of images, say images of zebras and unrelated images of horses, or photographs and unrelated paintings by Monet.

We want to construct translations between the two classes.



Cycle Gans

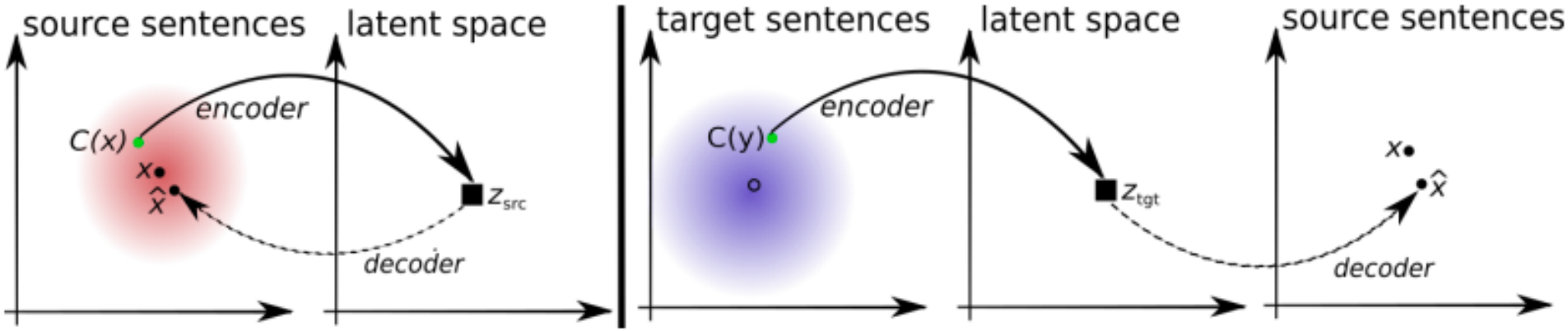

Input
Output
Input
Output
Input
Output
horse → zebra
zebra → horse

Cycle Gans

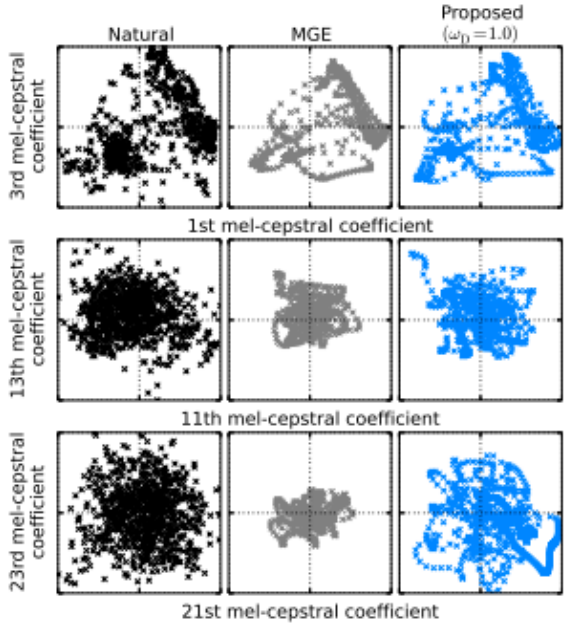
Horse → Zebra

Unsupervised Machine Translation (UMT)

Lample et al, Oct. 2017, also Artetxe et al., Oct. 2017



Text to Speech (Saito et al. Sept. 2017)

