On Mutual Information Maximization For Represntation Learning

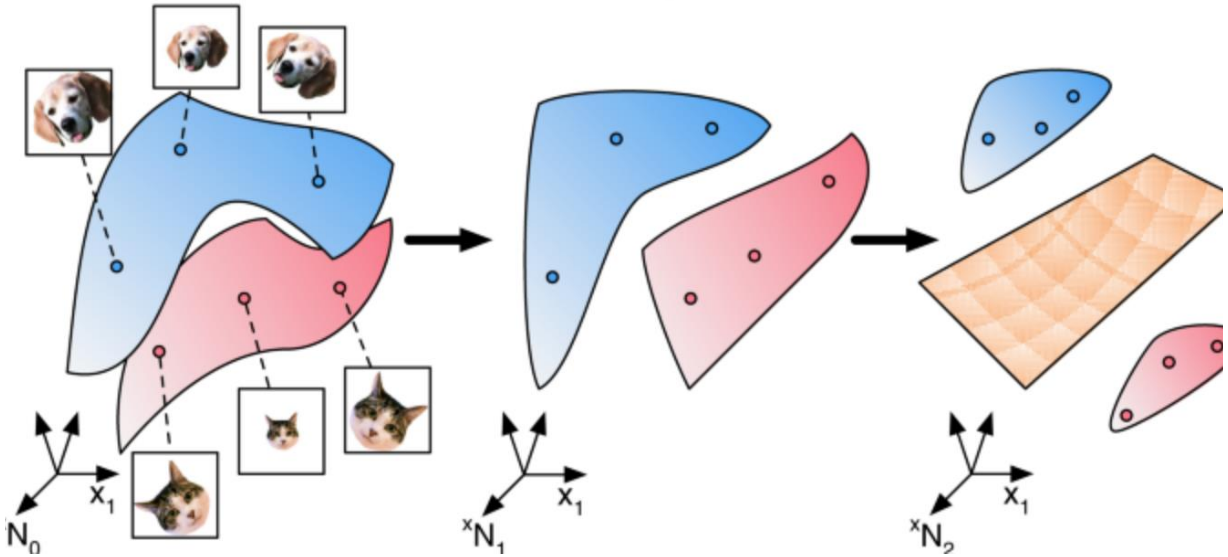
[Tschannen et al., ICLR 2020]

-Summary-

Introduction

- Dealing with unsupervised representation learning:
  - Goal : learn a function g which maps the data into lower-dimensional space

    (where we can solve some supervised tasks more efficiently)



Simple description of unsupervised representation learning

Introduction

Claim of this paper
: **Maximizing tighter lower bounds on MI can result in worse representations.**

- Recent approach : InfoMax principle (Linsker, 1998)
  - Choose a representation $g(x)$ maximizing mutual information (MI) between the input and its representation:

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

  - However, estimating MI in high-dimensional is notoriously difficult task
    - In practice, we usually maximizes a tractable variational lower bound of MI (Poole et al., 2019)
    - Using this method, several recent works have demonstrated promising empirical results in representation learning using MI maximization(ex : using I_{NCE}, I_{NWJ})

Background and related work

This paper focused on unsupervised **'image'** representation learning.

- Usual problem setup (~ Becker and Hinton, 1992) : Multi-view formulation
  - For a given image X, let $X^{(1)}$ and $X^{(2)}$ be different **views** of X.

    (ex : different cropped images, top and bottom halves of the image)
  - We focus below problem rather than original MI maximization :

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

where I_{EST} : samples-based estimator of the true MI

Note :

$\boldsymbol{I\left(g_1\left(X^{(1)}\right); g_2\left(X^{(2)}\right)\right) \leq I\left(X; g_1\left(X^{(1)}\right), g_2\left(X^{(2)}\right)\right) = I\left(X; g(X)\right)}$ ***by data-processing inequality.***

Thus, our problem can be interpreted as maximizing the lower bound of $\boldsymbol{I\left(X; g(X)\right)}$

Background and related work

- Q : Why we use multi-view formulations?

