Beyond I-Con: A Roadmap for Representation Learning Loss Discovery

The choice of optimization objective fundamentally determines the success of representation learning methods, yet the field has converged on a single statistical divergence measure without systematic exploration of alternatives. The Information Contrastive (I-Con) framework recently revealed that over 23 diverse representation learning methods all implicitly minimize KL divergence between data and learned distributions that encode similarities between data points [1].

This usage of KL divergence may lead to suboptimal optimization of representations. KL divergence is known to cause optimization issues due to properties such as asymmetry and the possibility of infinite values [2, 3]. Furthermore, since loss functions are only proxies for the actual goal of the task, optimizing for KL divergence may be misaligned with the true objective.

We present Beyond I-Con, making the following contributions: (1) We generalize I-Con by replacing KL divergence with alternative f-divergences, revealing that KL is not unique in enabling meaningful feature optimization; (2) We systematically explore combinations of f-divergence and similarity kernel, uncovering novel loss functions with superior performance on unsupervised clustering, supervised contrastive learning, and dimensionality reduction. Specifically, we demonstrate that total variation distance when paired with distance similarity kernels achieves superior performance compared to KL-based approaches.

We achieve the following results: (1) on unsupervised clustering of DINO-ViT embeddings, we achieve state-of-the-art results by modifying the PMI algorithm to use total variation (TV) distance; (2) on supervised contrastive learning, we outperform the standard approach by using TV and a distance-based similarity kernel instead of KL and an angular kernel; (3) on dimensionality reduction, we achieve superior qualitative results and better performance on downstream tasks than SNE by replacing KL with a bounded f-divergence. Our results highlight the importance of considering divergence and similarity kernel choices in representation learning optimization.

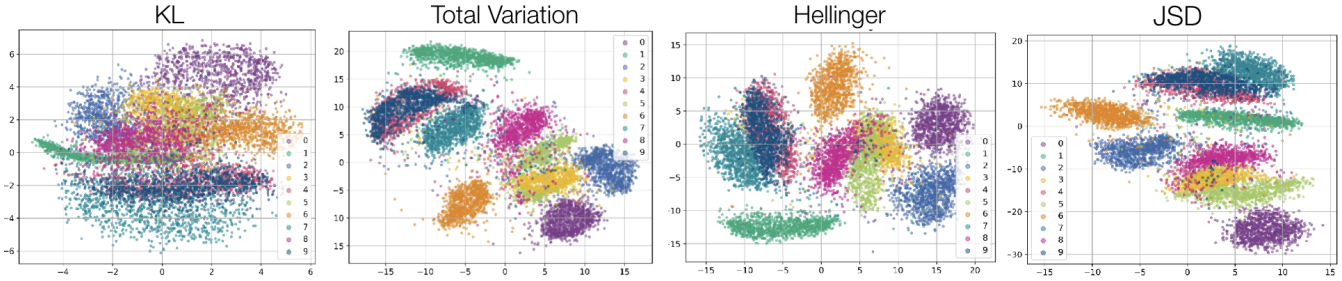

Figure 1: Results for running SNE on CIFAR-10 using different divergences, after 150 epochs with a CNN model architecture at learning rate 1e-3. Each color represents a class. KL divergence produces highly overlapping categories in the SNE visualization while other divergences achieve separation.

References

[1] Shaden Alshammari, Mark Hamilton, and William T. Freeman. Information contrastive learning: Unifying framework and survey. *arXiv preprint arXiv:2402.08254*, 2024.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[3] Nevena Lazic, Botao Hao, Yasin Abbasi-Yadkori, Dale Schuurmans, and Csaba Szepesvári. Optimization issues in kl-constrained approximate policy iteration. *CoRR*, abs/2102.06234, 2021.