TTIC 31230, Fundamentals of Deep Learning

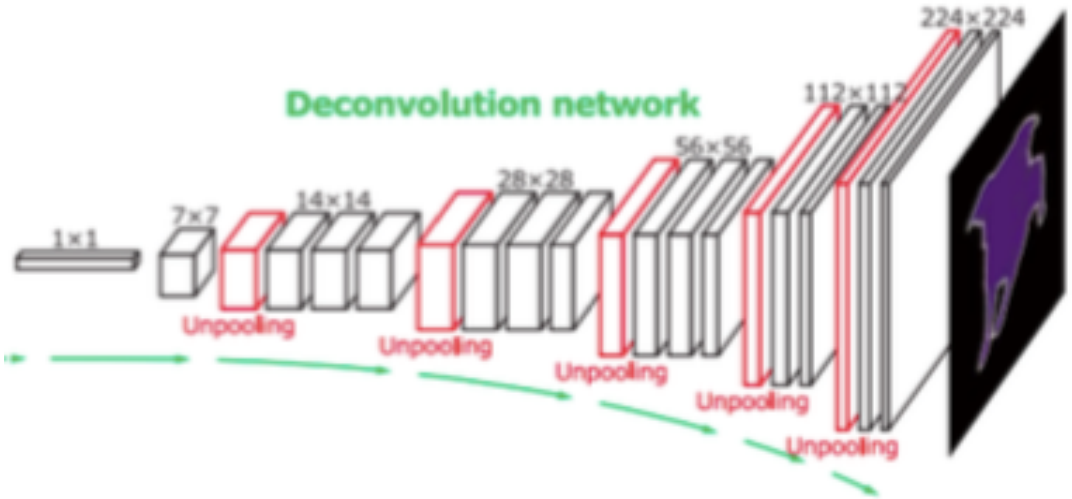
David McAllester, Winter 2019

Discrimination Loss

and Generative Adversarial Networks (GANs)

Representing a Distribution with a Generator

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



Generative Adversarial Nets

Goodfellow et al., June 2014

In a GAN the parameters Φ of a generator network are trained by the equation

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, \mathcal{L}_{\mathrm{Discr}}(\mathrm{pop}, p_\Phi).$$

Here $\mathcal{L}_{\mathrm{Discr}}(\mathrm{pop}, p_\Phi)$ is a "discrimination loss" — the ability of a discriminator to distinguish the population distribution pop from the generated model distribution p_Φ.

Classification Discrimination

In the original GAN paper (Goodfellow et al.), and in current practice, the discriminator is a classifier that classifies samples as either being from the population or being from the model.

Let $p \uplus q$ be the distribution defined by flipping an unbiased coin and, if heads, returning $(1, y)$ with $y \sim p$ and, if tails, returning $(-1, y)$ with $y \sim q$.

$$\mathcal{L}_{\text{Discr}}(p, q) = \max_{\Psi} \; E_{(i,y)\sim p \uplus q} \; \ln P_{\Psi}(i|y)$$

Assuming Universality of Ψ

$$\Phi^* = \operatorname*{argmin}_{\Phi} \mathcal{L}_{\text{Discr}}(\text{pop}, p_\Phi).$$

$$\mathcal{L}_{\text{Discr}}(\text{pop}, p_\Psi) = \max_{\Psi} E_{(i,y)\sim \text{pop} \uplus p_\Phi} \ln P_\Psi(i|y)$$

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y)\sim \text{pop} \uplus p_\Phi} - \ln P_\Psi(i|y)$$

Assuming Universality for Ψ, $P_{\Psi^*}(i|y)$ equals the true probability $P_\Phi(i|y)$ defined by the distribution $\text{pop} \uplus p_\Phi$.

$$P_{\Psi^*}(i|y) = P_\Phi(i|y)$$

Assuming Universality

$$\Phi^* = \operatorname*{argmin}_{\Phi} \mathcal{L}_{\mathrm{Discr}}(\mathrm{pop}, p_\Phi).$$

$$\mathcal{L}_{\mathrm{Discr}}(\mathrm{pop}, p_\Phi) = E_{(i,y)\sim \mathrm{pop}\uplus p_\Phi} \ln P_\Phi(i|y)$$

$$= -H(i|y)$$

The generator Φ is trying to maximize $H(i|y)$ which is achieved at $\ln 2$ (one bit) for $p_\Phi = \mathrm{pop}$.

Assuming Universality of both Φ and Ψ we have $p_{\Phi^*} = \mathrm{pop}$.

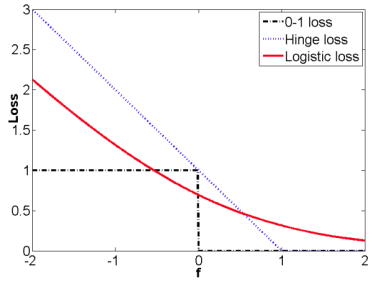
Review of Binary Classification

In the case of binary classification cross-entropy loss becomes the log loss of the margin

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

$$= \underset{\Psi}{\operatorname{argmin}} \, E_{(i,y)\sim(\mathrm{Pop} \uplus P_\Phi)} \quad \ln(1+e^{-m})$$

$$m = i \; s_\Psi(y)$$



Vanishing Gradients

For $i = 1$ and $y \sim$ pop:

$$\Psi \texttt{ += } \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_\Psi \, s_\Psi(y) \quad \approx 0 \text{ for } m >> 1$$

For $i = -1$ and $y \sim P_\Phi$:

$$\Psi \texttt{ -= } \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_\Psi \, s_\Psi(y_\Phi(z)) \quad \approx 0 \text{ for } m >> 1$$

$$\Phi \texttt{ += } \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_\Phi \, s_\Psi(y_\Phi(z)) \quad \approx 0 \text{ for } m >> 1$$

The gradients vanish when the discriminator achieves large margins.

