Learning with Mutual Information

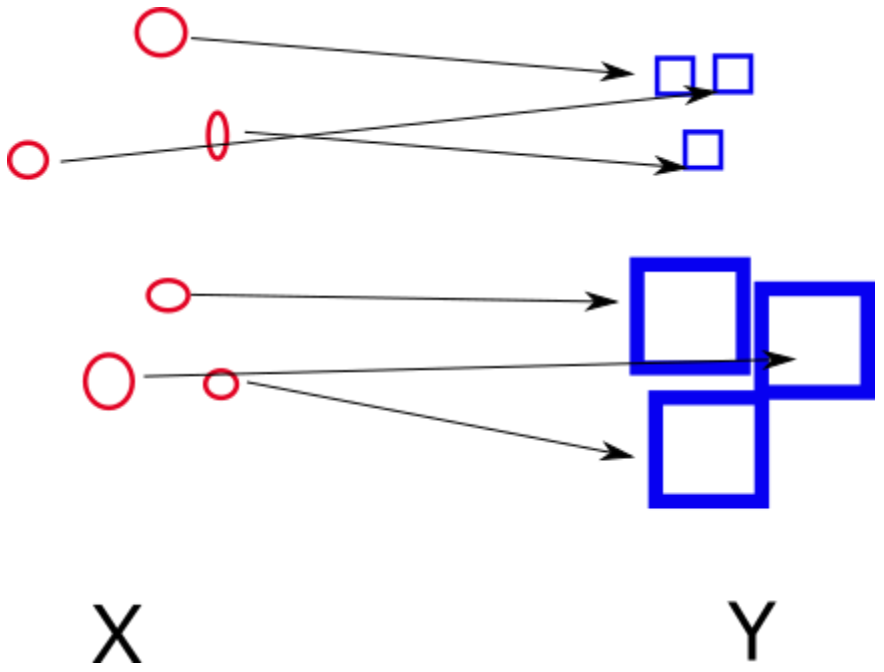
Dong Wang

2021/08/16

Content

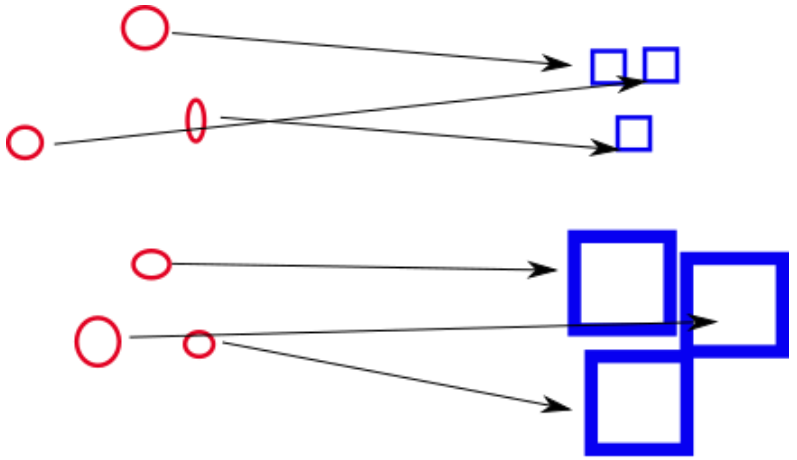
- Learning with Information
- MI estimation

How two variables are related?



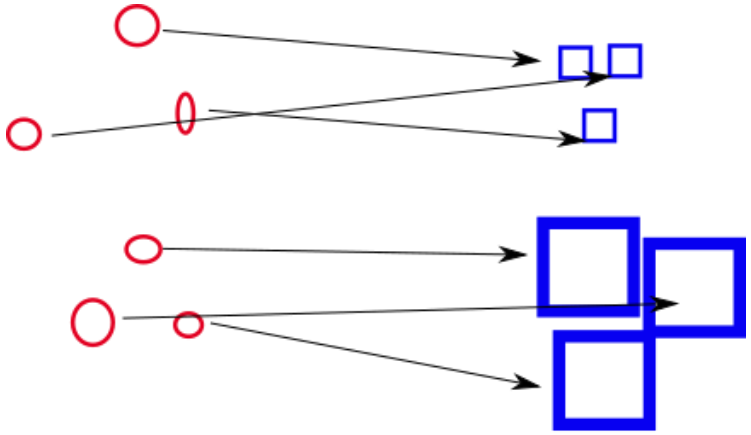
- Correlation (CCA)?
- Predictability?

