TTIC 31230, Fundamentals of Deep Learning

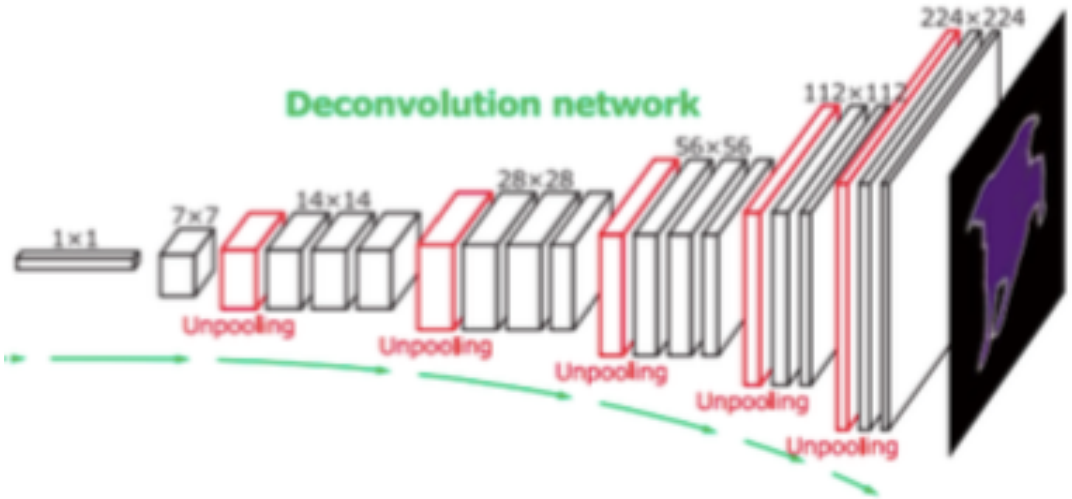
David McAllester, Winter 2019

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

Generative Adversarial Networks (GANs)

Let y range over images. We have a generator p_Φ. For $i \in \{-1, 1\}$ we define a probability distribution over pairs $\langle y, i \rangle$ by

$$\begin{aligned} \tilde{p}_\Phi(i = 1) &= 1/2 \\ \tilde{p}_\Phi(y|i = 1) &= \mathrm{pop}(y) \\ \tilde{p}_\Phi(y|i = -1) &= p_\Phi(y) \end{aligned}$$

We also have a discriminator $P_\Psi(i|y)$.

$$\Phi^* = \operatorname*{argmax}_\Phi \min_\Psi E_{\langle y, i \rangle \sim \tilde{p}_\Phi} - \ln P_\Psi(i|y)$$

Assuming universality of both p_Φ and P_Ψ we have $p_{\Phi^*} = \mathrm{pop}$.

The Discrimator Tends to Win

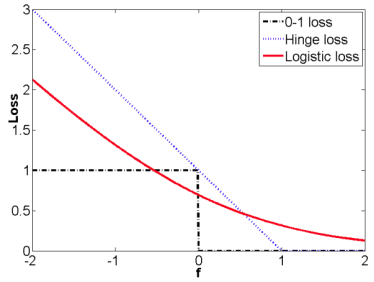
The log loss for the binary discrimination classifier is quickly driven to very near zero.

This causes the learning gradient to also become essentially zero and the learning stops.

Review of Binary Classification

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

$$\begin{aligned}
\Psi^* &= \operatorname*{argmin}_{\Psi} \; E_{(i,y)\sim\tilde{p}_\Phi} \; -\ln P_\Psi(i|y) \\
&= \operatorname*{argmin}_{\Psi} E_{(i,y)\sim\tilde{p}_\Phi} \quad \ln(1+e^{-m}) \\
m &= 2i s_\Psi(i|y) \quad \text{for} \quad \begin{aligned} s_\Psi(-1|y) &= -s_\Psi(1|y) \\ P_\Psi(i|y) &= \operatorname{softmax}_i \; s_\Phi(i|y) \end{aligned}
\end{aligned}$$



Vanishing Gradients

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

$$\Psi \mathbin{+}= \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_{\Psi} m \quad \approx 0 \text{ for } m >> 1$$

For $i = -1$ and $y \sim p_{\Phi}$:

$$\Psi \mathbin{+}= \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_{\Psi} m \quad \approx 0 \text{ for } m >> 1$$

$$\Phi \mathbin{-}= \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_{\Phi} m \quad \approx 0 \text{ for } m >> 1$$

The gradients vanish when the discriminator achieves large margins.

A Heuristic Patch

Replace

$$\Phi \texttt{ -= } \eta \, \frac{e^{-m}}{1+e^{-m}} \, \nabla_\Phi \, m \quad \approx 0 \text{ for } m >> 1$$

with

$$\Phi \texttt{ -= } \eta \, \nabla_\Phi \, m$$

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

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y)\sim\tilde{p}_\Phi} \ln \tilde{p}_\Phi(i|y)$$

$$\tilde{p}_\Phi(i|y) = \frac{\tilde{p}_\Phi(i \wedge y)}{\tilde{p}_\Phi(y)}$$

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

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

Assuming Universality of Ψ Only

$$E_{(i,y)\sim p_\Phi} \ \ln \tilde{p}_\Phi(i|y)$$

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

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

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

Contrastive GANs

A GAN can be built with a "contrastive" discriminator. Rather than estimate the probability that y is from the population, the discriminator must select which of $y_1, \ldots, y_N$ is from the population.