  1. **[fundamental reason]** the MI has to be estimated only between the learned representations of the two views.
  2. it give us various modeling flexibility.
     (① : how to choose the **objective** $\boldsymbol{I_{EST}}$, ② : how to define **two views** of an sample)

- For example :

  ① : choose $X^{(1)}$ = upper half of X , $X^{(2)}$ = lower half of X

  ② : choose variational lower bounds of MI (= $I_{NWJ}, I_{NCE}, \ldots$)

Background and related work

- If we assume usage of variational lower bounds such as I_{NCE}, I_{NWJ} as belows :

$$I_{NCE} := \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K}e^{f(x_i,y_j)}}\right], \qquad I_{NWJ} := \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}f(x,y) - e^{-1}\cdot\frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}e^{f(x,y)}\right]$$

where expectation is taken over a batch (= K independent samples $\{(x_i, y_i)\}_{i=1}^{K}$)

Note : lower bounds get tight when $f^*(x,y) = \log p(y|x)$ [I_{NCE}] or $f^*(x,y) = 1 + \log p(y|x)$ [I_{NWJ}]

- We train 'critic function' f to maximize I_{NCE} or I_{NWJ}. (= choosing $I_{EST} = \max_f I_{NCE(or\ \mathrm{NWJ})}$)

Note [Common architectures for f] :

① bilinear : $f(x,y) = x^T W y$, ② separable : $f(x,y) = \phi_1(x)^T \phi_2(y)$, ③ concatenated : $f(x,y) = \phi([x,y])$

where ϕ, ϕ_1, ϕ_2 : (typically) shallow multi-layer perceptrons (MLPs)

Arising question from InfoMax principle

- Intuitively, we discriminate two distributions $p(x, y)$ and $p(x)p(y)$ by maximizing mutual information $I(X; Y) = D_{KL}(p(x, y)|p(x)p(y))$ when we adopt InfoMax principle.

- However, it does not imply the learning of useful representations (Linsker, 1998)
  - Also, some issues occurred in clustering problem when we adopts MI criterion (Bridle et al., 1992)

- **<u>Then, what is the critical factor which leads to recent success of representation learning based on InfoMax principle</u>**? (candidates : InfoMax, architectures of encoder / critic, I_{est})

  (This paper argues that the connection between InfoMax and useful representations can be very loose.)

Claims and experiments

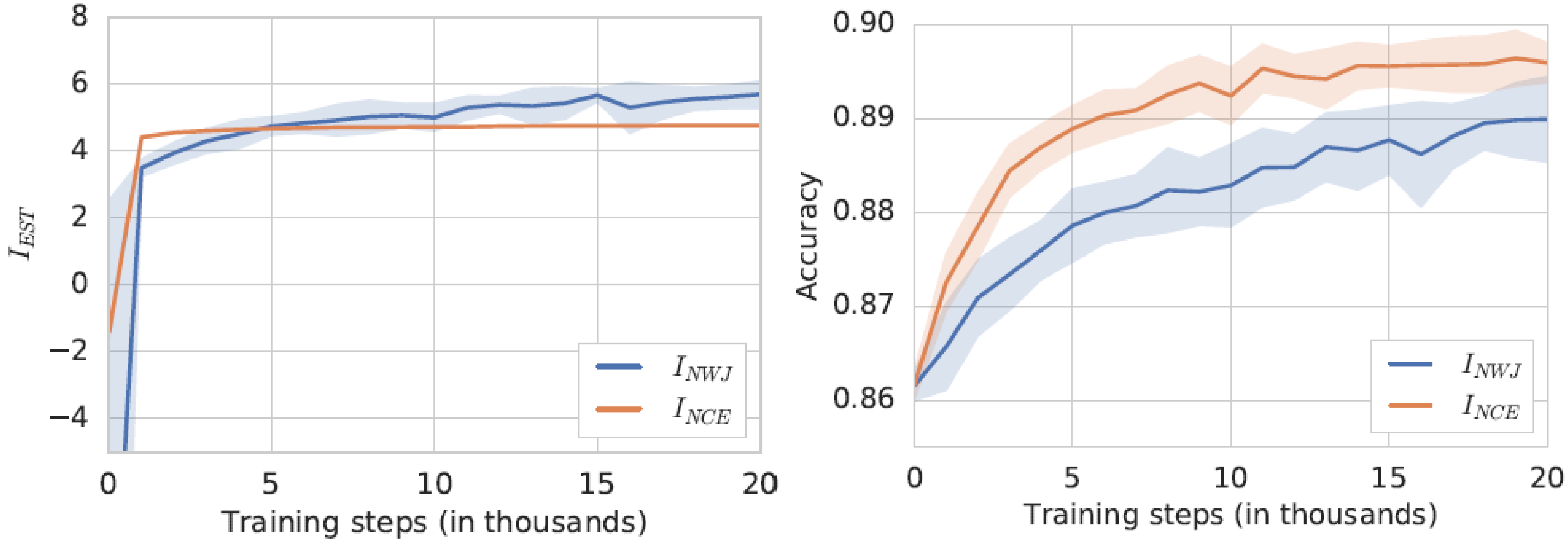
- Claim ① : Large MI is not predictive of good downstream performance.
  - Using invertible encoder (RealNVP, 2016), we can fix the MI as the constant as

