f −GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Least Squares Generative Adversarial Networks ICCV 2017

Energy-Based Generative Adversarial Networks ICLR 2017

The f-divergence family

Given two distributions P and Q, absolutely continue density function p and q, $x \in \mathcal{X}$, we define the f-*divergence*

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x,$$

f is convex
$f(1) = 0$

Name	$D_f(P\|Q)$	Generator $f(u)$	$T^*(x)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$	$u \log u$	$1 + \log \frac{p(x)}{q(x)}$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x$	$-\log u$	$-\frac{q(x)}{p(x)}$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \mathrm{d}x$	$(u-1)^2$	$2(\frac{p(x)}{q(x)} - 1)$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$	$(\sqrt{u} - 1)^2$	$(\sqrt{\frac{p(x)}{q(x)}} - 1) \cdot \sqrt{\frac{q(x)}{p(x)}}$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \mathrm{d}x$	$-(u+1) \log \frac{1+u}{2} + u \log u$	$\log \frac{2p(x)}{p(x)+q(x)}$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \mathrm{d}x - \log(4)$	$u \log u - (u+1) \log(u+1)$	$\log \frac{p(x)}{p(x)+q(x)}$

Variational Estimation of f-divergences

Nguyen et al. derive a general variational method to estimate f-divergences given only samples from P and Q.

Every convex, lower-simicontinuous function f has a *convex conjugate* function f^*:

$$f^*(t) = \sup_{u \in \mathrm{dom}_f} \{ut - f(u)\}.$$

Pair (f, f^*) is dual to another $\longrightarrow$ $f^{**} = f$, as

$$f(u) = \sup_{t \in \mathrm{dom}_{f^*}} \{tu - f^*(t)\}$$

Nguyen et al. leverage the above variational representation of f-divergence to obtain a lower bound on the divergence

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x,$$

$$\begin{aligned}
D_f(P\|Q) &= \int_{\mathcal{X}} q(x) \sup_{t\in\mathrm{dom}_{f^*}} \left\{t\frac{p(x)}{q(x)} - f^*(t)\right\} \mathrm{d}x \\
&\geq \sup\nolimits_{T\in\mathcal{T}} \left(\int_{\mathcal{X}} p(x)\, T(x)\, \mathrm{d}x - \int_{\mathcal{X}} q(x)\, f^*(T(x))\, \mathrm{d}x\right) \\
&= \sup_{T\in\mathcal{T}} \left(\mathbb{E}_{x\sim P}\left[T(x)\right] - \mathbb{E}_{x\sim Q}\left[f^*(T(x))\right]\right),
\end{aligned}$$

Variational Divergence Minimization(VDM)

Use the variational lower bound on the f-divergence to estimate a generative model Q given a true distribution P.

Generative model Q, parameters Q_θ. T is variational function, parameters T_ω.

f-GAN objective function, where minimize with respect to θ and maximize with respect to ω

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}\left[T_\omega(x)\right] - \mathbb{E}_{x \sim Q_\theta}\left[f^*(T_\omega(x))\right].$$

Representation for the Variational Function

To apply the variational objective for different f-divergence, we assume that $T_\omega(x) = g_f\left(V_\omega(x)\right)$ and rewrite the objective function:

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}\left[g_f(V_\omega(x))\right] + \mathbb{E}_{x \sim Q_\theta}\left[-f^*(g_f(V_\omega(x)))\right],$$

where $V_\omega \colon \mathcal{X} \to \mathbb{R}$ and $g_f \colon \mathbb{R} \to \mathrm{dom}_{f^*}$ is a *output activation function*

Name	Output activation g_f	dom_{f^*}	Conjugate $f^*(t)$	$f'(1)$
Kullback-Leibler (KL)	v	$\mathbb{R}$	$\exp(t-1)$	1
Reverse KL	$-\exp(-v)$	$\mathbb{R}_-$	$-1-\log(-t)$	-1
Pearson χ^2	v	$\mathbb{R}$	$\frac{1}{4}t^2 + t$	0
Squared Hellinger	$1-\exp(-v)$	$t < 1$	$\frac{t}{1-t}$	0
Jensen-Shannon	$\log(2) - \log(1+\exp(-v))$	$t < \log(2)$	$-\log(2-\exp(t))$	0
GAN	$-\log(1+\exp(-v))$	$\mathbb{R}_-$	$-\log(1-\exp(t))$	$-\log(2)$

GAN objective is a special case:

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}\left[\log D_\omega(x)\right] + \mathbb{E}_{x \sim Q_\theta}\left[\log(1 - D_\omega(x))\right],$$

where discriminator is the sigmoid $D_\omega(x) = 1/\left(1 + e^{-V_\omega(x)}\right)$

Least Squares Generative Adversarial Networks

- Propose LSGANs which adopt least squares loss function for the discriminator

- Show that minimizing objective function of LSGAN yield minimizing the Pearson χ^2 divergence

- Apply conditional LSGANs to the Chinese character generation.