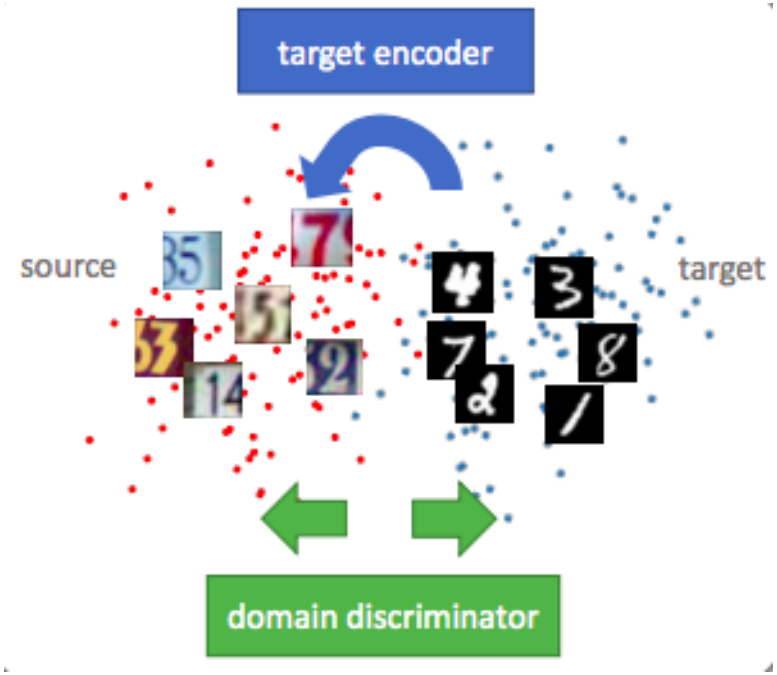


Minimum Generation Error (MGE) uses perceptual distortion — a distance between the feature vector of the generated sound wave and the feature vector of the original.

Perceptual Naturalness can be enforced by a discriminator.

Adversarial Discriminative Domain Adaptation

Tzeng et al. Feb. 2017



A GAN is used to map target images to the source feature distribution.