Mutual Information



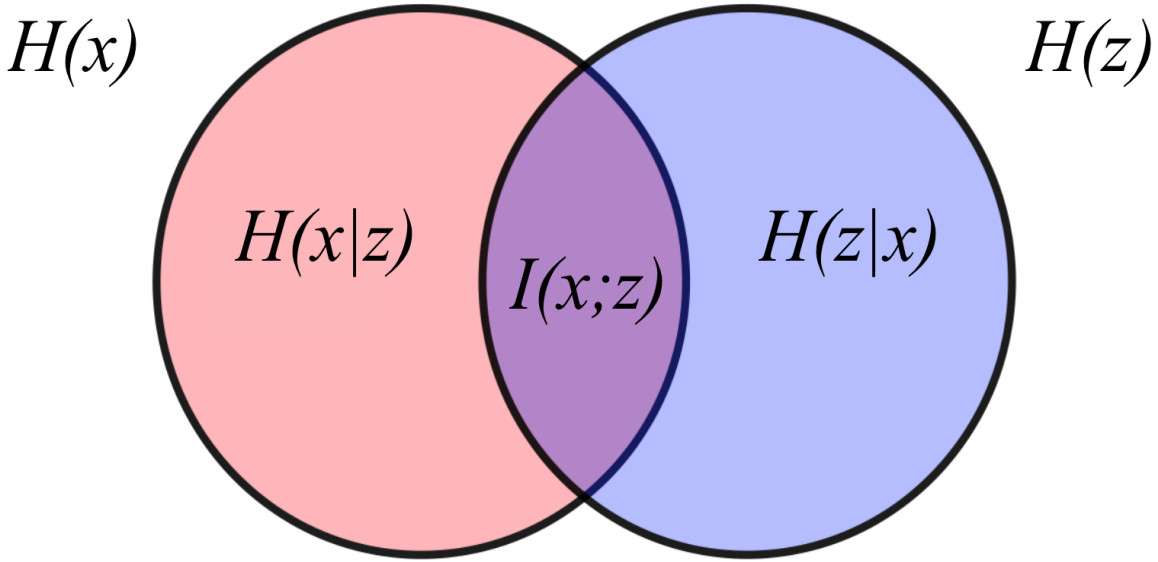
$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x|y)}{p(x)}\right] = \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)}{p(y)}\right]$$

