Unsupervised Learning of Visual Representation by solving Jigsaw Puzzles

Context-Free Prediction Network

1. Their network:
   - Spends the majority of computation for each tile independently of the context.
   - Delay the computations of statistics across tiles until the last fc^2 layer.
   - Use a siamese architecture where the shared weights are based on AlexNet.
   - They train the architecture from scratch to ensure similar perf. on ImageNet.

Jigsaw Puzzle Task

1. Design a set of Jigsaw puzzle permutations, e.g. a tile config $S=(3,1,2,\dots)$ and assign an index to each entry.
2. Randomly pick one permutation, rearrange the 9 input patches according to that permutation, and ask the CFN to return a vector with probability value for each index.
3. They found that the Hamming distance b/t permutation controls the difficulty of the Jigsaw task & correlates with ~~the~~ object detection performance.
4. They claim that:
   "A good self-supervised learning task is neither too ~~easy~~ simple nor ambiguous."

Preventing Short Cut

1. low-level statistics → normalize the mean & std of each patch independently.

2. Edge continuity → gap bt tiles

3. Chromatic aberration → . crop & resize.
   . train with both color & grayscale imgs.
   . spatially jitter the color channels of the color img.

4. c.a. is the relative spatial shift bt color channels that increases from the img centers to the borders.

Other interesting tidbits

1. Freezing the 1st conv layer & retraining the rest from scratch performs better than freezing the later conv layer & retraining fewer layers.

   ⇒ My conclusion: not a sig. amt of knowledge is transferred.

2. Training only takes 2.5 days on a Titan Xp, much faster than prev. methods.