More formally, for $N \geq 2$ let $\tilde{p}_\Phi^{(N)}$ be the distribution defined by drawing one "positive" from pop and $N-1$ IID negatives from p_Φ; then inserting the positive at a random position among the negatives; and returning $(i, y_1, \ldots, y_N)$ where i is the index of the positive.

Contrastive GANs

$$\Psi^*(\Phi) = \underset{\Psi}{\text{argmin}}\ E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} - \ln P_\Psi(i|y_1,\ldots,y_N)$$

$$\Phi^* = \underset{\Phi}{\text{argmax}}\ E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} - \ln P_{\Psi^*(\Phi)}(i|y_1,\ldots,y_N)$$

$\tilde{p}_\Phi^{(2)}(i|y_1,y_2)$ requires a choice between two y's while $\tilde{p}_\Phi(i|y)$ classifies a single y — these are different.

The discrimination gets more difficult as N gets larger.

Contrastive GANs

$$\Psi^*(\Phi) = \underset{\Psi}{\operatorname{argmin}} \; E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \; -\ln P_\Psi(i|y_1,\ldots,y_N)$$

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \; E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \; -\ln P_{\Psi^*(\Phi)}(i|y_1,\ldots,y_N)$$

Assuming universality

$$\mathcal{L}_{\text{Discr}}(\Psi^*(\Phi)) = H_\Phi(i|y_1,\ldots y_N)$$

$$p_{\Phi^*} = \text{pop} \qquad H_{\Phi^*}(i|y_1,\ldots,y_N) = \ln N$$

Noise Contrastive Estimation

Gutmann and Hyvärinen, 2010

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} - \ln P_\Psi(i|y_1,\ldots,y_N)$$

p_Φ is fixed "noise"

Assume p_Φ is both samplable and computable — we can sample from p_Φ and for any given y we can compute $p_\Phi(y)$.

Assume $P_\Psi(i|y_1,\ldots,y_N) = \text{softmax}_i\, s_\Psi(y_i)$

Assume Ψ universal

Noise Contrastive Estimation

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \; -\ln P_\Psi(i|y_1,\ldots,y_N)$$

p_Φ is fixed "noise"

Theorem: $\mathrm{pop}(y) = \mathrm{softmax}_y \quad s_{\Psi^*}(y) + \ln p_\Phi(y)$

We then have a computable score function (energy function) for the population. We do not have the partition function Z.

Noise Contrastive Estimation

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \; -\ln P_\Psi(i|y_1,\ldots,y_N)$$

p_Φ is fixed "noise"

Lemma: $P_{\Psi^*}(i|y_1,\ldots,y_N) = \mathrm{softmax}_i \;\; \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}$

Lemma Proof

$$\tilde{p}_{\Phi}^{(N)}(i \text{ and } y_1, \ldots, y_N) = \frac{1}{N} \operatorname{pop}(y_i) \prod_{j \neq i} p_{\Phi}(y_j)$$

$$= \alpha \frac{\operatorname{pop}(y_i)}{p_{\Phi}(y_i)}, \quad \alpha = \frac{1}{N} \prod_{i} p_{\Phi}(y_i)$$

$$\tilde{p}_{\Phi}^{(N)}(i \mid y_1, \ldots y_N) = \frac{\tilde{p}_{\Phi}^{(N)}(i \text{ and } y_1, \ldots, y_N)}{\sum_i \tilde{p}_{\Phi}^{(N)}(i \text{ and } y_1, \ldots, y_N)} = \frac{1}{Z} \frac{\operatorname{pop}(y_i)}{p_{\Phi}(y_i)}$$

$$= \operatorname*{softmax}_i \left(\ln \frac{\operatorname{pop}(y_i)}{p_{\Phi}(y_i)} \right)$$

Theorem Proof

$$\underset{i}{\operatorname{softmax}}\, s_{\Psi^*}(y_i) = \underset{i}{\operatorname{softmax}} \ln \frac{\operatorname{pop}(y_i)}{p_\Phi(y_i)}$$

is solved by

$$s_{\Psi^*}(y) = \ln \frac{\operatorname{pop}(y)}{p_\Phi(y)} - \ln Z$$

giving

$$\operatorname{pop}(y) = \frac{1}{Z} \exp(s_\Psi(y) + \ln p_\Phi(y))$$

Another Theorem

$$E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \; -\ln p_{\Psi*}(i|y_1,\ldots,y_N)$$

$$\geq \ln N - \frac{N-1}{N}(KL(\text{pop}, p_\Phi) + KL(p_\Phi, \text{pop}))$$

Note that the bound holds with equality for $p_\Phi = \text{pop}$.

This is analogous to the JSD expression for the optimal discriminator loss.

