Discriminability-Transferability Trade-Off: An Information-Theoretic Perspective

Accepted Paper @ ECCV 2022

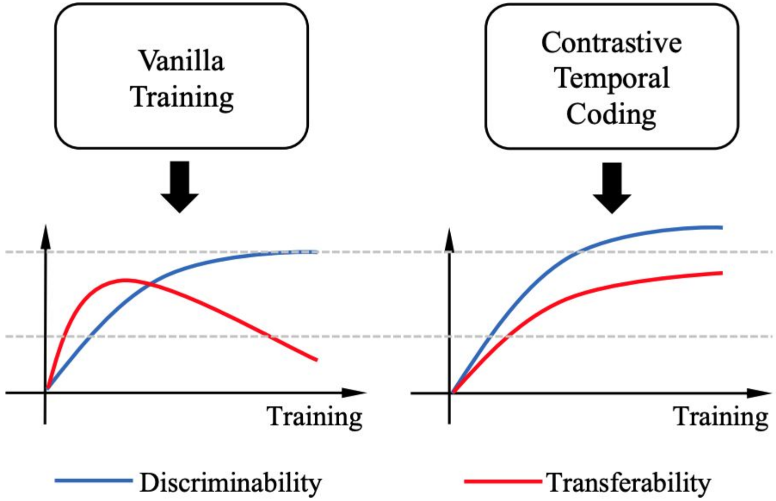
Quan Cui*, Bingchen Zhao*, Zhao-Min Chen, Borui Zhao, Renjie Song,
Boyan Zhou, Jiajun Liang, and Osamu Yoshie

*: Equal contribution

Motivation

We study two properties of a deep representation, namely: Discriminability and Transferrability.

- **Discriminability:** How well the representation performs the training task.
- **Transferrability:** How well the representation transfers to a new task.



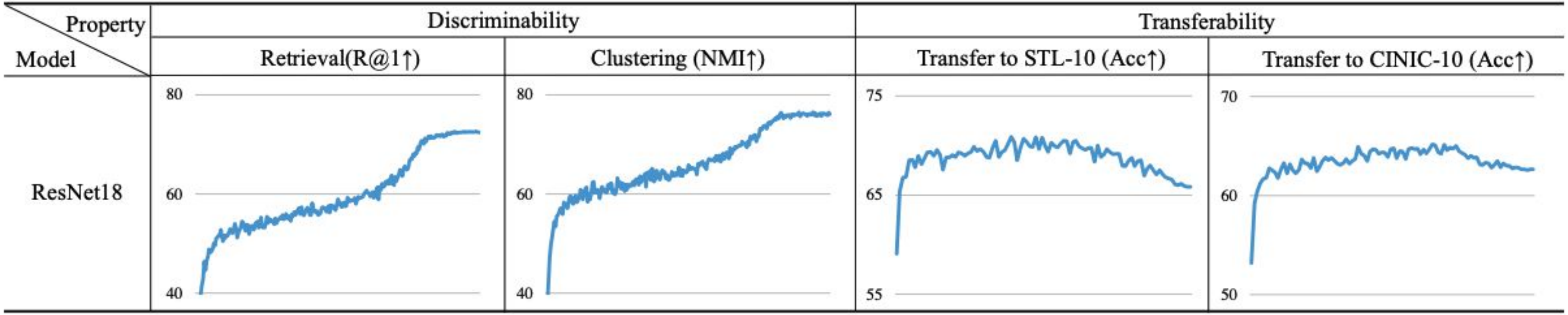
Temporal analyses of representations

- **Discriminability**

Recall@1 for retrieval and **NMI** for clustering on the training dataset. (*e.g.* CIFAR-100)

- **Transferability**

Top-1 Accuracy for linear probing a representation on out-of-sample datasets. (*e.g.* CIFAR->SVHN)



