Why latent representations in convolutional neural networks fall outside visual space

It is common to compare properties of visual information processing by artificial neural networks and the primate visual system. Some remarkable similarities were observed in the responses of neurons in IT cortex and units in higher layers of CNNs. Here I show that latent representations formed by weights in convolutional layers do not necessarily reflect visual domain. Instead, they are strongly dependent on a choice of training set and cost function.

Many of widely used computer-vision models (e.g. AlexNet, VGG16, Inception-v3, ResNet, etc.), were initially designed to solve classification tasks and trained on the ImageNet dataset, a common academic dataset for training image recognition systems. Specifically, the 2012 version (ILSVRC2012) serves as the baseline which consists of 1000 categories, including “library”, “dishwaher”, “jeep”, “landrover”, “zebra”, “sea urchin”, as well as 120 categories of dog breeds to showcase fine-grained classifications.

To measure the performance and adjust the weights during the training stage, cross-entropy loss function is often used. It computes the divergence of predicted probability from the actual label. It also assumes that classes are equally spaced, which means that a failure to distinguish a dog from a car is equivalent to a confusion among dog breeds. The more confidence a node has in predicting a class, the more significantly its weights will be adjusted in order to avoid further mistakes of that kind.

Figure 1. Examples of images, which evoke different level of response in a dog-selective neuron.

This strategy helps to successfully train models to differentiate categories, but it does not necessarily require visually similar images to be similar in the latent space of high-level layers of neural networks. Moreover, a neuron’s weights may be tuned to separate these resembling images. The most prominent example is when an individual unit, which is highly selective to some members of a category is, nevertheless, inhibited by visually similar objects of the same category, and this selectivity-profile cannot be attributed to incidental differences in low level statistics.

As the result, images from totally different categories, such as cars or cups, evoke higher activation in a “dog-selective” neuron compared to images of dogs of other breeds. These properties of neural networks may be the reason of unstable behaviour, when a model detect objects in their absence and fails to recognize obvious cases from a human observer’s point of view.