Mutual Information



$$MI(x,y) = H(x) - H(x|y) = H(y) - H(y|x)$$

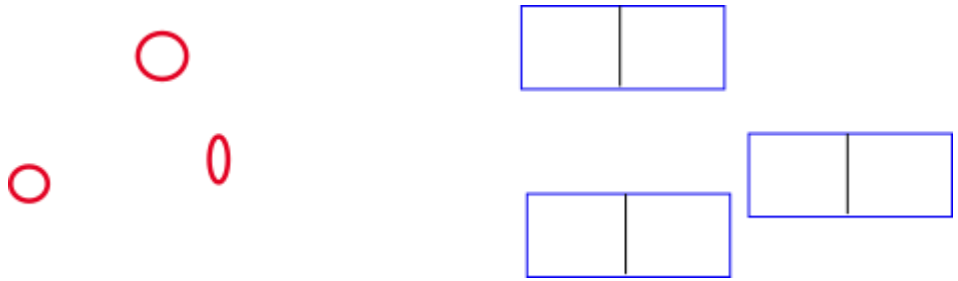
Graphical representation



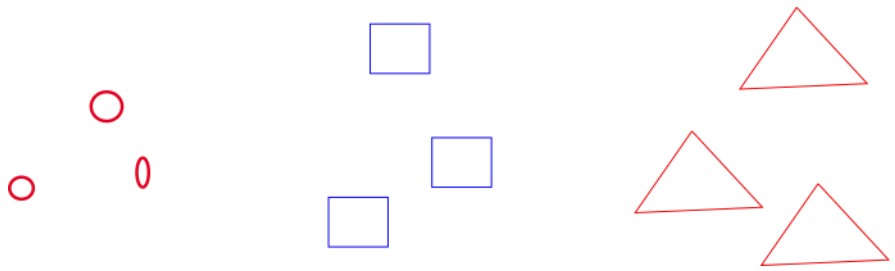
How if we know MI

- Make the representation more representative

- Make the representation more disentangled



- Make the representation more concise



How to estimate MI?

- Counting: count N(x,y) and N(x) and N(y)
- Kernel based: computing simility based on kernel
- likelihood: computing p(y|x) and p(y), by function approximation

Tsai et al., Neural Methods for Point-wise Dependency Estimation, NIPS 2020.

Estimate by maximize lower bound

- Design a function q(x|y), and then construct a lower bound of MI.

$$
\begin{aligned}
I(X;Y) &= \mathbb{E}_{p(x,y)}\left[\log \frac{q(x|y)}{p(x)}\right] \\
&\quad + \mathbb{E}_{p(y)}\left[KL(p(x|y)||q(x|y))\right] \\
&\geq \mathbb{E}_{p(x,y)}\left[\log q(x|y)\right] + h(X) \triangleq I_{\mathrm{BA}}
\end{aligned}
$$

Poole et al., On Variational Bounds of Mutual Information, ICML 2019.

Example in infoGAN

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c))$$

$$\begin{aligned} I(c; G(z,c)) &= H(c) - H(c|G(z,c)) \\ &= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\ &= \mathbb{E}_{x\sim G(z,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot|x) \parallel Q(\cdot|x))}_{\geq 0} + \mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\ &\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \end{aligned}$$

- Chen et al., InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, NIPS 2016.

Example in infoGAN

(a) Rotation

(b) Width

- Chen et al., InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, NIPS 2016.

Other bounds

$$\mathbb{E}_{p(x,y)}\left[f(x,y)\right] - \mathbb{E}_{p(y)}\left[\log Z(y)\right] \triangleq I_{\text{UBA}}$$

$$I_{\text{UBA}} \geq \mathbb{E}_{p(x,y)}\left[f(x,y)\right] - \log \mathbb{E}_{p(y)}\left[Z(y)\right] \triangleq I_{\text{DV}}$$

$$\mathbb{E}_{p(x,y)}\left[f(x,y)\right] - e^{-1}\mathbb{E}_{p(y)}\left[Z(y)\right] \triangleq I_{\text{NWJ}}$$

$$I(X;Y) \geq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K}e^{f(x_i,y_j)}}\right] \triangleq I_{\text{NCE}}$$

Poole et al., On Variational Bounds of Mutual Information, ICML 2019.

Point–wise estimation

- Having point–wise MI f(x,y)=log p(x,y)/p(x)p(y) or relevance r(x,y)=p(x,y)/p(x)p(y), it is possible to estimate the entire MI.

$$I(X;Y) = D_{\mathrm{KL}}(P_{X,Y} \parallel P_X P_Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \,\mathrm{d}x\mathrm{d}y = \mathbb{E}_{P_{X,Y}}[\log r(x,y)]$$
$$\approx \mathbb{E}_{P_{X,Y}}[\log \hat{r}_\theta(x,y)] \approx \mathbb{E}_{P_{X,Y}}[\hat{f}_\theta(x,y)],$$

Tsai et al., Neural Methods for Point-wise Dependency Estimation, NIPS 2020.