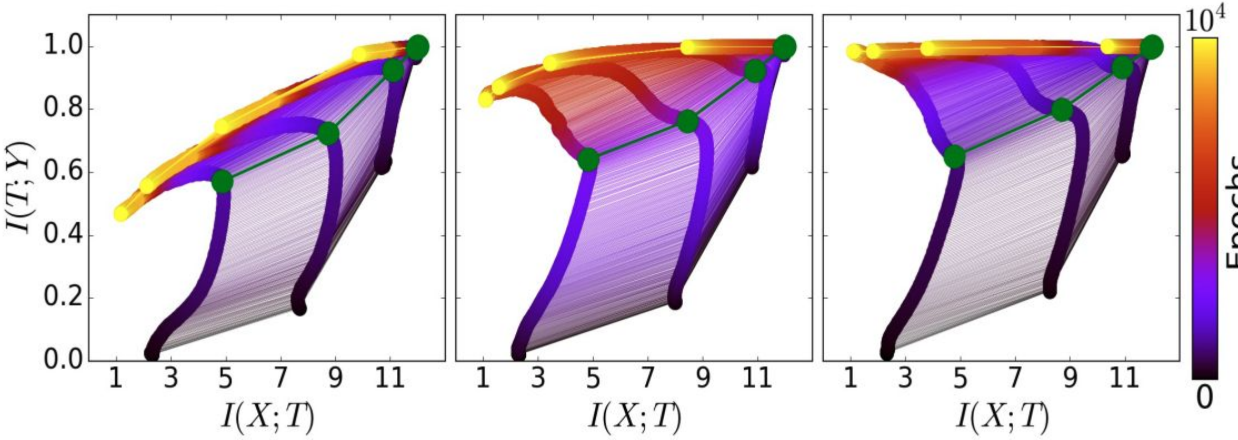
X-axis is the training epoch, with training progress, the **discriminability** keep climbing while **transferability** drops in later epochs.

Link to Information Bottleneck Trade-off

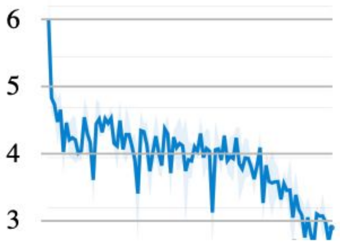
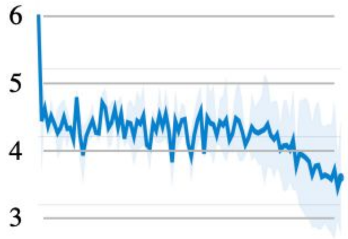
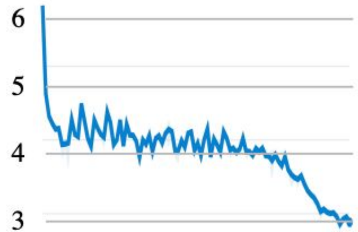
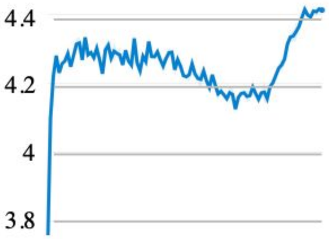
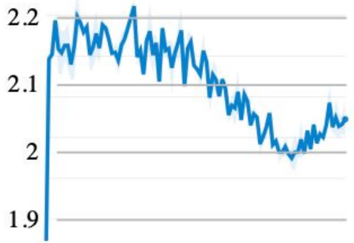
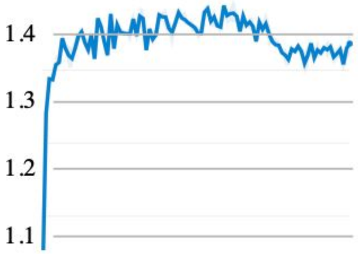
$$\min_{p(t|x),p(y|t),p(t)} \{I(X;T) - \beta I(T;Y)\} \;.$$