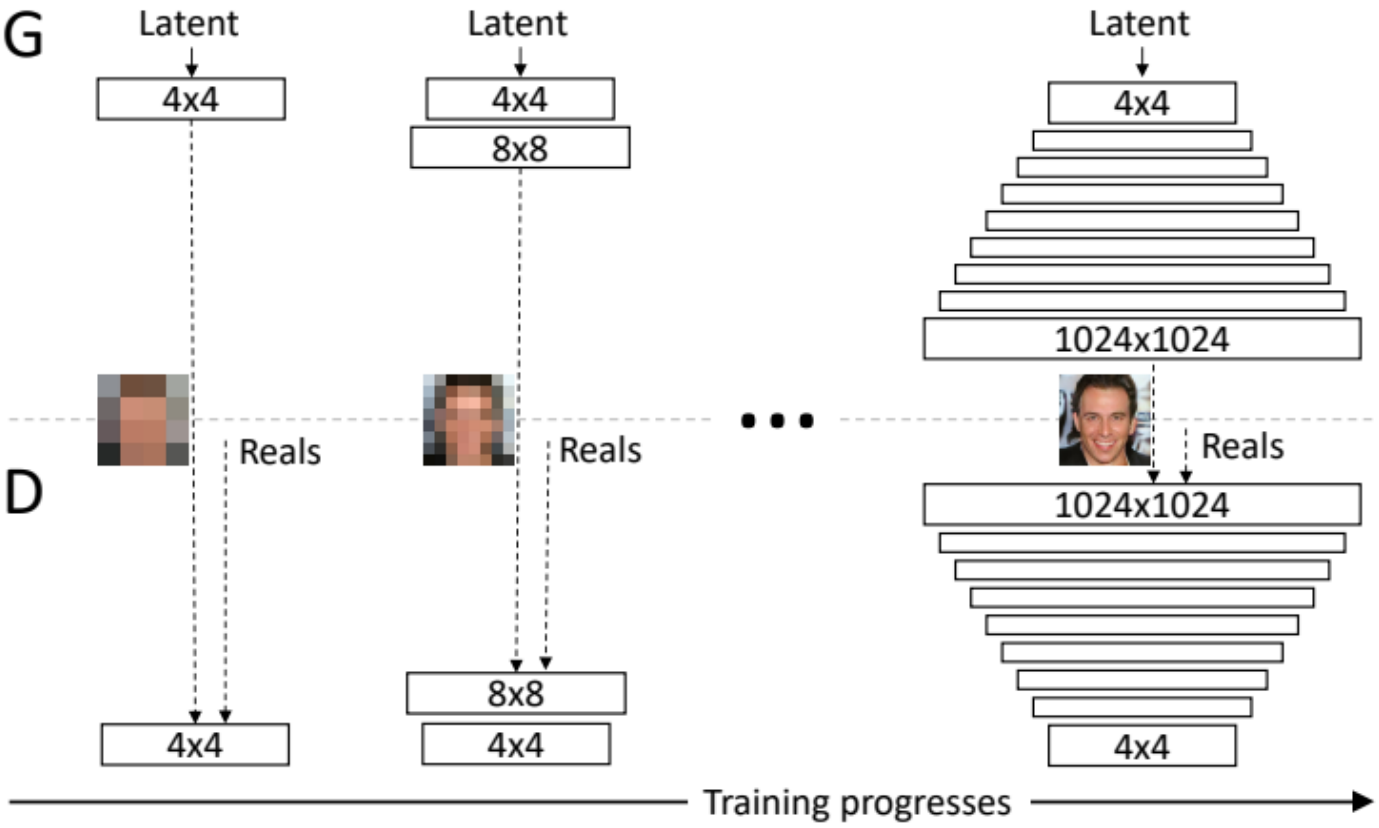
Progressive Growing of GANs

Karras et al., Oct. 2017

Figure 5: 1024 × 1024 images generated using the CELEBA-HQ dataset. See Appendix F for a larger set of results, and the accompanying video for latent space interpolations.