Ponit–wise estimation: Variational Bounds

- Use JS bound, once optimized, obtain the optimal PMI function

$$I_{\mathrm{JS}} := \sup_{\theta \in \Theta} \mathbb{E}_{P_{X,Y}} \Big[-\mathrm{softplus}\Big(-\hat{f}_\theta(x,y)\Big)\Big] - \mathbb{E}_{P_X P_Y} \Big[\mathrm{softplus}\Big(\hat{f}_\theta(x,y)\Big)\Big]$$

$$\hat{f}^*_\theta(x,y) = \log\left(p(x,y)/p(x)p(y)\right)$$

Tsai et al., Neural Methods for Point-wise Dependency Estimation, NIPS 2020.

Point–wise estimation: Density matching

- Matching p(x,y) in terms of KL distance.
- Once optimized, obtain the PMI function

$$\inf_{\theta\in\Theta} D_{\mathrm{KL}}(P_{X,Y} \,\|\, \hat{P}_{\theta X,Y}) := \inf_{\theta\in\Theta} I(X;Y) - \mathbb{E}_{P_{X,Y}}\left[\hat{f}_\theta(x,y)\right] \Leftrightarrow \sup_{\theta\in\Theta} \mathbb{E}_{P_{X,Y}}\left[\hat{f}_\theta(x,y)\right]$$

$$\hat{p}_\theta(x,y) := e^{\hat{f}_\theta(x,y)} p(x)p(y)$$

$$\max_{\theta\in\Theta} \mathbb{E}_{P_{X,Y}}[\hat{f}_\theta(x,y)], \quad \text{subject to } \mathbb{E}_{P_X P_Y}[e^{\hat{f}_\theta(x,y)}] = 1$$

$$\max_{\theta\in\Theta} \mathbb{E}_{P_{X,Y}}[\hat{f}_\theta(x,y)] - \lambda \cdot \left(\mathbb{E}_{P_X P_Y}[e^{\hat{f}_\theta(x,y)}] - 1\right)$$

Tsai et al., Neural Methods for Point-wise Dependency Estimation, NIPS 2020.

Point–wise estimation: Probabilistic Classifier Method

- Cast the problem of relevance computing to a discrimination task, to classify if (x,y) are from the true joint distribution.
- C denotes the class: C=1 means from true distribution, C=0 means from noise.

$$r(x,y) = \frac{p(x,y)}{p(x)p(y)} = \frac{p(x,y \mid C=1)}{p(x,y \mid C=0)} = \frac{p(C=0)}{p(C=1)} \frac{p(C=1 \mid x,y)}{p(C=0 \mid x,y)}$$

$$\hat{r}_\theta(x,y) = \frac{n_{P_X P_Y}}{n_{P_{X,Y}}} \frac{\hat{p}_\theta(C=1 \mid x,y)}{\hat{p}_\theta(C=0 \mid x,y)}, \quad \text{with } (x,y) \sim P_{X,Y} \text{ or } (x,y) \sim P_X P_Y$$

Tsai et al., Neural Methods for Point-wise Dependency Estimation, NIPS 2020.

Point–wise estimation: Density–Ratio Fitting Method

- Estimate the relevance directly

$$\inf_{\theta\in\Theta} \mathbb{E}_{P_X P_Y}\left[\left(r(x,y) - \hat{r}_\theta(x,y)\right)^2\right] \Leftrightarrow \sup_{\theta\in\Theta} \mathbb{E}_{P_{X,Y}}\left[\hat{r}_\theta(x,y)\right] - \frac{1}{2}\mathbb{E}_{P_X P_Y}\left[\hat{r}_\theta^2(x,y)\right]$$

Tsai et al., Neural Methods for Point-wise Dependency Estimation, NIPS 2020.

