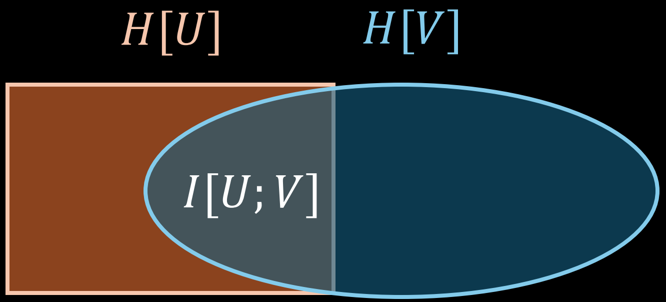
Connecting Jensen–Shannon and Kullback-Leibler Divergences: A New Bound for Representation Learning

Reuben Dorent[1], Polina Golland[2], William Wells III[2,3]
[1] Inria, Paris Brain Institute
[2] MIT
[3] Harvard Medical School

Mutual Information Maximization

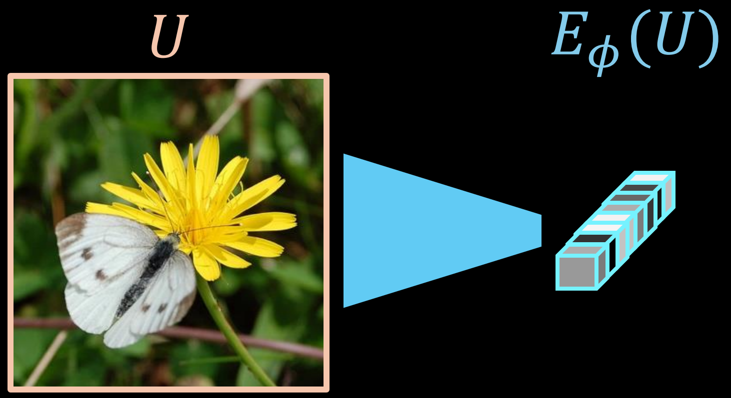
Mutual Information (MI) is a fundamental quantity for learning deep representations.

$$I[U;V] \coloneqq \mathrm{D}_{KL}[p_{UV} \,||\, p_U \otimes p_V]$$



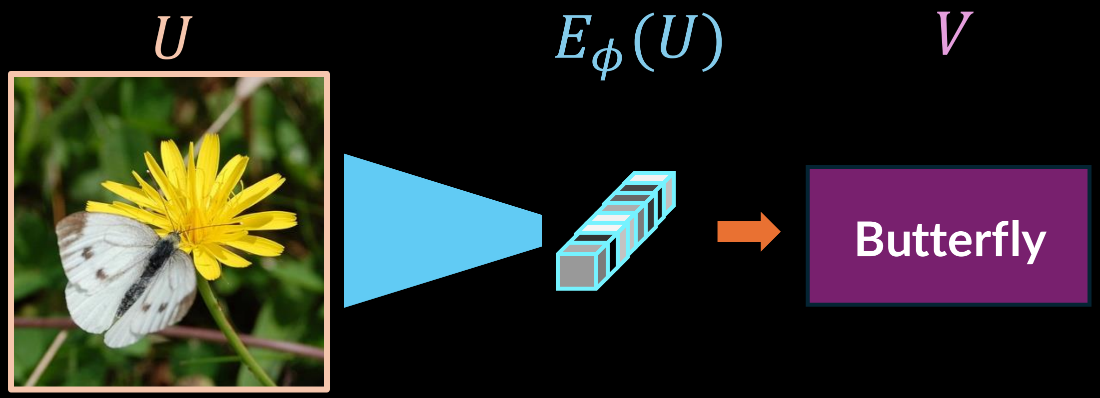
Representation learning

$$\max_{\phi} \; I[U;E_\phi(U)]$$



Information Bottleneck

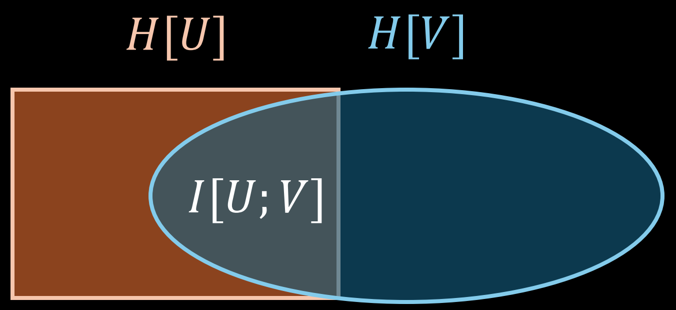
$$\max_{\phi} \; I\big[E_\phi(U);V\big] - \lambda \cdot I\big[U;E_\phi(U)\big]$$



