On Mutual Information Maximization For Representation Learning

An excursus into Deep InfoMax, its variants and their lacks

Matteo Tiezzi

References

- Learning deep representations by mutual information estimation and maximization
  - R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, Yoshua Bengio, ICLR 2019

- On Mutual Information Maximization for Representation Learning
  - Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, Mario Lucic, ICLR 2020

Unsupervised Representation Learning

- **Fundamental problem in ML:**
  - learn a function *g* which maps the data in a (usually lower-dim) space where hopefully is possible to solve more efficiently a supervised task

- **Recent revival of approaches inspired by the InfoMax principle**
  - choose representation *g(x)* maximizing the mutual information between the input and its representation
  - possibly subject to some structural constraints

Information Theory - Entropy

- A sender wishes to transmit the value of a random variable to a receiver
- average amount of info needed to specify the state of a random variable is called **Entropy**
- measure of the uncertainty of a probability distribution

$$H[X] = -\sum_{x} p(x) \log p(x)$$

Information Theory - Entropy and Relative Entropy (KL-div)

- **Relative Entropy**
  - We use an approximating distribution q(x) to model an unknown distribution p(x)
  - Transmit values of x to a receiver, using q(x) to construct a coding scheme (instead of the true p(x))
  - The Relative Entropy is the additional required amount of info

$$D_{KL}(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)}$$

- **Using the true distribution p(x) - code with average description length of H(p) bits (nats)**
  - Using q(x), the average required info is (measure of inefficiency)

$$H(p) + D_{KL}(p||q)$$

- **Interpretable as a distance measure between the two distribution**

$$D_{KL}(p||q) = 0 \iff p(x) = q(x)$$

Information Theory - Mutual Information

- Reduction of uncertainty of a r.v. ***x*** by virtue of being told the value of ***y***
  - hence, the amount of info that x contains about y

$$I[x, y] = H[x] - H[x|y]$$

- **Alternative formulation**
  - KL divergence between the joint density p(x,y) and the product of the marginals

$$I[x, y] = D_{KL}(p(x, y)||p(x)p(y))$$

  - if KL ≃ 0, almost independent -- low info contained about the other one

MI properties (Kraskov et al. 2004)

- **Invariant Under Reparametrization of the variables**

  - if $X' = f_1(X)$ and $Y' = f_2(Y)$ are homeomorphism (i.e. smooth invertible maps), then

$$I(X, Y) = I(X', Y')$$

  - Estimating Mi in high-dim is difficult
    - often one maximizes a tractable lower-bound

Recent Progress and InfoMax Principle

- **Usual problem setup (Becker and Hinton 1992)**
  - Given an image X, let $X^{(1)}$ and $X^{(2)}$ be different views of X (f.i. top and bottom halves of the image)
  - Encoders g_1 and g_2, maximize MI between the two representations, sample base estimator I_{EST}

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

  - Lower bound of the original InfoMax $\max_{g \in \mathcal{G}} I(X; g(X))$
  - Advantages of Multi-view formulation
    - estimate only between learned representation of the views (lower-dim space)
    - modeling flexibility - capture different aspects or modality of the data

Various approaches

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

- **DeepInfomax**
  - in the basic form, g_1 extracts global features from the entire image $X^{(1)}$ and g_2 local features from image patches $X^{(2)}$
- **Contrastive multiview Coding**
- **Contrastive Predictive Coding, etc.**

Lower Bounds on MI

I_{EST} is a critical choice. Idea:

- If a classifier can accurately distinguish between samples drawn from the joint p(x,y) and those drawn from the marginal p(x)p(y), then X and Y have high MI
  - DV (Donsker and Varadhan, 1983)

$$\mathcal{I}(X;Y) := \mathcal{D}_{KL}(\mathbb{J}||\mathbb{M}) \geq \widehat{\mathcal{I}}_{\omega}^{(DV)}(X;Y) := \mathbb{E}_{\mathbb{J}}[T_{\omega}(x,y)] - \log \mathbb{E}_{\mathbb{M}}[e^{T_{\omega}(x,y)}].$$

  - InfoNCE (van den Oord et al. 2018)