Experiment1: MI approximation

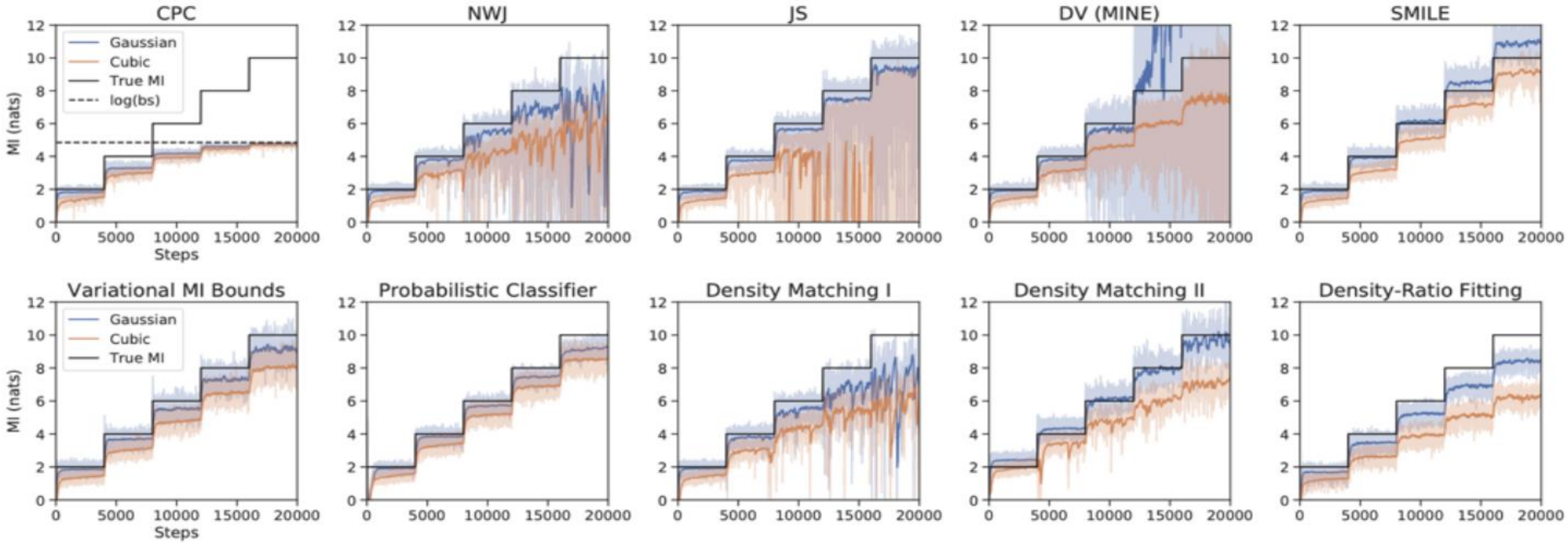


Figure 1: **Gaussian** and **Cubic** task for correlated Guassians with tractable ground truth MI. The upper row are the baselines and the lower row are our methods. Network, learning rate, optimizer, and batch size are fixed for all MI neural estimators. The only differences are the learning and inference objectives shown in Table 1.

Tsai et al., Neural Methods for Point-wise Dependency Estimation, NIPS 2020.

Experiment 2: Self supervised representation learning

Connection between Contrastive Learning and PD Our goal is to show that our learning objectives resemble contrastive learning. We first take the *Probabilistic Classifier* approach as an example and incorporate the learning of F/G, which we name it as *Probabilistic Classifier Coding* (PCC):

$$\sup_{F,G}\sup_{\theta\in\Theta}\mathbb{E}_{P_{\mathcal{V}_1,\mathcal{V}_2}}[\log\hat{p}_\theta(c=1|(F(v_1),G(v_2)))]+\mathbb{E}_{P_{\mathcal{V}_1}P_{\mathcal{V}_2}}[\log\Big(1-\hat{p}_\theta(c=1|(F(v_1),G(v_2')))\Big)], \quad (10)$$

which aims at learning F/G to better classify (i.e., differentiate) between similar or random data pairs. Next, we consider the *Density-Ratio Fitting* approach, which we refer to the objective as *Density-Ratio Fitting Coding* (D-RFC):

$$\sup_{F,G}\sup_{\theta\in\Theta}\mathbb{E}_{P_{\mathcal{V}_1,\mathcal{V}_2}}[\hat{r}_\theta(F(v_1),G(v_2))]-\frac{1}{2}\mathbb{E}_{P_{\mathcal{V}_1}P_{\mathcal{V}_2}}[\hat{r}^2_\theta(F(v_1),G(v_2'))], \quad (11)$$

which aims at learning F/G to maximize $\hat{r}_\theta(F(v_1),G(v_2))$ and minimize $\hat{r}_\theta(F(v_1),G(v_2'))$. We leave the discussion for the adaptations of Variational MI Bounds, Density Matching I ,and Density Matching II in Supplementary.