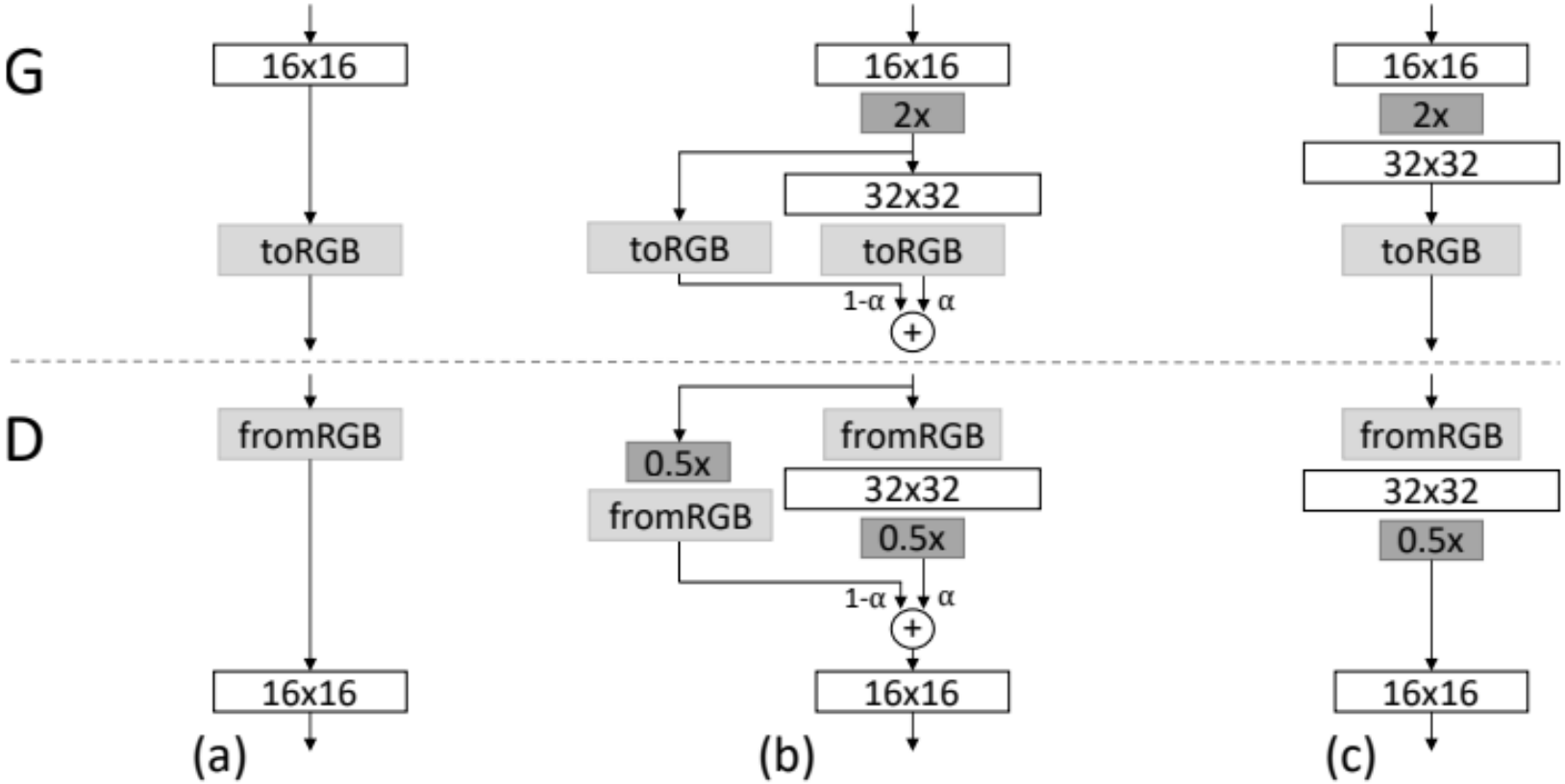
Progressive Growing of GANs

Karras et al., Oct. 2017



Progressive Growing of GANs

Karras et al., Oct. 2017



Large Scale Gan Training

Brock et al., Sept. 2018

Figure 1: Class-conditional samples generated by our model.

This is a conditional GAN conditioned on the imagenet class label.

This generates 512 X 512 images without using using progressive training.

Issues

Jensen-Shannon Divergence

Vanishing Gradients

Unstable Training

Mode Collapse

Measuring Perfomance

Discrimination: Jensen-Shannon Divergence

Assuming Universality:

$$E_{(i,y)\sim(\mathrm{Pop}\uplus Q)} \; - \ln P_{\Phi^*}(i|y)$$

$$= \; E_{(i,y)\sim(\mathrm{Pop}\uplus Q)} \; - \ln P(i|y)$$

$$= \; \frac{1}{2}E_{y\sim\mathrm{Pop}} - \ln \; \frac{\mathrm{Pop}(y)}{\mathrm{Pop}(y)+Q(y)} + \frac{1}{2}E_{y\sim Q} \; - \ln \; \frac{Q(y)}{\mathrm{Pop}(y)+Q(y)}$$

$$= \; (\ln 2) - \frac{1}{2}\left(\; KL\left(\mathrm{Pop}, \frac{\mathrm{Pop}+Q}{2}\right), \; KL\left(Q, \frac{\mathrm{Pop}+Q}{2}\right)\right)$$

Discrimination: Jensen-Shannon Divergence

$$E_{(i,y)\sim(\mathrm{Pop}\uplus Q)} \; -\ln P_{\Psi^*}(i|y)$$

$$= (\ln 2) - \frac{1}{2}\left(KL\left(\mathrm{Pop}, \frac{\mathrm{Pop}+Q}{2}\right),\ KL\left(Q, \frac{\mathrm{Pop}+Q}{2}\right)\right)$$

$$= (\ln 2) + JSD(\mathrm{Pop}, Q)$$

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}}\ JSD(\mathrm{Pop}, P_\Phi)$$

$$0 \leq JSD(P,Q) \leq \ln 2$$

Converting to Cross Entropy

Goodfellow, 2014

In Goodfellow's original paper he expressed a preference for cross entropy loss (the fundamental equation) over Jensen-Shannon loss.

$$\Phi^* = \underset{\Phi}{\text{argmin}} \; JSD(\text{Pop}, P_\Phi)$$

vs.

$$\Phi^* = \underset{\Phi}{\text{argmin}} \; H(\text{Pop}, P_\Phi)$$

He presented a modification to the GAN adversarial objective that yields cross-entropy loss rather than Jensen-Shanon loss.

Converting to Cross Entropy

Goodfellow, 2014

$$\Psi^*(\Phi) = \underset{\Psi}{\operatorname{argmin}} \; E_{(i,y)\sim(\mathrm{Pop} \uplus P_\Phi)} \; -\ln P_\Psi(i|y)$$

$$\text{Assume:} \quad P_{\Psi^*}(1|y) = \frac{\mathrm{Pop}(y)}{\mathrm{Pop}(y) + P_\Phi(y)}$$

$$\text{Define:} \quad f_{\Psi^*}(y) \doteq \frac{P_{\Psi^*}(1|y)}{P_{\Psi^*}(-1|y)}$$