- **Fast ERM** phase
  I(T;Y) increases

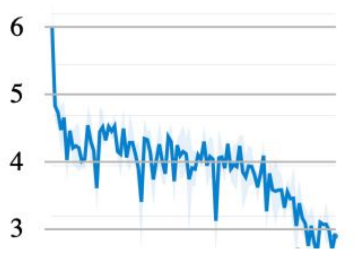
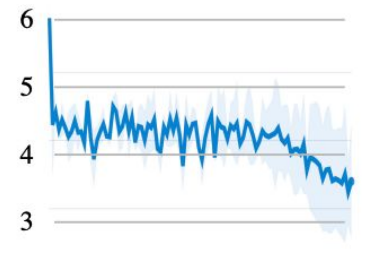
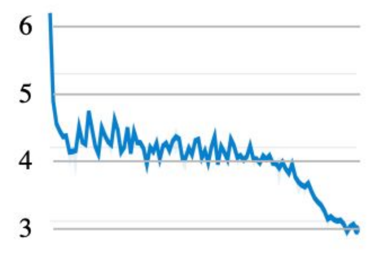
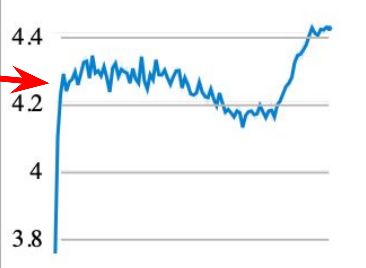
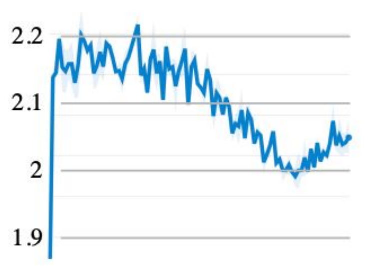
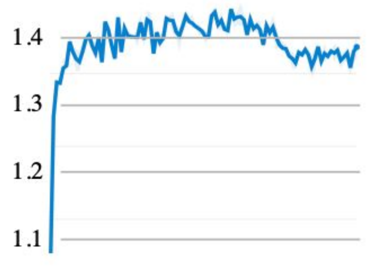
- **Compression** phase
  I(X;T) decreases



Mutual Informations

Metric	Source Dataset CIFAR-100	Target Dataset STL-10	Target Dataset CINIC-10
$I(X;T)$			
$I(T;Y)$			

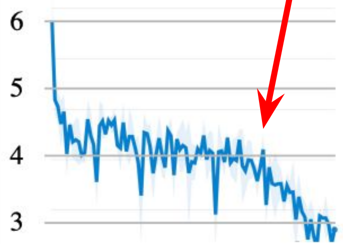
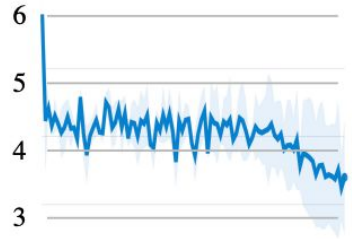
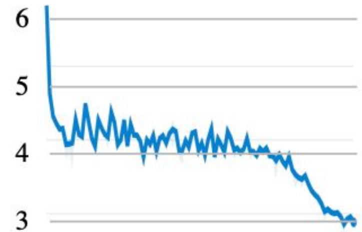
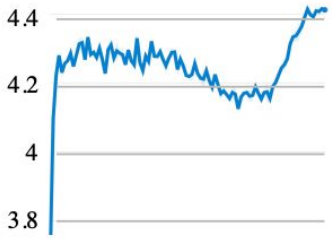
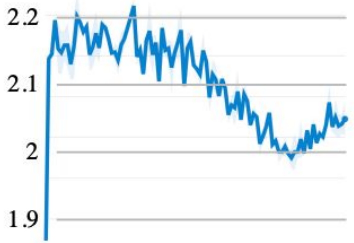
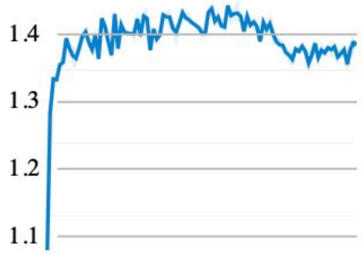
Mutual Informations

Metric	Source Dataset CIFAR-100	Target Dataset STL-10	Target Dataset CINIC-10
$I(X;T)$			
$I(T;Y)$			