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

    - Expectation is over K independent samples $\{(x_i,y_i)\}_{i=1}^{K}$ from the joint (Monte Carlo estimate averaging over multiple batches of samples)

Lower Bounds on MI

- Intuitively, the **critic** function f tries to predict for each x_i which of the K samples y_1, ..., y_K it was jointly drawn with.
  - Various types:
    - bilinear
    - separable
    - concatenated

$$f(x, y) = x^\top W y$$
$$f(x, y) = \phi_1(x)^\top \phi_2(y)$$
$$f(x, y) = \phi([x, y])$$

  - ϕ, ϕ_1, ϕ_2 are typically shallow multi-layer perceptrons

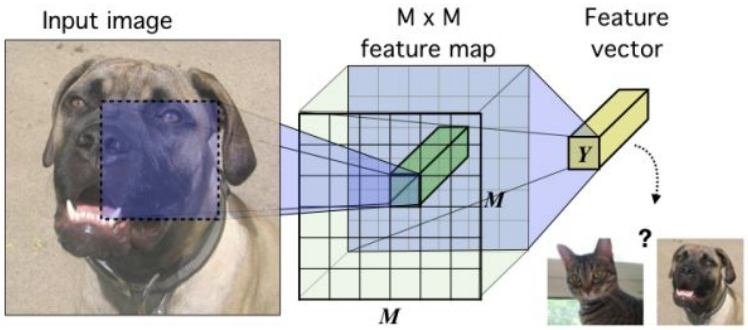
DeepInfomax (Hjelm et al. 2019)

- **Maximize global MI**, between an input data and the resulting global feature vector produced by the encoder E_ψ
  - often insufficient for learning useful representation

- **Maximize local MI**, the MI between the local and global features produced by the encoder.

- **Enforce a statistical constraint,** to avoid a trivial solution to the MI maximization objective
  - usefulness of a representation is not just a matter of information content
  - marginal distribution of the encoded features must be close to a statistical prior

DeepInfomax - Base Encoder



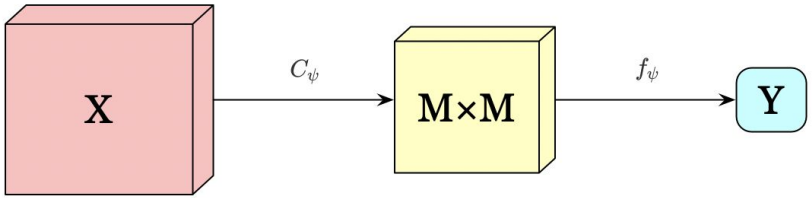


Figure 1: **The base encoder model in the context of image data.** An image (in this case) is encoded using a convnet until reaching a feature map of $M \times M$ feature vectors corresponding to $M \times M$ input patches. These vectors are summarized into a single feature vector, Y. Our goal is to train this network such that useful information about the input is easily extracted from the high-level features.

$$E_\psi = f_\psi \circ C_\psi$$

DeepInfomax - Global objective

- Estimate the MI training a classifier to distinguish between samples coming from the joint J and the product of marginals M

