TTIC 31230, Fundamentals of Deep Learning

David McAllester, April 2017

Generative Adversarial Networks (GANs)

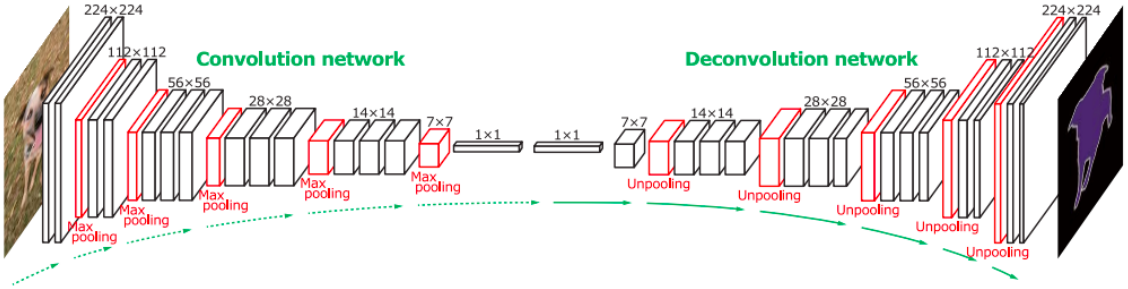
Modeling Distributions (Cross Entropy)

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{y \sim \mathrm{Pop}} \; -\log Q_\Phi(y)$$

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{(x,y) \sim \mathrm{Pop}} \; -\log Q_\Phi(y|x)$$

Variational AutoEncoders

$$Q_\Psi(\hat{z}|y) \qquad Q_\Phi(\hat{z}) \qquad \hat{y}_\Phi(\hat{z})$$



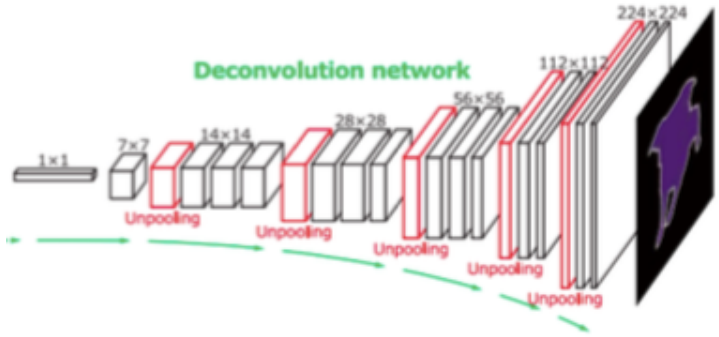
[Hyeonwoo Noh et al.]

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}}\, E_{y\sim \mathrm{Pop}}\ \log Q_\Phi(y) \quad \text{(Cross Entropy)}$$

$$\log Q_\Phi(y) \geq E_{\hat{z}\sim P_\Psi(\hat{z}|y)}\ \ln Q_\Phi(\hat{z}, y) + H(P_\Psi(\hat{z}|y)) \quad \text{(ELBO)}$$

Generative Adversarial Networks (GANs)

$\epsilon \sim$ noise $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\hat{y}_\Phi(\epsilon)$



[Hyeonwoo Noh et al.]

In a GAN a distribution is modeled by a **generator** — there is no encoder.

GANs, Goodfellow et al., 2014 (also Schmidhuber 1992)

Cross entropy loss on Q_Φ is replaced by

$$\Phi^* = \operatorname*{argmax}_{\Phi} \min_{\Psi} E_{(y,s)\sim(\mathrm{Pop} \uplus Q_\Phi)} - \log Q_\Psi(s|y)$$

Φ is the generator.

y is either drawn from Q_Φ or from Pop (with equal probabiity) and s is a flag telling which.

Ψ is the discriminator — the descriminator must predict s from y.

The Theorem

$$\Phi^* = \operatorname*{argmax}_{\Phi} \min_{\Psi} E_{(y,s)\sim(\text{Pop} \uplus Q_\Phi)} - \log Q_\Psi(s|y)$$