A Heuristic Patch

Replace

$$\Phi \texttt{ += } \eta \, \frac{e^{-m}}{1+e^{-m}} \, \nabla_\Phi \, s_\Psi(y_\Phi(z)) \quad \approx 0 \text{ for } m >> 1$$

with

$$\Phi \texttt{ += } \eta \, \nabla_\Phi \, s_\Psi(y_\Phi(z))$$

This allows the generator to recover.

Generative Adversarial Nets

Goodfellow et al., June 2014

The rightmost column (yellow boarders) gives the nearest neighbor in the training data to the adjacent column.

Assuming Universality of Ψ Only

$$-H(i|y) = E_{(i,y)\sim(\mathrm{pop}\uplus p_\Phi)} \ln P(i|y)$$

$$P(i|y) = \frac{p(i \wedge y)}{p(y)}$$

$$P(1|y) = \frac{\frac{1}{2}\mathrm{pop}(y)}{\frac{1}{2}\mathrm{pop}(y) + \frac{1}{2}p_\Phi(y)}$$

Assuming Universality of Ψ Only

$$E_{(i,y)\sim(\mathrm{pop}\uplus p_\Phi)}\ \ln P(i|y)$$

$$= \frac{1}{2}E_{y\sim\mathrm{pop}}\ \ln\frac{\frac{1}{2}\mathrm{pop}(y)}{\frac{1}{2}\mathrm{pop}(y)+\frac{1}{2}p_\Phi(y)} + \frac{1}{2}E_{y\sim p_\Phi}\ \ln\frac{\frac{1}{2}p_\Phi(y)}{\frac{1}{2}\mathrm{pop}(y)+\frac{1}{2}p_\Phi(y)}$$

$$= \frac{1}{2}\left(\ KL\left(\mathrm{pop}, \frac{\mathrm{pop}+p_\Phi}{2}\right),\ KL\left(p_\Phi, \frac{\mathrm{pop}+p_\Phi}{2}\right)\right) - \ln 2$$

$$= \mathrm{JSD}(\mathrm{pop}, p_\Phi) - \ln 2$$

Assuming Universality of Ψ Only

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \ \operatorname{JSD}(\text{pop}, p_\Phi)$$

Contrastive Discrimination

Gutmann and Hyvärinen, 2010

Let $\text{pop} \hookrightarrow p_\Phi^N$ be the distribution defined by drawing one "positive" from pop and k IID negatives from p_Φ; then inserting the positive at a random position among the negatives; and returning $(i, y_1, \ldots, y_{N+1})$ where i is the index of the positive.

$$\mathcal{L}_{\text{Discr}}(\text{pop}, p_\Phi) = \max_\Psi \; E_{(i,y_1,\ldots,y_{N+1})\sim(\text{pop}\hookrightarrow p_\Phi^N)} \; \ln P_\Psi(i|y_1, \ldots, y_{N+1})$$

Contrastive Discrimination

$$\Psi^* = \operatorname*{argmin}_{\Psi} \; E_{(i,y_1,\ldots,y_{N+1})\sim(\mathrm{pop}\hookrightarrow p_\Phi^N)} \; -\ln P_\Psi(i|y_1,\ldots,y_{N+1})$$

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

The contrastive task gets more difficult as N gets larger.

Assuming Universality

$$\Psi^* = \operatorname*{argmin}_{\Psi} \; E_{(i,y_1,\ldots,y_{N+1})\sim(\mathrm{pop}\hookrightarrow p_\Phi^N)} \; -\ln P_\Psi(i|y_1,\ldots,y_{N+1})$$

Assuming universality of Ψ we get

$$\mathcal{L}_{\mathrm{Discr}}(\Phi) = -H(i|y_1,\ldots y_{N+1})$$

The generator Φ is trying to maximize $H(i|y_1,\ldots,y_{N+1})$.

The maximum is achieved at $\ln(N+1)$ for $p_\Phi = \mathrm{pop}$.

Assuming Universality of a score function s_Ψ

Gutmann and Hyvärinen, 2010

$$\Psi^* = \underset{\Psi}{\operatorname{argmin}}\, E_{(i,y_1,\ldots,y_{N+1})\sim(\mathrm{pop}\hookrightarrow p_\Phi^N)} -\ln P_\Psi(i|y_1,\ldots,y_{N+1})$$

$$\text{Assume}: \quad P_\Psi(i|y_1,\ldots,y_{N+1}) \doteq \underset{i}{\operatorname{softmax}}\, s_\Psi(y_i)$$

Theorem: Assuming universality:

$$P_{\Psi^*}(i|y_1,\ldots,y_{N+1}) = \underset{i}{\operatorname{softmax}} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}$$

And therefore:

$$\mathrm{pop}(y) = \underset{y}{\operatorname{softmax}} \ \ s_{\Psi^*}(y) - \ln p_\Phi(y) \ \ Z \text{ must be estimated}$$

Proof

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

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

Assuming Universality of s_Ψ

Theorem:

$$P_{\Psi^*}(i|y_1,\ldots,y_{N+1}) = \operatorname*{softmax}_i \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}$$

implies

$$E_{(i,y_1,\ldots,y_{N+1})\sim \mathrm{pop}\hookrightarrow p_\Phi^N} \ln p_{\Psi^*}(i|y_1,\ldots,y_{N+1})$$

$$\leq \frac{N}{N+1}(KL(\mathrm{pop},p_\Phi)+KL(p_\Phi,\mathrm{pop}))+\ln\frac{1}{N+1}$$

Note that this upper bound holds with equality for $p_\Phi = \mathrm{pop}$.

