ETH zürich

On Variational Bounds of Mutual Information

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, George Tucker

Presented by **Doruk Çetin**

Oct 29, 2019

Outline

- Why estimate Mutual Information (MI)?
- Review of variational bounds of MI
- Experimental analysis of lower bounds
- Discussion and future work

Unsupervised representation learning

- Automatically discovering useful low-dimensional features to summarize high-dimensional data
- Disentanglement: separating the distinct, informative factors of variations in the data

Samples

3D disentanglement

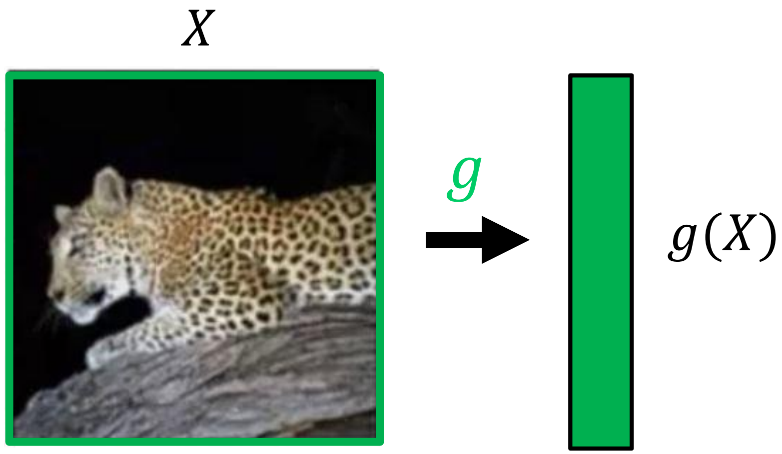
- Red
- Ferrari Testarossa
- Seen from right-side

Images taken from http://von.csail.mit.edu/ and https://preowned.ferrari.com

Why estimate Mutual Information (MI)?

- InfoMax (Linsker, 1988; Bell & Sejnowski, 1995)
  - Maximize the MI between inputs and outputs of an encoder (possibly subject to some structural constraints)

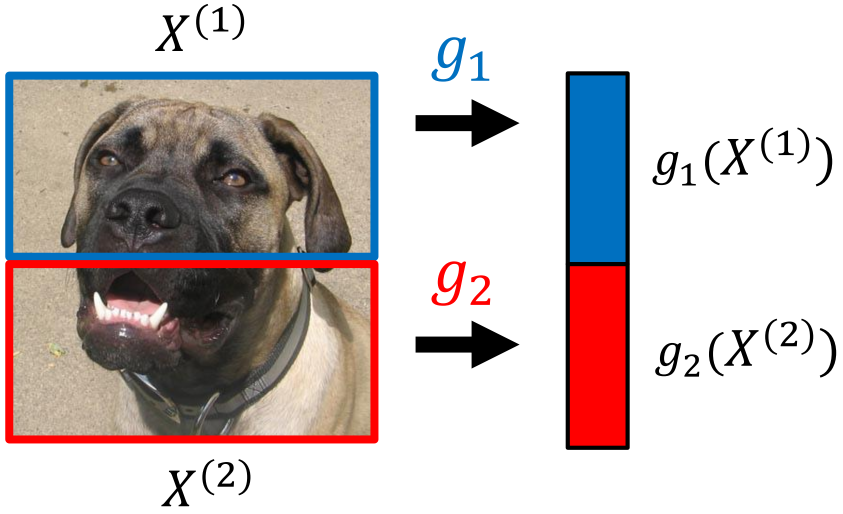
$$\max_{g \in \mathcal{G}} I(X; g(X))$$



  - Multi-view formulation, lower-bound the original objective

$$I(g_1(X^{(1)}); g_2(X^{(2)})) < I(X; g_1(X^{(1)}), g_2(X^{(2)}))$$

by data processing inequality

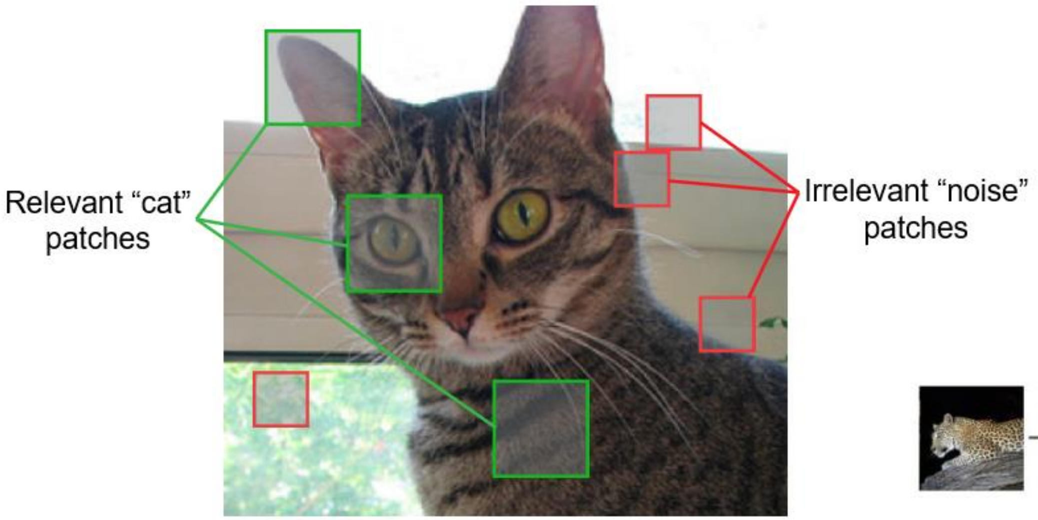


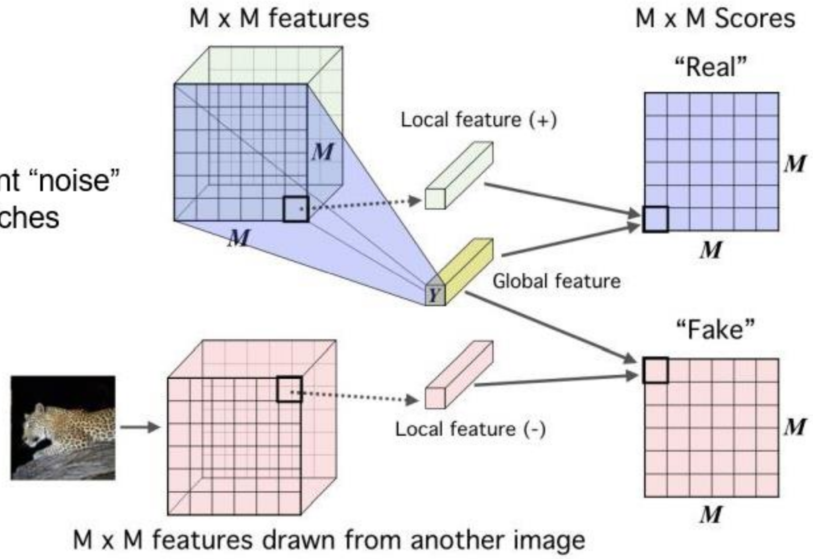
Images taken from http://www.image-net.org

Representation learning with Deep InfoMax

- Deep InfoMax (Hjelm et al., 2019)
  - Maximize MI between global and local features

$$\max_{g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2} I(g_1(X^{(1)}); g_2(X^{(2)}))$$