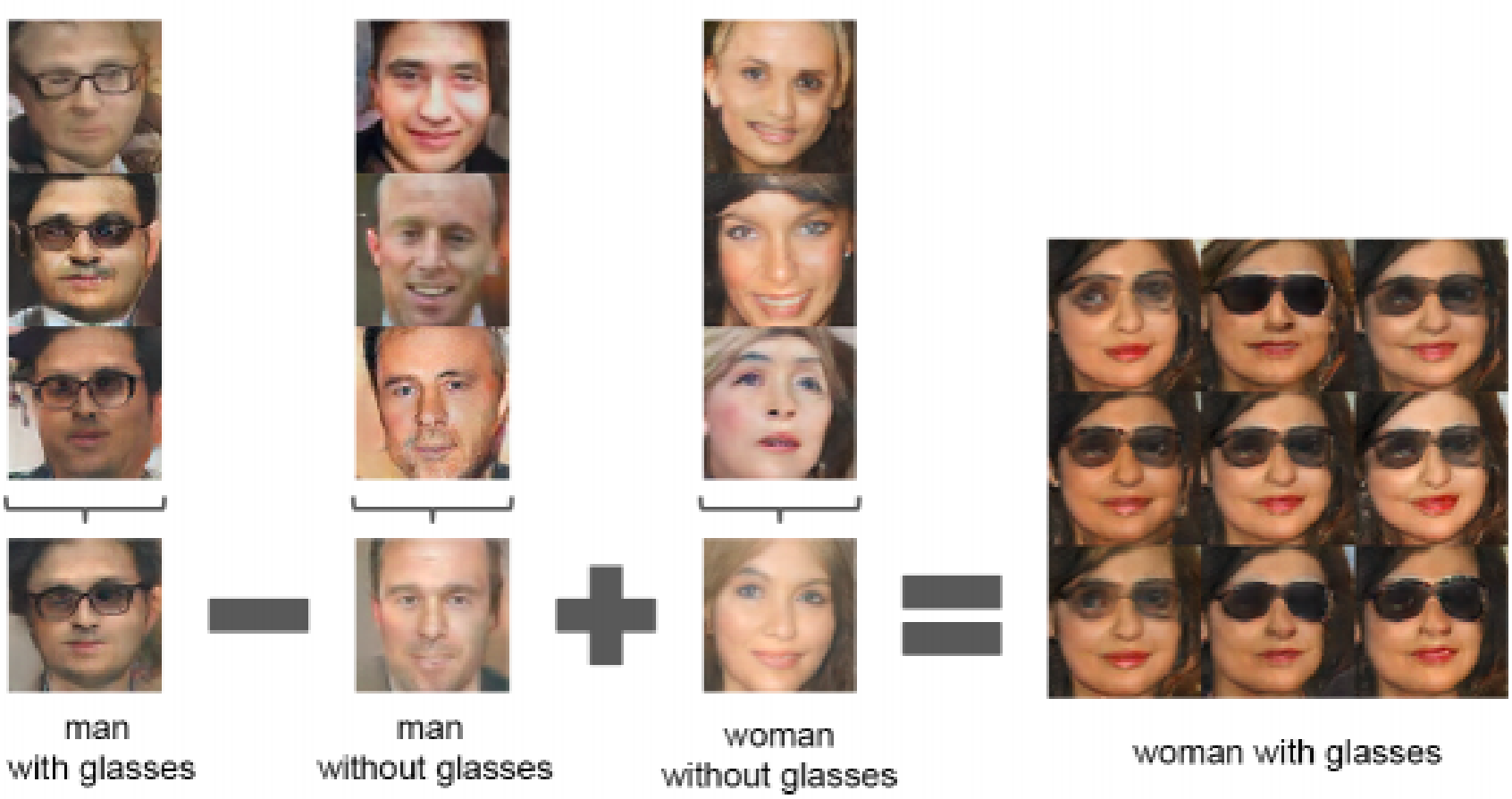
Theorem: If $Q_\Phi(y)$ and $Q_\Psi(s|y)$ are universally expressive (can represent any distribution) then $Q_{\Phi^*} = \text{Pop}$.

Generated Bedrooms(DC GANS, Radford et al., ICLR 2016)

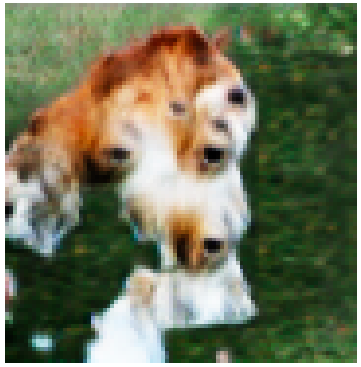
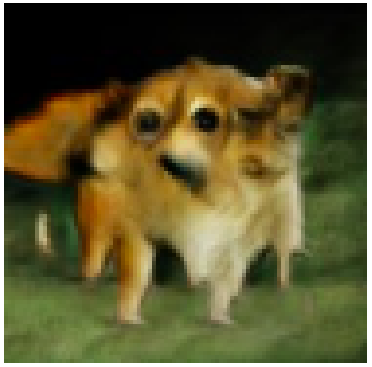
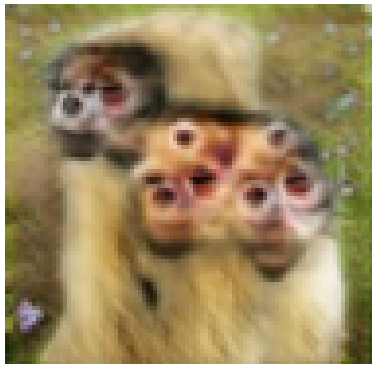
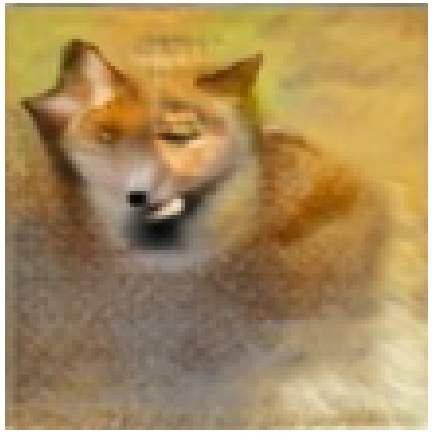
Interpolated Faces