Proof Part A.

$$
\begin{aligned}
& E_{(i,y_1,\ldots,y_{N+1})\sim \mathrm{pop}\hookrightarrow p_\Phi^N} \ \ln p_{\Psi^*}(i|y_1,\ldots,y_{N+1}) \\
= \ & E_{(i,y_1,\ldots,y_{N+1})\sim \mathrm{pop}\hookrightarrow p_\Phi^N} \ \ln \left(\operatorname*{softmax}_i \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} \right)[i] \\
= \ & E_{(i,y_1,\ldots,y_{N+1})\sim \mathrm{pop}\hookrightarrow p_\Phi^N} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln \left(\sum_i \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} \right) \\
= \ & E_{(i,y_1,\ldots,y_{N+1})\sim \mathrm{pop}\hookrightarrow p_\Phi^N} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln \left(\frac{1}{N+1} \sum_i \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} \right) - \ln (N+1)
\end{aligned}
$$

Proof Part B.

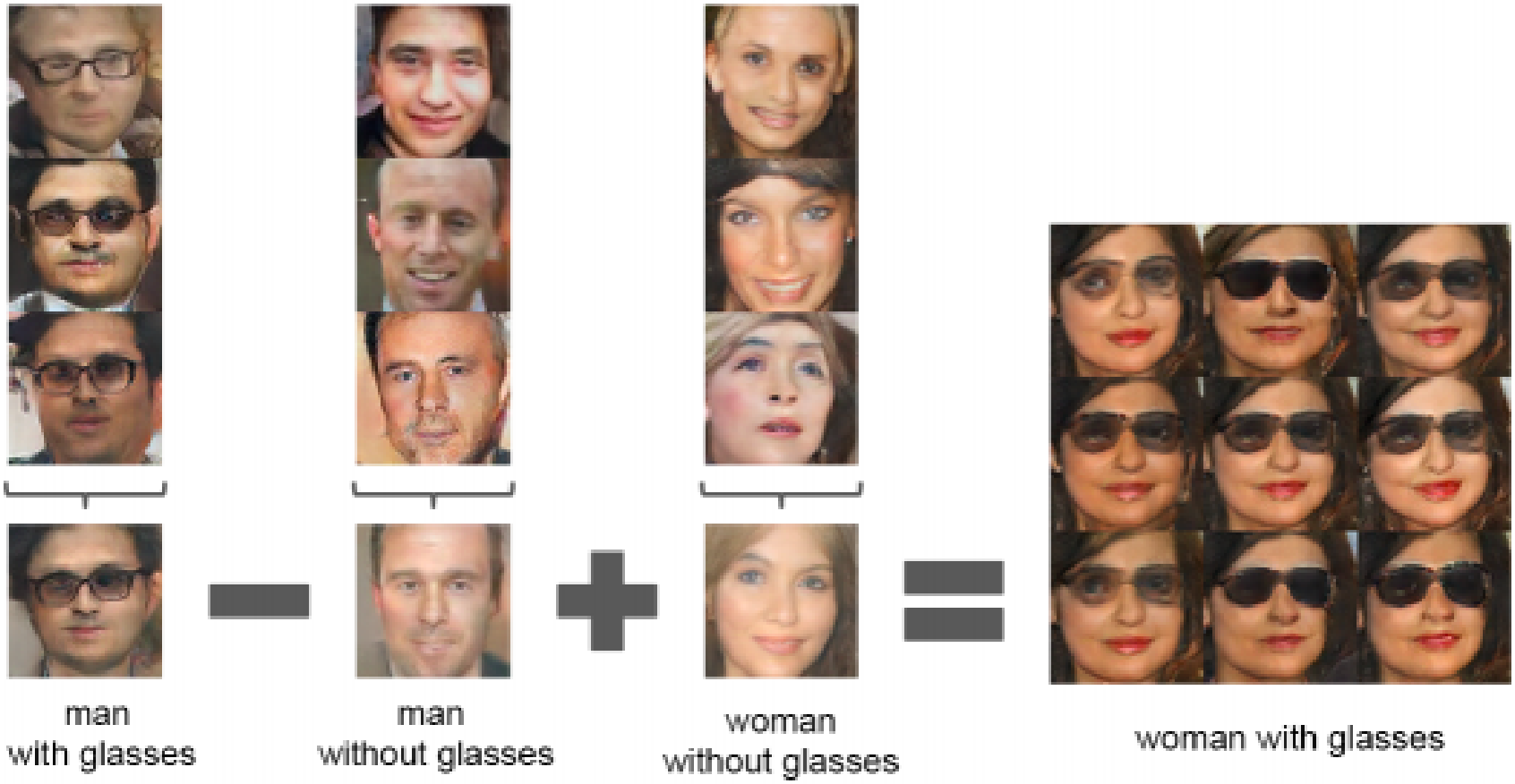
$$
\begin{aligned}
& E_{(i,y_1,\ldots,y_{N+1})\sim \mathrm{pop}\hookrightarrow p_\Phi^N} \ \ln p_{\Psi^*}(i|y_1,\ldots,y_{N+1}) \\
= \ & E_{(i,y_1,\ldots,y_{N+1})\sim \mathrm{pop}\hookrightarrow p_\Phi^N} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln\left(\frac{1}{N+1}\sum_i \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}\right) - \ln\,(N+1) \\
\leq \ & E_{(i,y_1,\ldots,y_{N+1})\sim \mathrm{pop}\hookrightarrow p_\Phi^N} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \frac{1}{N+1}\sum_i \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln\,(N+1) \\
= \ & \frac{N}{N+1} E_{y\sim \mathrm{pop}} \ln \frac{\mathrm{pop}(y)}{p_\Phi(y)} - \frac{N}{N+1} E_{y\sim p_\Phi} \ln \frac{\mathrm{pop}(y)}{p_\Phi(y)} - \ln\,(N+1) \\
= \ & \frac{N}{N+1}(KL(\mathrm{pop}, p_\Phi) + KL(p_\Phi, \mathrm{pop})) - \ln\,(N+1)
\end{aligned}
$$