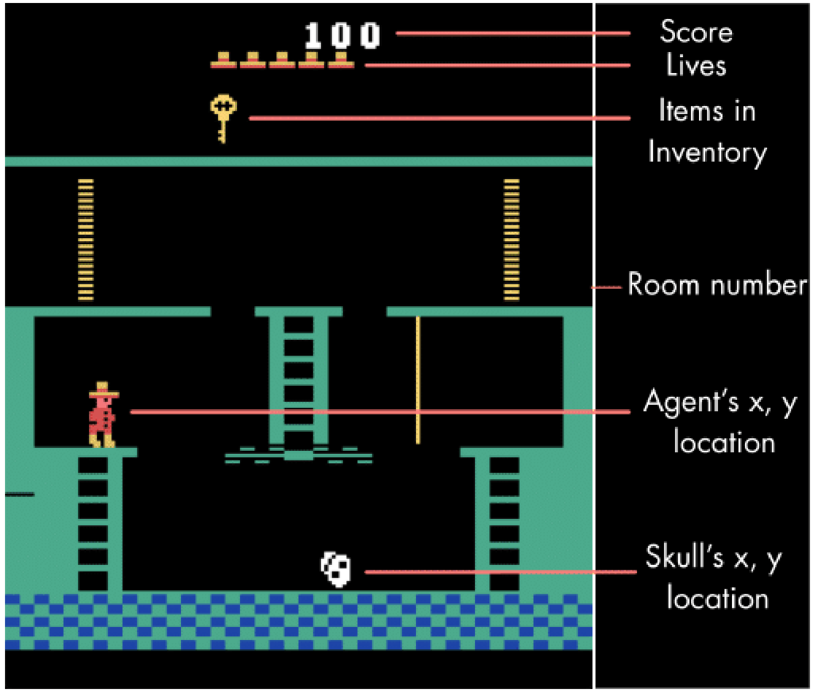
InfoMax

Deep InfoMax

Image taken from https://www.microsoft.com

Unsupervised State Representation Learning in Atari

- Spatiotemporal Deep Infomax (ST-DIM) (Anand et al., 2019)
  - Maximize MI across both spatial and temporal axes



Montezuma's Revenge

Probe F1 scores averaged across categories (data collected by random agents)

GAME	MAJ-CLF	RANDOM-CNN	VAE	PIXEL-PRED	CPC	ST-DIM	SUPERVISED
MONTEZUMAREVENGE	0.08	0.68	0.69	0.74	0.75	**0.78**	0.87
MSPACMAN	0.10	0.48	0.38	**0.74**	0.65	0.70	0.87
PITFALL	0.07	0.34	0.56	0.44	0.46	**0.60**	0.83
PONG	0.10	0.17	0.09	0.70	0.71	**0.81**	0.87
PRIVATEEYE	0.23	0.70	0.71	0.83	0.81	**0.91**	0.97
QBERT	0.29	0.49	0.49	0.52	0.65	**0.73**	0.76
MEAN	0.14	0.44	0.39	0.58	0.60	**0.68**	0.83

Recent work on MI estimation

- We need scalable, tractable and differentiable objectives

$$\max_{g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2} I(g_1(X^{(1)}); g_2(X^{(2)}))$$

- Combine variational lower bounds with deep learning

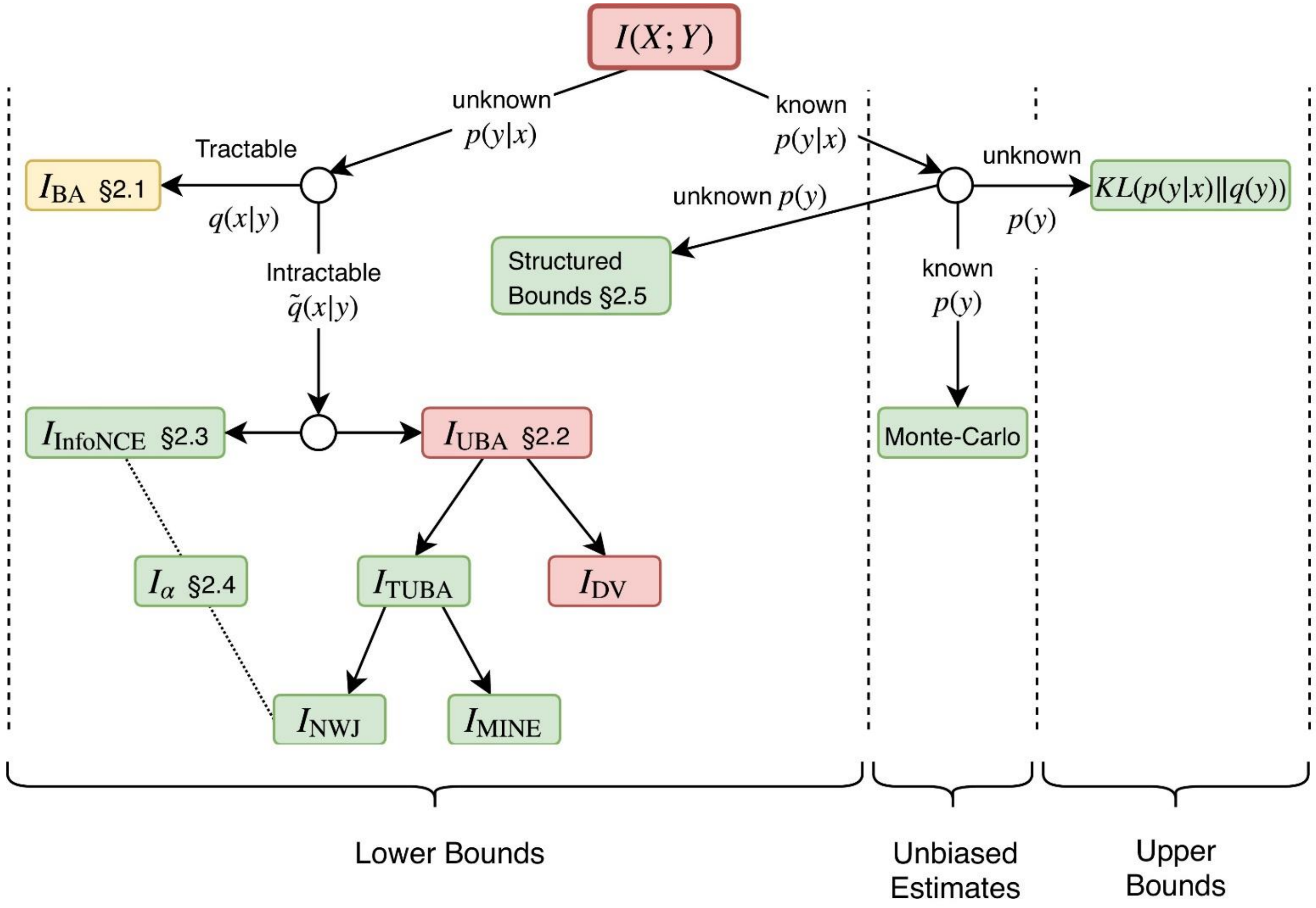
$$\max_{g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2} I_{\mathrm{EST}}(g_1(X^{(1)}); g_2(X^{(2)}))$$

- Use “critic” function to distinguish samples from joint and product of marginals, if classifier can distinguish, then *X* and *Y* have a high MI

$$I(X;Y) = D_{KL}(p(x,y) \parallel p(x)p(y))$$

Review of existing estimators

- **Contribution:** unify recent developments in a single framework



Variational bounds of MI

- Classical (normalized) lower bound (Barber & Agakov, 2003)