Proof Part A.

$$
\begin{aligned}
& E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \ \ln p_{\Psi^*}(i|y_1,\ldots,y_N) \\
= \ & E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \ \ln \left(\operatorname*{softmax}_i \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} \right) [i] \\
= \ & E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln \left(\sum_j \frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)} \right) \\
= \ & E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln \left(\frac{1}{N} \sum_j \frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)} \right) - \ln \ N
\end{aligned}
$$

Proof Part B.

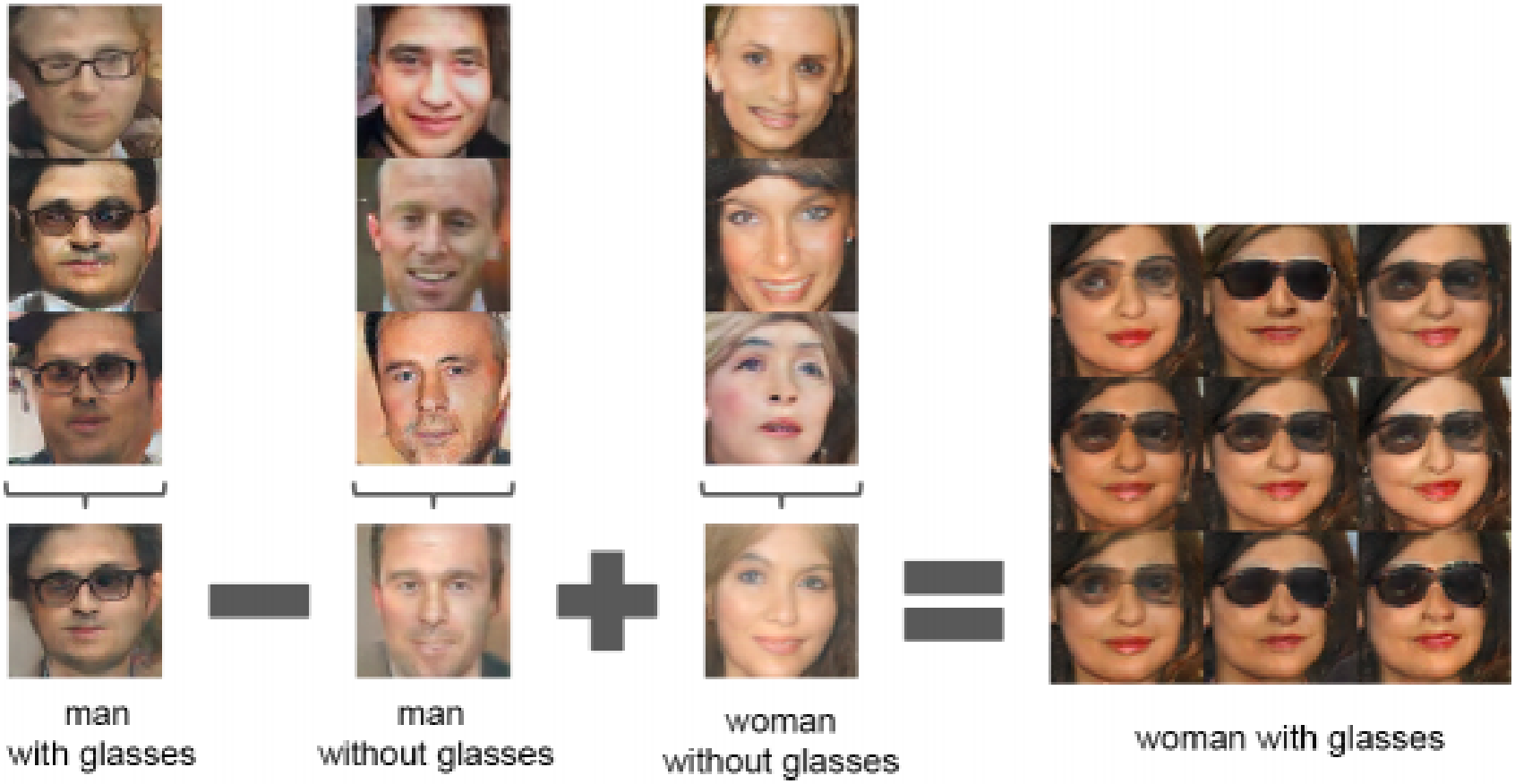
$$
\begin{aligned}
& E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln \left(\frac{1}{N} \sum_j \frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)} \right) - \ln \ N \\
\leq \ & E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \ \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \frac{1}{N} \sum_j \ln \frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)} - \ln \ N \\
= \ & E_{y\sim\mathrm{pop}} \ln \frac{\mathrm{pop}(y)}{p_\Phi(y)} - E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \frac{1}{N} \sum_j \ln \frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)} - \ln N \\
= \ & \frac{N-1}{N} (KL(\mathrm{pop}, p_\Phi) + KL(p_\Phi, \mathrm{pop})) - \ln \ N
\end{aligned}
$$

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]