$$= \frac{\mathrm{Pop}(y)}{P_\Phi(y)}$$

Converting to Cross Entropy

$$
\begin{aligned}
\Phi^* &= \operatorname*{argmax}_{\Phi} \; E_{y \sim \mathrm{Pop}} \; f_{\Psi^*}(y) \\
\nabla_\Phi \; E_{y \sim P_\Phi} \; f_{\Psi^*}(y) &= \nabla_\Phi \sum_y \; P_\Phi(y) f_{\Psi^*}(y) \\
&= \sum_y \; P_\Phi(y) f_{\Psi^*}(y) \nabla_\Phi \ln P_\Phi(y) \\
&= \sum_y \; \mathrm{Pop}(y) \nabla_\Phi \ln P_\Phi(y) \\
&= E_{y \sim \mathrm{Pop}} \; \nabla_\Phi \ln P_\Phi(y) \\
&= \nabla_\Phi \; E_{y \sim \mathrm{Pop}} \; \ln P_\Phi(y)
\end{aligned}
$$

Vanishing Gradients

The discriminator typically "wins".

The log loss goes to zero (becomes exponentially small) and there is no gradient to guide the generator.

In this case the learning stops and the generator is blocked from minimizing JSD(Pop, P_Φ).

A Heuristic Fix

We continue to use

$$\Psi^*(\Phi) = \underset{\Psi}{\operatorname{argmin}}\ E_{(y,s)\sim(\mathrm{Pop} \uplus P_\Phi)}\ -\ln P_\Psi(s|y)$$

But switch the optimization for Φ from

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}}\, E_{y\sim P_\Phi}\ -\ln P_\Psi(-1|y)$$

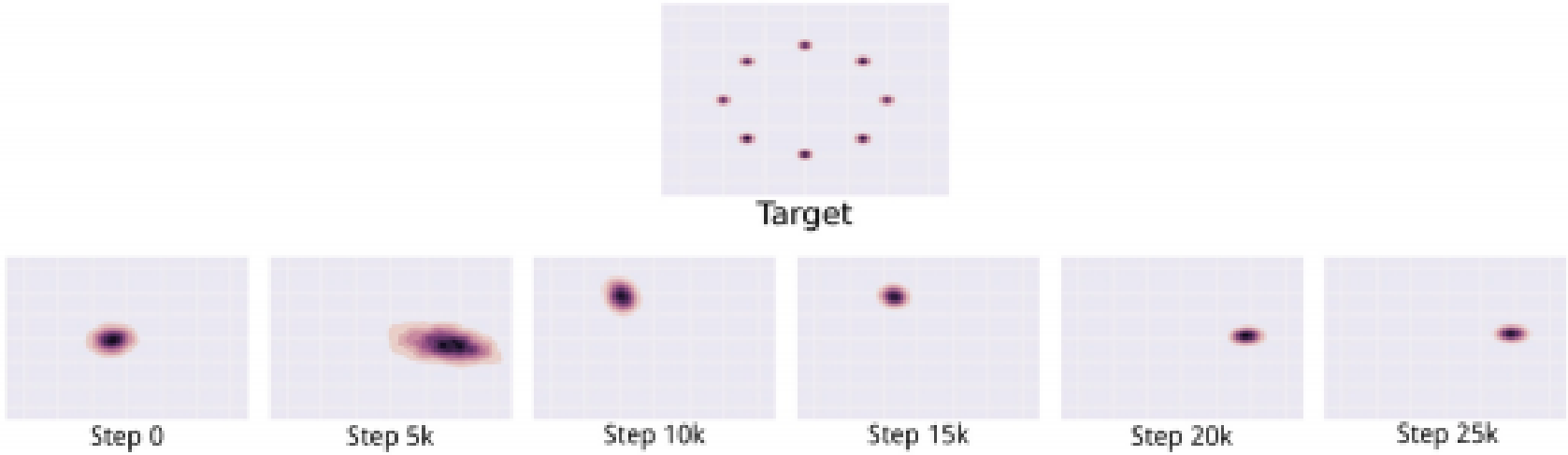
to

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, E_{y\sim P_\Phi}\ -\ln P_\Psi(1|y)$$

It can be shown that $-\ln P_\Psi(1|y)$ is essentially the margin of the binary classifier Ψ.

Mode Collapse a.k.a Mode Dropping

The generator distribution drops portions of the population.