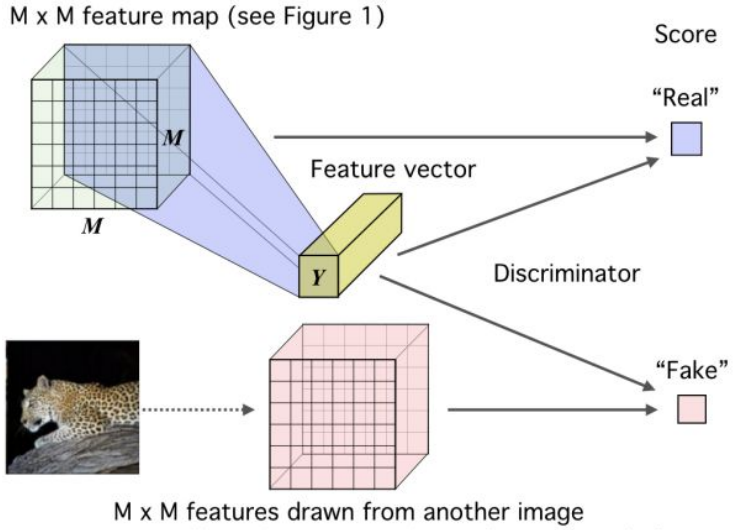
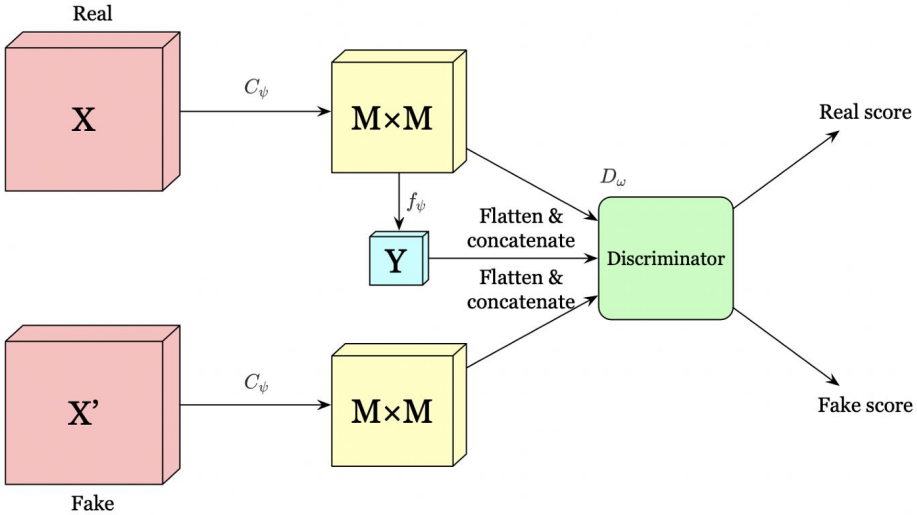


Figure 2: **Deep InfoMax (DIM) with a global MI$(X;Y)$ objective.** Here, we pass both the high-level feature vector, Y, and the lower-level $M \times M$ feature map (see Figure 1) through a discriminator to get the score. Fake samples are drawn by combining the same feature vector with a $M \times M$ feature map from another image.

$$(\hat{\omega}, \hat{\psi})_G = \underset{\omega,\psi}{\operatorname{argmax}} \widehat{\mathcal{I}}_\omega(X; E_\psi(X))$$

DeepInfomax - Global objective

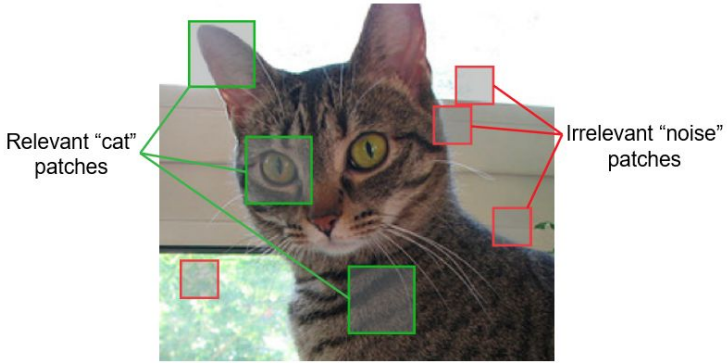
A better visualization:



$$(\hat{\omega}, \hat{\psi})_G = \underset{\omega,\psi}{\operatorname{argmax}} \widehat{\mathcal{I}}_\omega(X; E_\psi(X))$$

Lacks of Global MI Maximization

- An encoder maximizing the MI between the input and output yields representations that contain trivial or “noisy” information from the input.
- global DIM biased toward learning unrelated features, as their sum has more unique information than redundant locations.