Unsupervised Representation Learning ... (DC GANS)

Radford et al., Nov. 2015

Unsupervised Representation Learning ... (DC GANS)

Radford et al., Nov. 2015



Interpolated Faces

[Ayan Chakrabarti, January 2017]

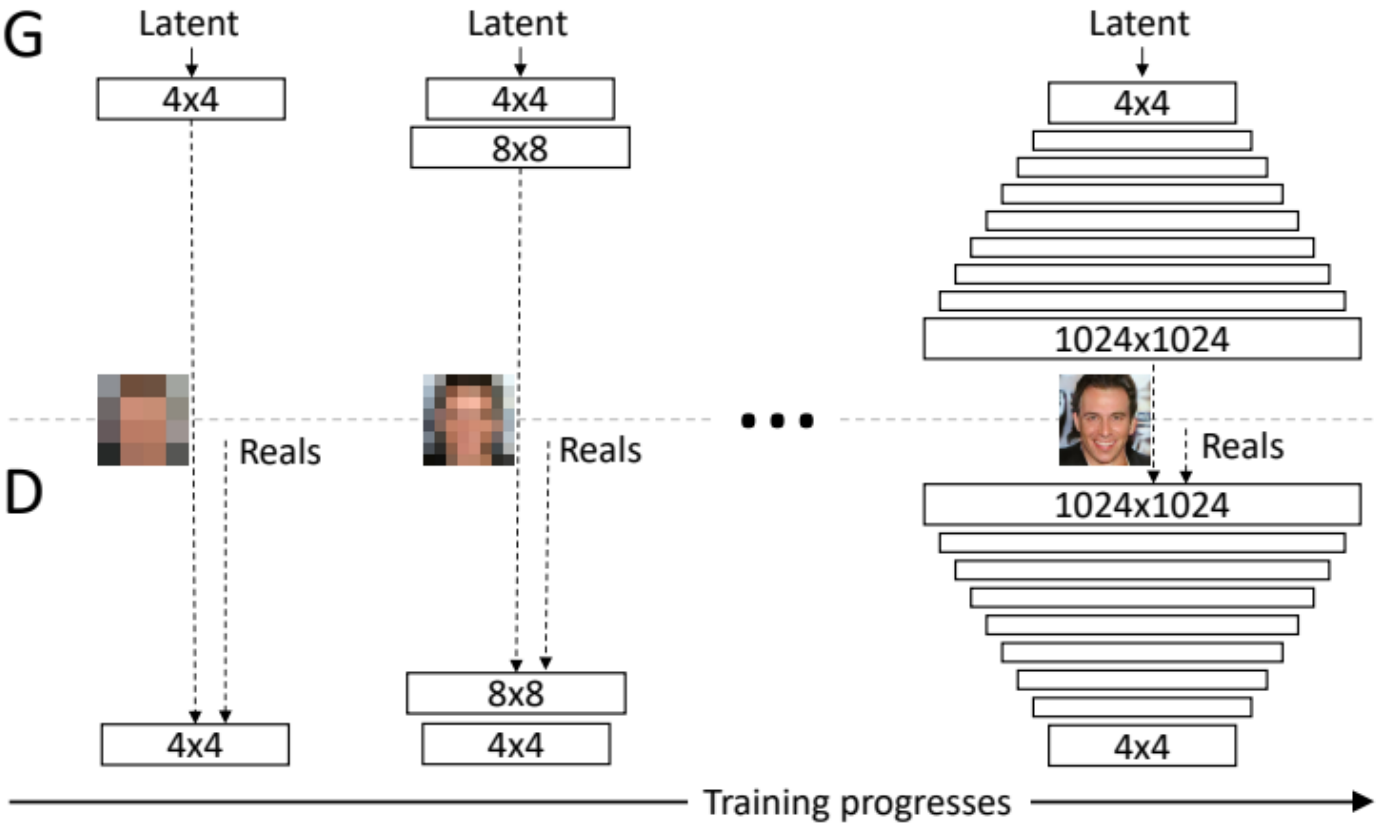
Progressive Growing of GANs

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.