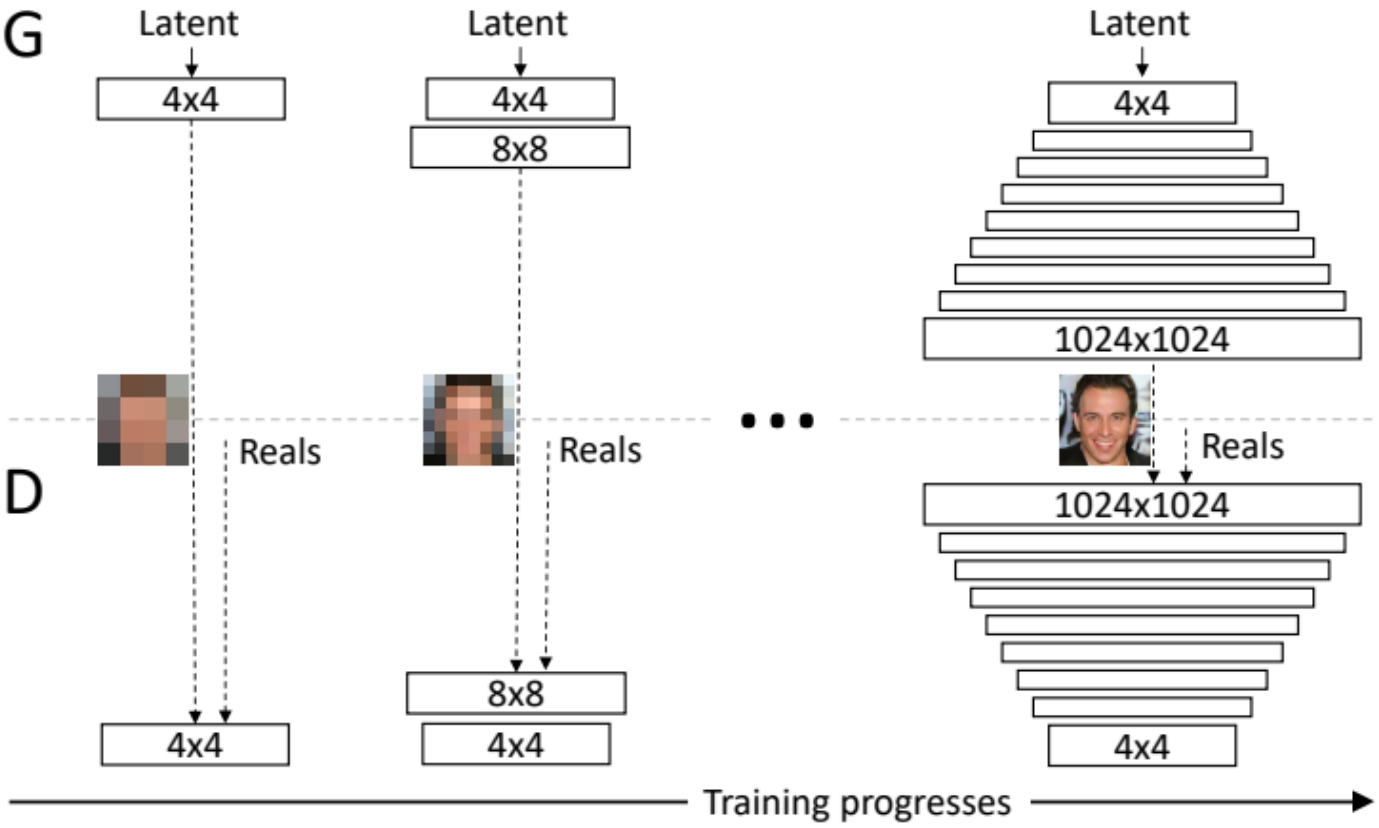
Progressive Growing of GANs

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.

