Decomposed **M**utual **I**nformation Estimation for Contrastive Representation Learning (DEMI)

Alessandro Sordoni*, Nouha Dziri*, Hannes Schulz*, Geoff Gordon, Phil Bachman, Remi Tachet

ICML 2021

Context

- **Self-supervised learning:**
  - Estimate representations without task labels
- **Contrastive learning:**
  - Representations of views **x** and **y** of the same input closer than *K* random negative samples
- **Maximizing mutual information (MI)**, *InfoNCE* [1]
  - Estimating MI is hard, InfoNCE is biased (< log *K*)

[1] Representation Learning with Contrastive Predictive Coding, Oord, Li, Vinyals, 2019.

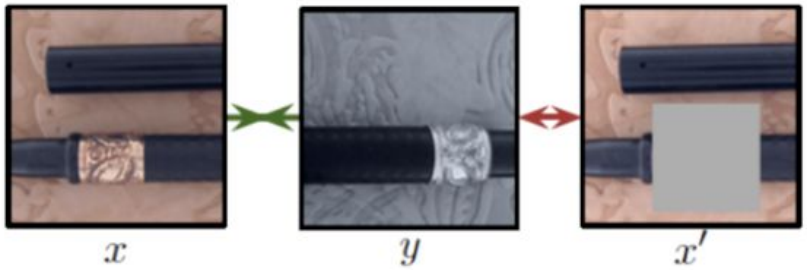
DEMI, idea

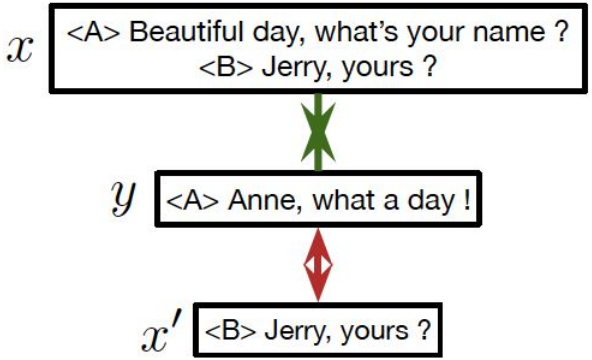
Chunk a hard estimation problem into smaller subproblems with less bias

Given **x, y** two views of the same input datum:

1. Generate *n* sub-views of **x** , e.g. **x', x'', …**
    a. say *n* = 2, **x'** = **cutout(x), x'' = x**
2. Write **I(x; y) = I(x'; y) + I(x; y | x')** ***(via chain rule on MI)***
3. Maximize each term in the sum

DEMI





- **I(x', y)** (standard InfoNCE)
  - Representations of **x'** and **y** more similar than **easy negative ~y** random negatives
- **I(x; y | x')**
  - Representations of **x** and **y** more similar than **hard negatives ~y** sampled from p(**y** | **x'**)
  - Representations encouraged to capture the embossed detailing

Conditional MI

- **I(x; y) >= I(x'; y) + I(x; y | x')**
- **I(x'; y)** can be estimated using InfoNCE
- What about **I(x; y | x')** ?

Takeaway 1 (Conditional InfoNCE): InfoNCE with a negative sampling distribution $p(\mathbf{y} \mid \mathbf{x'})$ (instead of $p(\mathbf{y})$) is a lower-bound on **I(x; y | x')**

But $p(\mathbf{y} \mid \mathbf{x'})$ is not known ! :-(

I_*VAR*

Approximate the unknown p(y | x’) with a distribution q(y | x’)

Takeaway 2: Sampling negative examples from a variational $q(\mathbf{y} \mid \mathbf{x'})$ provides a lower-bound on conditional MI.

Train $q(\mathbf{y} \mid \mathbf{x'})$ by maximum-likelihood.

It’s still kind of expensive to train/sample q (e.g. train cond. flow on pixel data)

[2] A Unified Stochastic Gradient Approach to Designing Bayesian-Optimal Experiments, Foster et al., 2020

I_*IS*

Don’t need p(y|x’), only samples from it!

Takeaway 3: Draw approximate samples from p(**y** | **x’**) by importance resampling using the optimal NCE critic ***s*** estimated from maximizing **I(x’, y)**

Basically, reweight a negative sample **y_k ~ p(D)** by **w_k ~ *s(x’, y_k)***

$$I_{IS}(x, y|x', \phi, K) =$$
$$\mathbb{E}\left[\log \frac{e^{\phi(x', x, y_1)}}{\frac{1}{K}(e^{\phi(x', x, y_1)} + (K-1)\sum_{k=2}^{K} w_k e^{\phi(x', x, y_k)})}\right]$$

I_*BO*

You need no samples from p(y | x')! (with a hiccup)

Takeaway 4: If you have **the optimal NCE critic I(x', y)**, you can skip sampling from p(**y** | **x**') altogether and opt for a *'boosted estimation'*.