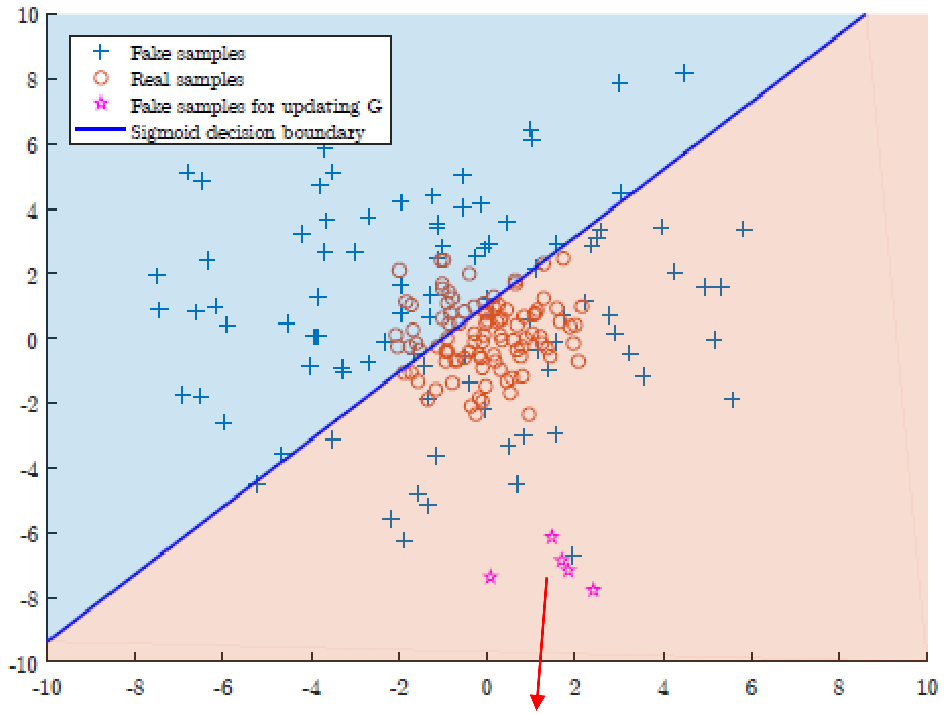
Motivation

Use the such fake samples to update generator

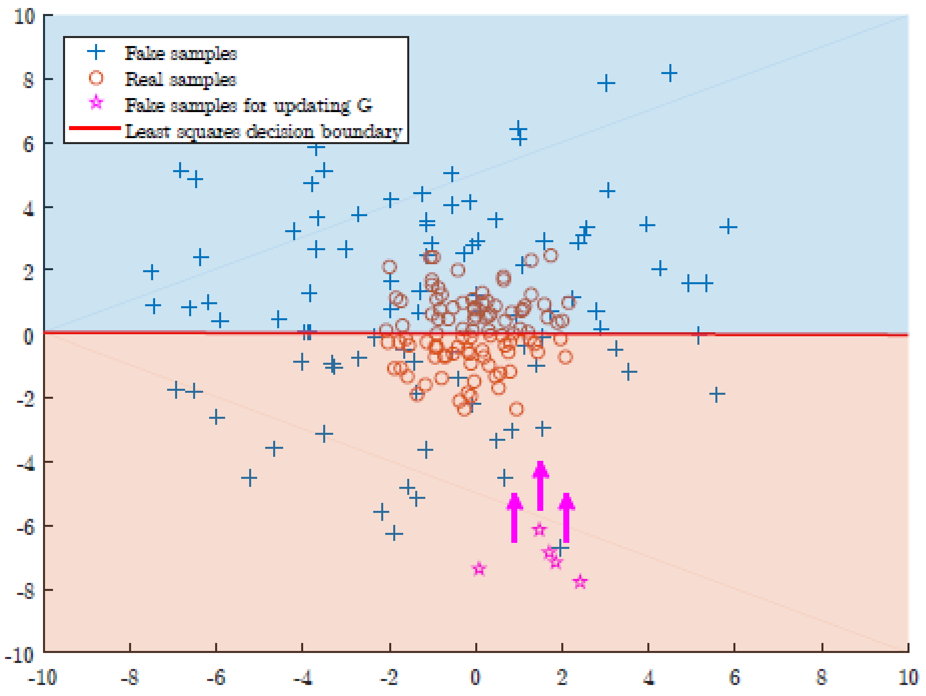
↓

Problem of vanishing gradient

idea: adopt the least squares loss function for the discriminator



on the correct side of decision boundary, but are still far from the real data



The least square function penalize samples that lie in a long way on the correct side of decision boundary