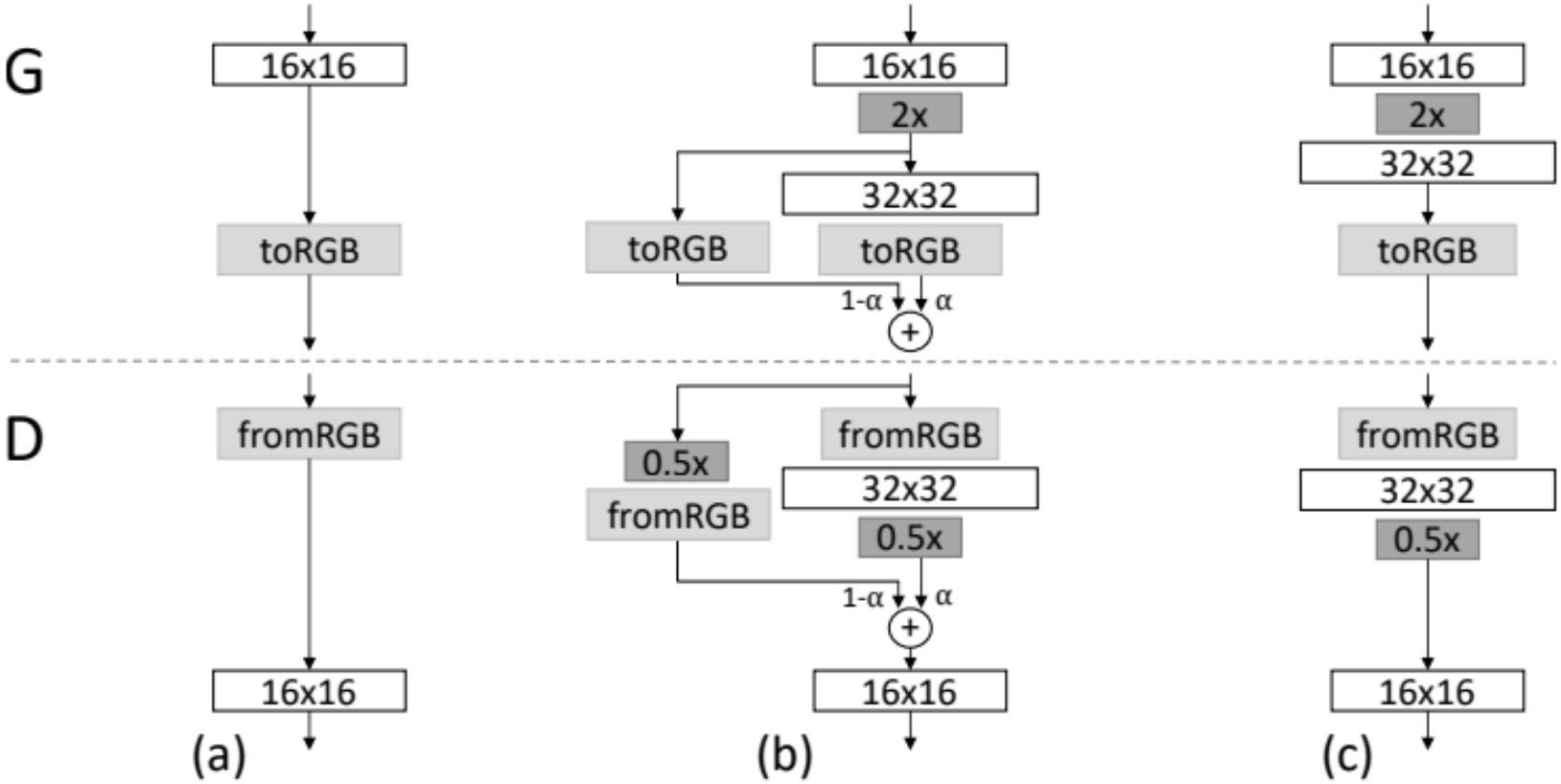
Progressive Growing of GANs

Karras et al., Oct. 2017

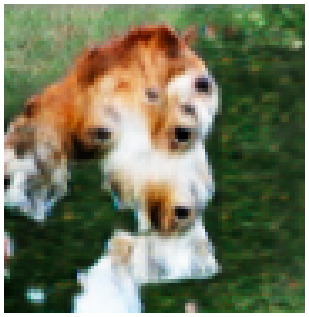
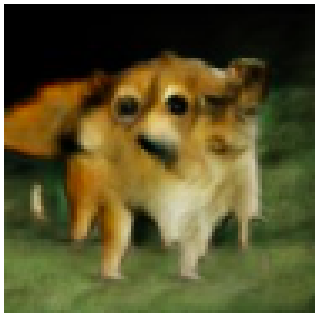
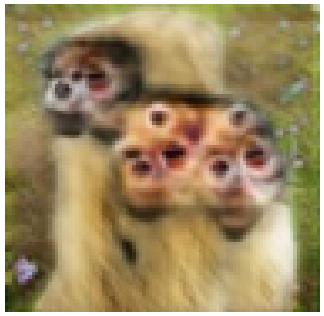
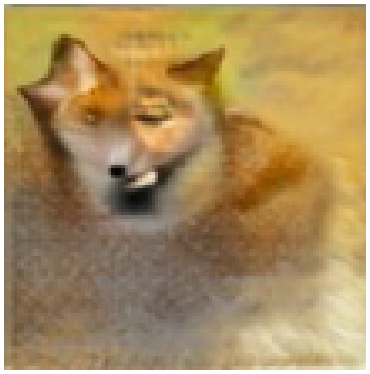