Experiment 2: Self supervised representation learning

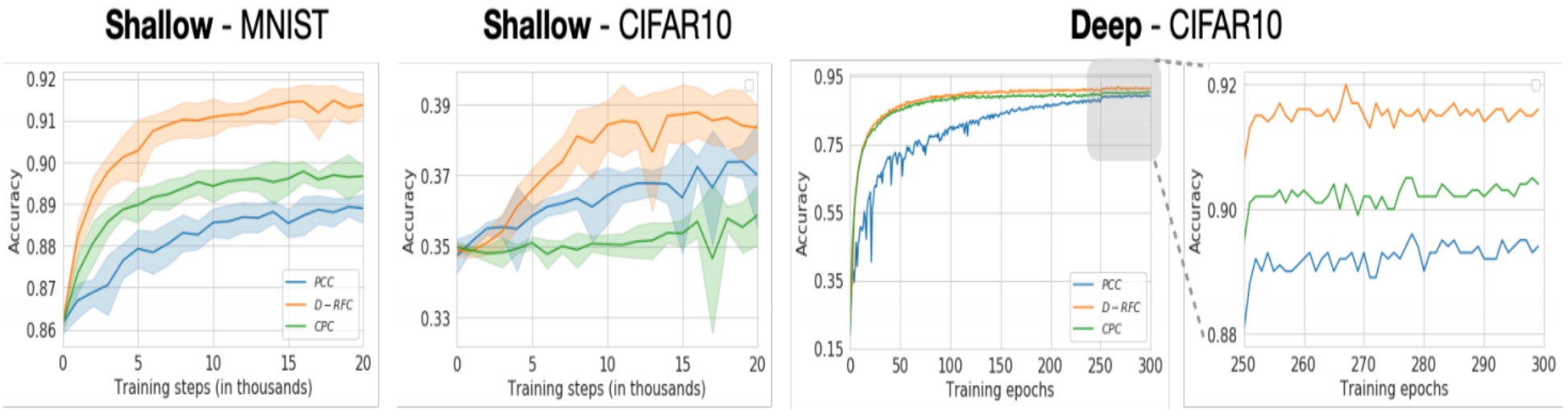


Figure 2: **Shallow** [48] and **Deep** [5] task for self-supervised visual representation learning using *downstream linear evaluation protocol*. We compare the presented Probabilistic Classifier Coding (PCC) and Density-Ratio Fitting Coding (D-RFC) with baseline Contrastive Predictive Coding (CPC). Network, learning rate, optimizer, and batch size are fixed for all the methods. The only differences are the learning objectives.

Experiment 3: Cross–modal Learning

- Match audio feature and video feature, using probabilistic classifier

Correct Audio-Textual Retrieval Examples (Top-1 Accuracy: 96.24%)					
Audio Feature	Textual Features (Ranked by logarithm of point-wise dependency)				
depths	**depths (15.22)**	mildewed (-58.62)	lugged (-92.24)	alison (-108.02)	raffleshurst (-161.74)
receptacle	**receptacle (1.32)**	bloated (-15.41)	recreate (-39.77)	sting (-90.51)	pity (-104.44)
frontiers	**frontiers (3.36)**	institution (-31.01)	laterally (-54.17)	pretends (-105.11)	vibrating (-124.88)

Incorrect Audio-Textual Retrieval Examples					
Audio Feature	Textual Features (Ranked by logarithm of point-wise dependency)				
cos	tortoise (-2.33)	**cos (-10.72)**	tickling (-12.53)	undressed (-18.11)	cromwell's (-44.31)
elbowing	itinerary (-6.51)	**elbowing (-8.22)**	swims (-12.98)	rigid (-24.14)	integrity (-39.76)
alma's	roughness (-3.11)	**alma's (-3.67)**	montreal (-11.81)	tuneful (-12.22)	levant (-18.26)