$$I\left(g_1\left(X^{(1)}\right); g_2\left(X^{(2)}\right)\right) = I\left(X^{(1)}; X^{(2)}\right)$$

  - **1st experiment performs training via InfoMax principle on invertible encoders**
    - Although the true MI is fixed, I_{EST} and downstream performance get increased during the training
    - This confirms that the **MI estimator biases the encoders towards solutions suitable to solve the downstream linear classification task (despite of fixed MI).**

Claims and experiments

Data : MNIST / $x^{(1)}$: upper half image / $x^{(2)}$: lower half image / Critic f : bilinear
Note : Use only $g_1(x^{(1)})$ as the representation for the linear evaluation



Left : Maximizing I_{EST} over invertible models

Right : Downstream classification performance (by linear evaluation protocol)

Claims and experiments

- Claim ① : Large MI is not predictive of good downstream performance. (InfoMax)
    - **2nd experiment performs adversarial training of encoder and classifier**
        - By doing so, the encoder is trained to make the classifier to predict as hard as possible.
        - Here, training of encoder is not done by InfoMax principle, but by adversarial stage to deliberately make poor quality encoder.

Details of adversarial training :

Adversarial stage : encoder + (temporary) classifier

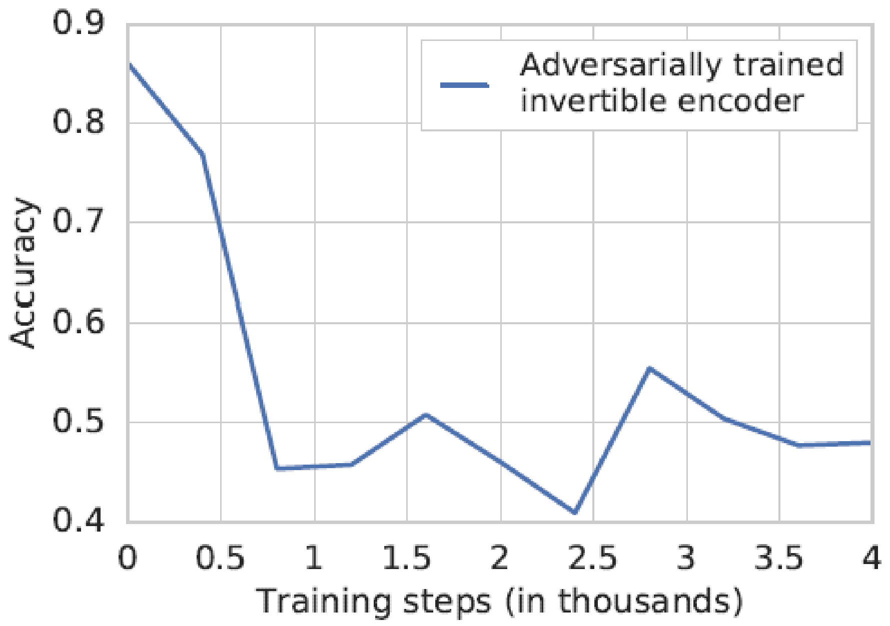
-> encoder : minimizing CE loss with uniform label / classifier : minimizing CE loss with true label

Linear evaluation stage : encoder + (new) classifier

-> encoder : brought and fixed from above stage / classifier : minimizing CE loss with true label

Claims and experiments

- Claim ① : Large MI is not predictive of good downstream performance.



Downstream classification accuracy of a adversarially trained invertible encoder

Note : this demonstrates the existence of encoders that maximize MI yet have bad downstream performance.