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x|y)}{p(x)}\right] = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x|y)q(x|y)}{q(x|y)p(x)}\right]$$
$$= \mathbb{E}_{p(x,y)}\left[\log \frac{q(x|y)}{p(x)}\right] + \mathbb{E}_{p(y)}[D_{KL}(p(x|y) \parallel q(x|y))] \geq \underline{\mathbb{E}_{p(x,y)}[\log q(x|y)] + h(X) \triangleq I_{BA}}$$

  - Requires a tractable decoder $q(x|y)$
  - Intractable because $h(X)$ is often unknown
  - We can compare the amount of information different variables (e.g. Y_1 and Y_2) carry about X

Variational bounds of MI

- Unnormalized lower bound
  - Choose a variational family that uses critic $f(x, y)$ and substitute into I_{BA}

$$q(x|y) = \frac{p(x)}{Z(y)} e^{f(x,y)}, \; Z(y) = \mathbb{E}_{p(x)}\left[e^{f(x,y)}\right] \longrightarrow I_{BA}$$

  - Unnormalized version of the classical bound, intractable because of $\log Z(y)$

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

- Critic functions revisited:

$$(x, y) \text{ drawn from joint } p(x,y) \implies f(x,y) \text{ is high}$$

$$(x, y) \text{ drawn from product of marginals } p(x)p(y) \implies f(x,y) \text{ is low}$$

Variational bounds of MI

- Tractable unnormalized lower bound
  - Upper bound the log partition to form a tractable bound

$$\log Z(y) \leq \frac{Z(y)}{a(y)} + \log(a(y)) - 1, \forall Z(y), a(y) \longrightarrow I_{UBA}$$

  - Holds for any $a(y) > 0$, maximize w.r.t. variational parameter $a(y)$ and f to tighten the bound

$$I \geq I_{UBA} \geq \mathbb{E}_{p(x,y)}[f(x,y)] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}[e^{f(x,y)}]}{a(y)} + \log(a(y)) - 1\right] \triangleq I_{TUBA}$$

- Bound of Nguyen, Wainwright and Jordan (Nguyen et al., 2010)

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

  - Also known as *f*-GAN KL (Nowozin et al., 2016) and MINE-*f* (Belghazi et al., 2018)

Variational bounds of MI

- InfoNCE (multi-sample unnormalized bound) (Oord et al., 2018)
  - Idea is to use multiple samples to lower the variance of unnormalized bounds

1. Get a random minibatch of size K

$$(x_1, y_1) \quad (x_2, y_2) \quad (x_3, y_3) \quad \cdots \quad (x_K, y_K)$$

2. For each x_i predict which of the K samples $y_1, y_2, \ldots, y_K$ it was jointly drawn with

$$\underset{\sim p(x,y)}{\underset{\text{Positive example}}{\boxed{(x_1, y_1)}}} \quad \underbrace{\boxed{(x_1, y_2)} \; \boxed{(x_1, y_3)} \; \cdots \; \boxed{(x_1, y_K)}}_{\substack{K-1 \text{ negative examples} \\ \sim p(x)p(y)}}$$

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

- InfoNCE loss the categorical cross-entropy of classifying the positive example
- InfoNCE is upper bounded by $\log K$, becomes loose when $I(X;Y) > \log K$

Nonlinearly interpolated lower bounds

- **Contribution:** a new continuum of multi-sample lower bounds

$$1+\mathbb{E}_{p(x_{1:K})p(y|x_1)}\left[\log \frac{e^{f(x_1,y)}}{\alpha m\left(y; x_{1:K}\right)+(1-\alpha)q(y)}\right]-\mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{e^{f(x_1,y)}}{\alpha m\left(y; x_{1:K}\right)+(1-\alpha)q(y)}\right] \triangleq I_\alpha$$

