Invariant Information Clustering
for Unsupervised Image Classification and Segmentation

Xu Ji, Joao Henriques, Andrea Vedaldi – VGG @ University of Oxford

github.com/xu-ji/IIC

Key Features
• Learn a classifier without labels
• Novel application of information theory to deep clustering
• No need for post-processing or fine-tuning
• Automatically prevents degenerate solutions
• 9 state-of-the-art classification and segmentation results

Theory
Mutual information breaks down into entropy and conditional entropy.

\[ I(x, x') = I(x) - \sum_{z, z'} P(z) \log \frac{P(z|x)}{P(z|x')} \]

Thus IIC optimises for classifications that are invariant to nuisance perturbations, with even spread, and certain prediction.

Results

Results on fully unsupervised classification and segmentation (left), and semi-supervised classification (right).

Networks trained with IIC clustering STL10 (top), segmenting Potsdam-3 (middle) and segmenting COCO-Stuff-3 over successive epochs (bottom).

Training with IIC on unlabelled MNIST in successive epochs from random initialisation (left). There is no cherry-picking as the entire dataset is shown in every snapshot. Predictions are rendered as points; ground truth labelling (unseen by model) is given by colour. With neither labels nor heuristics, the clusters discovered by IIC correspond perfectly to unique digits, with one-hot certain prediction (right).

Given an image (or patch) \( x \), and randomly transformed version \( g(x) \), maximise mutual information between their predictions, across the batch.

\[ \max I(\Phi(x), \Phi(g(x))) = \max I(x, x') \]