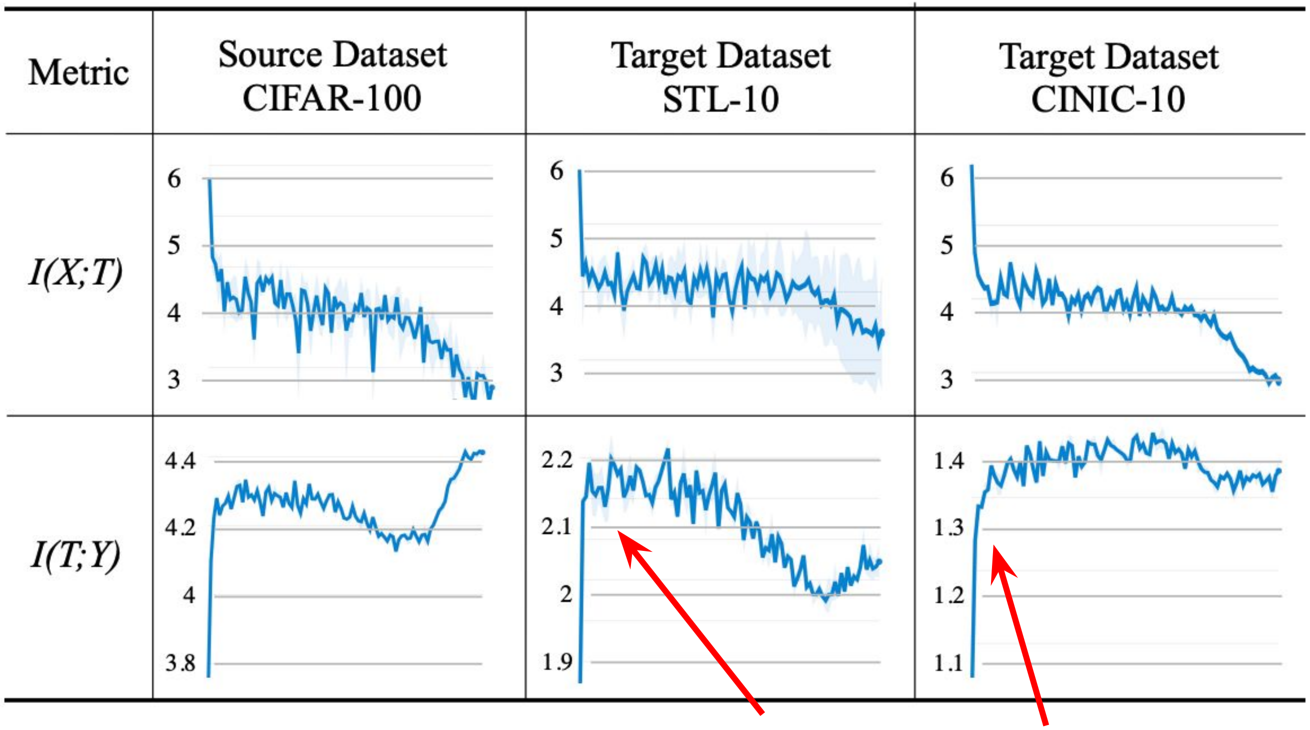
Fast ERM

Mutual Informations

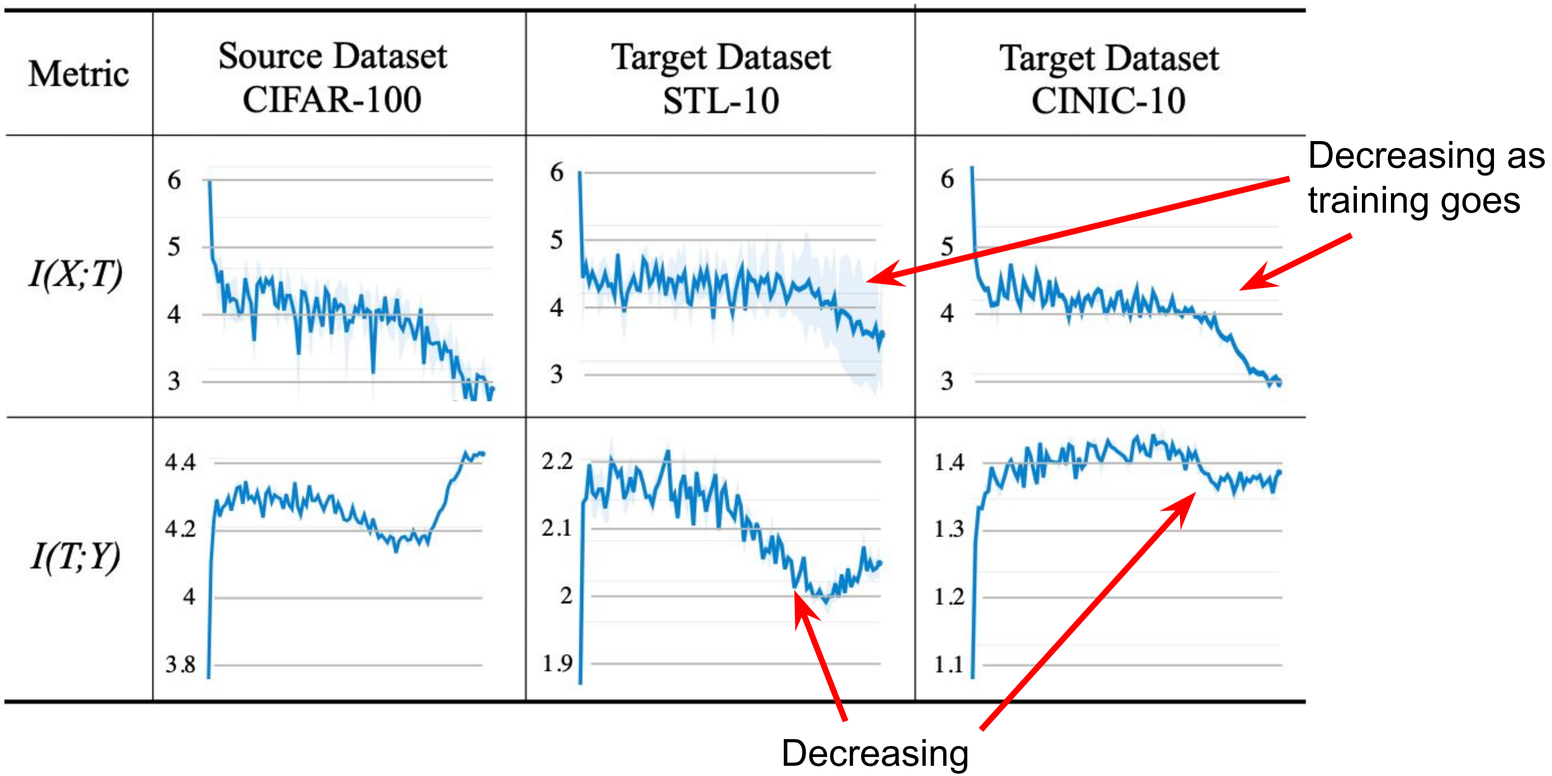
Compression

Metric	Source Dataset CIFAR-100	Target Dataset STL-10	Target Dataset CINIC-10
I(X;T)			
I(T;Y)			