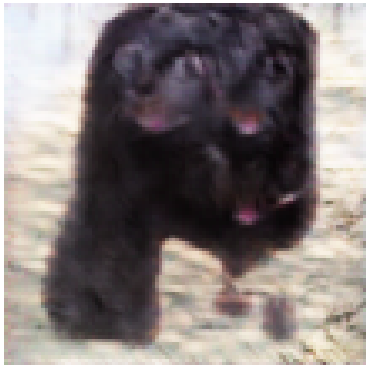
Progressive Growing of GANs

Karras et al., Oct. 2017



Early 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

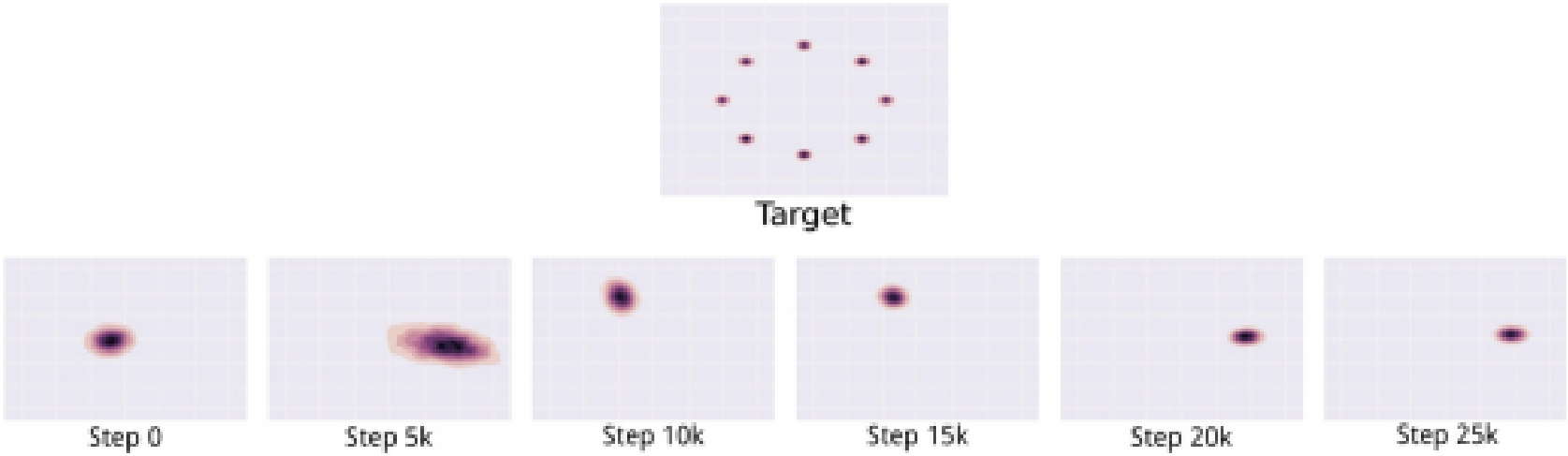
Mode Collapse

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 $x = y = 0$.

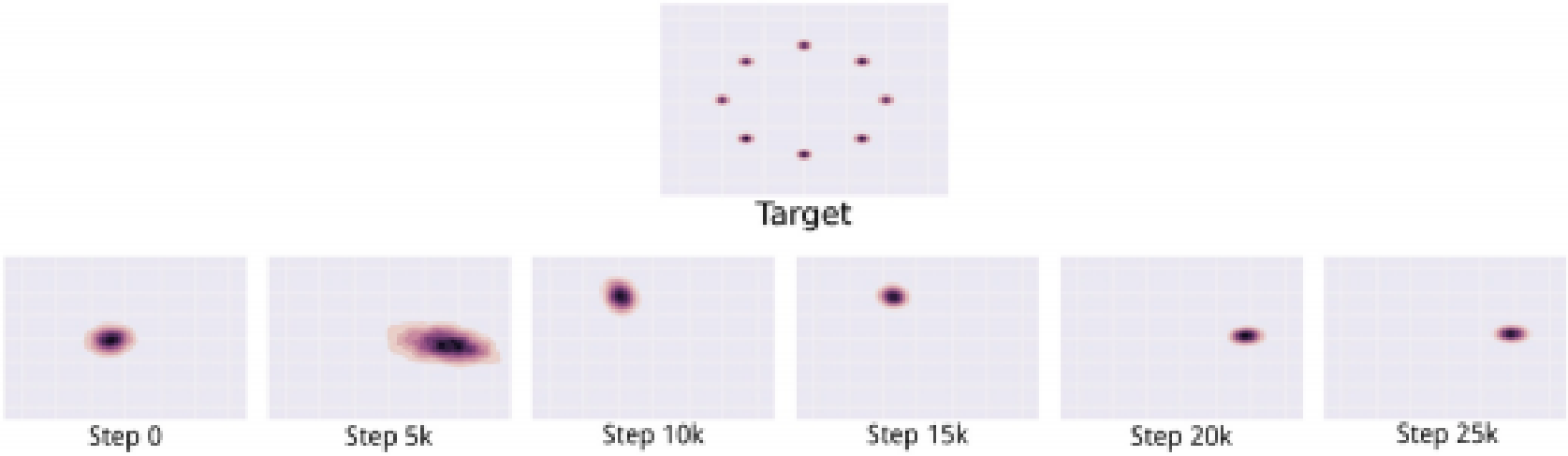
Simultaneous gradient flow yields

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

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.