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

Claims and experiments

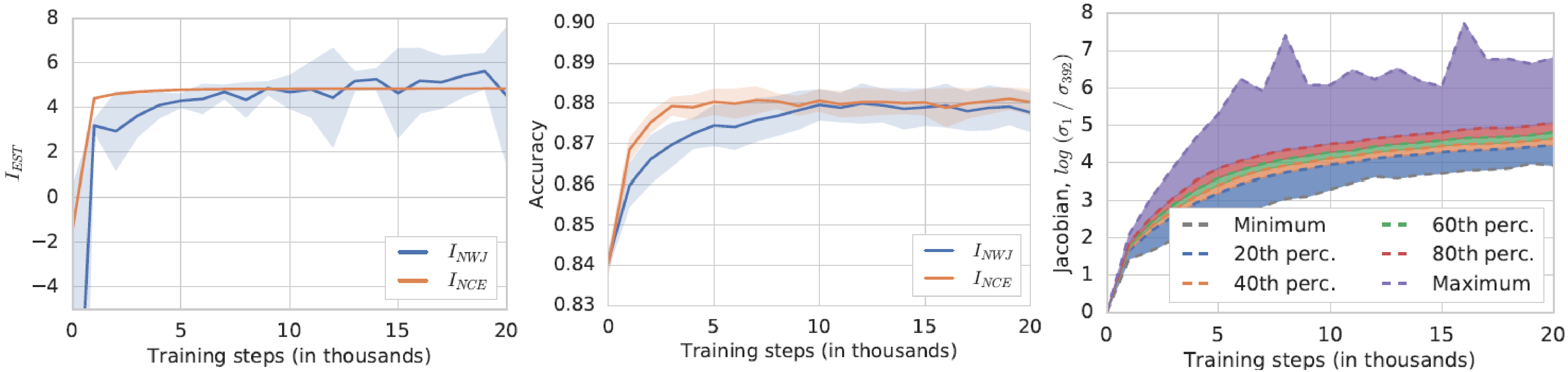
Condition number of matrix $A := \frac{\sigma_{max}(A)}{\sigma_{min}(A)} = \|A\| \cdot \|A^{-1}\|$

- Claim ① : Large MI is not predictive of good downstream performance. (InfoMax)
  - **3rd experiment shows training with InfoMax principle biases model towards hard-to-invert encoders**
  - Here, we use MLP architecture encoder which can be both invertible / non-invertible (adding skip connection added tot each layer)
  - Recall that function is invertible ⇔ input Jacobian is invertible (Implicit function theorem)
    - To quantifying the 'invertibility' of jacobian, we use condition number of Jacobian (Higher condition number of jacobian => Harder to invert the jacobian)

Claims and experiments

Condition number of matrix $A := \frac{\sigma_{max}(A)}{\sigma_{min}(A)} = \|A\| \cdot \|A^{-1}\|$

- Claim ① : Large MI is not predictive of good downstream performance. (InfoMax)
  - Even though encoder is initialized to very close to identify function, the condition number of its Jacobian evaluated at randomly sampled inputs deteriorates over times.
  - This implies the objective (I_{NWJ}, I_{NCE}) biases the encoder towards hard-to-invert models.



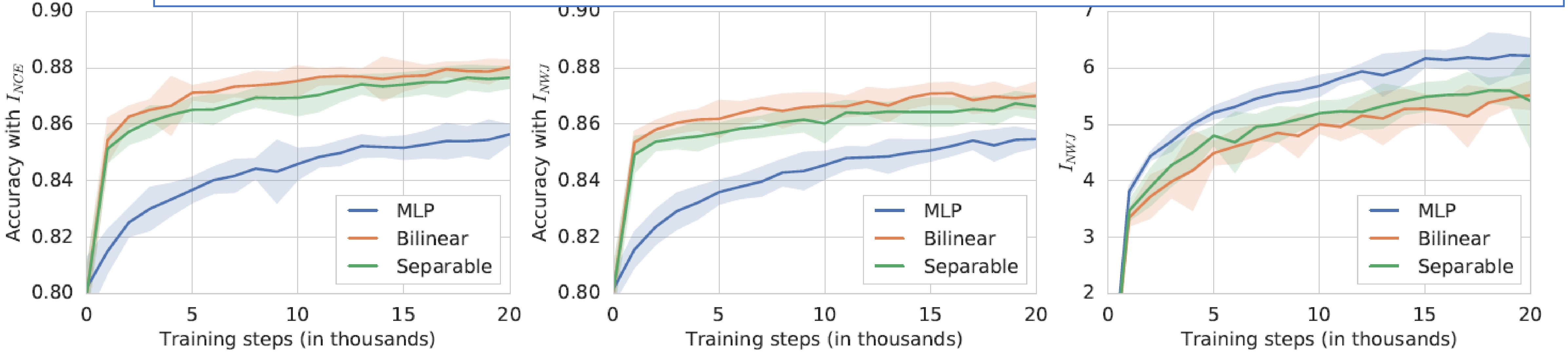
Left : I_{EST} during training / Middle : Downstream accuracy with I_{NWJ}, I_{NCE} / Right : log($\kappa(Jacobian)$) during training