Mutual Informations on Target Datasets

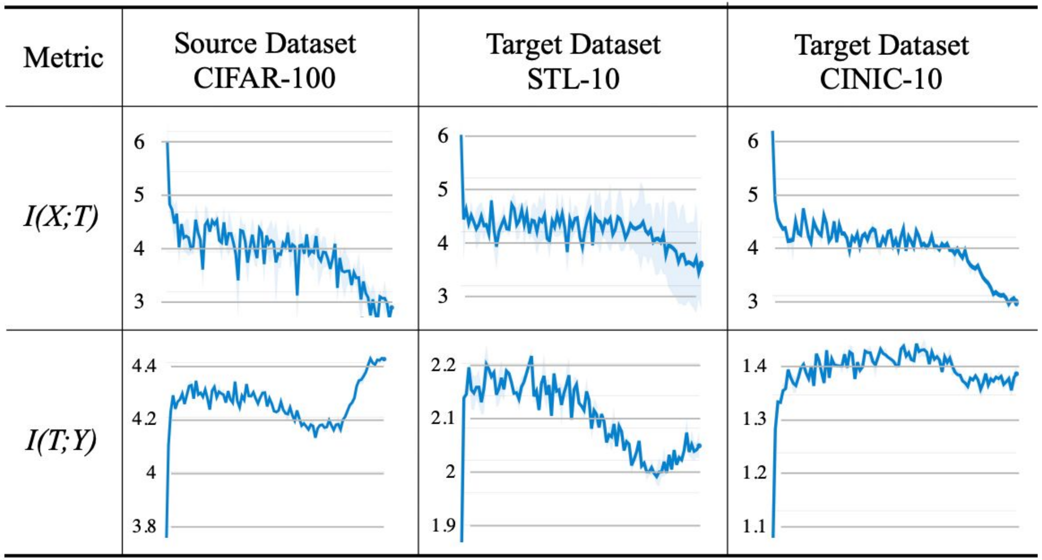


Overcompression

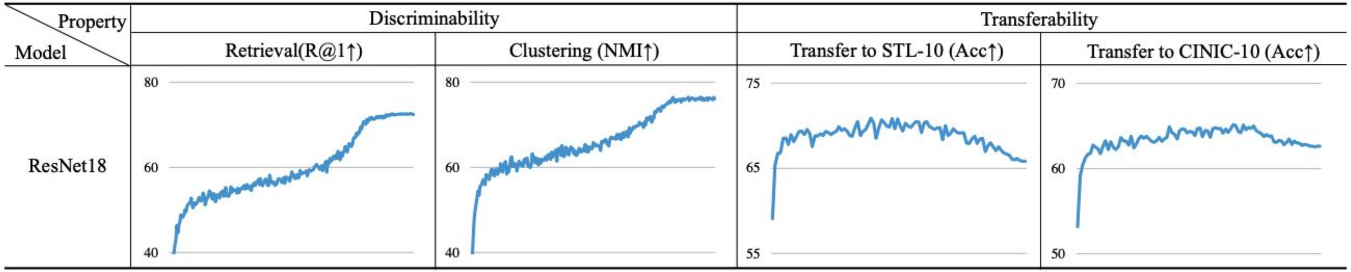


Overcompression

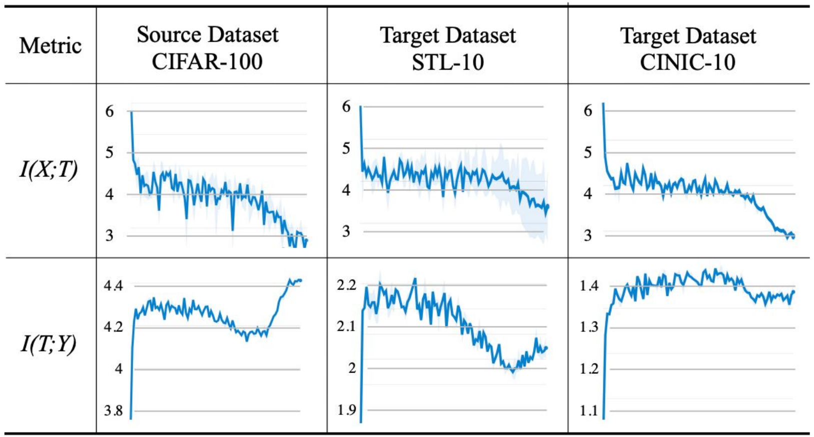
As training on source progress, *I(X;T)* and *I(T;Y)* on the target dataset both **degrades,** while *I(T;Y)* on source dataset keeps climbing



Overcompression



Overcompression correspond to the Discriminability-Transferability trade-off we discovered before.



Can Alleviating Over-compression help Transferability?

- **Transferability** depends on the model learn common representations between source dataset and target dataset.
- **Over-compression** happens when *I(X;T)* on target dataset decreases.
- Since target dataset is not available, we can alleviate the decrease of *I(X;T)* instead.

