## Generative: AR

$$P_\theta(X) = \prod_{t=1}^{T} P_\theta(x_t \mid x_{1:t-1})$$

$$= P_\theta(x_2 \mid x_{1:1})\, P_\theta(x_3 \mid x_{1:2})\, P_\theta(x_4 \mid x_{1:3}) \cdots P(x_T \mid x_1 \ldots x_{T-1})$$

● ○     ● ● ○     ● ● ● ○

$$\log P_\theta(X) = \sum_{t=1}^{T} \log P_\theta(x_t \mid x_{1:t-1})$$

## Generative: flow

$$X = g(z), \quad z = f(X)$$

$$P(X) = \tilde{P}(f(X)) \left| \det Dg(f(X)) \right|^{-1} = \tilde{P}(f(X)) \left| \det Df(X) \right|$$

$$Dg(z) = \frac{\partial g}{\partial z}$$

$$\boxed{\begin{aligned} \det Df(X) &= \prod_{i=1}^{N} \det Df_i(x_i) \\ & \hookrightarrow x_i = g_i \circ \cdots \circ g_1(z) \\ & \qquad = f_{i+1} \circ \cdots f_N(X) \end{aligned}}$$

$$z \to g_1 \to g_2 \to \cdots \to g_i \to x_i \leftarrow f_N \leftarrow f_{N-1} \leftarrow \cdots \leftarrow f_{i+1} \leftarrow X$$

$$z \to g_1 \to g_2 \to \cdots \to g_N \to X$$

$$z \leftarrow f_N \leftarrow f_{N-1} \leftarrow \cdots \leftarrow f_1 \leftarrow X$$

# Generative : AE

$$x \to h(\cdot) \to \text{sigm}(\cdot) \to \hat{x}$$

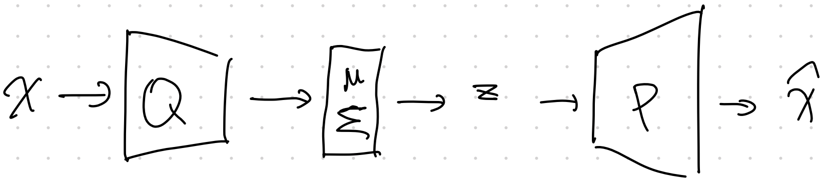
$$x \to \boxed{h(\cdot)} \to \overset{h(x)}{\boxed{\phantom{x}}} \to \boxed{\text{sign}(\cdot)} \to \hat{x}$$

$h(x) = g(b + W \cdot x)$ $\quad$ $\hat{x} = \text{sigm}(c + V \cdot h(x))$

Loss

- e.g. binary, cross entropy.

# Generative : VAE




$$\ln p_\theta(x) = \ln \int p_\theta(x|z)\, p(z)\, dz$$

$$= \ln \int \frac{q_\phi(z|x)}{q_\phi(z|x)}\, p_\theta(x|z)\, p(z)\, dz$$

reconstruction error

$$\geq -\underbrace{\left[ D_{KL}\left( q_\phi(z|x) \,||\, p(z) \right) - E_q\left( \ln p_\theta(x|z) \right) \right]}_{F(x)}$$

$F(x)$ ← free energy

$-F(x)$ ← ELBO

$\theta$ : model for inference

$\phi$ : variational approximation

<u>Reparametrization trick:</u> reparametrization of latent var

e.g.,

$$q_\phi(z) = \mathcal{N}(z|\mu, \sigma^2), \quad \phi = \{\mu, \sigma^2\}$$

# VAE loss

construct $z$.

$$-\ln P(x) + D[q(z|x) \| p(z|x)] = -E_{z\sim q_\phi}[\ln P(x|z)] + D[q(z|x) \| p(z)]$$

bits for constructing $X$ using ideal coding

penalty

as $q$ and $p$ are not necessarily the same.

i.e., $q$ is only sub-optimal

information to reconstruct $X$ from $z$ using ideal coding

extra information about $X$ if we get $X$ using $z$ from $q(z|x)$ instead of $p(z)$

Contrastive　　learn to compare

NCE : Noise Contrastive Estimation

$$\mathcal{L} = E_{x, x^+, x^-}\left[ -\log \frac{C(x, x^+)}{C(x, x^+) + C(x, x^-)} \right]$$

$x^+$ : similar to $x$

$x^-$ : dissimilar to $x$

e.g. $$\mathcal{L} = E_{x, x^+, x^-}\left[ -\log \left( \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}} \right) \right]$$

$f$ : encoder

Contrastive: Context Instance

## Predict Spatial Relation

- jigsaw
- angle of image

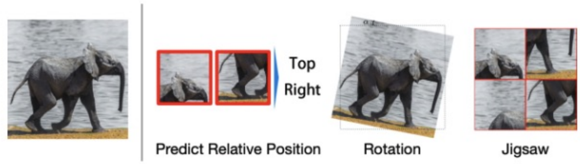



Fig. 8: Three typical methods for spatial relation contrast: predict relative position [37], rotation [43] and solve jigsaw [67], [87], [92], [141].