↓

Move the fake samples toward decision boundary

Objective Function

The standard minimax objective for GANs:

$$\min_G \max_D V_{\text{GAN}}(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Use the $a - b$ coding scheme for the discriminator, where a and b are the labels for fake data and real data, respectively.

The objective function for LSGANs :

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x}) - b)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})) - a)^2\right]$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})) - c)^2\right],$$

the value G wants D to believe for fake sample

Relation to Pearson χ^2 Divergence

Minimizing the original GAN objective yields minimizing the JS divergence

$$C(G) = KL\left(p_{\text{data}} \left\| \frac{p_{\text{data}} + p_g}{2}\right.\right) + KL\left(p_g \left\| \frac{p_{\text{data}} + p_g}{2}\right.\right) - \log(4).$$

Explore the relation between LSGANs and f-divergence.

Consider the following extension of LSGAN objective function:

$$\begin{aligned}
\min_D V_{\text{LSGAN}}(D) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x}) - b)^2\right] \\
&+ \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})) - a)^2\right] \\
\min_G V_{\text{LSGAN}}(G) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x}) - c)^2\right] \\
&+ \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})) - c)^2\right].
\end{aligned}$$

For a fix G, we have the optimal discriminator D:

$$D^*(\boldsymbol{x}) = \frac{bp_{\text{data}}(\boldsymbol{x}) + ap_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}.$$

$$\begin{aligned}
2C(G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{d}}}\left[(D^*(\boldsymbol{x}) - c)^2\right] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}\left[(D^*(G(\boldsymbol{z})) - c)^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{d}}}\left[(D^*(\boldsymbol{x}) - c)^2\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[(D^*(\boldsymbol{x}) - c)^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{d}}}\left[\left(\frac{bp_{\text{d}}(\boldsymbol{x}) + ap_g(\boldsymbol{x})}{p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} - c\right)^2\right] \\
&\quad + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\left(\frac{bp_{\text{d}}(\boldsymbol{x}) + ap_g(\boldsymbol{x})}{p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} - c\right)^2\right] \\
&= \int_{\mathcal{X}} p_{\text{d}}(\boldsymbol{x})\left(\frac{(b-c)p_{\text{d}}(\boldsymbol{x}) + (a-c)p_g(\boldsymbol{x})}{p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right)^2 \mathrm{d}x \\
&\quad + \int_{\mathcal{X}} p_g(\boldsymbol{x})\left(\frac{(b-c)p_{\text{d}}(\boldsymbol{x}) + (a-c)p_g(\boldsymbol{x})}{p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right)^2 \mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\left((b-c)p_{\text{d}}(\boldsymbol{x}) + (a-c)p_g(\boldsymbol{x})\right)^2}{p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\left((b-c)(p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x})) - (b-a)p_g(\boldsymbol{x})\right)^2}{p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \mathrm{d}x.
\end{aligned}$$

Reformulate $V_{LSGAN}(G)$ by $D^*(x)$:

Set $b - c = 1$ and $b - a = 2$, then:

$$\begin{aligned}
2C(G) &= \int_{\mathcal{X}} \frac{\left(2p_g(\boldsymbol{x}) - (p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x}))\right)^2}{p_{\text{d}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \mathrm{d}x \\
&= \chi^2_{\text{Pearson}}(p_{\text{d}} + p_g \| 2p_g),
\end{aligned}$$

Parameter Selection

Minimize the Pearson χ^2 divergence between $p_d + p_g$ and $2p_g$:

set $a = -1$, $b = 1$ and $c = 0$

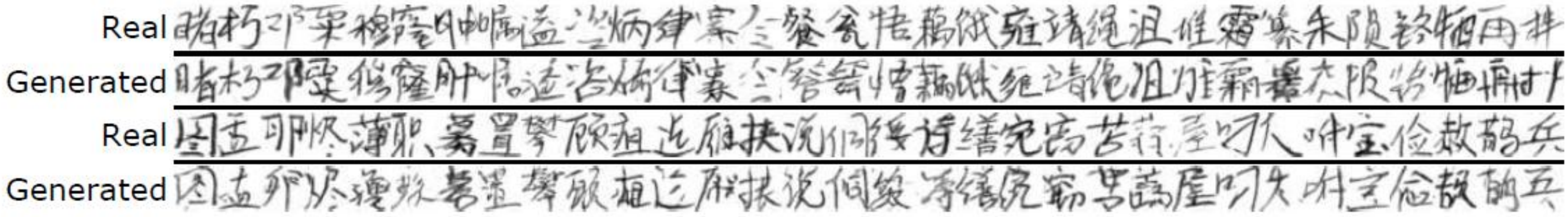
$$\begin{aligned}\min_D V_{\text{LSGAN}}(D) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x})-1)^2\right]\\ &+\frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))+1)^2\right]\\ \min_G V_{\text{LSGAN}}(G) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})))^2\right].\end{aligned}$$

Make G generate samples as real as possible by setting $c = b$

$$\begin{aligned}\min_D V_{\text{LSGAN}}(D) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x})-1)^2\right]\\ &+\frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})))^2\right]\\ \min_G V_{\text{LSGAN}}(G) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))-1)^2\right].\end{aligned}$$

Experiment

Handwritten Chinese Characters



Train a conditional LSGAN on a handwritten Chinese characters dataset which contains 3740 classes.

Energy-Based Generative Adversarial Network

- View the discriminator as an energy function with lower energies for realistic samples.

- An EBGAN framework with the discriminator using an auto-encoder(AE) architecture.

- EBGAN framework generates reasonable-looking high-resolution images from the ImageNet dataset at 256×256 pixel resolution

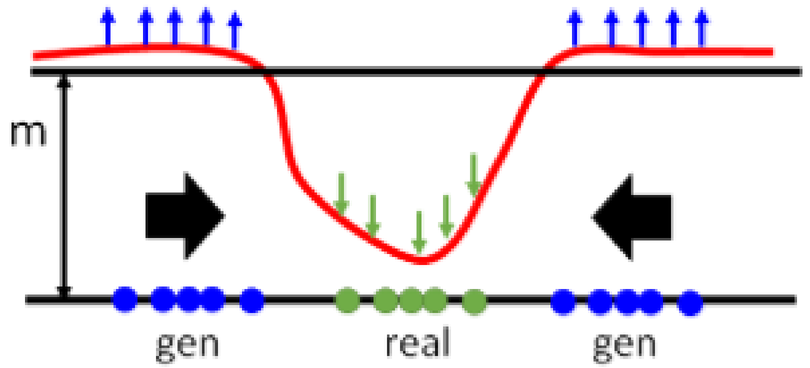
Objective Functional

Generator G, Discriminator D, random vector $z \backsim \mathcal{N}(0,1)$, the discriminator loss and generator loss:

$[\cdot]^+ = max(0,\cdot)$

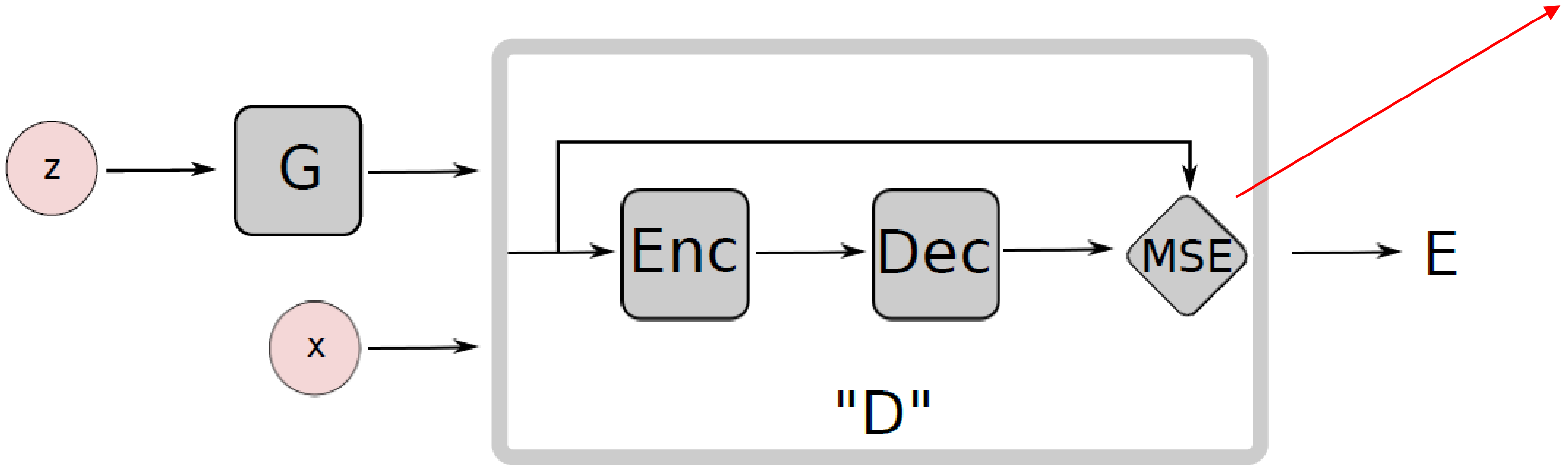
$$\mathcal{L}_D(x,z) = D(x) + [m - D(G(z))]^+$$
$$\mathcal{L}_G(z) = D(G(z))$$