  - Upper bounded by $\log \frac{K}{\alpha}$

$\alpha = 0$ $\alpha = 1$

I_{NWJ} Low bias
High variance

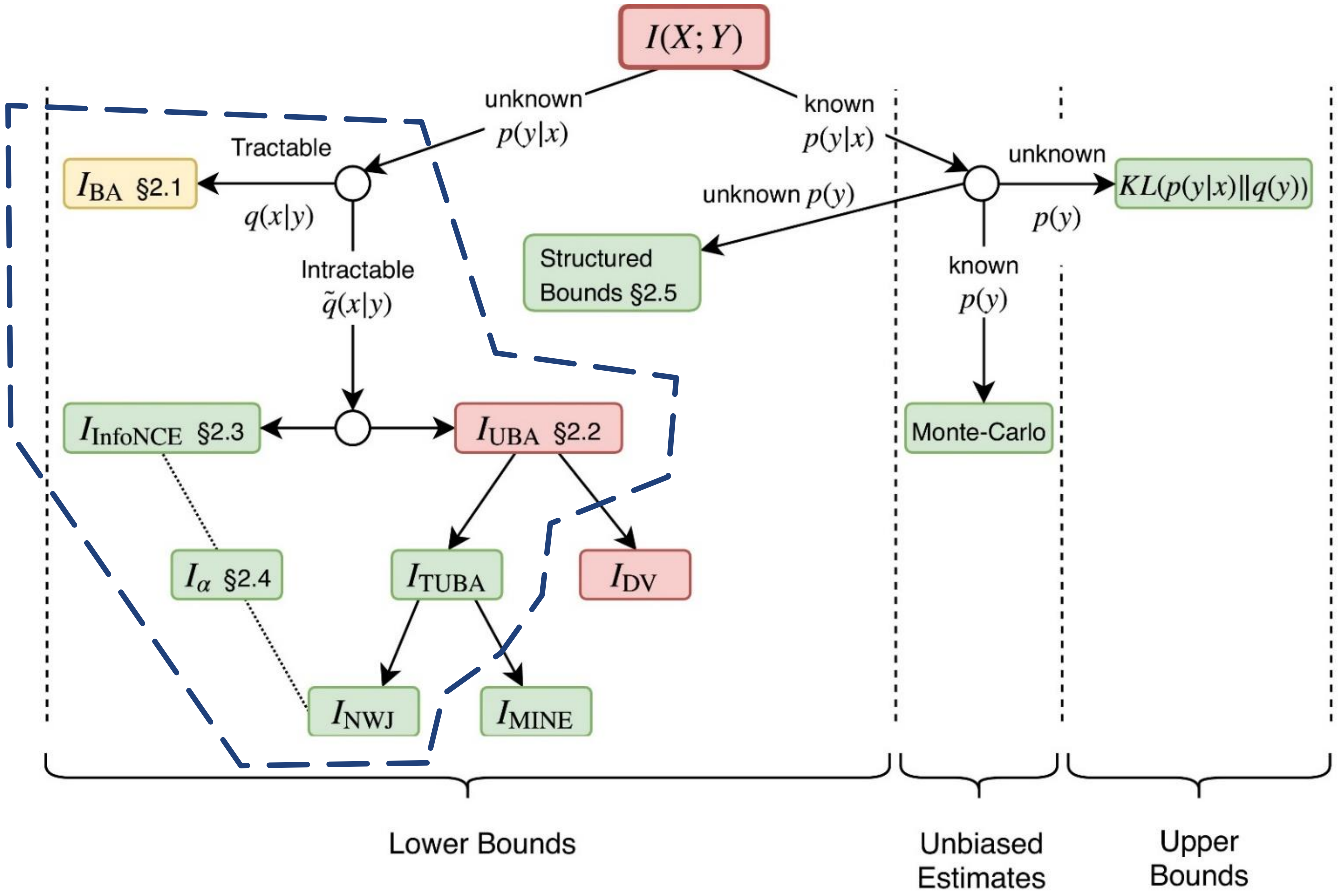
I_α

I_{NCE} Low variance
High bias

Flexibly tradeoff bias and variance

Review of existing estimators

- A single framework of variational bounds



Tractability for optimization and estimation

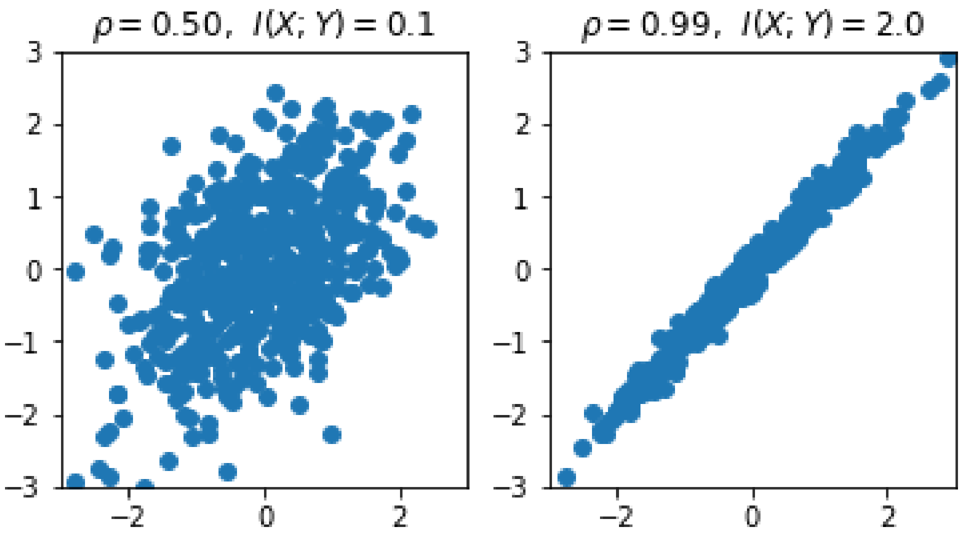
Both optimization and estimation

Optimization but not estimation

Neither optimization, nor estimation

Experiments

- Correlated 20-dim Gaussian problem (Belghazi et al., 2018)



- Two toy problems

1. Sampling from (x, y)
2. Sampling from $(x, (Wy)^3)$

- For full rank linear transformations:

$$I(x, y) = I(x, (Wy)^3)$$

- Vary correlation coefficient ρ for different values of MI
- Can compute true MI: $I(x, y) = -\frac{d}{2}\log(1-\rho^2)$

Experiments

- Two critic architectures
  - Both are fully connected networks with ReLU activations

1. Separable critic (van den Oord et al., 2018)
   - Map x, y to embedding space and take inner product

$$f(x, y) = h(x)^T g(y)$$

   - Requires $2N$ forward passes for a batch-size of N

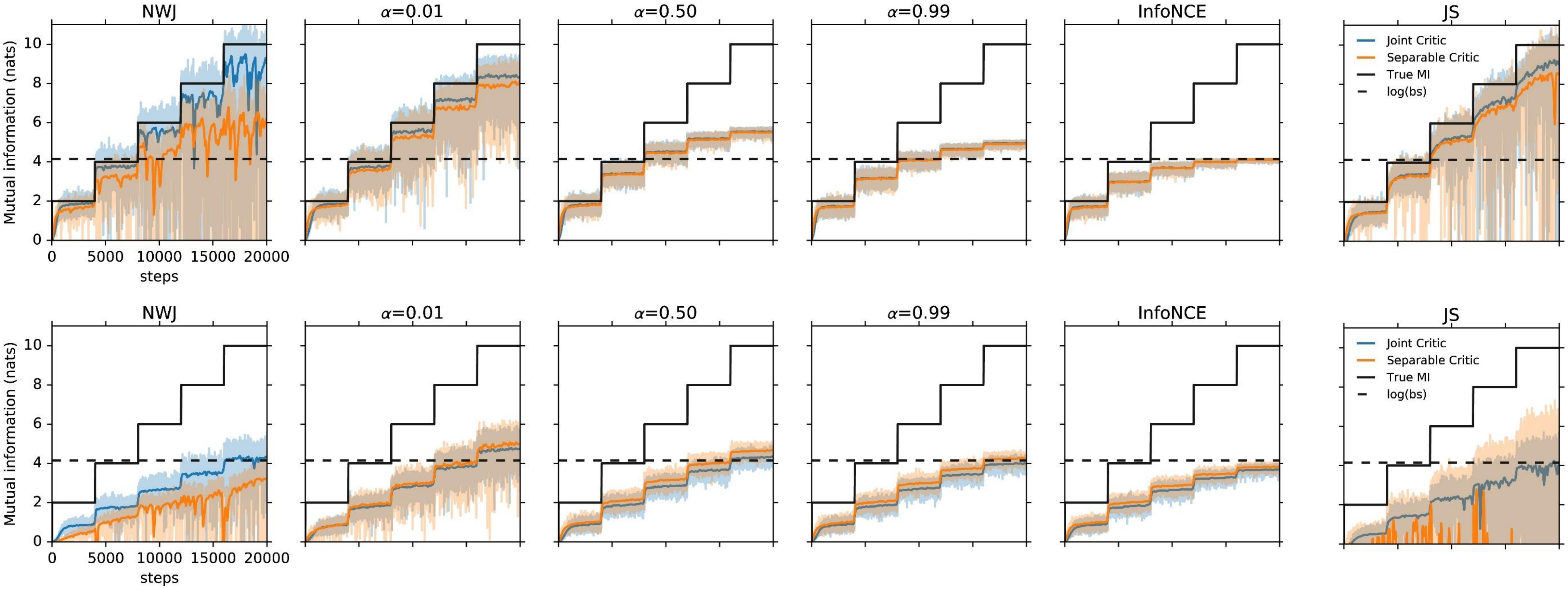
2. Joint critic (Belghazi et al., 2018)
   - Concatenate x, y before feeding it into network