Claims and experiments

- Claim ② : Higher capacity critics can lead to worse downstream performance. (critic archit.)
  - Higher capacity critic should allow for a tighter lower-bound on MI (Belghazi et al., 2018)
  - Here, we compare bilinear / separable / concatenate(MLP) critic f architecture

    (Note : # of parameters : bilinear (=10k) << separable = concatenate (= 40k))

Although the MLP estimate true MI better, it shows worse downstream performance than simpler model (Bilinear)



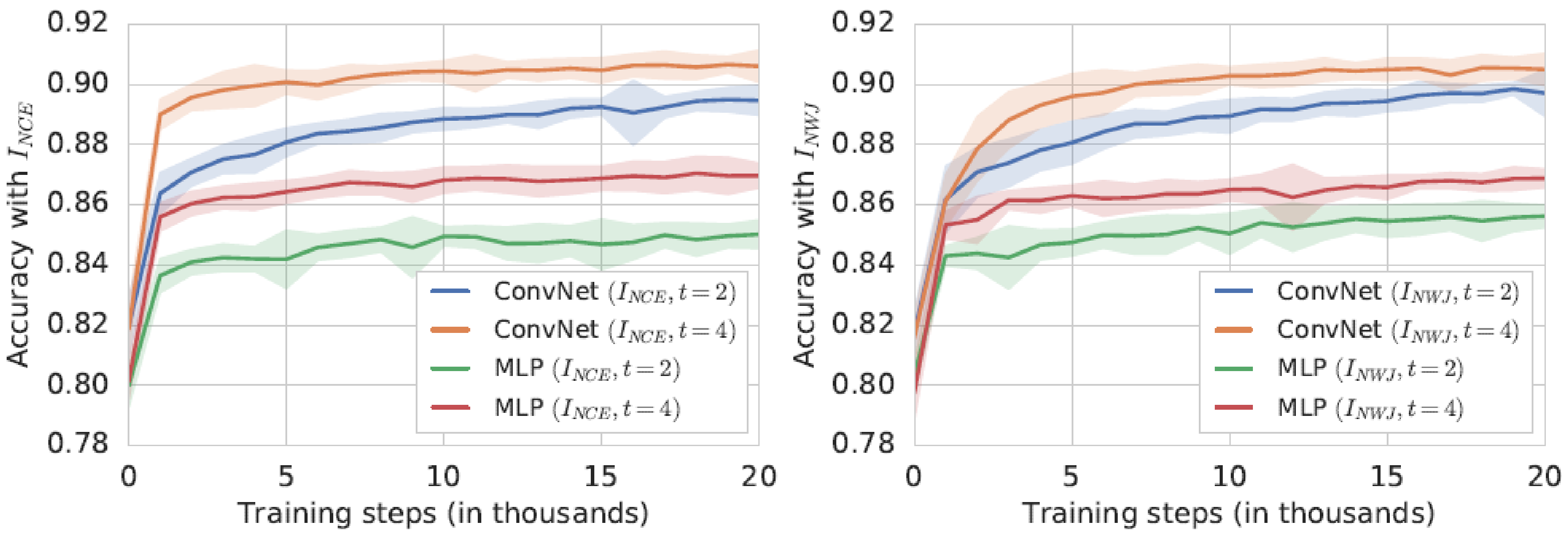
Left : Downstream accuracy with I_{NCE} / Middle : Downstream accuracy with I_{NWJ} / Right : I_{NWJ} value

Claims and experiments

- Claim ③ : Encoder architecture can be more important than the specific estimator.
    - Here, we compare downstream performance from MLP / ConvNet encoder when they have the same estimate of MI (I_{NCE}, I_{NWJ}).