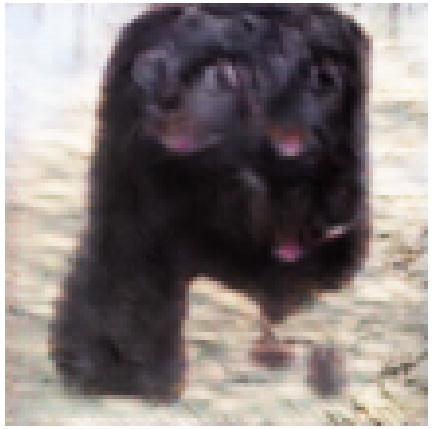
[Ayan Chakrabarti]

Image Arithmetic (DC GANS, Radford et al., ICLR 2016)



GANs on Imagenet

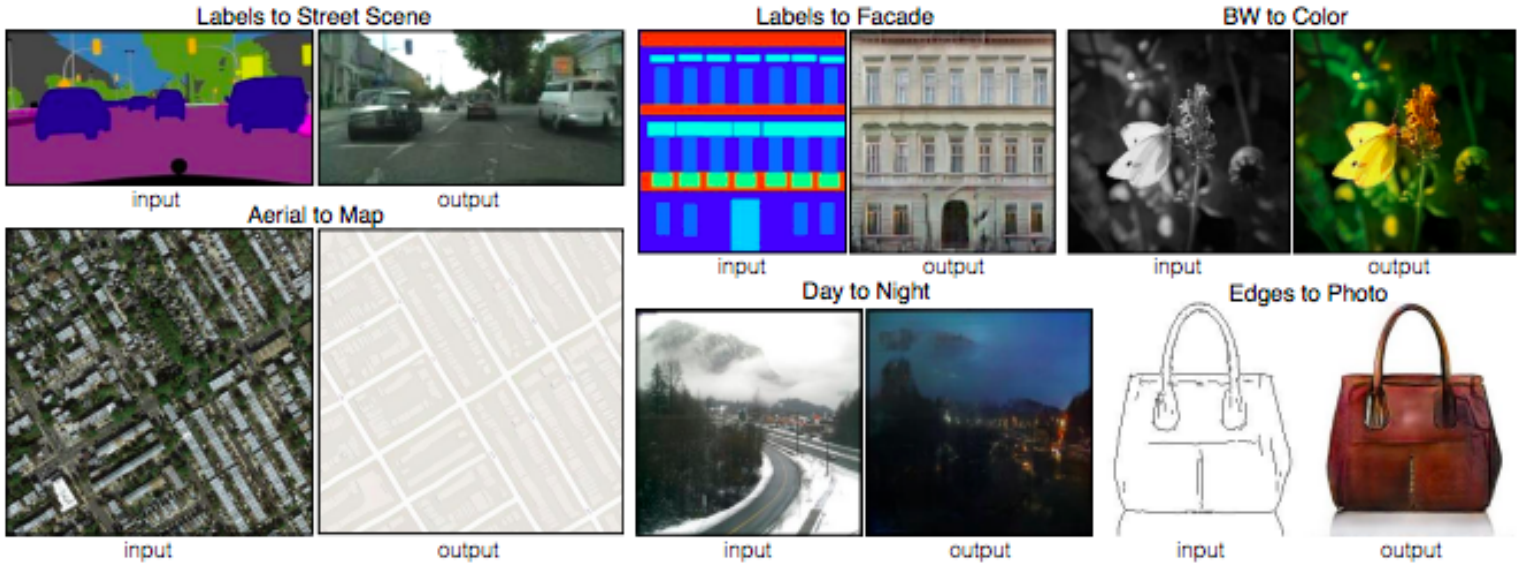


Conditional GANs

All distribution modeling methods apply to conditional distributions.

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \ \min_{\Psi} \ E_{(x,y,s)\sim(\mathrm{Pop} \uplus \mathrm{Pop}(x)Q_\Phi(y|x))} \ -\log Q_\Psi(s|x,y)$$

Image-to-Image Translation (Isola et al., 2016)



U-Nets (Ronnenberger et al. 2015)