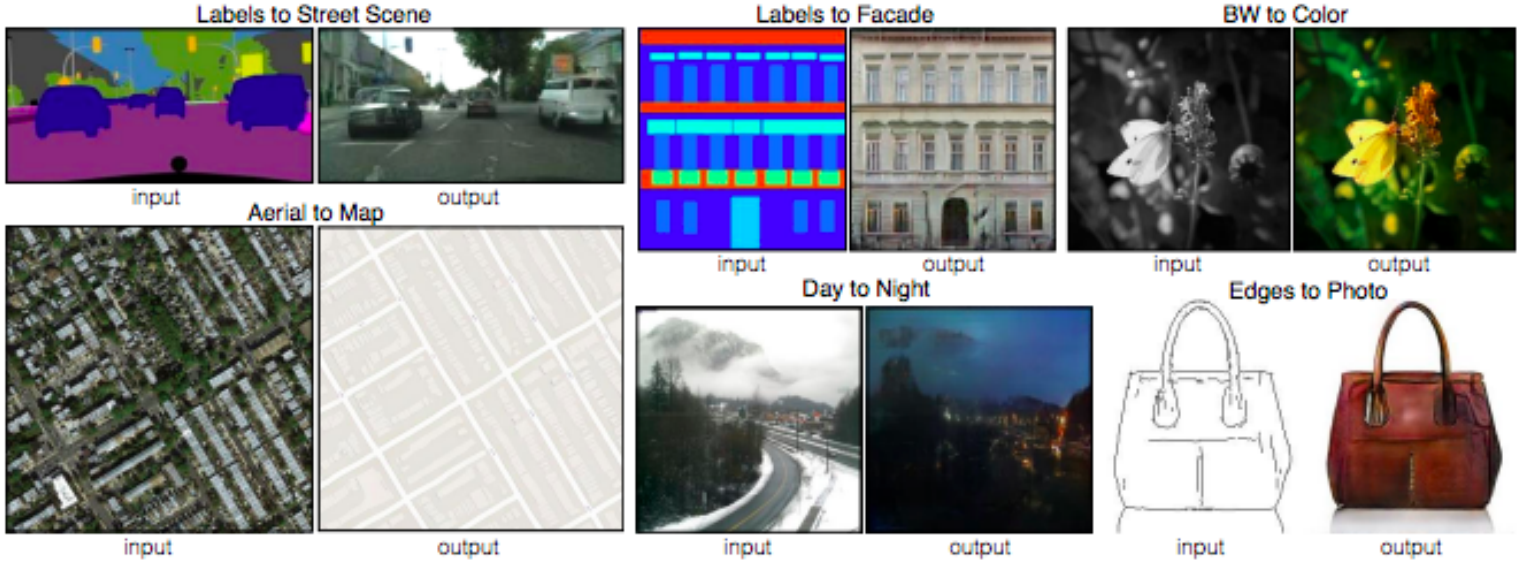
Sometimes automated metrics are also used.

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

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.



Conditional GANS

All unconditional distribution modeling methods apply to conditional distribution modeling.

Let y range over images. We have a generator p_Φ. For $i \in \{-1, 1\}$ we define a probability distribution over triple $\langle x, y, i\rangle$ by

$$\begin{aligned}\tilde{p}_\Phi(i = 1) &= 1/2\\ \tilde{p}_\Phi(y|i = 1) &= \mathrm{pop}(y|x)\\ \tilde{p}_\Phi(y|i = -1) &= p_\Phi(y|x)\end{aligned}$$

We also have a discriminator $P_\Psi(i|y)$.

$$\Phi^* = \operatorname*{argmax}_\Phi \min_\Psi E_{\langle x, y, i\rangle \sim \tilde{p}_\Phi} - \ln P_\Psi(i|x, y)$$

Conditional GANs

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \ \underset{\Psi}{\max} \ E_{x,y,i\sim\tilde{p}_\Phi(y|x)} \ \ln P_\Psi(i|x,y)$$

Adversarial Discrimination as an Additional Loss

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

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

Discrimination as an Additional Loss

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

$$\text{cGAN}: \quad \Phi^* = \text{argmin}_\Phi \; \mathcal{L}_{\text{Discr}}(\Phi)$$