- Want to maximize information that is shared across the input—in this case, across relevant locations.
- To accomplish this, maximize the mutual information between high-level representation and local patches

Local MI Maximization

- The high-level representation *Y* is encouraged to have high mutual information with all patches
- This favors encoding aspects of the data that are shared across patches

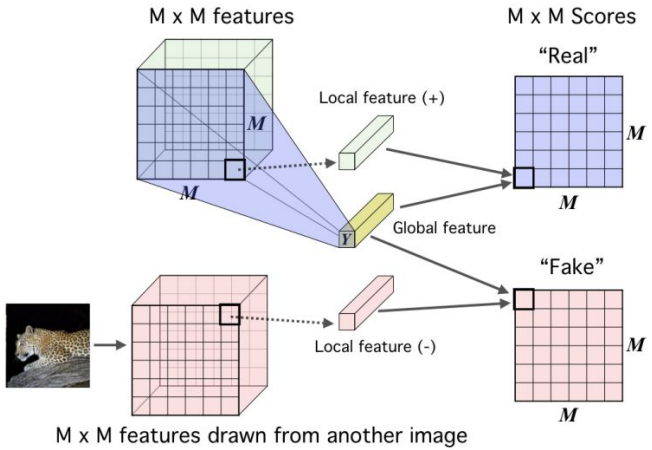
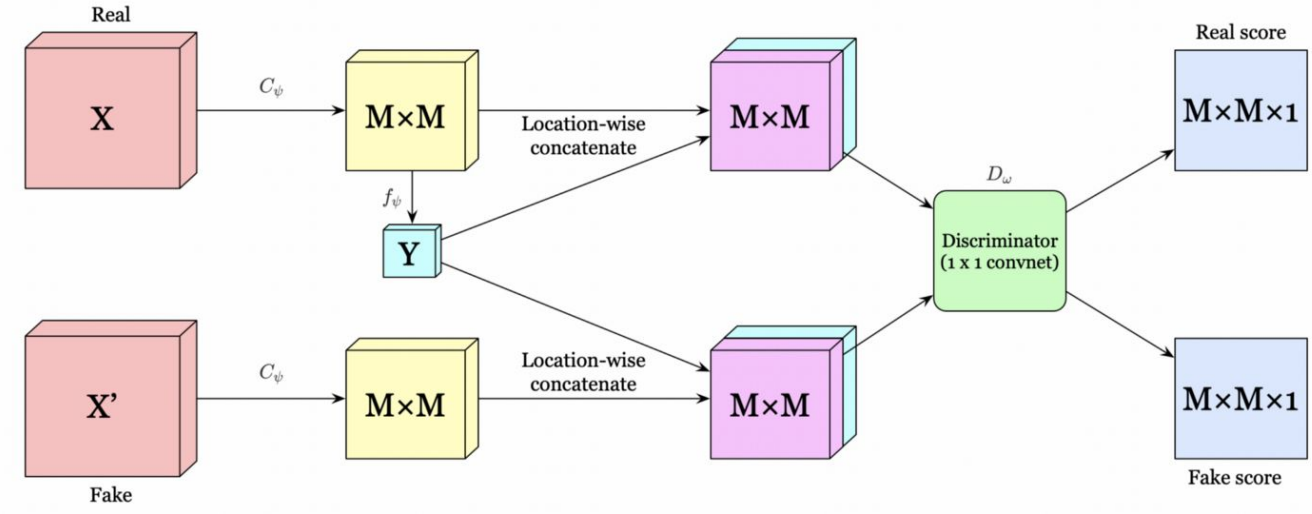


Figure 3: **Maximizing mutual information between local features and global features.** First we encode the image to a feature map that reflects some structural aspect of the data, e.g. spatial locality, and we further summarize this feature map into a global feature vector (see Figure 1). We then concatenate this feature vector with the lower-level feature map *at every location*. A score is produced for each local-global pair through an additional function (see the Appendix A.2 for details).