Experiment 4: Disentangled representation

- Train a Max MI system with some constraints (important!)
- Let the code sets share no information

$$\underset{p(y|x)}{\text{maximize}}\ I(X;Y)$$

$$\text{subject to}\ \ TC(Y) = \sum_{i=1}^{K} I(X;Y_i) - I(X;Y) \leq \delta$$

$$\mathbb{E}_{p(x)p(\epsilon)}\left[KL(p(y|x)\|p(y|x+\epsilon))\right] \leq \gamma$$

Poole et al., On Variational Bounds of Mutual Information, ICML 2019.

Experiment 4: Disentangled representation

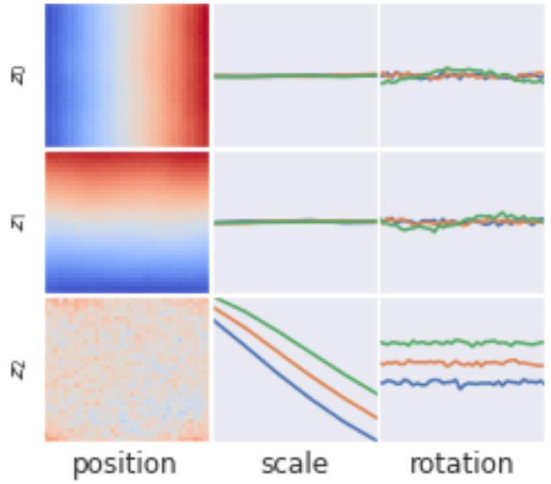


Figure 5. **Feature selectivity on dSprites.** The representation learned with our regularized InfoMax objective exhibits disentangled features for position and scale, but not rotation. Each row corresponds to a different active latent dimension. The first column depicts the position tuning of the latent variable, where the x and y axis correspond to x/y position, and the color corresponds to the average activation of the latent variable in response to an input at that position (red is high, blue is low). The scale and rotation columns show the average value of the latent on the y axis, and the value of the ground truth factor (scale or rotation) on the x axis.

Poole et al., On Variational Bounds of Mutual Information, ICML 2019.

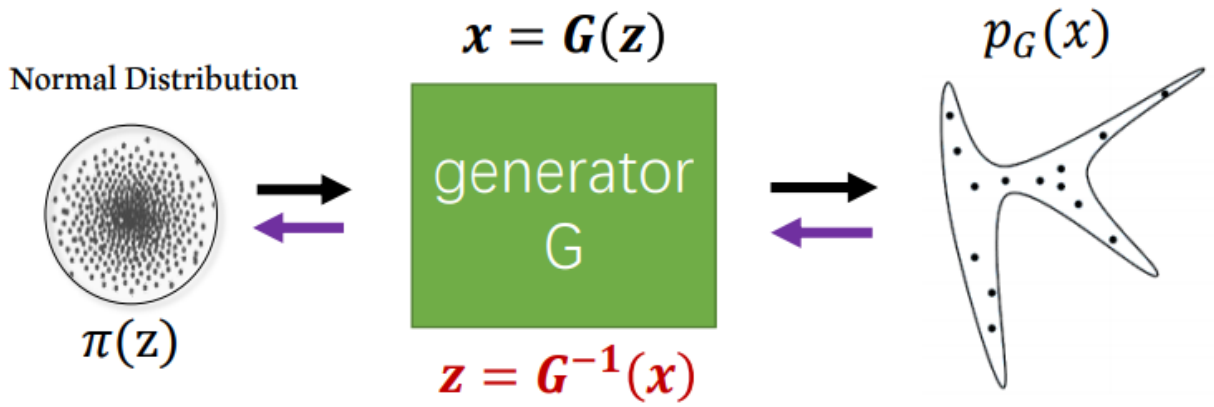
Discussion: MI and supervised learning

- Learning q(y|x) that maximize the lowerbound of MI.
- The optimal q(y|x) is p(y|x), by which the lower bound is tight.
- This is equal to the CE loss if y is the target label and x is the input feature.
- This means that CE can be used to perform MI estimation, as in infoGAN.
- This also explains the revers gradient in domain advesarial training (e.g., making the code less sensitive to domains). Once the classifier is well trained by CE on the domain label, it estimates MI of the code and the domain label, hopefully. Fixing the classifier and reversing the gradient makes the code and the domain label has a lower MI, i.e., less dependent. Note that since the reverse gradient ‘decrease the LOWER bound parameterized by q that was estimated with the old data’, it is not necessarily decrease the true MI really, though there is possibility.