We minimize the loss $L_t(g_1, g_2) = |I_{EST}(g_1(X^{(1)}; g_2(X^{(2)}) - t|$

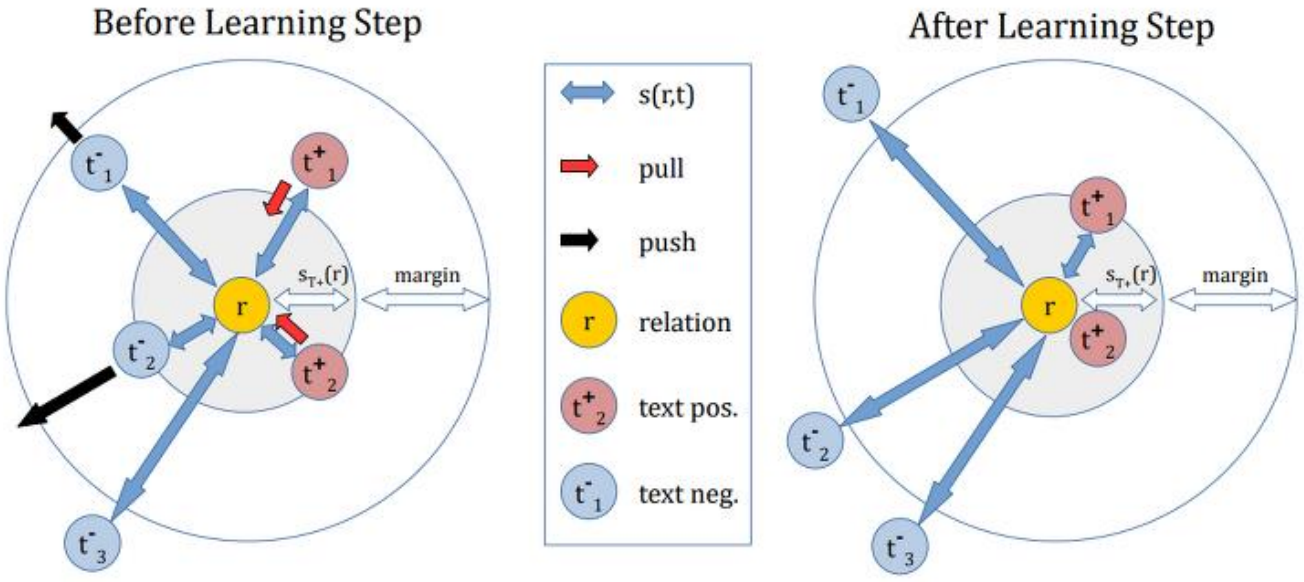
    - Despite of matching estimates of MI, ConvNet performs superiorly than MLP.



Left : Downstream accuracy with I_{NCE} / Right : Downstream accuracy with I_{NWJ}

Claims and experiments

- Claim ④ : Big connection to deep metric learning, which does not use notion of MI.
  - [Metric learning] : Given set of triplets (x, y, z) = (anchor, positive, negative), we want to learn g such that the distance between $g(x)$ and $g(y)$ becomes smaller and $g(x)$ and $g(z)$ becomes larger.



Brief depiction of Metric learning

Claims and experiments

- Claim ④ : Big connection to deep metric learning, which does not use notion of MI.
  - Although there is a loose connection between InfoMax and representation performance, why many recent works have applied I_{NCE} and achieved good performance?
  - Recall that I_{NCE} can be written as follows :

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

Claims and experiments

- Claim ④ : Big connection to deep metric learning, which does not use notion of MI.
  - One famous loss proposed in metric learning (multi-class-K-pair loss , 2016) :

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

which is the same as maximizing I_{NCE} with separable critic $f(x, y) = \phi(x)^T\phi(y)$.

  - Hence, the success of InfoMax principle with I_{NCE} can be attributed to it’s connection to metric learning. (So, many recent paper may call this method as ‘Contrastive Learning’)

Claims and experiments - Summary

- Claim ① : Large MI is not predictive of good downstream performance.

- Claim ② : Higher capacity critics can lead to worse downstream performance.

- Claim ③ : Encoder architecture can be more important than the specific estimator.

- Claim ④ : There is a big connection to deep metric learning, which does not use notion of MI.