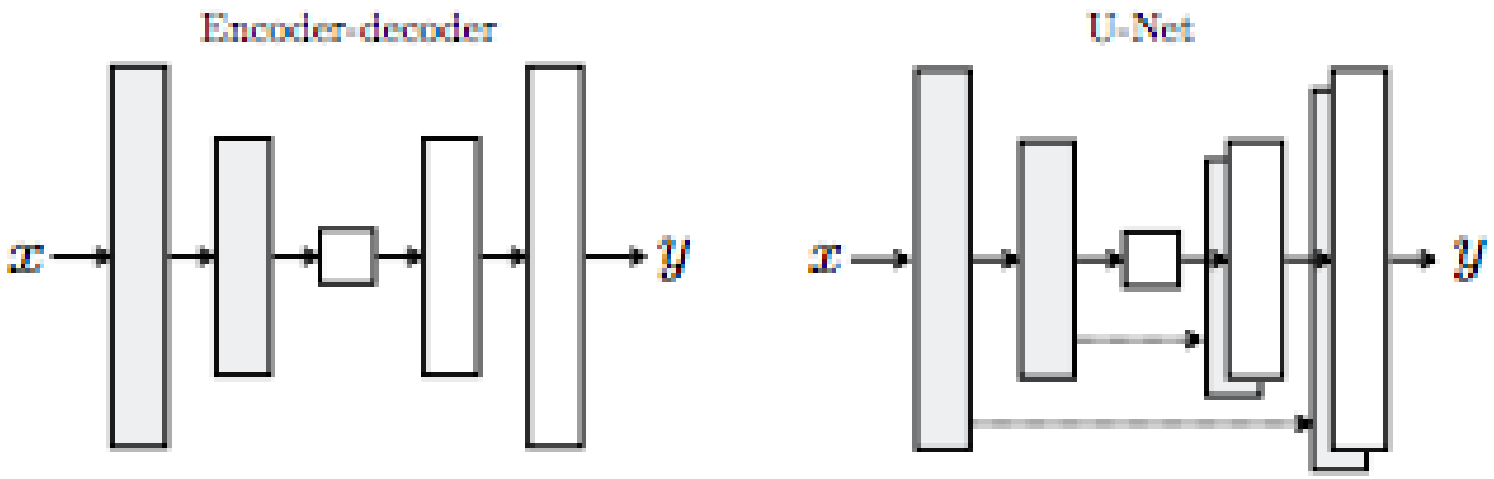
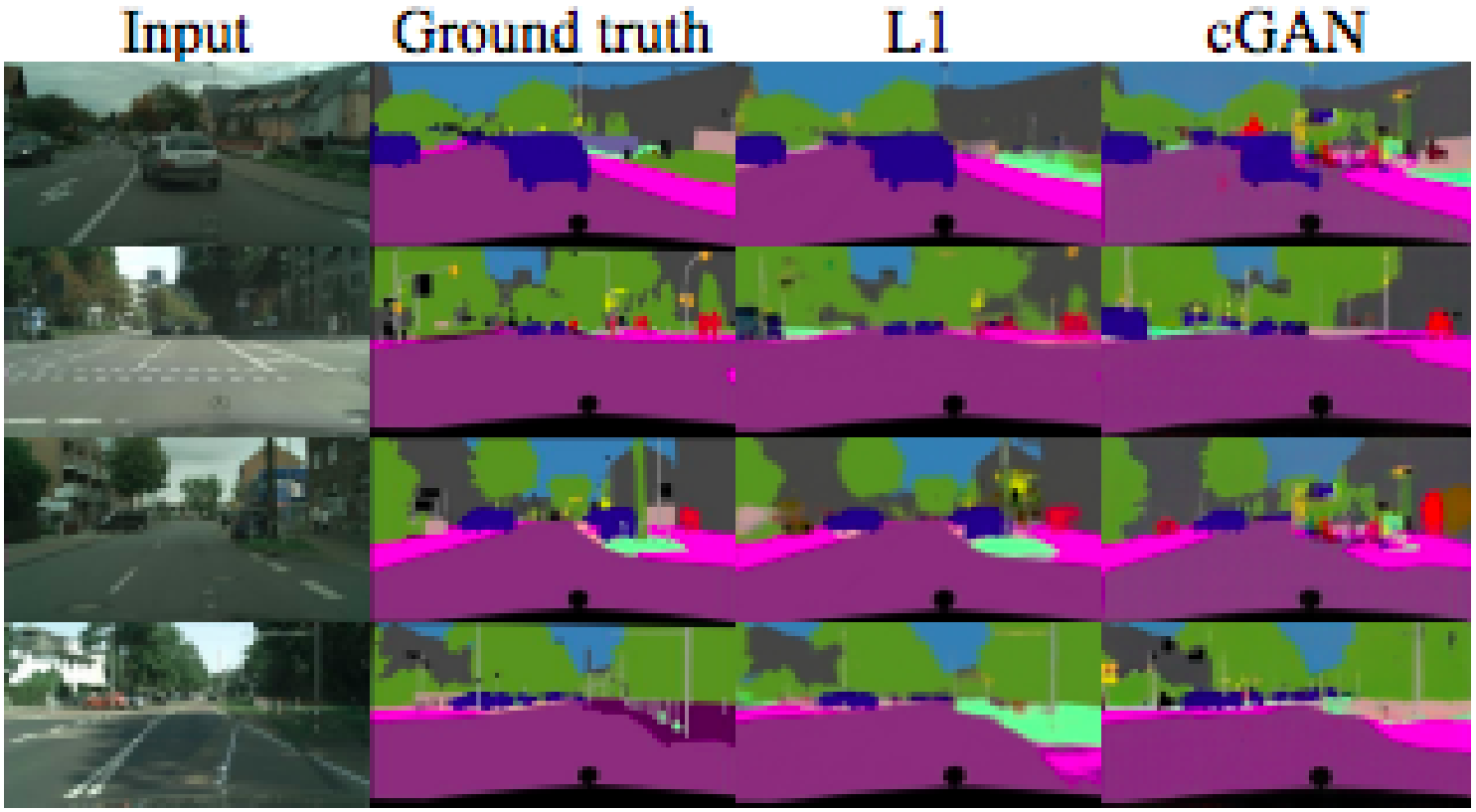


