Can visual object representations in the human brain be modelled by untrained convolutional neural networks with random weights?

Anna Truzzi
Trinity College Dublin, Dublin, Ireland

Rhodri Cusack
Trinity College Dublin, Dublin, Ireland

Abstract

Convolutional neural networks (CNNs) have proven effective as models of visual semantic responses in the inferior temporal cortex (IT). The belief has been that training a network for visual recognition leads it to represent visual features in a way similar to those the brain has learned. However, a CNNs response is affected by its architecture and not just its training. We therefore explicitly measured the effect of training different CNN architectures on their representational similarity with IT. We evaluated two versions of AlexNet and two training regimes, supervised and unsupervised. Surprisingly, we found that the representations in an untrained (random-weight) variant of AlexNet, reflected brain representations in IT better than the benchmark supervised AlexNet and also better than the corresponding network trained in either a supervised or unsupervised manner. These results require a re-evaluation of the explanation of why CNNs act as an effective model of visual representations.