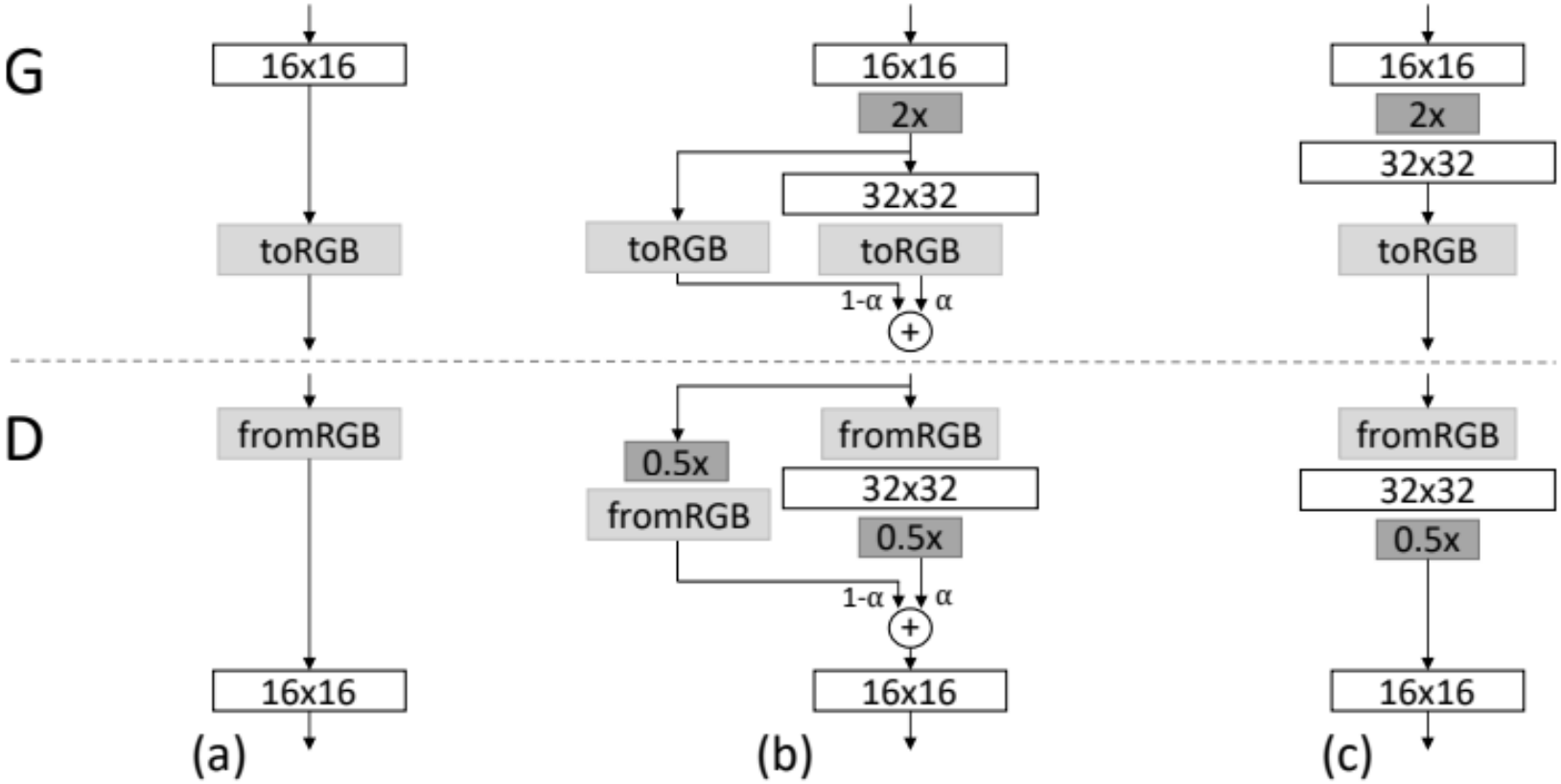
Progressive Growing of GANs

Karras et al., Oct. 2017



Progressive Growing of GANs

Karras et al., Oct. 2017



Weakly Conditional GANs

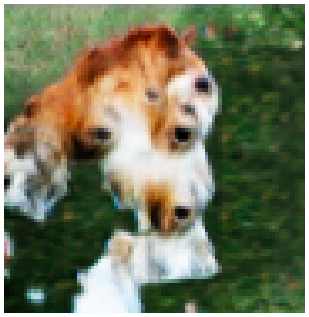
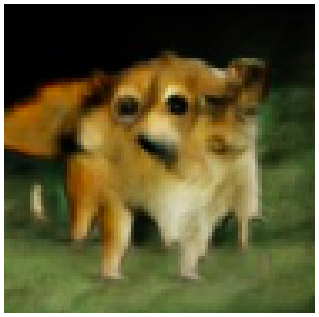
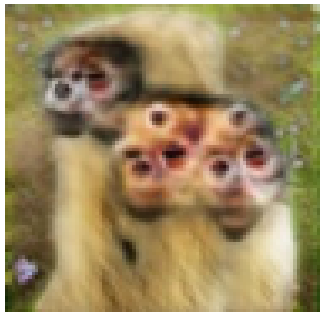
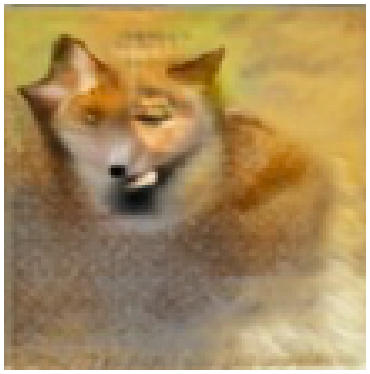
All unconditional distribution modeling methods apply to conditional distribution modeling.

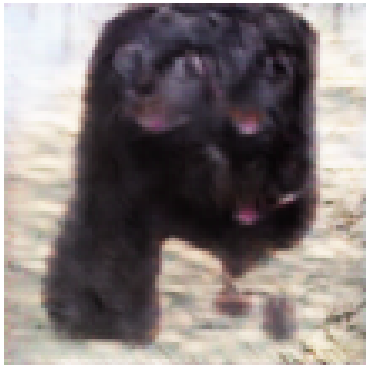
$$\Phi^* = \operatorname*{argmin}_{\Phi} \max_{\Psi} E_{i,x,y\sim(\text{pop} \uplus \text{Pop}(x)p_\Phi(y|x))} \ln P_\Psi(i|x,y)$$

By "weakly conditional" we mean that x is discrete and carries little information ($H(x)$ is small).