Mutual Information Maximization

Mutual Information (MI) is a fundamental quantity for learning deep representations.

$$I[U;V] \coloneqq \mathrm{D}_{KL}[p_{UV} \,||\, p_U \otimes p_V]$$



Variational Lower Bounds

$$VLB \leq \mathrm{D}_{KL}[p\,||q]$$

MINE [1], NWJ [2]
→ **Unstable estimation** (high variance)

CPC (InfoNCE) [3]
→ **Biased estimates**
→ **Upper-bounded** by $\log(b)$ (b=batch size)

Two-step Estimators

Step 1: Optimizing a discriminator D_{θ^*}
Step 2: Estimating $\mathrm{I}[\mathrm{U};\mathrm{V}]$ using D_{θ^*}

NJEE [4], DEMI [5], PCM[6], f-DIME [7]:
→ Must be **retrained** when the underlying distribution of U or V varies
→ **Impractical** for representation learning

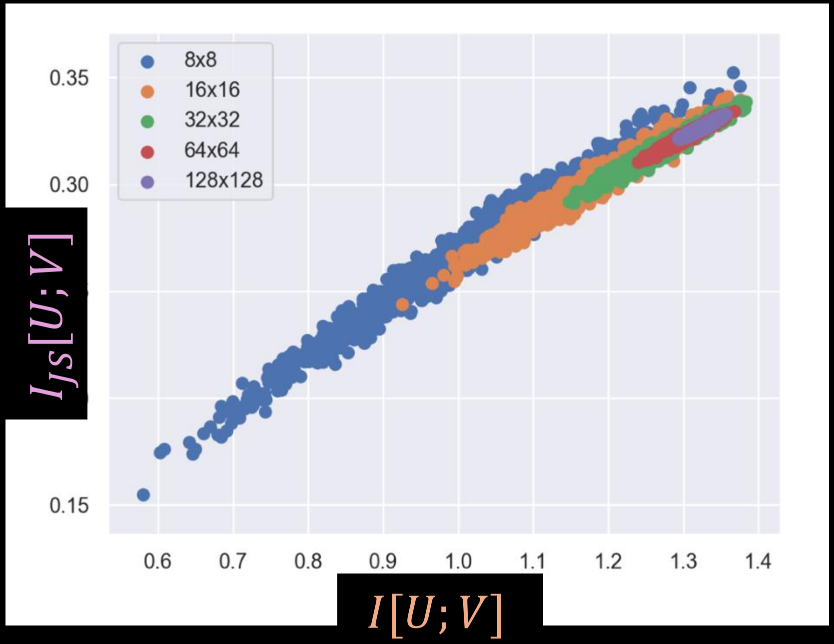
Mutual Information Maximization

Maximizing **Jensen-Shannon-based Mutual Information:**

$$I[U;V] \coloneqq D_{KL}[p_{UV} \,||\, p_U \otimes p_V]$$

⇓

$$I_{JS}[U;V] \coloneqq D_{JS}[p_{UV} \,||\, p_U \otimes p_V]$$



Hjelm, *et al.* ICLR (2019)

Key advantages:
- **Stable optimization (bounded + symmetric)**
- Empirically **correlates with true MI**

Maximizing $I_{JS}[U;V]$ → maximizing a lower-bound on MI $I[U;V]$.

