Unsupervised Visual Representation Learning by Context Prediction

Proposed by Kilian Weinberger

Recently, new computer vision methods have leveraged large datasets of millions of labeled examples to learn rich, high-performance visual representations. Yet efforts to scale these methods to truly Internet-scale datasets (i.e. hundreds of billions of images) are hampered by the sheer expense of the human annotation required. A natural way to address this difficulty would be to employ unsupervised learning, which aims to use data without any annotation.

This work explores the use of spatial context as a source of free and plentiful supervisory signal for training a rich visual representation. Given only a large, unlabeled image collection, we extract random pairs of patches from each image and train a convolutional neural net to predict the position of the second patch relative to the first. We argue that doing well on this task requires the model to learn to recognize objects and their parts. We demonstrate that the feature representation learned using this within-image context indeed captures visual similarity across images. For example, this representation allows us to perform unsupervised visual discovery of objects like cats, people, and even birds from the Pascal VOC 2011 detection dataset. Furthermore, we show that the learned ConvNet can be used in the RCNN framework and provides a significant boost over a randomly-initialized ConvNet, resulting in state-of-theart performance among algorithms which use only Pascalprovided training set annotations.

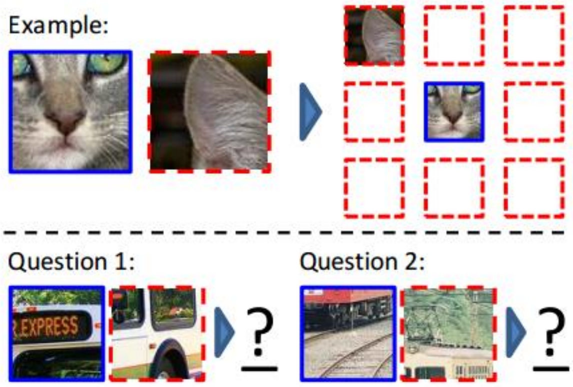

Figure 1. Our task for learning patch representations involves randomly sampling a patch (blue) and then one of eight possible neighbors (red). Can you guess the spatial configuration for the two pairs of patches? Note that the task is much easier once you have recognized the object!