For example x might be the class label associated with an image.

Early Unconditional GANs on ImageNet



Large Scale GAN Training

Brock et al., Sept. 2018

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

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

This generates 512 X 512 images without using progressive training.

Issues

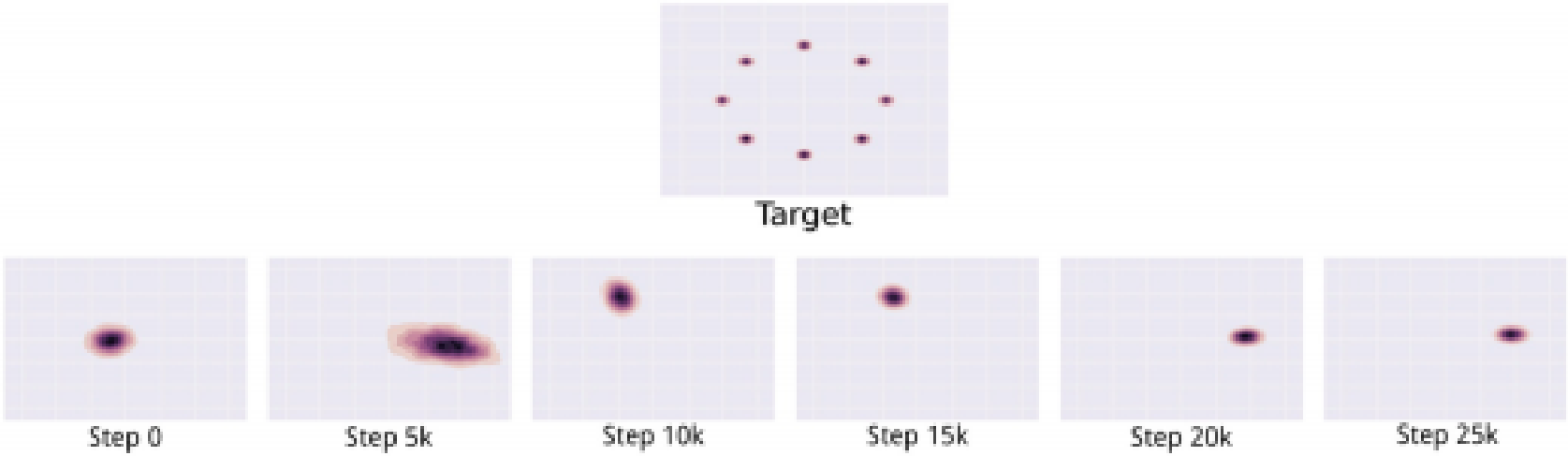
Mode Collapse

Unstable Training

Measuring Perfomance

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 $x = y = 0$.

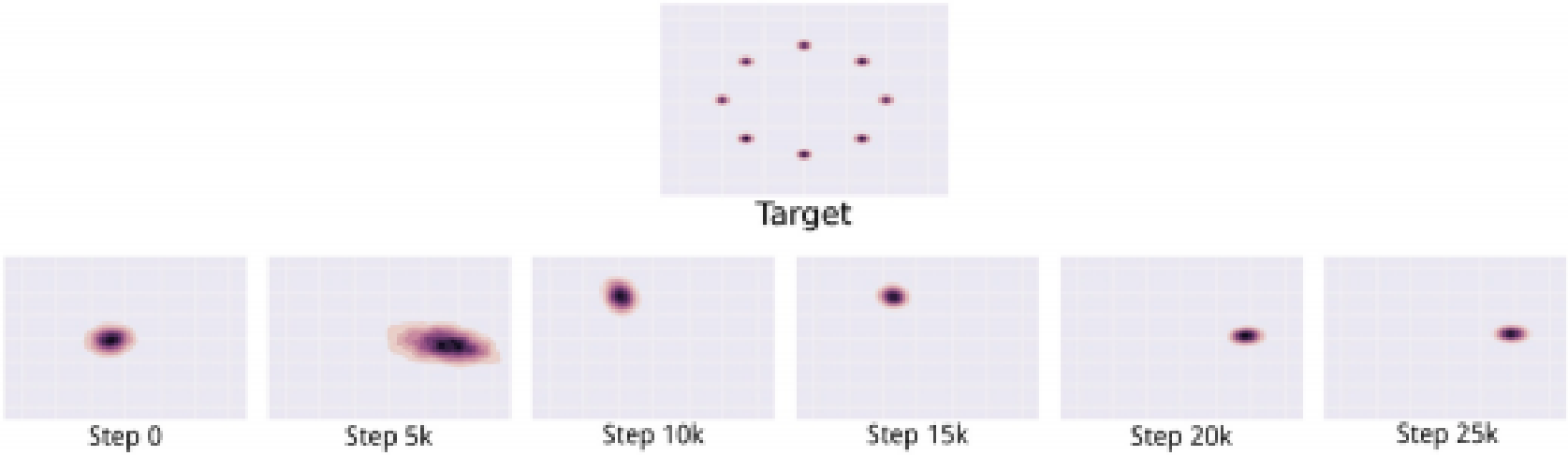
Simultaneous gradient flow yields

$$\frac{dx}{dt} = y \qquad \frac{dy}{dt} = -x$$

This goes in a circle.

Unstable Training

The generator distribution drifts as the discriminator follows.



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.

