Information Preserving Dimensionality Reduction for Mutual Information Analysis of Deep Learning

Shizuma Namekawa* and Taro Tezuka†

*University of Tsukuba
kirua515@gmail.com

†University of Tsukuba
tezuka@iit.tsukuba.ac.jp

Mutual information has been actively investigated as a tool for analyzing neural networks' behavior, most notably the information bottleneck theory. However, estimating mutual information is a notoriously tricky task, especially for high-dimensional stochastic variables. Recently, mutual information neural estimation (MINE) was proposed as a non-parametric method to estimate mutual information for continuous variables without discretization. Unfortunately, MINE also produces significant errors for high-dimensional variables. Analyzing the activity of neural networks requires a dimensionality reduction mechanism, with the resulting low-dimensional representations retaining as much information as possible. We investigated different dimensionality reduction methods to determine their capabilities in estimating mutual information regarding the activity of trained neural networks. We combined MIME with principal component analysis (PCA), a convolutional neural network (CNN), and global average pooling (GAP). The experiments showed that introducing dimensionality reduction provides more stable results than the baseline method. In terms of stability, PCA-MINE and GAP-MINE performed better than CNN-MINE. They also require much less computation time than CNN. Another advantage of GAP-MINE is that it requires no hyperparameter optimization. However, the results suggested that GAP-MINE may underestimate mutual information.

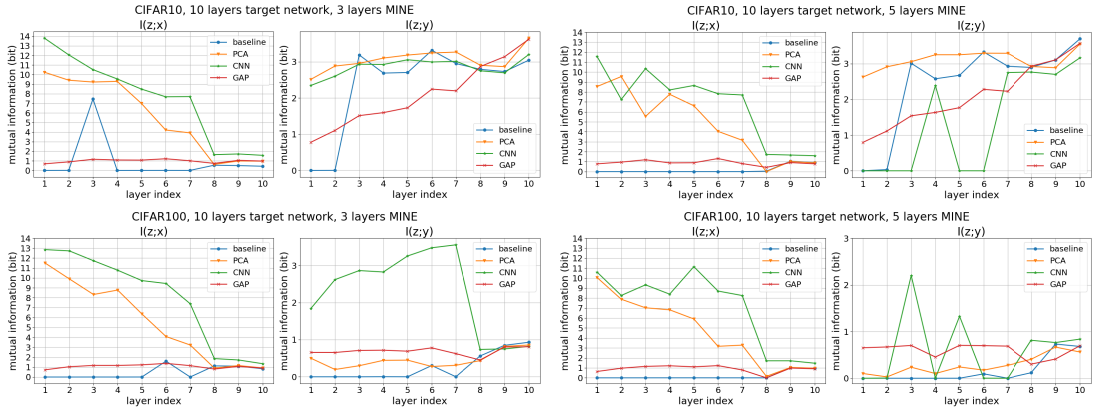

Figure 1: Comparison of the baseline, PCA-MINE, CNN-MINE, and GAP-MINE using $5,000$ samples from the CIFAR-10 and CIFAR-100 with minor random transformations.