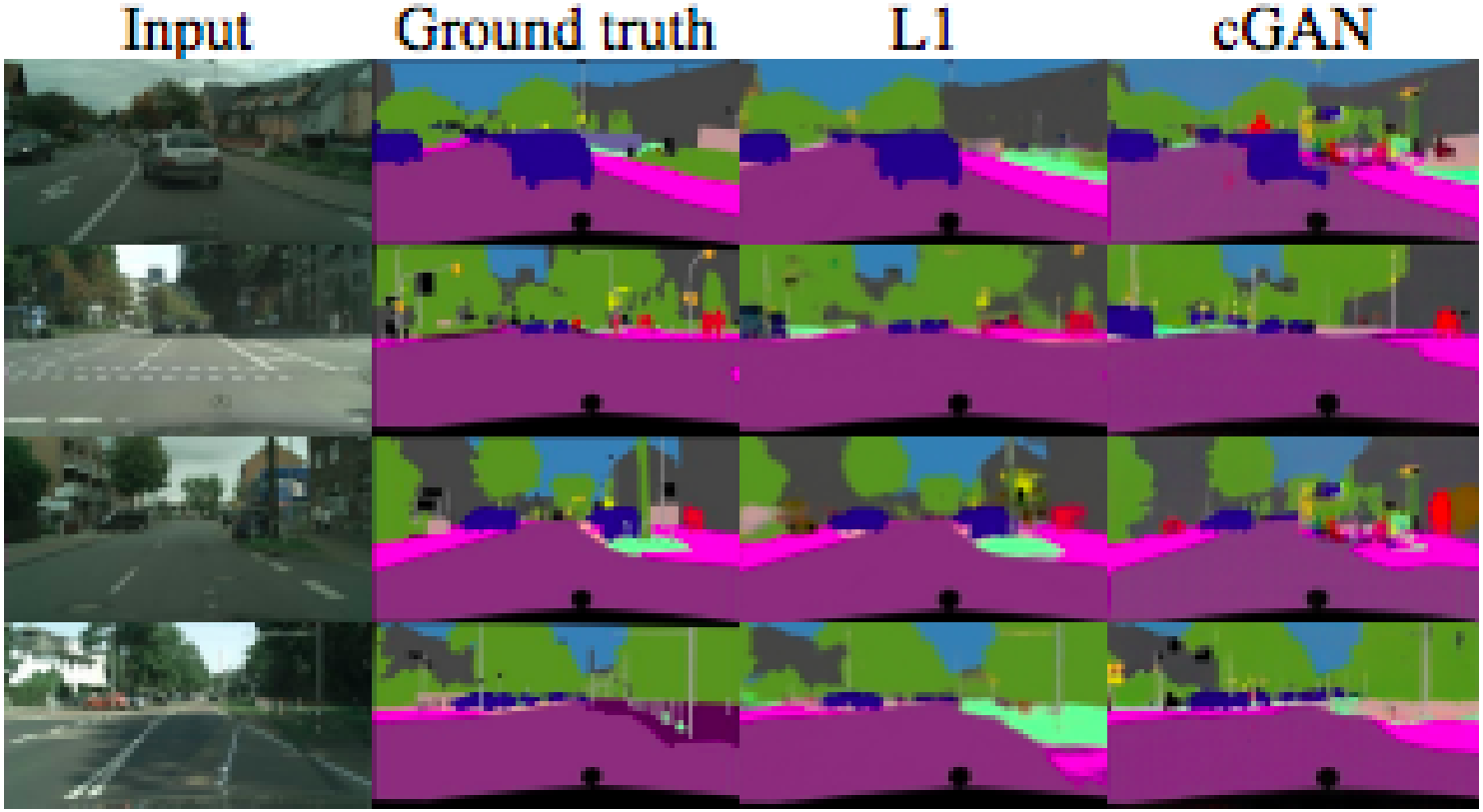
$$\text{L1} + \text{cGAN}: \quad \Phi^* = \text{argmin}_\Phi \; E_{(x,y)\sim\text{pop}} \; ||y - y_\Phi(x)||_1 \; + \; \lambda \, \mathcal{L}_{\text{Discr}}(\Phi)$$

Image-to-Image Translation (Pix2Pix)

Isola et al., Nov. 2016

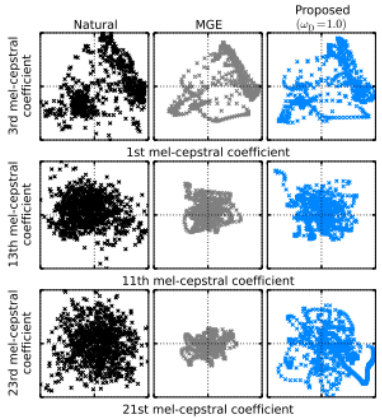
Arial Photo to Map and Back

Semantic Segmentation



Feature Alignment by Discrimination

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

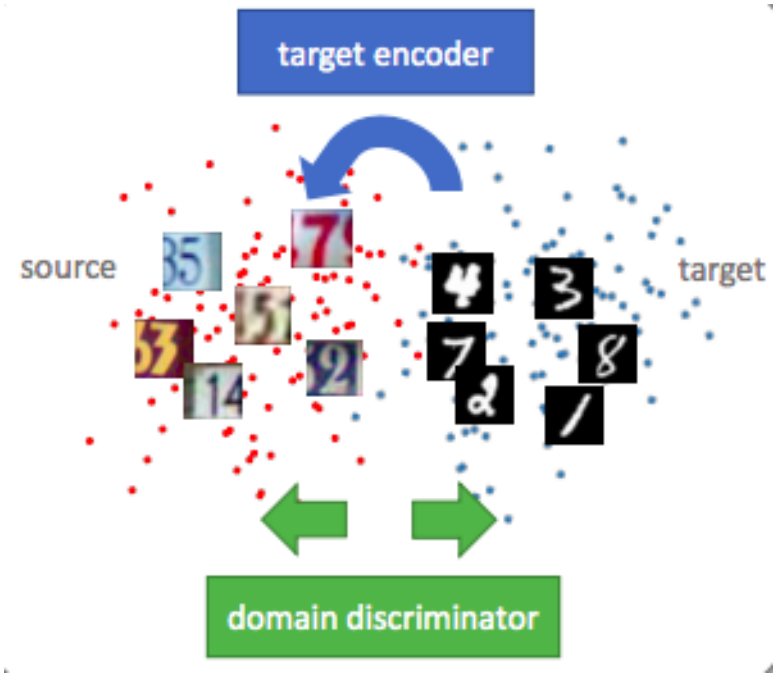


Minimum Generation Error (MGE) uses perceptual distortion — a distance between the feature vector of the generated sound wave and the feature vector of the original.

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. We will see that cooperative discrimination has been effective in pretraining.

END