Adversarial Discrimination as an Additional Loss

Adversarial Discrimination as an Additional Loss

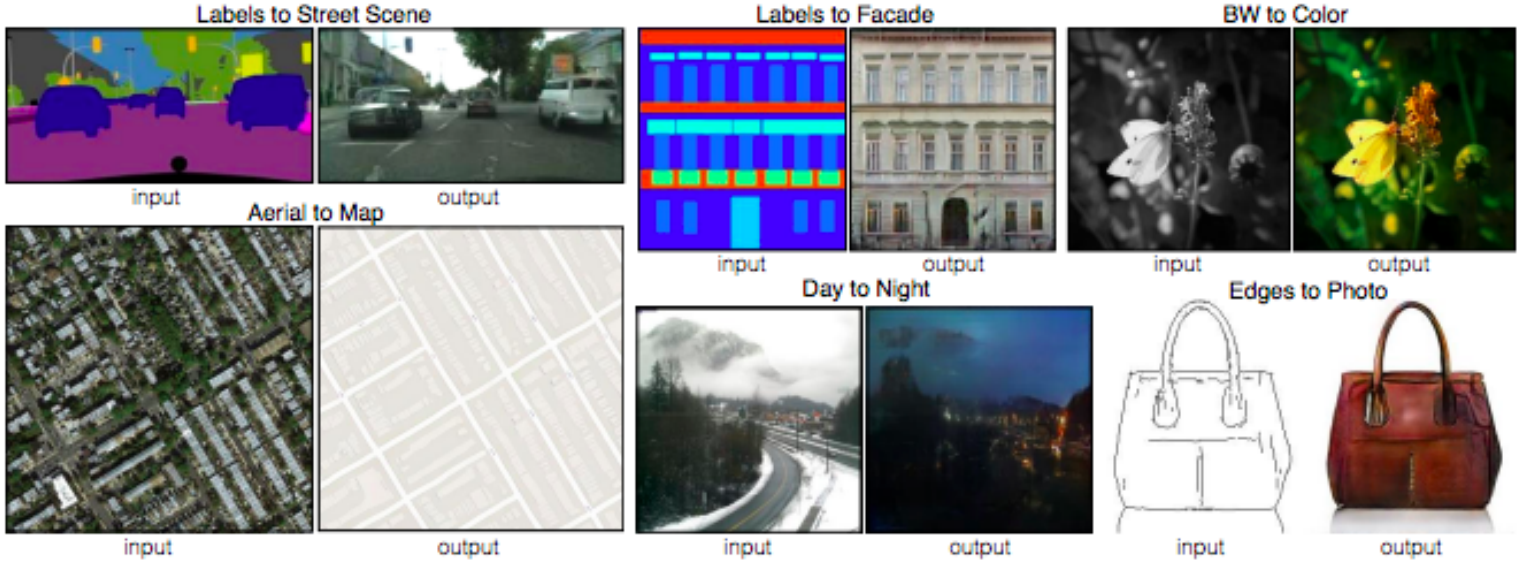
$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \; E_{(x,y)\sim \mathrm{pop}} \; ||y - \hat{y}(x)||^2 \; + \; \lambda \, \mathcal{L}_{\mathrm{Discr}}(y, \hat{y}(x), x)$$

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

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.



Discrimination as an Additional Loss

$$\text{L1}: \qquad \Phi^* \;=\; \text{argmin}_\Phi \;\; E_{(x,y)\sim\text{pop}} \;\; ||y - \hat{y}(x)||_1$$

$$\text{cGAN}: \qquad \Phi^* \;=\; \text{argmin}_\Phi \;\; E_{(x,y)\sim\text{pop}} \;\; \mathcal{L}_{\text{Discr}}(y, \hat{y}(x), x)$$

$$\text{L1} + \text{cGAN}: \;\; \Phi^* \;=\; \text{argmin}_\Phi \;\; E_{(x,y)\sim\text{pop}} \;\; ||y - \hat{y}(x)||_1 \;+\; \lambda \, \mathcal{L}_{\text{Discr}}(y, \hat{y}(x), x)$$