$$f(x, y) = h([x, y])$$

   - Requires N^2 forward passes for a batch-size of N

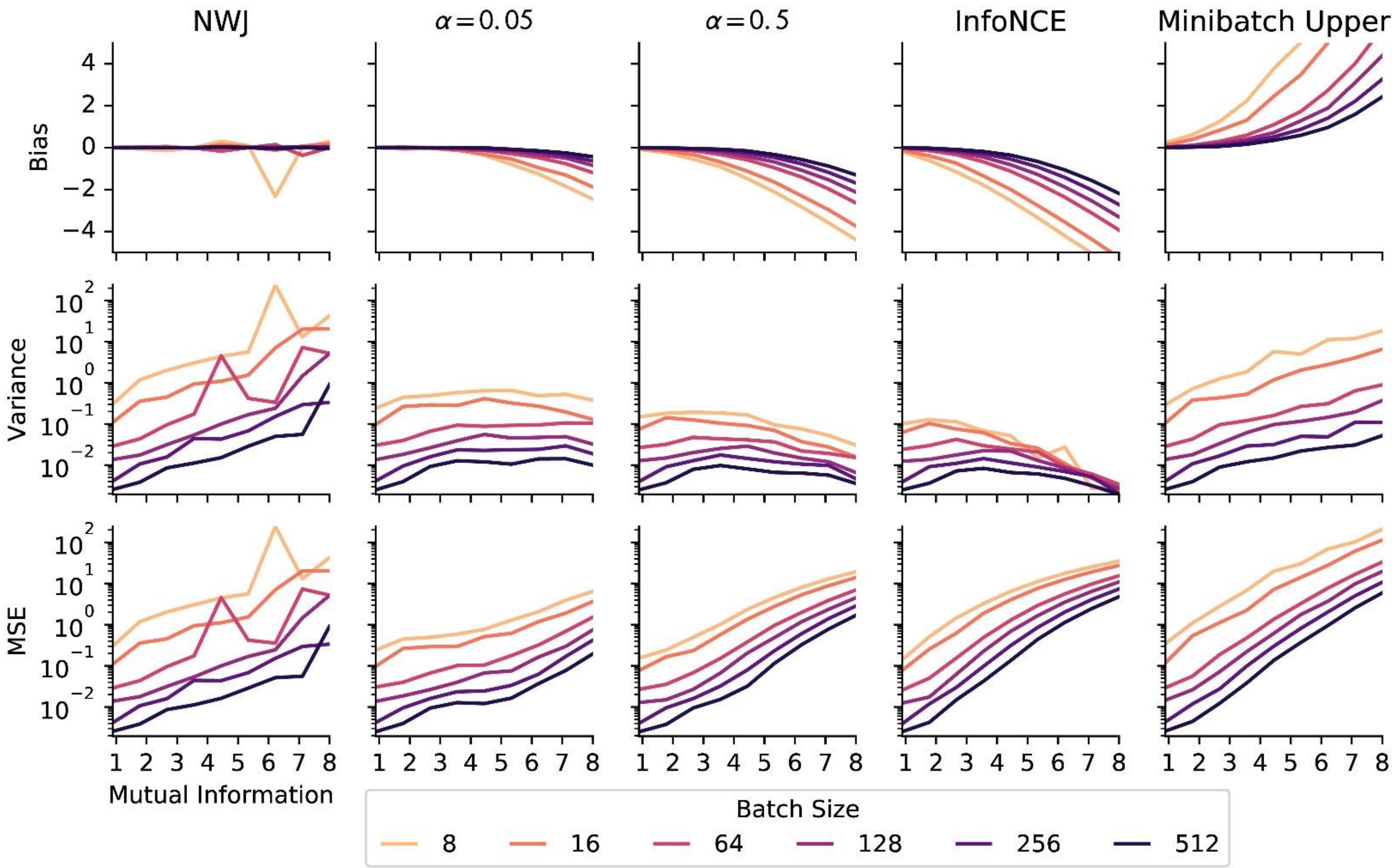
Experiments

- Efficiency-accuracy tradeoffs for critic architectures



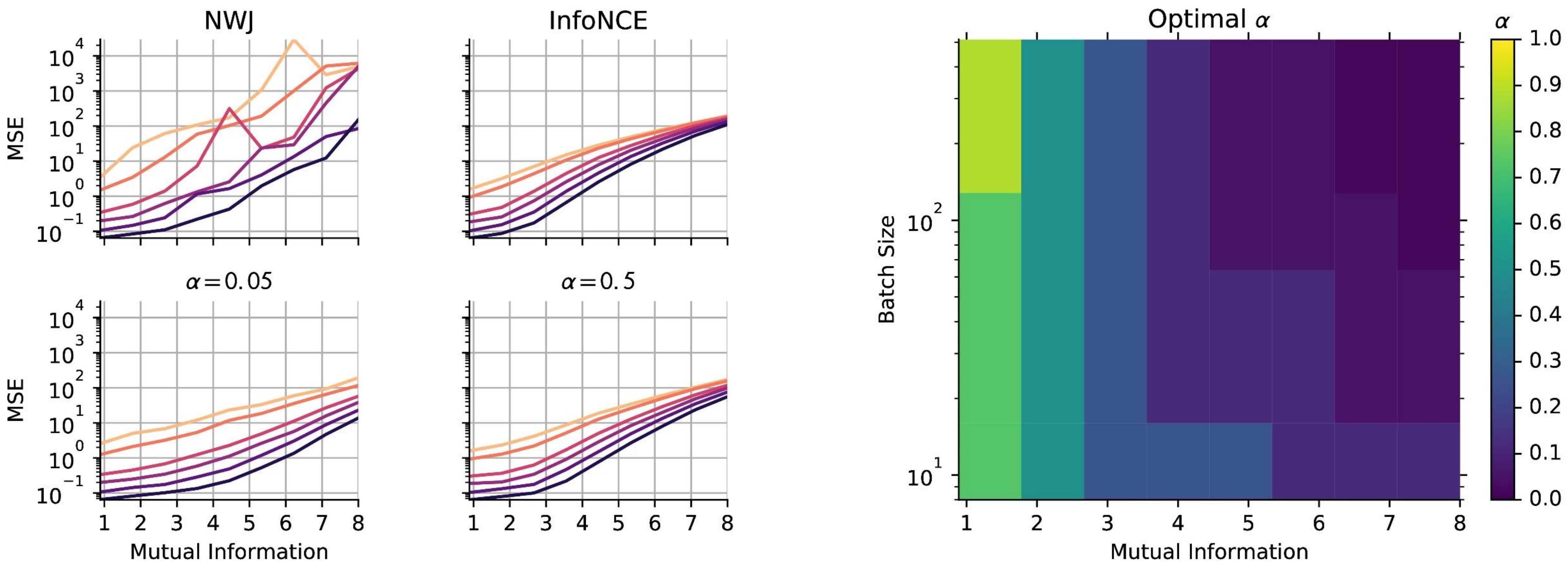
Experiments

- Bias-variance tradeoff for optimal critics



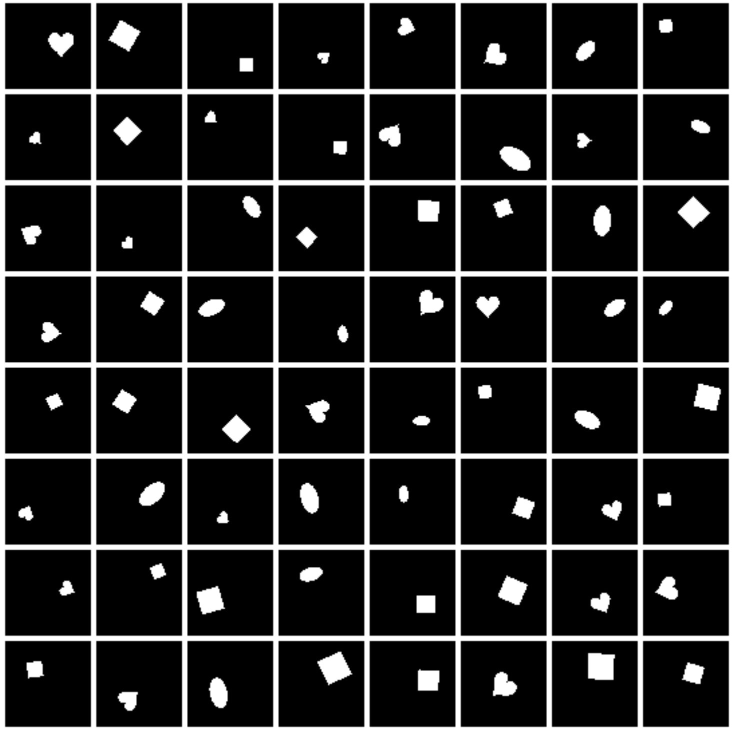
Experiments

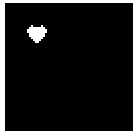
- Bias-variance tradeoffs for representation learning