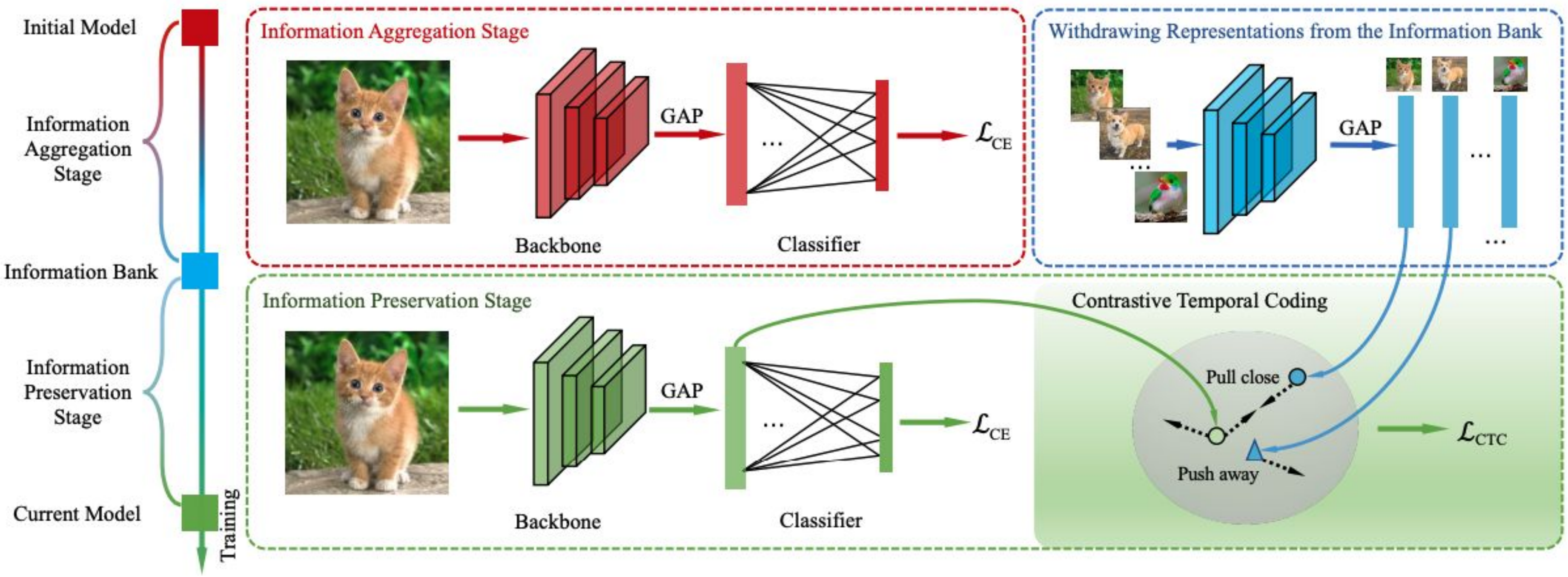
$$I(X;T) = H(T) - H(T|X),$$

Neural network is a deterministic mapping, so the conditional entropy is 0, thus:

$$I(X;T) = H(T),$$

Thus, a way to improve transferability via alleviating over-compression is to alleviating the decrease of *H(T)*

Alleviating over-compression via InfoNCE



How this works

InfoNCE loss between two variables has been shown to be optimizing for:

$$I(T_1; T_2) \geq \log(N) - \mathcal{L}_{\text{InfoNCE}},$$

Using the definition of mutual information:

$$\begin{aligned} I(T_1; T_2) &= H(T_1) + H(T_2) - H(T_1, T_2) \\ &\leq H(T_1) + H(T_2) - \max(H(T_1), H(T_2)), \\ &= \min(H(T_1), H(T_2)), \end{aligned}$$

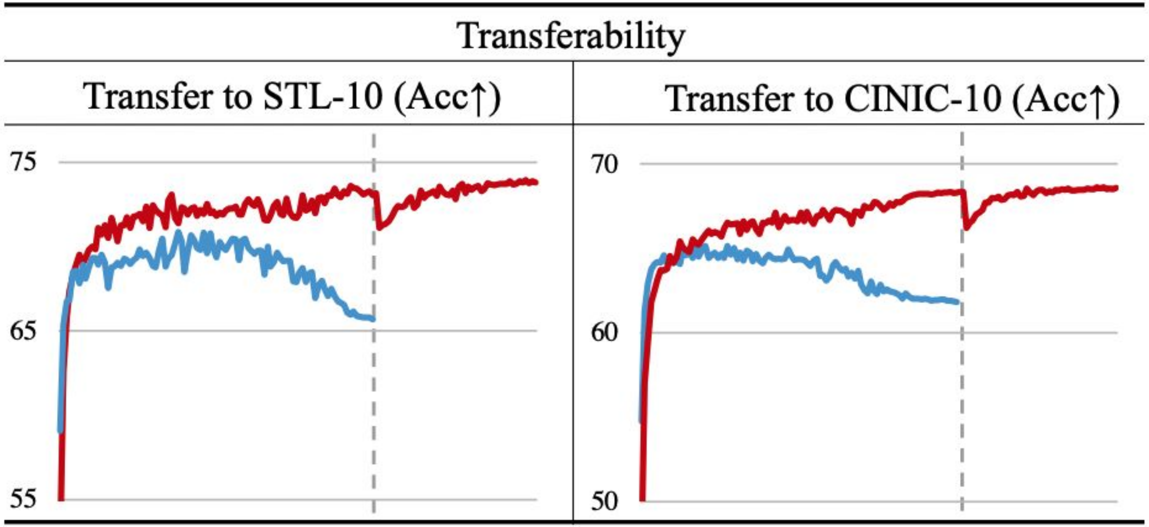
Thus:

$$\log(N) - \mathcal{L}_{\text{InfoNCE}} \leq \min(H(T_1), H(T_2)).$$

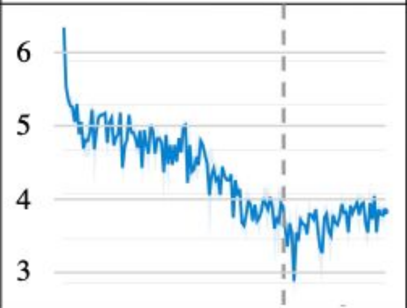
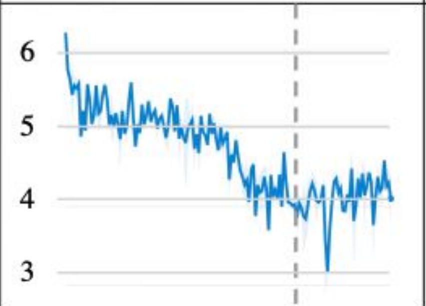
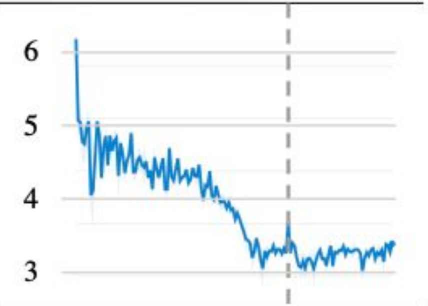
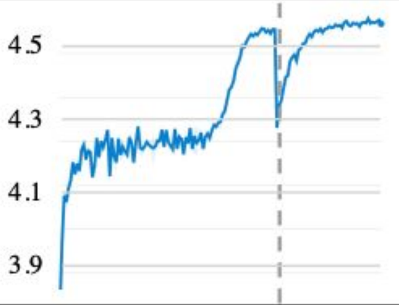
Here, T_1 is the fixed representation and T_2 is the learning representation, so the objective of improving $H(T)$ is reached.

Experiments - Improved Transferrability

- Ours
- Vanilla



Experiments - Improved Mutual Information

Metric	Source Dataset CIFAR-100	Target Dataset STL-10	Target Dataset CINIC-10
$I(X;T)$			
$I(T;Y)$			

Continue improving

Experiments - Larger datasets

pre-training method	CUB200 top-1 acc. (%)	Aircraft top-1 acc. (%)
Res50+CosLr†	62.5	27.8
Res50+CTC(Ours)†	**63.7**	**28.2**
Res50+AA+CosLr†	64.8	31.2
Res50+AA+CTC(Ours)†	**66.1**	**32.1**
Res50+CosLr‡	80.1	82.5
Res50+CTC(Ours)‡	**81.7**	**84.1**
Res50+AA+CosLr‡	81.3	83.4
Res50+AA+CTC(Ours)‡	**83.5**	**85.6**

pre-training method	iNat-18 top-1 acc. (%)
Res50+CosLr	66.1
Res50+MoCo v1 (IN-1M) [14]	65.6
Res50+MoCo v1 (IG-1B) [14]	65.8
Res50+CTC (ours)	**66.4**
Res50+AA+CosLr	66.3
Res50+AA+CTC (ours)	**66.7**

Experiments - Other tasks

pre-training method	Performance					
	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Res50 random init.	30.2	48.9	32.7	28.6	46.6	30.7
Res50+MoCo v2 [6]	38.5	58.3	41.6	33.6	54.8	35.6
Res50+InfoMin [41]	39.0	58.5	42.0	34.1	55.2	**36.3**
Res50+CosLr	38.2	58.2	41.2	33.3	54.7	35.2
Res50+CTC(Ours)	**39.5**	**58.7**	**42.0**	**34.2**	**55.4**	36.2

Table 5. COCO object detection and instance segmentation based on Mask-RCNN-FPN with 1× schedule.

Our code for reproducing is on GitHub

https://github.com/DTennant/dt-tradeoff

Thanks for Listening!