Unstable Training

Joint SGD is not the same as nested max-min.

Consider

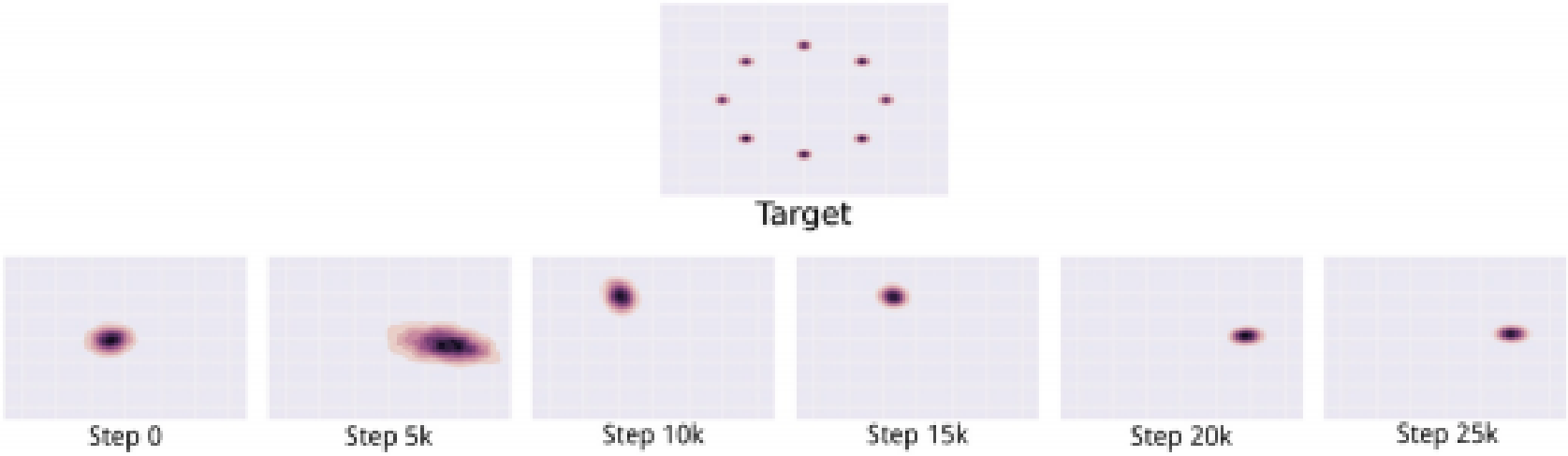
$$\max_x \min_y xy$$

A Nash equilibrium is $x = y = 0$.

Simultaneous gradient flow yields a circle.

Mode Collapse a.k.a Mode Dropping

The generator distribution drops portions of the population.



Pros and Cons of GAN Evaluation Measures

Borji, Oct 2018

We would like a rate-distortion metric on distribution models.

This has not yet been achieved for GANs.

Evaluation of GANs always involves, at least in part, subjective judgments of naturalness.

Sometimes automated metrics are also used.

The above paper discusses various proposed automated metrics of GAN performance. Current automated metrics are questionable.

Contrastive GANS (TZ)

The Discriminator also Models the Population

Fix the generator at a "noise distribution" Q where $Q(y)$ is computable, and train a discriminator.

$$\Psi^* = \underset{\Psi}{\operatorname{argmin}}\, E_{(i,y)\sim(\mathrm{Pop}\uplus Q)} \; -\ln P_\Psi(i|y)$$

The Discriminator also Models the Population

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, E_{(i,y)\sim(\mathrm{Pop}\uplus Q)} \; -\ln P_\Phi(i|y)$$

Assuming Universality

$$P_{\Phi^*}(1|y) = P(1|y) = \frac{P(1 \text{ and } y)}{P(y)} = \frac{\frac{1}{2}\mathrm{Pop}(y)}{\frac{1}{2}\mathrm{Pop}(y) + \frac{1}{2}Q(y)}$$

$$P_{\Phi^*}(1|y)(\mathrm{Pop}(y) + Q(y)) = \mathrm{Pop}(y)$$

$$\frac{\mathrm{Pop}(y)}{Q(y)} = \frac{P_{\Phi^*}(1|y)}{1 - P_{\Phi^*}(1|y)}$$

Discrimination

The discrimination estimate of the population distribution is poor when the discrimination problem is easy — when the discrimination loss can be driven close to zero.