Proposition 4 (Boosted Critic Estimation). *Assuming* $\psi^* = \arg\sup_\psi I_{NCE}(x', y)$*, the following holds, with:*

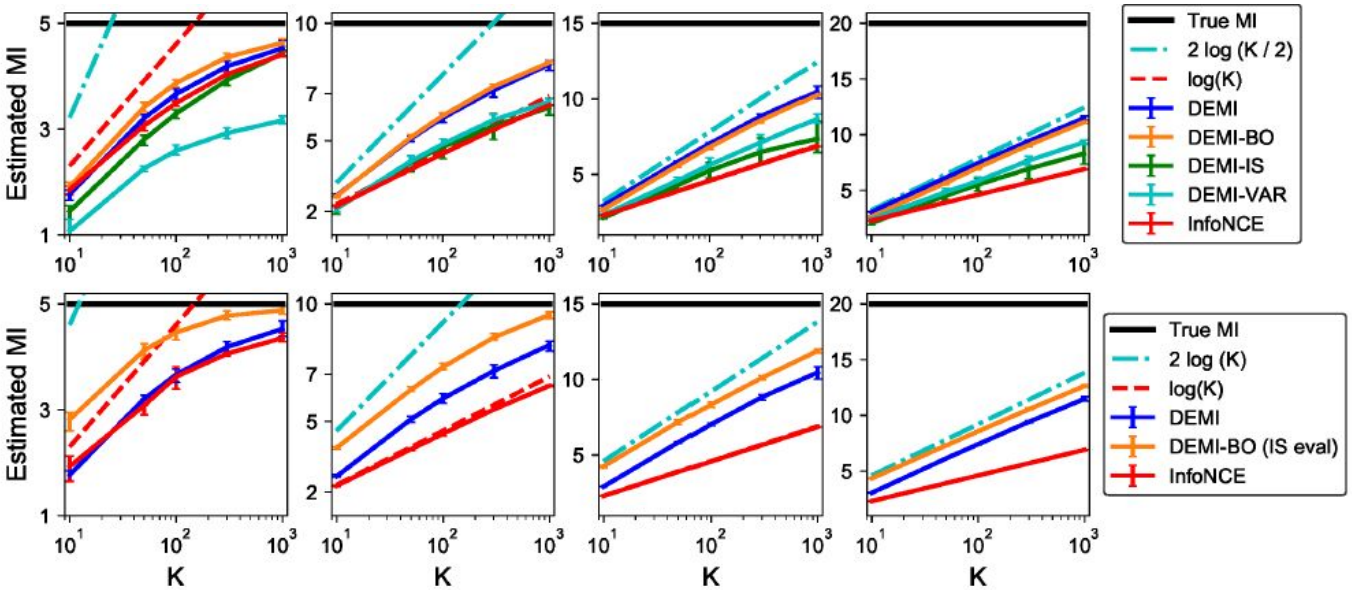
$$I_{BO}(x, y|x', \phi, K) = \mathbb{E}\left[\log \frac{e^{\psi^*(x', y_1)+\phi(x', x, y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\psi^*(x', y_k)+\phi(x', x, y_k)}}\right], \tag{10}$$

1. $I_{BO} \leq I(x, x'; y)$,
2. $\phi^* = \arg\sup_\phi I_{BO} = \log \frac{p(y|x', x)}{p(y|x')} + c(x, x')$.

This gives you the conditional log-ratio but not an estimator of conditional MI.

Synthetic Results

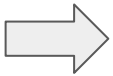
- DEMI captures more MI than InfoNCE for the same number of K



Empirical Results: Imagenet / Dialogue

- We use three views: **x, y, x'**, where **x'** is either **cutout(x)** or **multicrop(x)**
- We use the InfoMin [3] architecture but augment the loss function with conditional MI maximization across views
- I_IS ~~ I_BO in practice

No conditional MI



Model	Views	IN100	IN1K	STL10	C10	C100	CARS	CUB	FLOWERS
SimCLR (Chen et al., 2020a)	$x \leftrightarrow y$	-	66.6	-	90.6	71.6	50.3	-	91.2
MocoV2 (Chen et al., 2020b)	$x \leftrightarrow y$	-	67.5	-	-	-	-	-	-
InfoMin (Tian et al., 2020)	$x \leftrightarrow y$	74.9	70.1	96.2	92.0	73.2	48.1	41.7	93.2
InfoMin (multi)	$x, x' \leftrightarrow y$	77.2	70.2	95.9	92.6	74.5	49.2	42.1	94.7
DEMI	$x, x' \leftrightarrow y$	**78.6**	**70.8**	**96.4**	**92.8**	**75.0**	**51.8**	**43.6**	**95.0**

Model	ppl	BLEU	H-rel	H-hum	H-int
GPT2	19.21	0.78	✓	✓	✓
TransferTransfo	19.32	0.75	✓	✓	✓
GPT2-MMI	19.30	0.65	✓	✓	✓
InfoNCE	18.85	0.80	=	✓	✓
DEMI	**18.70**	**0.82**	=	=	=
Human	–	–	✗	✗	✗

[3] What Makes for Good Views for Contrastive Learning? Tian et al. ICML 2021

The end - Thanks