## Max Mutual Info

MI: $I(X;Y) = E_{P_{X,Y}} \log \frac{P_{XY}}{P_X \; P_Y}$

$I \sim D_{KL}(P_{XY}(x,y) \| P_X(x) P_Y(y))$

$I(X;Y) = H(X) - H(X|Y)$

reduction of uncertainty of X after observing Y

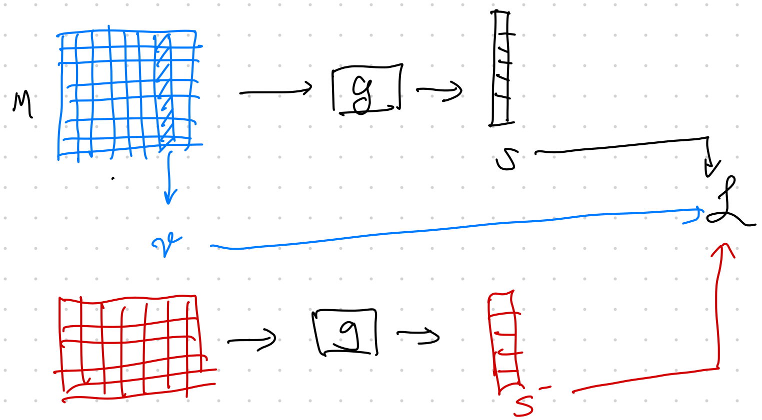
Max MI models:

$$\max_{g_1 \in G_1, g_2 \in G_2} I(g_1(x_1), g_2(x_2))$$

# Max MI : DeepInfoMax

$f(x) = R^{M \times M \times d}$ . a feature vector $v \in R^d$ from $\underbrace{f(x)}_{\text{encoding of img}}$

$\uparrow$ image




$f(x) \rightarrow \boxed{g} \rightarrow$ $s$: context of $v$

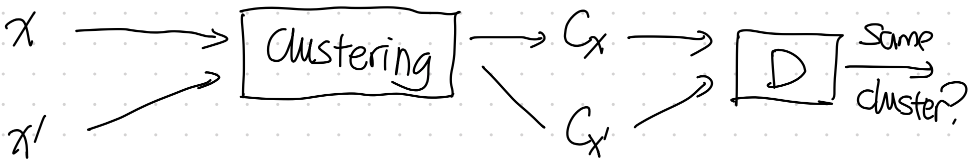
$f(x^-) \rightarrow \boxed{g} \rightarrow$ $s^-$ context of $v^-$

$$\mathcal{L} = E_{v,x}\left[-\log \frac{e^{v^T \cdot s}}{e^{v^T \cdot s} + e^{v^T s^-}}\right]$$

Contrastive : Instance Instance

## Cluster Discrimination

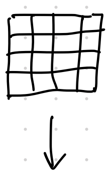
Deep Cluster




## Instance Discrimination

MoCo

Encoder $\boxed{q = f_q(x)}$



$f(\,\cdot\,) \to q$

loss :

$k_+ = f_q(x) \quad k_i = f_q(x_i)$

$$\mathcal{L} = -\log \frac{\exp(q \cdot \frac{k_+}{\tau})}{\sum_{i=0}^{K} \exp(q \cdot \frac{k_i}{\tau})}$$

all negative samples

two encoders
- query : $\to q$
- key : $\to k$ , queue of data

# SimCLR

$$\text{mini-batch } (N)\ \Big\downarrow \begin{array}{l}\text{augment data}\\ \text{to } 2N\end{array}$$

$$N \begin{pmatrix} \vdots & \vdots \\ \hat{x}_i & \hat{x}_{i'} \\ \vdots & \vdots \end{pmatrix} \begin{array}{l} \}\ \text{negative} \\ \leftarrow \text{positive} \\ \}\ \text{negative} \end{array}$$

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} \left[ l_{2i-1,2i} \, , \, l_{2i,2i-1} \right]$$

$$l_{i,i'} = -\log \frac{\exp\left(\text{sim}(\hat{x}_i, \hat{x}_{i'})/\tau\right)}{\sum_{\substack{k=1 \\ \{k \neq i\}}}^{2N} \exp\left(\text{sim}(\hat{x}_i, \hat{x}_k)/\tau\right)}$$

# Adversarial: Complete Input

## GAN

task: $\xi, Y \to X$

noise class feature

minmax game:

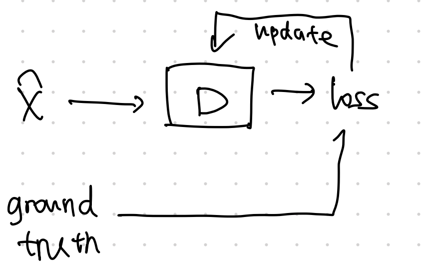
two players $G, D$: $\min_G \max_D V(D, G)$ { worst case $\hat{G}$ min $V$ | then find $\hat{D}$ that max $V$ }

G: fool D  D: min discrimination error

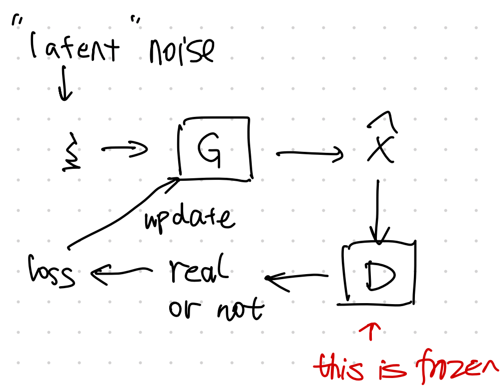
$$\min_G \max_D E_{x \sim P_{data}}[\log D(x)] + E_{z \sim P_z}[\log(1 - D(G(z)))]$$

## Alternating training

### Train discrimination




### Train Generation



# BiGAN & ALI

$z \longrightarrow \boxed{G} \longrightarrow x'$

$x \longrightarrow \boxed{E} \longrightarrow z'$

$$\left.\begin{matrix} (z,\ G(z)) \\ (E(x),\ x) \end{matrix}\right\} \longrightarrow \boxed{D} \xrightarrow[\text{from the real data}]{\text{which group is}} \text{loss}$$

$\Rightarrow \quad E \to G^{-1}$