Main contributions

Joint range [8] between JSD and KLD

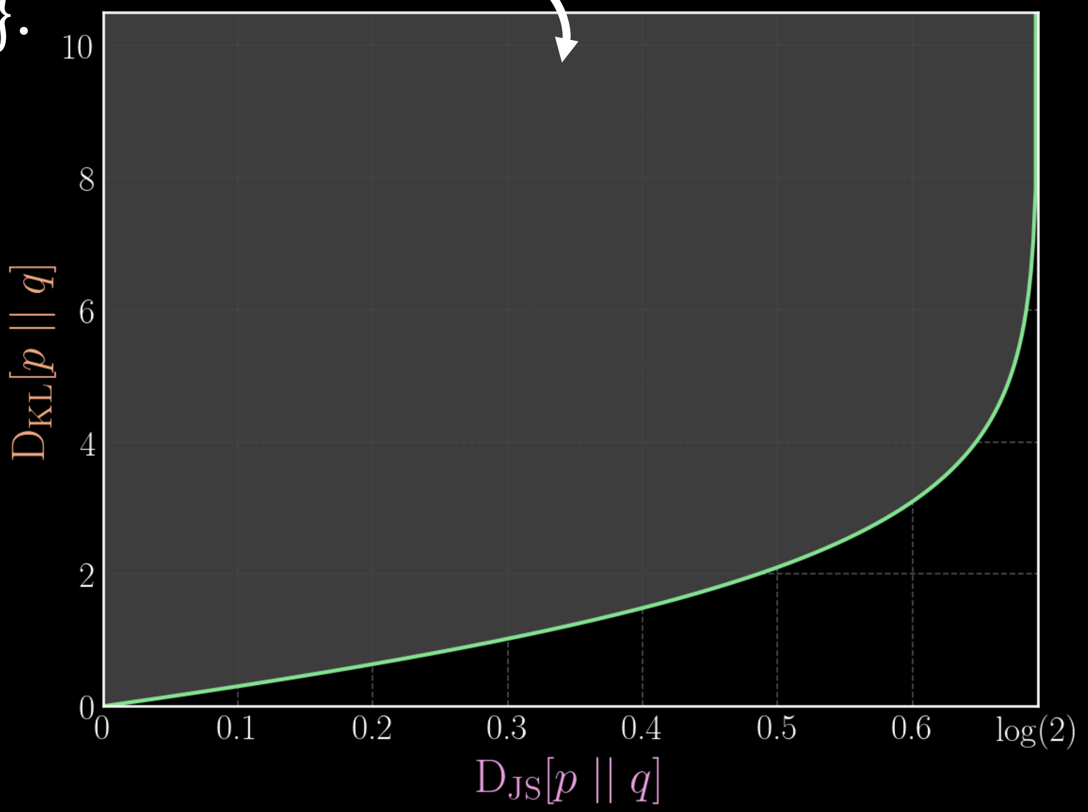
$$\mathcal{R}_{JS,KL} = \{ (\mathrm{D}_{JS}[p \,||\, q], \mathrm{D}_{KL}[p \,||\, q]) \; : \; p, q \in \mathcal{P} \}.$$

Theorem 1:

There exists a strictly increasing function Ξ such that for any pair of distributions p, q :

$$\Xi(\mathrm{D}_{JS}[p \,||\, q]) \leq \mathrm{D}_{KL}[p \,||\, q]$$

Optimal lower bound between JSD and KLD



Main contributions

Theorem 1:

There exists a strictly increasing function Ξ such that for any pair of distributions p, q :

$$\Xi\left(\mathrm{D}_{JS}[p \,||\, q]\right) \leq \mathrm{D}_{KL}[p \,||\, q]$$

New lower bound between JSD-based Mutual Information and Mutual Information:

$$\Xi\Big(\underbrace{\mathrm{D}_{JS}[p_{UV} \,||\, p_U \otimes p_V]}_{= I_{JS}\,[U;V]}\Big) \leq \underbrace{\mathrm{D}_{KL}[p_{UV} \,||\, p_U \otimes p_V]}_{= I[U;V]}$$

Maximizing $I_{JS}[U;V]$ → maximizing a lower-bound on MI $I[U;V]$.

Main contributions

Optimizing the **cross-entropy loss** $\mathcal{L}_{CE}$ of a discriminator to distinguish dependent and independent pairs

increases a **lower bound on MI**, using the following inequality chain:

$$\Xi(\log 2 - \mathcal{L}_{CE}) \leq \Xi\left(I_{JS}[U;V]\right) \leq I[U;V]$$

Experiments

Goal: Validate the theoretical results and assess both the **tightness** of our new bound and its **practical usefulness**.

Synthetic Experiments:

- Both Mutual Information and JS-based Mutual Information can be computed exactly.
- Comparison with other variational lower bounds (VLBs).
- Our JSD-based lower bound proves to be **tight**, **stable**, and has **lower variance.**

Information Bottleneck:

- We replaced the standard MI term with our discriminative lower bound.
- Achieved **SOTA performance** on MNIST:
  - Improved generalization
  - Stronger adversarial robustness
  - Better out-of-distribution detection

Takeaway message

If you aim to **maximize mutual Information**, our work provides a principled justification for using **discriminative approaches**.

Acknowledgments

Marie Skłodowska-Curie grant No 101154248
(project: SafeREG)

P41EB028741 and R01EB032387

[1] Belghazi, *et al.* 2018. Mutual information neural estimation. *ICML*.
[2] Nguyen, *et al.* 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*.
[3] Oord, *et al.* 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint*.
[4] Shalev, *et al.* 2022. Neural joint entropy estimation. *IEEE Transactions on Neural Networks and Learning Systems*.
[5] Liao, *et al.* 2020. DEMI: Discriminative Estimator of Mutual Information. *arXiv preprint*.
[6] Tsai, *et al.* 2020. Neural methods for point-wise dependency estimation. *NeurIPS*.
[7] Letizia, *et al.* 2024. Mutual Information Estimation via f-Divergence and Data Derangements. *NeurIPS*.
[8] Harremoës, *et al.* 2011. On pairs of f-divergences and their joint range. *IEEE Transactions on Information Theory*.