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

Local MI Maximization

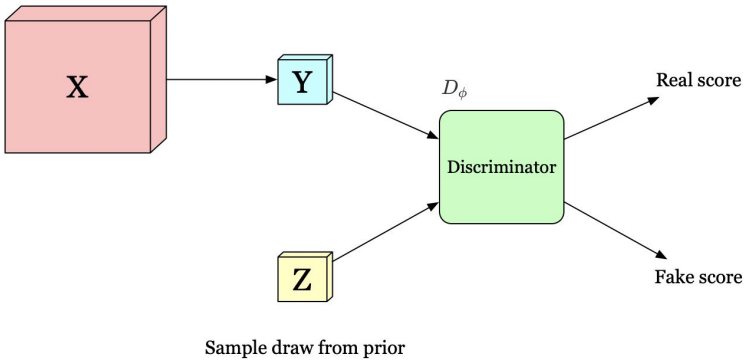
A better visualization:



$$(\hat{\omega}, \hat{\psi})_L = \underset{\omega,\psi}{\operatorname{argmax}} \frac{1}{M^2} \sum_{i=1}^{M^2} \widehat{\mathcal{I}}_{\omega,\psi}(C_\psi^{(i)}(X); E_\psi(X))$$

Match the representation to a Prior

- Impose structural constraint to the obtained representation
  - expect E_ψ to learn a representation with some desirable properties
    - independence, disentanglement
- Adversarial learning - try to fool a discriminator which distinguishes if the input distribution is from the output of the encoder or from the prior distribution



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

Some results

Table 1: Classification accuracy (top 1) results on CIFAR10 and CIFAR100. DIM(L) (i.e., with the local-only objective) outperforms all other unsupervised methods presented by a wide margin. In addition, DIM(L) approaches or even surpasses a fully-supervised classifier with similar architecture. DIM with the global-only objective is competitive with some models across tasks, but falls short when compared to generative models and DIM(L) on CIFAR100. Fully-supervised classification results are provided for comparison.

Model	CIFAR10			CIFAR100		
	conv	fc (1024)	$Y(64)$	conv	fc (1024)	$Y(64)$
Fully supervised		75.39			42.27	
VAE	60.71	60.54	54.61	37.21	34.05	24.22
AE	62.19	55.78	54.47	31.50	23.89	27.44
β-VAE	62.4	57.89	55.43	32.28	26.89	28.96
AAE	59.44	57.19	52.81	36.22	33.38	23.25
BiGAN	62.57	62.74	52.54	37.59	33.34	21.49
NAT	56.19	51.29	31.16	29.18	24.57	9.72
DIM(G)	52.2	52.84	43.17	27.68	24.35	19.98
DIM(L) (DV)	**72.66**	**70.60**	**64.71**	**48.52**	**44.44**	**39.27**
DIM(L) (JSD)	**73.25**	**73.62**	**66.96**	**48.13**	**45.92**	**39.60**
DIM(L) (infoNCE)	**75.21**	**75.57**	**69.13**	**49.74**	**47.72**	**41.61**

Biases in approximate Information Maximization

- Folklore knowledge that maximizing MI does not necessarily lead to useful representations
  - Linsker (1988)
  - Bridle et al. (1992)

- To what can we attribute the recent success of several works? (DeepInfomax, CMC, CPC)
  - Loose connection to the InfoMax principle
  - want to show they counter-intuitive behave if one equates them to MI maximization
  - performance depend on bias encoded by **encoders** and **estimators**

Setup of Tschannen et al. (2020)

- Learning a representation of the top half of MNIST images (CIFAR10)
  - x_{top} (corresponding to $X^{(1)}$), x_{bottom} (corresponding to $X^{(2)}$)

- Downstream linear evaluation protocol
  - train a linear classifier for digit classification on the learned representation using all train labels
- Train g_1,g_2, f using ADAM
  - use a bilinear critic for $f(x, y) = x^\top W y$
- Baseline
  - linear classifier on pixel space on x_{top} (test accuracy 85%)

Remember:

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

1) Large MI is not predictive of downstream performance

- Consider bijective encoders (RealNVP, Dinh et al. 2016) g_1 and g_2
  - Remember that if $X' = f_1(X)$ and $Y' = f_2(Y)$ are homeomorphism (i.e. smooth invertible maps), then

$$I(X, Y) = I(X', Y')$$

- Hence for any choice of the encoder parameters, the MI is constant

$$I(g_1(X^{(1)}); g_2(X^{(2)})) = I(X^{(1)}; X^{(2)}) \text{ for all } g_1, g_2.$$