Image-to-Image Translation (Isola et al., 2016)

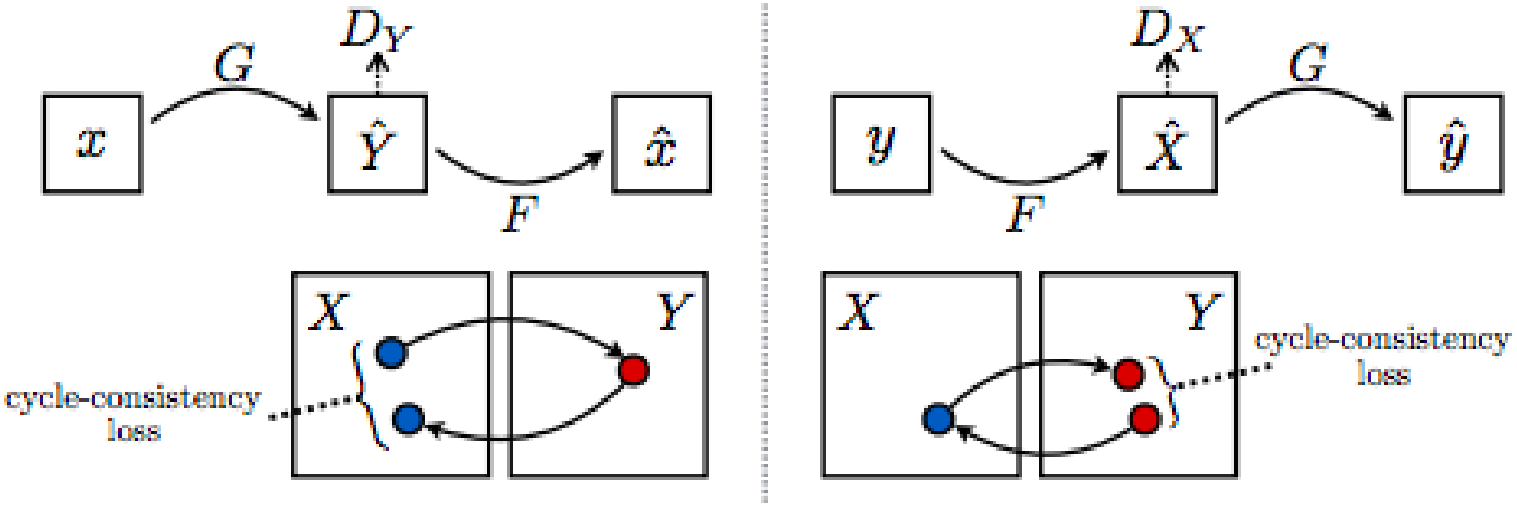
Arial Photo to Map and Back

Colorization

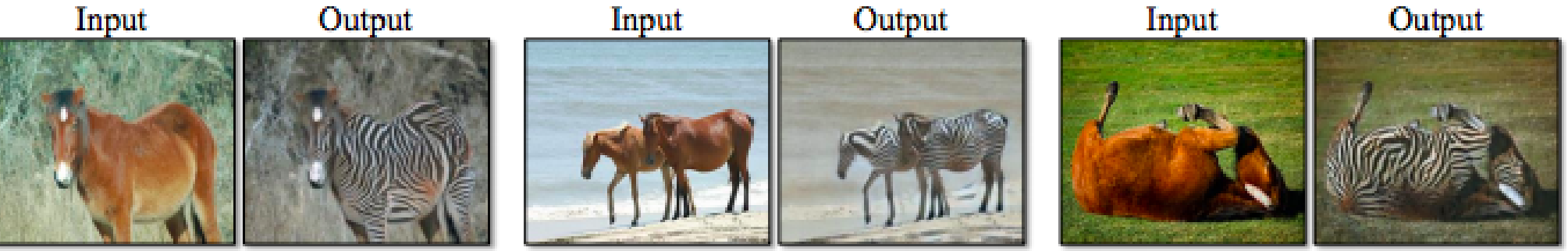
Semantic Segmentation



Cycle Gans (Zhu et al., 2017)