$$
\begin{aligned}
MI(X,Y) &= \mathbb{E}_{p(x,y)} \ln q(y|x)/p(y) + \ln p(y|x)/q(y|x) \\
&= \mathbb{E}_{p(x,y)} \ln q(y|x) + H(y) + \mathbb{E}_{p(x)} KL(p(y|x)||q(y|x))
\end{aligned}
$$

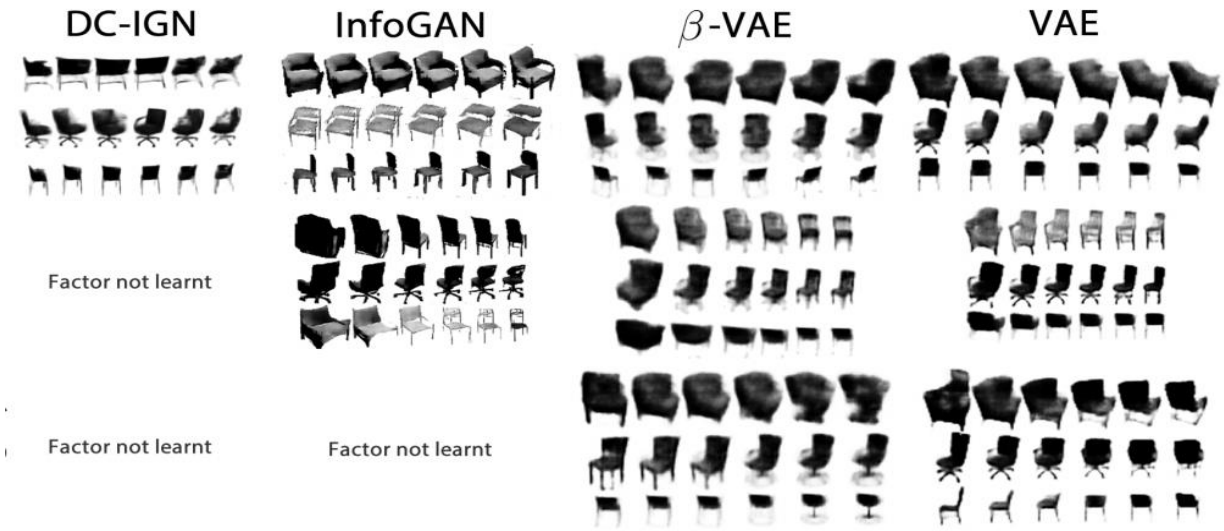
Discussion: MI in representation learning

- We certainly want to learn representation y that can ‘represent’ x.
- How it means? By maximizing MI, it means y can be predicted from x with less uncertainty. Note MI(X,Y) = H(Y) – H(Y|X)
- But only MI does not mean a better representation. E.g., a simple invertible function leads to perfect representation, and MI(X,Y)=H(Y).



Discussion: MI in representation learning

- Bias is required to generate reasonable reprentation
  - VAE: Bottleneck, discard some trivial information
  - beta–VAE: control the strength of the information propagated to the code
  - DC–IGN: clamp certain factors
  - InfoGAN: Maximum MI on partial codes
- The information flow should be carefully designed (dimension control is not enough)
- Using multiple objectives is important



- Kulkarni et al., Deep Convolutional Inverse Graphics Network, 2015.
- Higgins, β-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK, 2017

Conclusions

- Mutual information is an important measure/criterion in learning good representations.
- MI can be computed in a point–wise.
- MI is closely related to ML and contrastive learning.