Experiments

- dSprites dataset for disentanglement testing



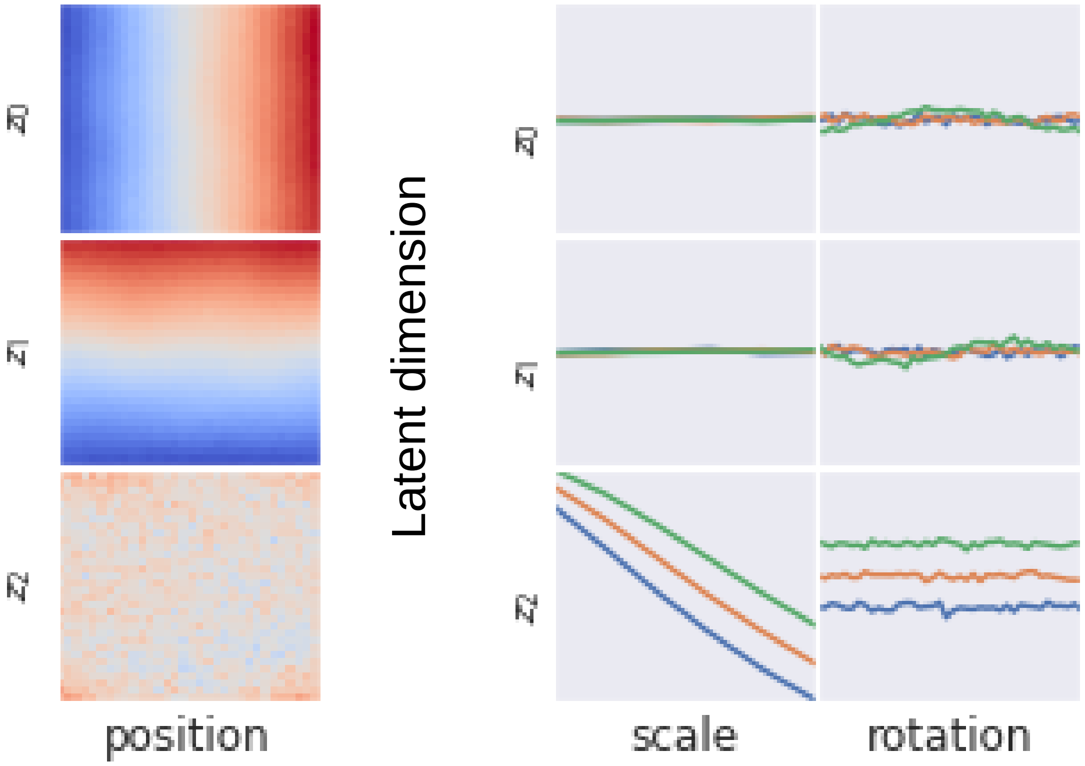


- Color: white
- Shape: square, ellipse, heart
- Scale: 6 values linearly spaced in [0.5, 1]
- Orientation: 40 values in [0, 2 pi]
- Position X: 32 values in [0, 1]
- Position Y: 32 values in [0, 1]

Images taken from https://github.com/deepmind/dsprites-dataset and http://blog.adeel.io

Experiments

- Decoder-free repr. learning on dSprites
  - Objective includes three terms:
  1. Mutual information maximization
  2. Statistical dependency minimization
  3. Smoothness regularization
  - Use I_{JS} lower bound for the estimation



- axes: x/y position
- color: average activation of the latent variable

- y-axis: avg. value of the latent variable
- x-axis: value of the ground truth factor

Discussion

- Unify recent developments in a single framework
  - Proof that I_{NCE} loss is indeed a lower bound on MI
- New interpolated bounds to tradeoff bias and variance
  - No low-variance, low-bias estimator for large MI and small batch size
- Systematic evaluation of estimators
  - Study is limited to infinite dataset and no overfitting setting, not realistic
- An open question
  - Is mutual information maximization more useful for representation learning than other unsupervised and self-supervised approaches?

Should we use MI maximization?

- Maximizing MI does not necessarily lead to useful representations
  - Invariances under arbitrary invertible transformations, need for regularization

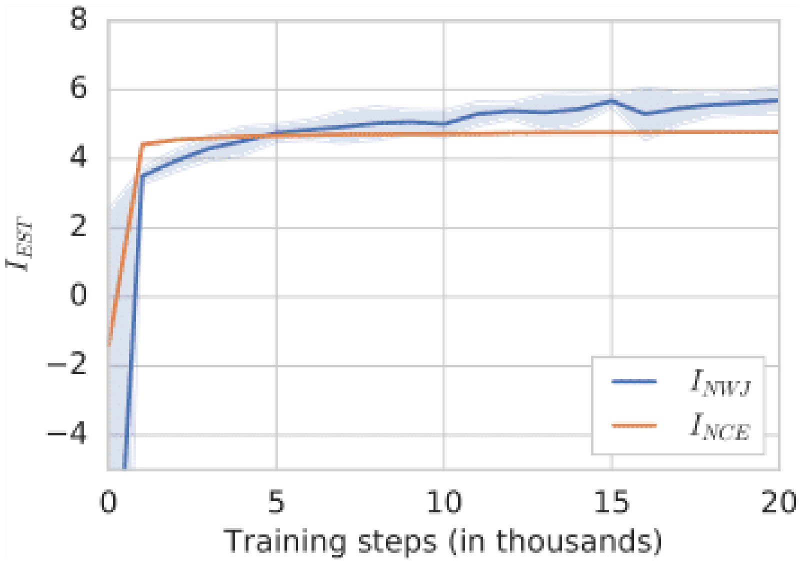
- Yet, many promising results using InfoMax
  - Image and video classification, natural language understanding…

- On Mutual Information Maximization for Representation Learning
  (Tschannen et al., 2019)

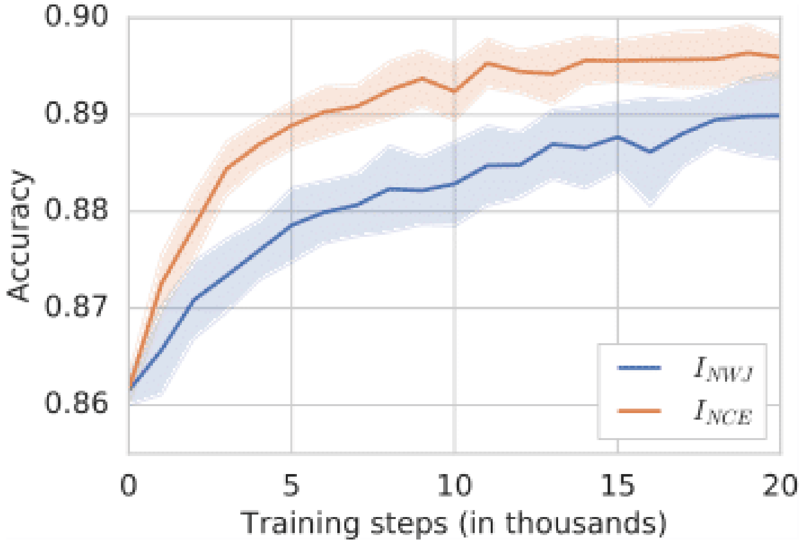
  - *“Success of these methods might be loosely attributed to the properties of MI”*

Large MI is not predictive of downstream performance

- Encoders g_1 and g_2 are parameterized to be always invertible
- MI is constant for any choice of parameters: $I(g_1(X^{(1)}); g_2(X^{(2)})) = I(X^{(1)}; X^{(2)})$