Cycle Gans

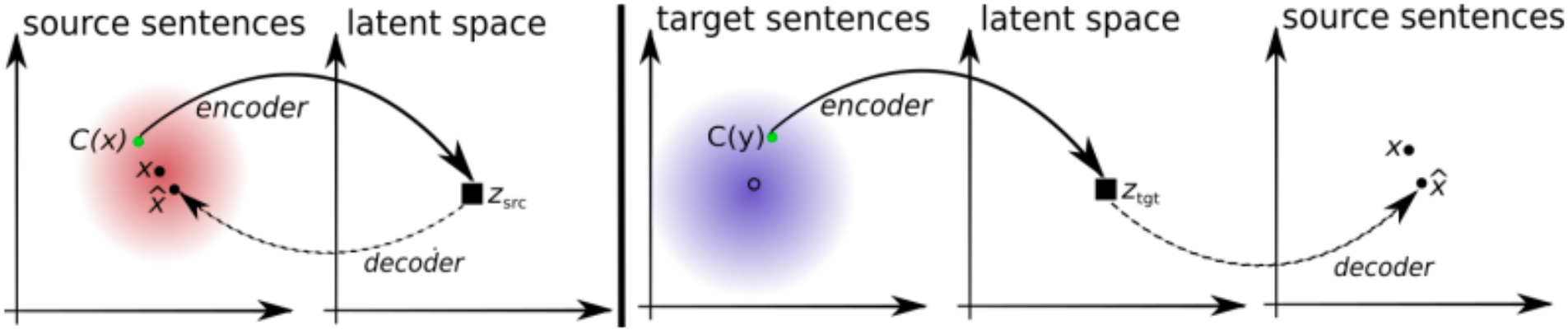

Input
Output
Input
Output
Input
Output
horse → zebra
zebra → horse

Cycle Gans

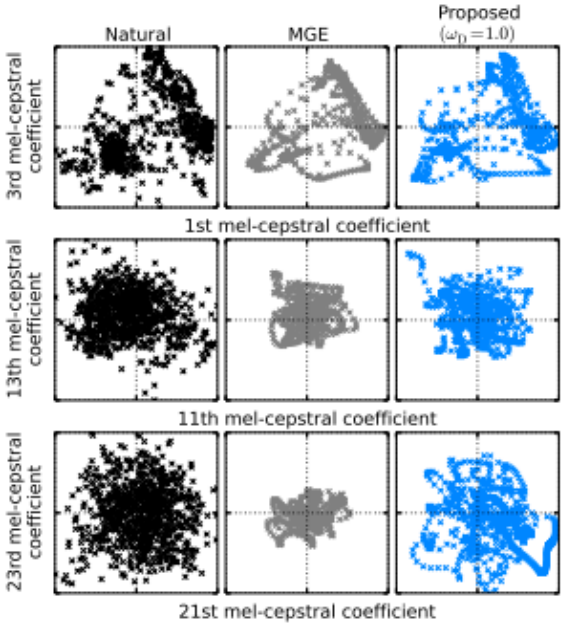
Horse → Zebra

Cycle Training of Machine Translation

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



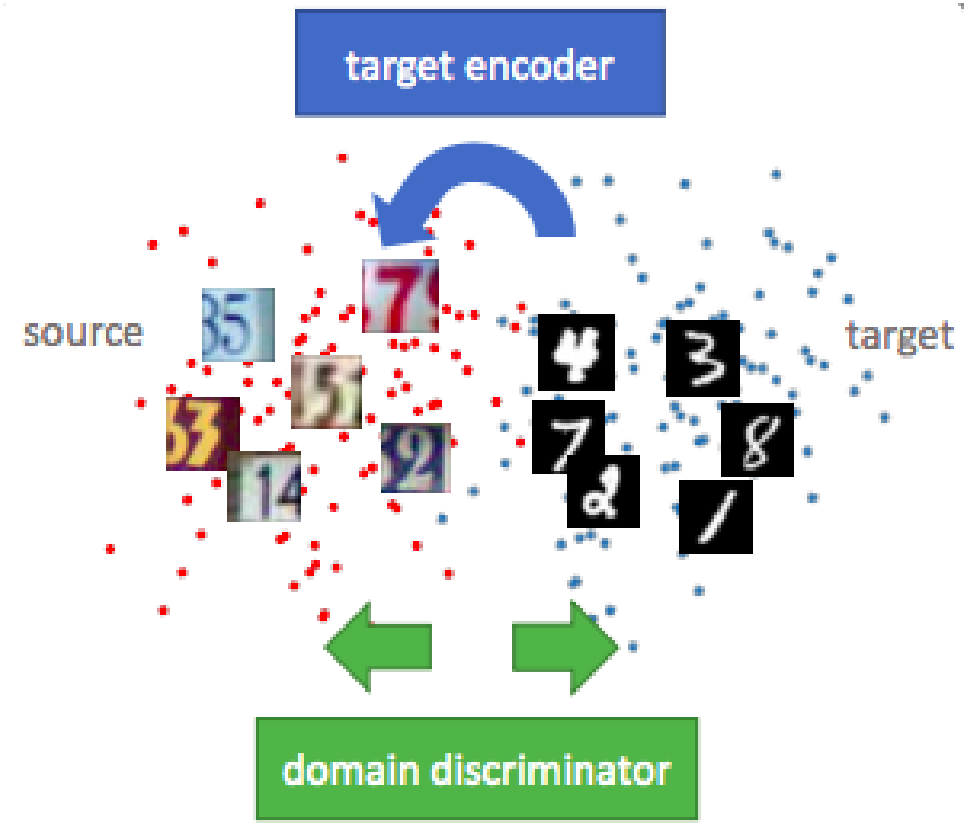
Text to Speech (Saito et al. 2017)