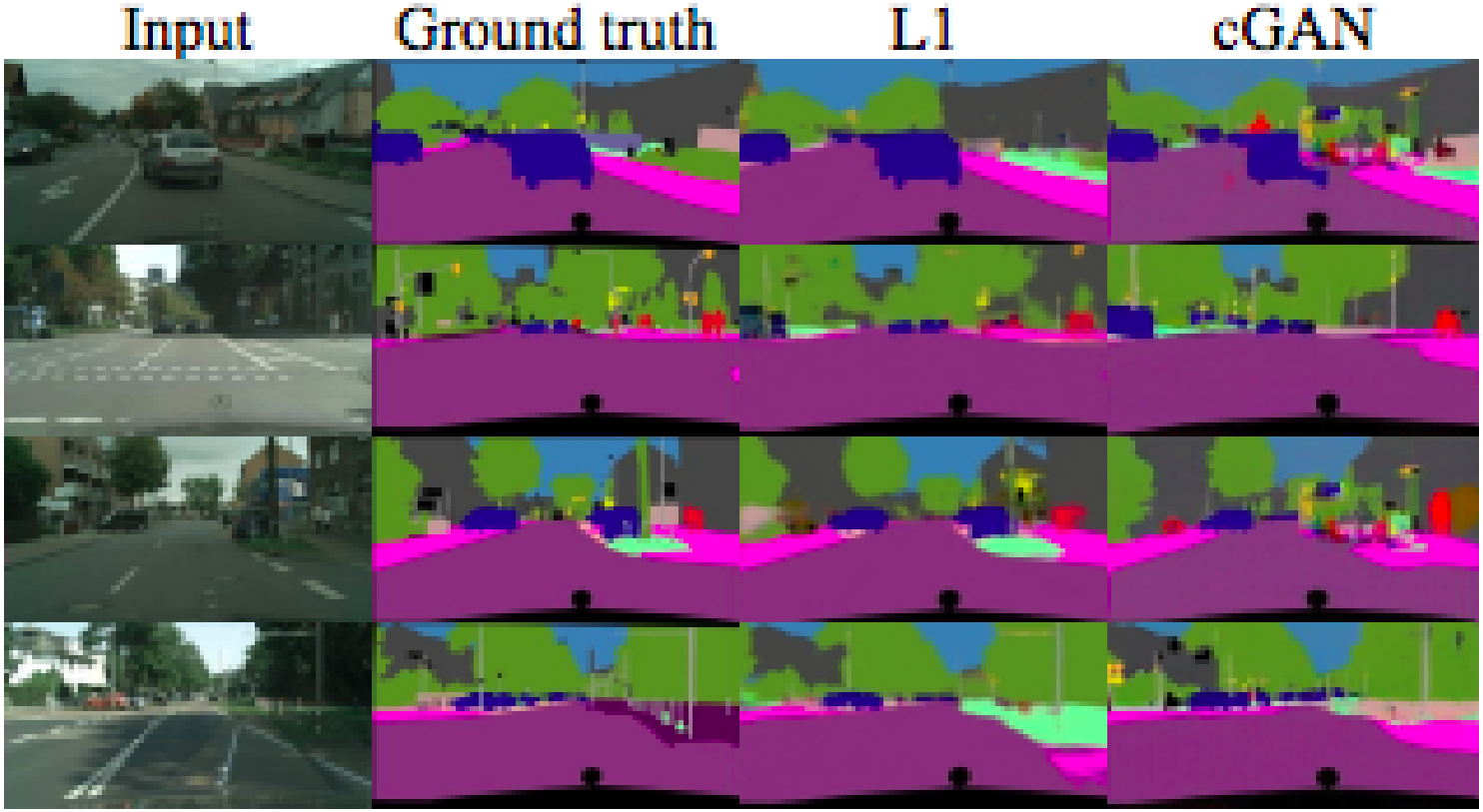
$$\mathcal{L}_{\text{Discr}}(y, \hat{y}, x) = \max_\Psi \;\; E_{i,y' \sim \{y\} \uplus \{\hat{y}\}} \;\; \ln P_\Psi(i | y', x)$$

Image-to-Image Translation (Pix2Pix)

Isola et al., Nov. 2016

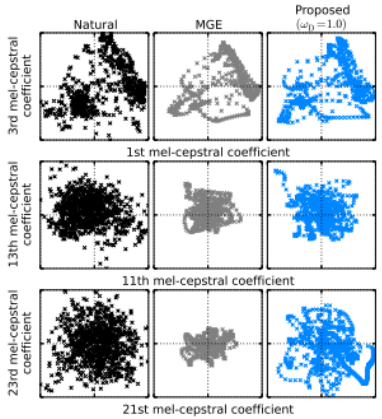
Arial Photo to Map and Back

Semantic Segmentation



Feature Alignment by Discrimination

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

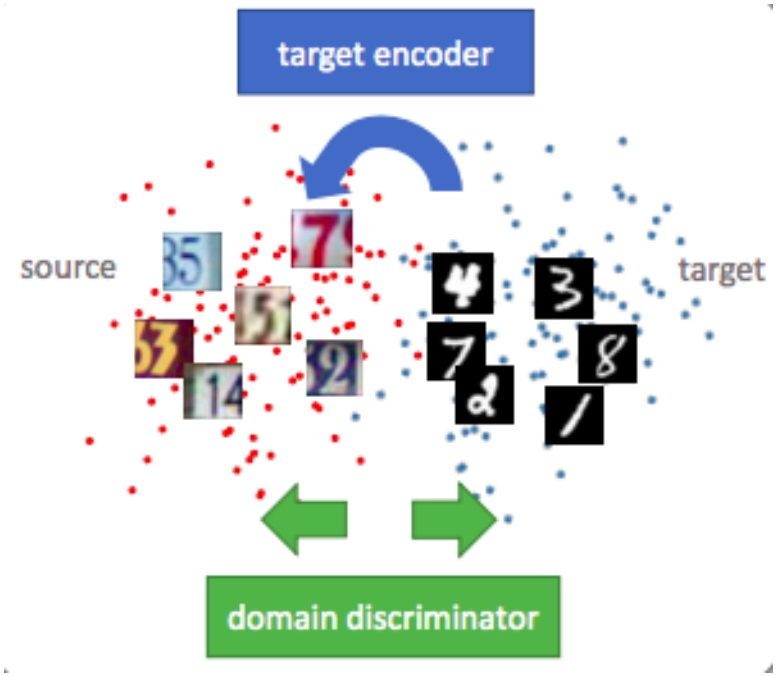


Minimum Generation Error (MGE) uses perceptual distortion — a distance between the feature vector of the generated sound wave and the feature vector of the original.

Perceptual Naturalness can be enforced by a feature discrimination loss.

Adversarial Discriminative Domain Adaptation

Tzeng et al. Feb. 2017



A feature discrimination loss can be used to align source and target features.

Comments

I predict that in a few years adversarial discrimination will be limited to enforcing perceptual naturalness in the generation of sounds and images.

Cooperative discrimination seems more useful for predictive tasks. Cooperative discrimination has been effective in pretraining. We consider pretraining in the next lecture.

END