(a)



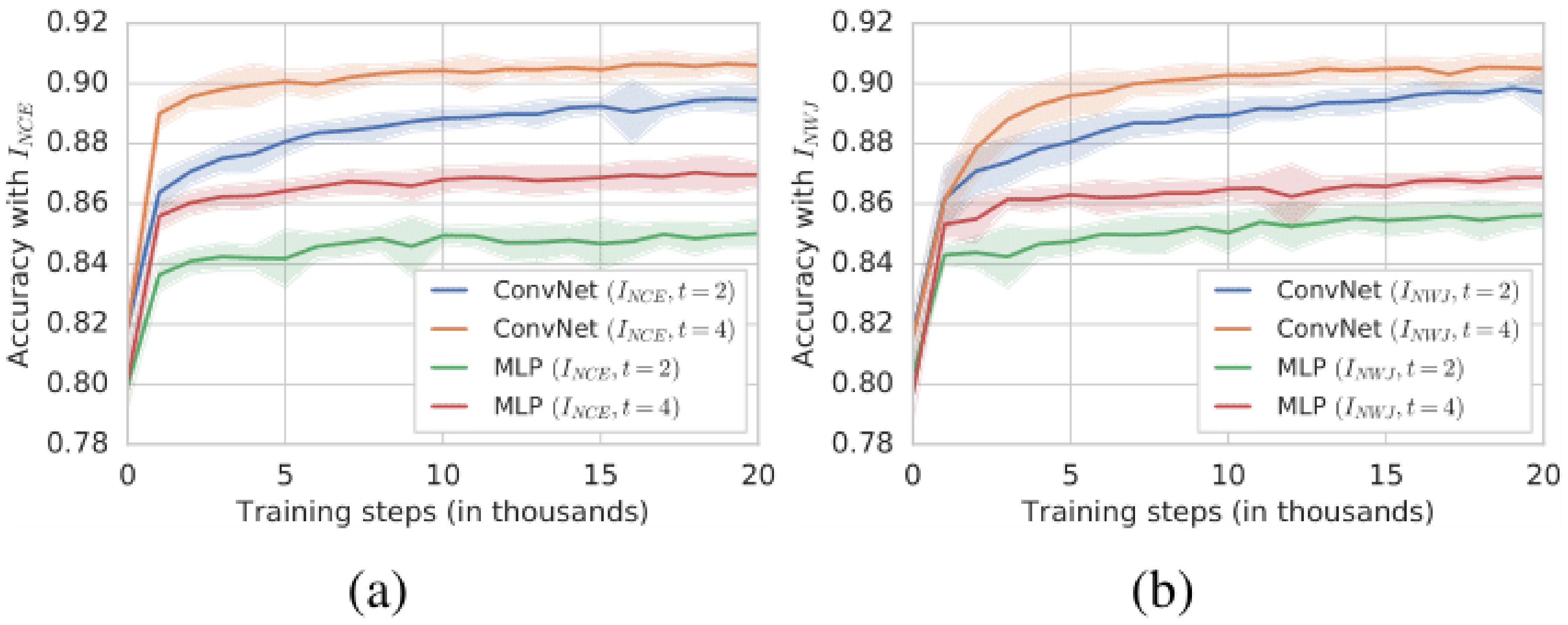
(b)

Thought experiment:
- Pixel space
- PNG compressed bit stream

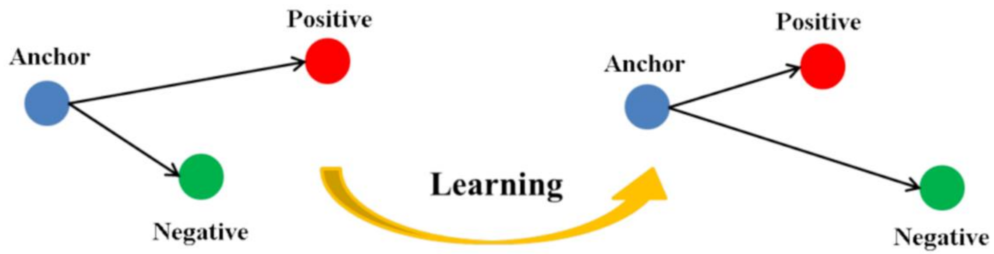
- Estimators bias the encoders towards solutions suitable for the downstream task

Encoder architecture can be more important than the specific estimator

- All configurations are ensured to achieve same lower bound
- Despite matching bounds ConvNets have better results than MLPs



Connection to deep metric learning and triplet losses



Triplet example

$$
\begin{aligned}
I_{\mathrm{NCE}} &= \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K}e^{f(x_i,y_j)}}\right] \\
&= \log K - \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\left(1+\sum_{j\neq i}e^{f(x_i,y_j)-f(x_i,y_i)}\right)\right]
\end{aligned}
$$

InfoNCE objective rewritten

$$
L_{\mathrm{K-pair-mc}}\left(\{(x_i,y_i)\}_{i=1}^{K},\phi\right) = \frac{1}{K}\sum_{i=1}^{K}\log\left(1+\sum_{j\neq i}e^{\phi(x_i)^{\top}\phi(y_j)-\phi(x_i)^{\top}\phi(y_i)}\right)
$$

Multi-class k-pair loss

Image taken from http://cs231n.stanford.edu

Conclusion

- Maximizing MI is not always a good idea
- Common estimators and architectures have strong inductive biases
- Triplet-based metric learning may serve plausible explanations

References

A. Anand, E. Racah, S. Ozair, Y. Bengio, M.-A. Côté, and R. D. Hjelm. Unsupervised State Representation Learning in Atari. June 19, 2019. arXiv: 1906.08226 [cs, stat].

D. Barber and F. Agakov. The IM Algorithm: A Variational Approach to Information Maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems* (Whistler, British Columbia, Canada), NIPS'03, pages 201–208, Cambridge, MA, USA. MIT Press, 2003.

M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. MINE: Mutual Information Neural Estimation. Jan. 12, 2018. arXiv: 1801.04062 [cs, stat].

A. J. Bell and T. J. Sejnowski. An Information-maximization Approach to Blind Separation and Blind Deconvolution. *Neural Comput.*, 7(6):1129–1159, Nov. 1995. ISSN: 0899-7667. DOI: 10.1162/neco.1995.7.6.1129.

R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. Aug. 20, 2018. arXiv: 1808.06670 [cs, stat].

R. Linsker. An Application of the Principle of Maximum Information Preservation to Linear Systems. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 1*, pages 186–194. Morgan-Kaufmann, 1989.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, Nov. 2010. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.2010.2068870.

B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker. On Variational Bounds of Mutual Information. May 16, 2019. arXiv: 1905.06922 [cs, stat].

M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On Mutual Information Maximization for Representation Learning. July 31, 2019. arXiv: 1907.13625 [cs, stat].

A. van den Oord, Y. Li, and O. Vinyals. Representation Learning with Contrastive Predictive Coding. July 10, 2018. arXiv: 1807.03748 [cs, stat].

Questions?