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 Domain Adaptation (Tzeng et al. 2017)



Issues

Jensen-Shannon Divergence

Vanishing Gradients

Unstable Training

Mode Collapse

Measuring Perfomance

Jensen-Shannon Divergence

$$\Phi^* = \operatorname*{argmax}_{\Phi} \min_{\Psi} E_{(y,s)\sim(\mathrm{Pop} \uplus Q_\Phi)} - \log Q_\Psi(s|y)$$

$$\Psi^*(\Phi) = \operatorname*{argmin}_{\Psi} E_{(y,s)\sim(\mathrm{Pop} \uplus Q_\Phi)} - \log Q_\Psi(s|y)$$

$$Q_{\Psi^*(\Phi)}(s=1|y) = \frac{P(y,s=1)}{P(y)} = \frac{\mathrm{Pop}(y)}{\mathrm{Pop}(y) + Q_\Phi(y)}$$

$$\Phi^* = \operatorname*{argmax}_{\Phi} E_{(y,s)\sim(\mathrm{Pop} \uplus Q_\Phi)} - \log Q_{\Psi^*(\Phi)}(s|y)$$

$$= \operatorname*{argmax}_{\Phi} \frac{1}{2} E_{(y,1)\sim\mathrm{Pop}} - \log \frac{\mathrm{Pop}(y)}{\mathrm{Pop}(y) + \pi(y|\Phi)}$$
$$+ \frac{1}{2} E_{(y,-1)\sim Q_\Phi} - \log \frac{Q_\Phi(y)}{\mathrm{Pop}(y) + Q_\Phi(y)}$$

$$= \operatorname*{argmax}_{\Phi} 1 - \frac{1}{2} KL(\mathrm{Pop}, A) - \frac{1}{2} KL(Q_\Phi, A)$$

$$A(y) = \frac{1}{2}(\mathrm{Pop}(y) + Q_\Phi(y))$$

Jensen-Shannon Divergence (JSD)

We have arrived at the Jensen-Shannon divergence.

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

$$\mathrm{JSD}(P, Q) = \frac{1}{2} KL\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} KL\left(Q, \frac{P+Q}{2}\right)$$

$$0 \leq \mathrm{JSD}(P, Q) = \mathrm{JSD}(Q, P) \leq 1 \quad \text{(in bits)}$$

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, Q_Φ).

A Heuristic Fix

We continue to use

$$\Psi^*(\Phi) = \operatorname*{argmin}_{\Psi} \; E_{(y,s)\sim(\mathrm{Pop} \uplus Q_\Phi)} \; -\log Q_\Psi(s|y)$$

But switch the optimization for Φ from

$$\Phi^* = \operatorname*{argmax}_{\Phi} E_{y\sim Q_\Phi} \; -\log Q_\Psi(-1|y)$$

to

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{y\sim Q_\Phi} \; -\log Q_\Psi(1|y)$$

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

Converting to Cross Entropy (Goodfellow)

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

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

$$\begin{aligned} \text{Define:} \quad f_{\Psi^*}(y) &\doteq \frac{Q_{\Psi^*}(1|y)}{Q_{\Psi^*}(-1|y)} \\ &= \frac{\mathrm{Pop}(y)}{Q_\Phi(y)} \end{aligned}$$

Converting to Cross Entropy

$$
\begin{aligned}
\nabla_\Phi \, E_{y \sim Q_\Phi} \, f_\Psi(y) &= \nabla_\Phi \sum_y \, Q_\Phi(y) f_\Psi(y) \\
&= \sum_y \, Q_\Phi(y) f_\Psi(y) \nabla_\Phi \ln Q_\Phi(y) \\
&= \sum_y \, \mathrm{Pop}(y) \nabla_\Phi \ln Q_\Phi(y) \\
&= E_{y \sim \mathrm{Pop}} \, \nabla_\Phi \ln Q_\Phi(y) \\
&= \nabla_\Phi \, E_{y \sim \mathrm{Pop}} \, \ln Q_\Phi(y)
\end{aligned}
$$