- If we can compute the exact MI, any parameter choice would be a global maximizer
  - gradients vanish everywhere
  - any instantiation of g_1 and g_2

However, the representation quality improves during training

Biased estimators

1) Large MI is not predictive of downstream performance

a) Maximized MI and improved downstream performance

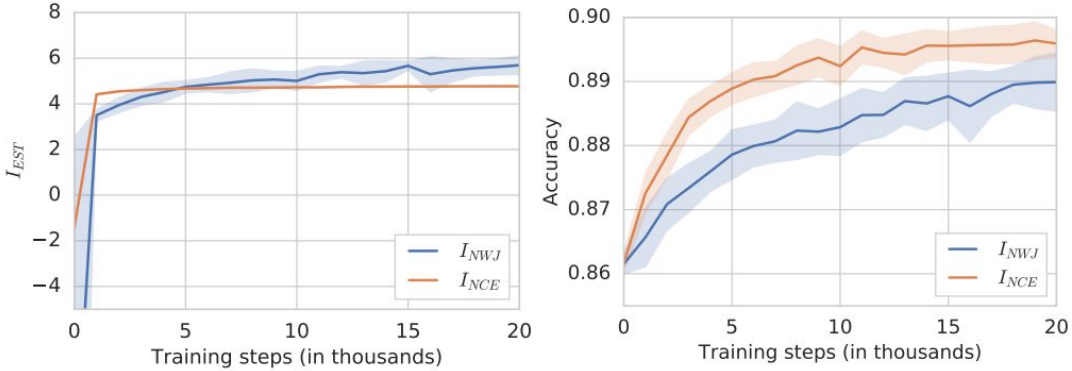


Figure 1: (a, b) Maximizing I_{EST} over a family of invertible models. We can see that during training the downstream classification performance improves (and the testing I_{EST} value increases), even though the true MI remains constant throughout.

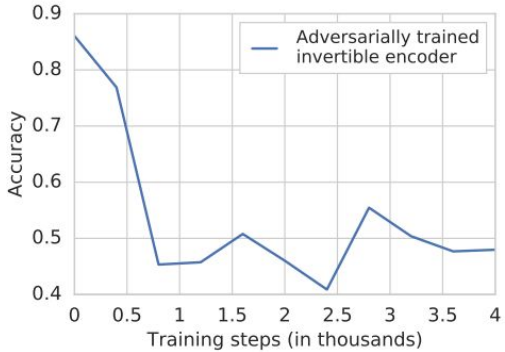
- Despite the fact that MI is maximized for every instantiation of g_1 and g_2 , **I_{EST} and downstream accuracy increase**
- Estimators provide gradient leading to a representation useful for linear classification
  - estimators biases encoder towards solution suitable to the downstream task
- Among many invertible encoders (all globally optimal MI maximizers), some give rise to improved linear classification performance

1) Large MI is not predictive of downstream performance

b) Maximized MI and worsened downstream performance

- **For the same invertible encoders, there are parameters for which linear classification is worse than using raw pixels**
  - despite also being globally optimal MI maximizers
- Achieve this by adversarial training (encoder vs a linear classifier)
  - train the encoder to make hard the classification task for a linear classifier
  - a separate classifier is trained for the downstream evaluation

(c) Downstream classification accuracy of a different invertible encoder (with the same architecture) trained to have poor performance. This demonstrates the existence of encoders that provably maximize MI yet have bad downstream performance.



(c)

2) Bias Towards hard-to-invert encoders

- Use a network architecture that can model both invertible and non-invertible functions
  - I_{EST} prefers the net to remain bijective (thus maximizing the true MI - initialized as identity) or to ignore part of input signal?
  - To quantify invertibility, analyze the condition number of the Jacobian of g_1