Image taken from https://www.linkedin.com

Summary of mutual information lower bounds

- Characterization of mutual information lower bounds

	Lower Bound	L	∇L	$\perp$ BS	Var.	Norm.
I_{BA}	Barber & Agakov (2003)	✗	✓	✓	✓	✗
I_{DV}	Donsker & Varadhan (1983)	✗	✗	–	–	–
I_{NWJ}	Nguyen et al. (2010)	✓	✓	✓	✗	✓
I_{MINE}	Belghazi et al. (2018)	✗	✓	✓	✗	✓
I_{NCE}	van den Oord et al. (2018)	✓	✓	✗	✓	✓
I_{JS}	Appendix D	✓	✓	✓	✗	✓
I_{α}	Eq. 11	✓	✓	✗	✓	✓

Table 1. Characterization of mutual information lower bounds. Estimators can have a tractable (✓) or intractable (✗) objective (L), tractable (✓) or intractable (✗) gradients (∇L), be dependent (✗) or independent (✓) of batch size ($\perp$ BS), have high (✗) or low (✓) variance (Var.), and requires a normalized (✗) vs unnormalized (✓) critic (Norm.).

Summary of mutual information lower bounds

- Parameters and objectives for mutual information estimators

Lower Bound	Parameters	Objective
I_{BA}	$q(x\|y)$ tractable decoder	$\mathbb{E}_{p(x,y)}\left[\log q(x\|y) - \log p(x)\right]$
I_{DV}	$f(x,y)$ critic	$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \log\left(\mathbb{E}_{p(x)p(y)}\left[f(x,y)\right]\right)$
I_{NWJ}	$f(x,y)$	$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \frac{1}{e}\mathbb{E}_{p(x)p(y)}\left[f(x,y)\right]$
I_{MINE}	$f(x,y)$, $\text{EMA}(\log f)$	I_{DV} for evaluation, $I_{\text{TUBA}}(f, \text{EMA}(\log f))$ for gradient
I_{NCE}	$f(x,y)$	$\mathbb{E}_{p^K(x,y)}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{f(y_i,x_i)}{\frac{1}{K}\sum_{j=1}^{K}f(y_i,x_j)}\right]$
I_{JS}	$f(x,y)$	I_{NWJ} for evaluation, f-GAN JS for gradient
I_{TUBA}	$f(x,y)$, $a(y) > 0$	$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}[f(x,y)]}{a(y)} + \log(a(y)) - 1\right]$
I_{TNCE}	$e(y\|x)$ tractable encocder	I_{NCE} with $f(x,y) = e(y\|x)$
I_{α}	$f(x,y)$, α, $q(y)$	$1 + \mathbb{E}_{p(x_{1:K},y)}\left[\log\frac{e^{f(x_1,y)}}{\alpha m(y;x_{1:K}) + (1-\alpha)q(y)}\right]$ $-\mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{e^{f(x_1,y)}}{\alpha m(y;x_{1:K}) + (1-\alpha)q(y)}\right]$

Table 2. Parameters and objectives for mutual information estimators.

More on Deep InfoMax

- Complete DIM objective
  - Local MI max.

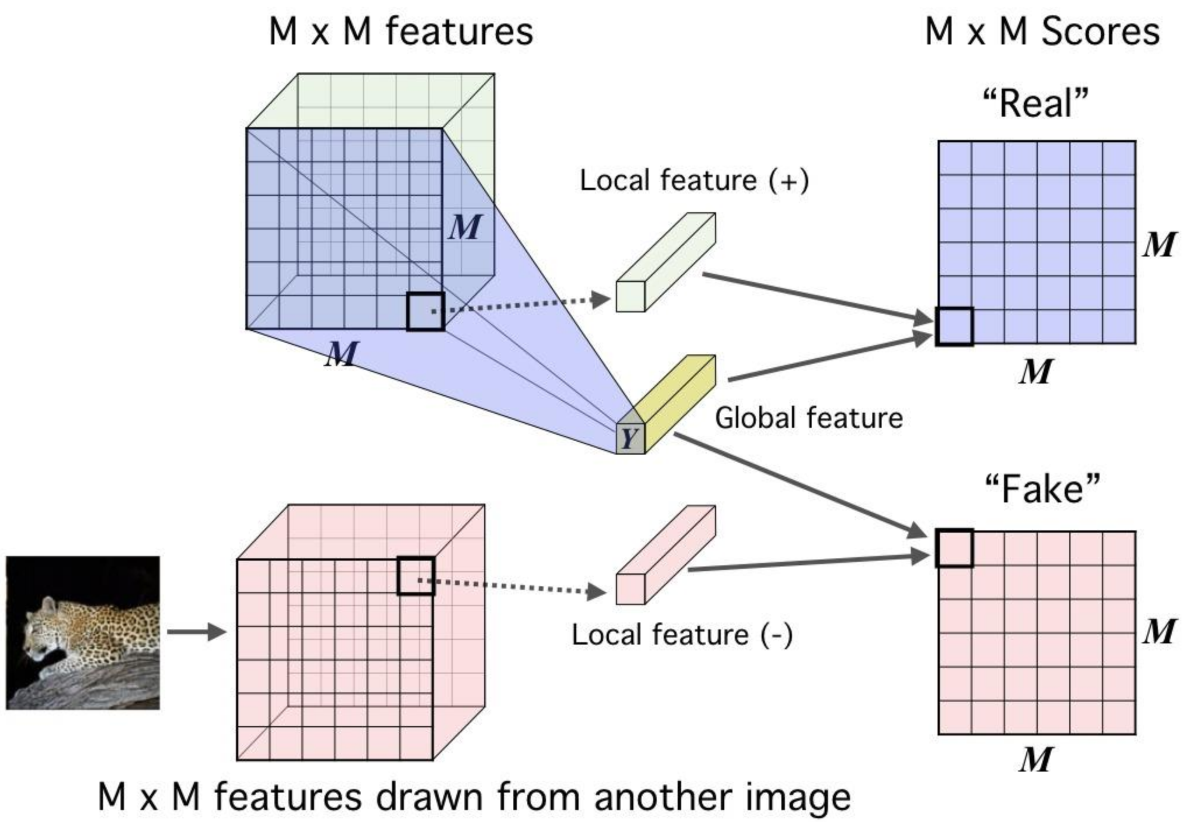
$$(\hat{\omega}, \hat{\psi})_L = \arg\max_{\omega,\psi} \frac{1}{M^2} \sum_{i=1}^{M^2} \widehat{\mathcal{I}}_{\omega,\psi} \left(C_\psi^{(i)}(X); E_\psi(X) \right)$$

  - Global MI max.

$$(\hat{\omega}, \hat{\psi})_G = \arg\max_{\omega,\psi} \widehat{\mathcal{I}}_\omega \left(X; E_\psi(X) \right)$$

  - Prior matching

$$(\hat{\omega}, \hat{\psi})_P = \arg\min_{\psi} \arg\max \widehat{\mathcal{D}}_\phi \left(\mathbb{V} \| \mathbb{U}_{\psi,\mathrm{P}} \right) = \mathbb{E}_{\mathbb{V}} \left[\log D_\phi(y) \right] + \mathbb{E}_{\mathbb{P}} \left[\log \left(1 - D_\phi \left(E_\psi(x) \right) \right) \right]$$



MINE: Mutual Information Neural Estimation (Belghazi et al., 2018)

- Produces estimates that are neither an upper or lower bound on MI

$$I \geq I_{UBA} \geq \mathbb{E}_{p(x,y)}[f(x,y)] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}[e^{f(x,y)}]}{a(y)} + \log(a(y)) - 1\right] \triangleq I_{TUBA}$$

- Improved MINE gradient estimator
  - Sound justification for the heuristic optimization procedure through I_{TUBA}
  - Set $a(y)$ to be the scalar exponential moving average (EMA) of $e^{f(x,y)}$ across minibatches