Unstable Training

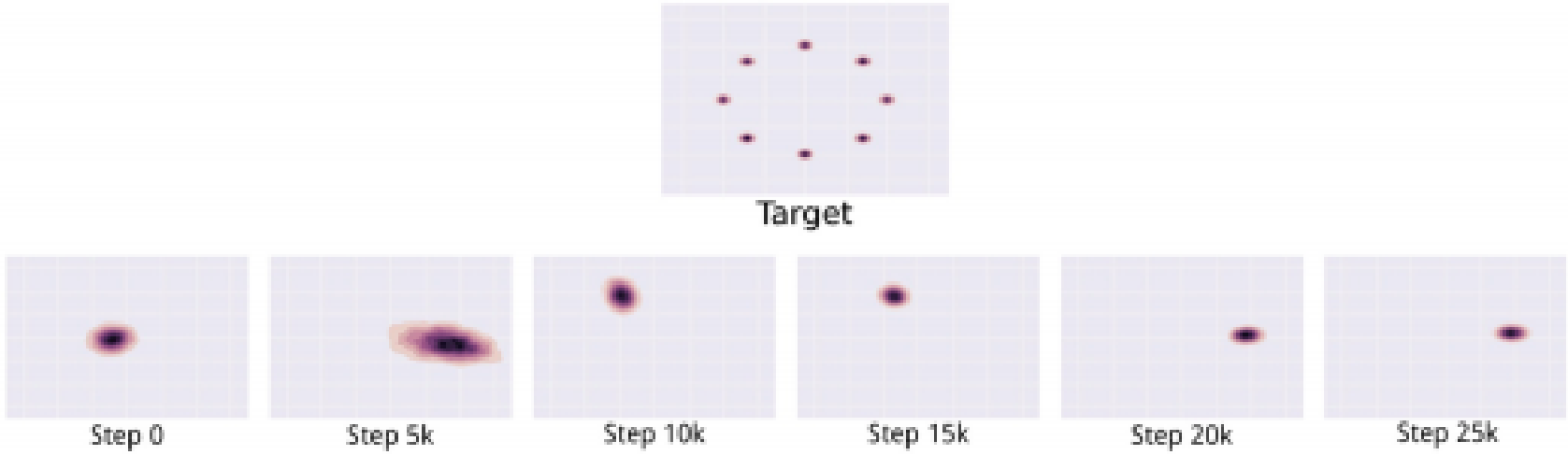
Simultaneous gradient descent is not the same as nested max-min.

$$\max_{\Phi} \min_{\Psi} \; E_{(y,s)\sim(\mathrm{Pop} \uplus Q_\Phi)} \; -\log Q_\Psi(s|y)$$

vs.

$$\min_{\Psi} \max_{\Phi} \; E_{(y,s)\sim(\mathrm{Pop} \uplus Q_\Phi)} \; -\log Q_\Psi(s|y)$$

A Synthetic Example



Another Example

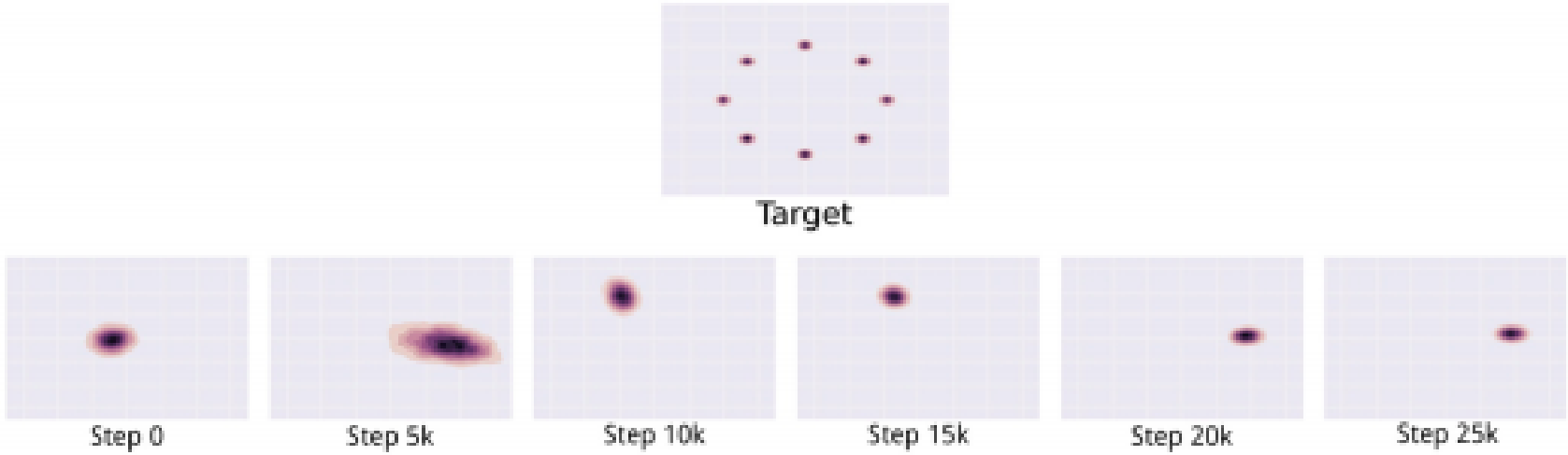
$$\min_x \max_y xy$$

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.



Measuring Performance

Most evaluation of GANs is based on subjective judgments of naturalness.

This is in contrast to language modeling where performance is directly measured by cross-entropy (bits per character or perplexity).

Summary

GANs have not generally proved useful in 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