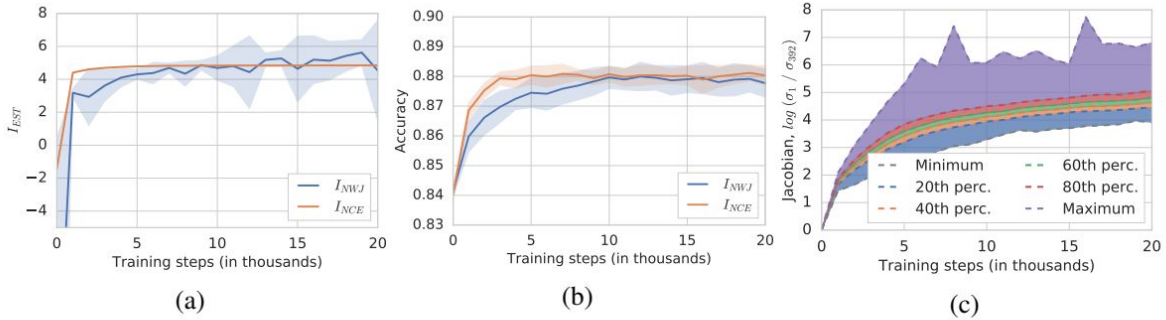


Figure 2: Maximizing I_{EST} using a network architecture that can realize both invertible and non-invertible functions. (a, b) As I_{EST} increases, the linear classification testing performance increases. (c) Meanwhile, the condition number of Jacobian evaluated at inputs randomly sampled from the data distribution deteriorates, i.e. g_1 becomes increasingly ill-conditioned (lines represent 0th, 20th, . . . , 100th percentiles for I_{NCE}, the corresponding figure for I_{NWJ} can be found in Appendix F; the empirical distribution is obtained by randomly sampling 128 inputs from the data distribution, computing the corresponding condition numbers, and aggregating them across runs).

2) Bias Towards hard-to-invert encoders

- Proved that:
  - during training, inverting the model becomes increasingly hard

Hence

- the bound prefer hard-to-invert encoders, which heavily attenuate part of the noise
  - they do not maximize the true MI

- well conditioned encoders which preserve the noise are not preferred
  - preserve the noise, hence, the entropy of the data

- **MI and downstream performance are only loosely connected**

3) Loose bounds can led to better representations

- **How the critic architecture impacts the quality of the learned representation?**

  - Remind the role of the critic f: distinguish between samples from joint distr. and product of marginals
    - determines the tightness of the lower bound

  - an higher capacity critic should allow for a tighter lower-bound on MI (Belghazi et al. 2018)

  - f is a neural net, provides gradient feedback to g_1 and g_2
    - shapes the learned representation

3) Loose bounds can led to better representations

Simple bilinear critic leads to better downstream performances

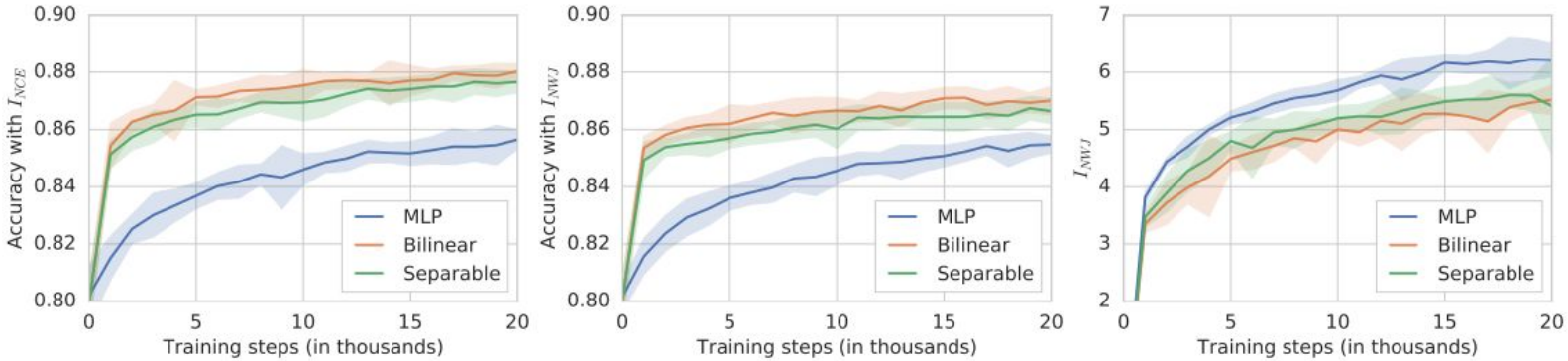


Figure 3: Downstream testing accuracy for I_{NCE} and I_{NWJ}, and testing I_{NWJ} value for MLP encoders g_1, g_1 and different critic architectures (the testing I_{NCE} curve can be found in Appendix F). Bilinear and separable critics lead to higher downstream accuracy than MLP critics, while reaching lower I_{NWJ}.