Noise Contrastive Estimation

Gutmann and Hyvärinen, 2010

Consider a population distribution Pop and a fixed generating distribution Q where $Q(y)$ is computable.

Define the distribution $\text{Pop} \hookrightarrow Q^k$ to be the result of drawing one "positive" from Pop and k IID negatives from Q; then inserting the positive at a random position among the negatives; and returning $(i, y_1, \ldots, y_{k+1})$ where i is the index of the positive.

$$\Psi^* = \underset{\Psi}{\text{argmin}} \; E_{(i,y_1,\ldots,y_{k+1})\sim(\text{Pop}\hookrightarrow Q^k)} \; -\ln P_\Psi(i|y_1, \ldots, y_{k+1})$$

Contrastive Estimation

$$\Psi^* = \underset{\Psi}{\operatorname{argmin}} \; E_{(i,y_1,\ldots,y_{k+1})\sim(\mathrm{Pop}\hookrightarrow Q^k)} \; -\ln P_\Psi(i|y_1,\ldots,y_{k+1})$$

Note that $\mathrm{Pop} \hookrightarrow Q^1$ requires a choice between two y's while $\mathrm{Pop} \uplus Q$ classifies a single y — these are different.

The contrastive task gets more difficult as k gets larger.

Constrastive Estimation

$$\Psi^* = \underset{\Psi}{\operatorname{argmin}}\, E_{(i,y_1,\ldots,y_{k+1})\sim(\mathrm{Pop}\hookrightarrow Q^k)} \; -\ln P_\Psi(i|y_1,\ldots,y_{k+1})$$

$$P_\Psi(i|y_1,\ldots,y_{k+1}) \doteq \underset{i}{\operatorname{softmax}}\, s_\Psi(y_i)$$

Theorem: Assuming universality:

$$s_{\Psi^*}(y) = \left(\ln \frac{\mathrm{Pop}(y)}{Q(y)}\right) + C \quad \text{for arbitrary } C$$

$$\mathrm{Pop}(y) = \underset{y}{\operatorname{softmax}} \quad s_{\Psi^*}(y) - \ln Q(y) \quad Z \text{ must be estimated}$$

Proof

$$
\begin{aligned}
P(i \text{ and } y_1, \ldots, y_{k+1}) &= \frac{1}{k+1} \operatorname{Pop}(y_i) \prod_{j \neq i} Q(y_j) \\
&= \alpha \frac{\operatorname{Pop}(y_i)}{Q(y_i)}, \quad \alpha = \frac{1}{k+1} \prod_i Q(y_i)
\end{aligned}
$$

$$
\begin{aligned}
P(i \mid y_1, \ldots y_{k+1}) &= \frac{P(i \text{ and } y_1, \ldots, y_{k+})}{\sum_i P(i \text{ and } y_1, \ldots, y_{k+})} \\
&= \operatorname*{softmax}_i \left(\ln \frac{\operatorname{Pop}(y_i)}{Q(y_i)} \right) + C
\end{aligned}
$$

Constrastive Estimation

Like the discrimination estimate, the contrastive estimate of Pop is poor when the contrastive task is easy — when the contrastive loss can be driven near zero.

However, the contrastive task can be made more difficult by increasing k.

Contrastive GANs (TZ)

Discriminative GAN:

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \min_{\Psi} \; E_{(i,y)\sim(\mathrm{Pop}\uplus P_\Phi)} - \ln P_\Psi(i|y)$$

Contrastive GAN:

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \min_{\Psi} \; E_{(i,y_1,\ldots,y_{k+1})\sim(\mathrm{Pop}\hookrightarrow P_\Phi^k)} - \ln P_\Psi(i|y_1,\ldots,y_{k+1})$$

Assuming Universality:

$$P_{\Phi^*} = \mathrm{Pop}$$

Summary

GANs have not generally proved useful for representation learning for discriminative tasks such as image segmentation, speech recognition, or machine translation.

I predict that there will ultimately be better ways to model distributions (as in language modeling).

I predict that in a few years discriminators will be limited to enforcing perceptual naturalness in applications such as text to speech and image decompression.

END