positive margin



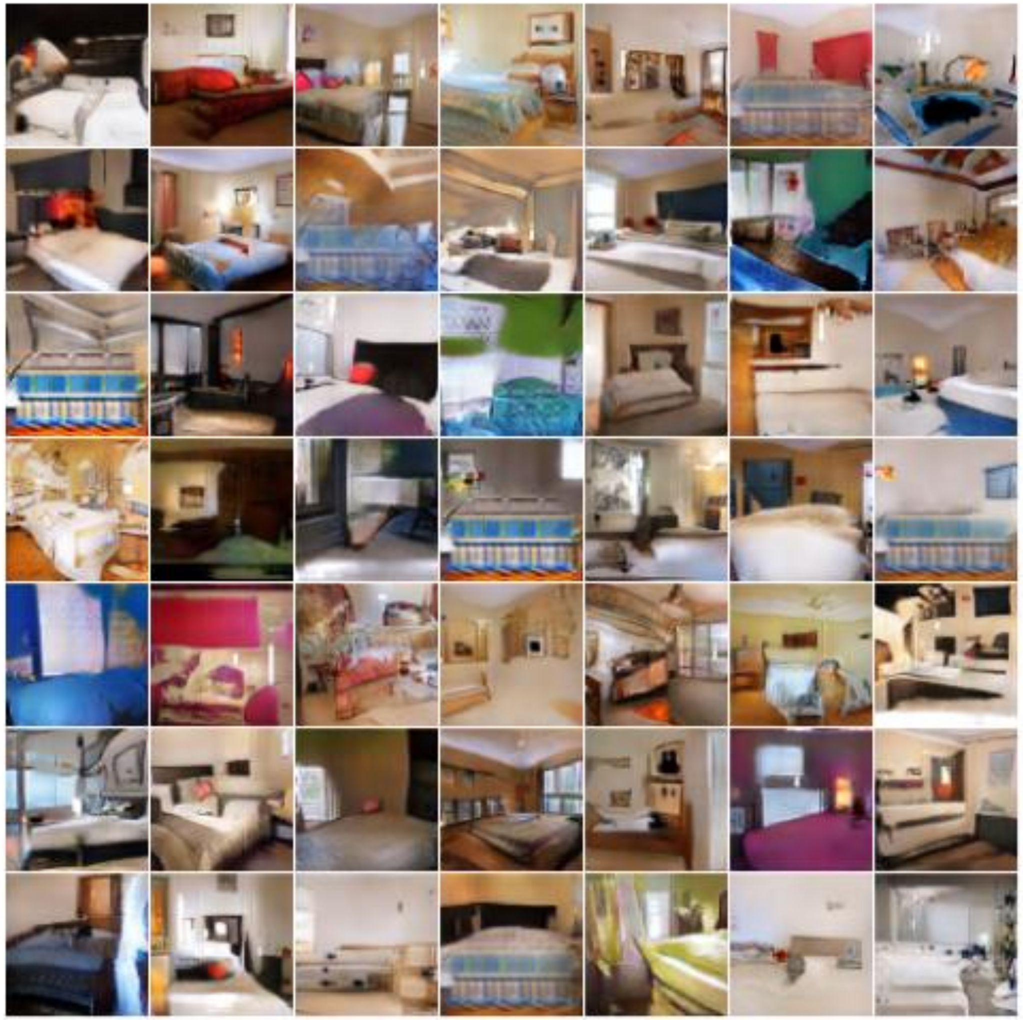
Using Auto-Encoder

$$D(x) = ||Dec(Enc(x)) - x||.$$



- Rather than using a single bit of target information to train the model, the reconstruction-based outputs offers a diverse targets.
- Auto-encoders have traditionally been used to represent energy-based model, which have the ability to learn energy manifold.

Experiment



DCGAN

EBGAN

Figure 8: ImageNet 256×256 generations using an EBGAN-PT.