4) Representation quality impacted more by the choice of encoder than the estimator

- Optimize the estimators to the same MI lower bound
  - with different encoder architectures (MLP, ConvNet)

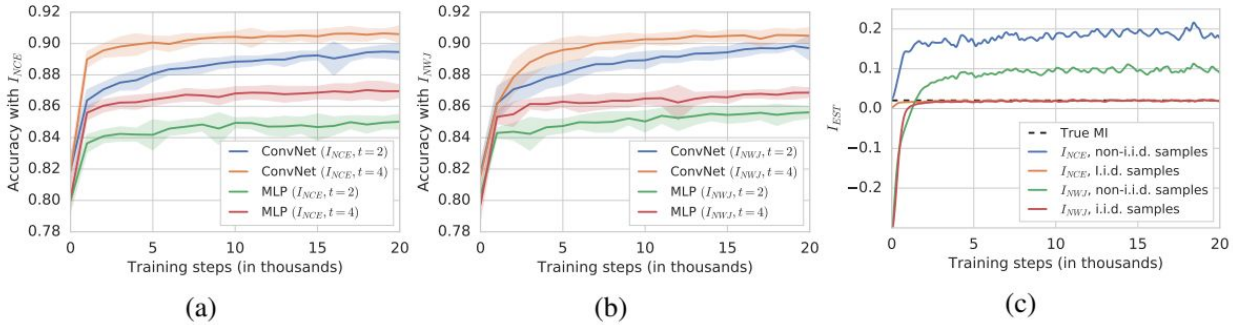


Figure 4: (a, b) Downstream testing accuracy for different encoder architectures and MI estimators, using a bilinear critic trained to match a given target I_{EST} of t (we minimize $L_t(g_1, g_2) = |I_{\mathrm{EST}}(g_1(X^{(1)}); g_1(X^{(2)})) - t|$; loss curves can be found in Appendix F). For a given estimator and t, ConvNet encoders clearly outperform MLP encoders in terms of downstream testing accuracy. (c) Estimating MI from *i.i.d.* and non-*i.i.d.* samples in a synthetic setting (Section 4). If negative samples are not drawn *i.i.d.*, both I_{NCE} and I_{NWJ} estimators can be *greater* than the true MI. Despite being commonly justified as a lower bound on MI, I_{NCE} is often used in the non-*i.i.d.* setting in practice.

Conclusions

- Is MI maximization a good objective for learning good representation in unsupervised fashion?
    - possibly, but clearly not sufficient

- Estimators have strong inductive biases
- looser bounds on MI can lead to better representations
- unclear whether the connection to MI is a sufficient/necessary component for powerful unsupervised representation learning

Suggestions

- **Alternative measures of information**
  - MI is not sufficient for representation learning (hard to estimate, invariant to bijections,..)
  - use a notion of information accounting both the amount of stored info and the **geometry** of the induced space
    - F-information Xu et al. (2020)
    - other statistical divergences to measure discrepancy between p(x,y) and p(x)p(y)
      - Wasserstein distance forces smoothness in encoders

Suggestions

- **Holistic view**
  - Downstream performance depends on intricate balance between choices of
    - critic used to measure info
    - encoders
    - downstream models/evaluation protocol
  - Might be possible to rely on weaker assumptions (i.e. invariances relevant for the downstream tasks)
- **Go beyond widely used linear evaluation protocol**
- **Investigation into design decision that matters**
  - new methods that take away from goal of estimating MI and place more weight on aspects having stronger effects on performances
    - negative sampling strategy

THANKS

On Mutual Information Maximization For Representation Learning

An excursus into Deep InfoMax, its variants and their lacks

Matteo Tiezzi