Disentangled Latent Representations with CapsNet

Chuanquan (Charles) Shu
Massachusetts Institute of Technology
Cambridge, MA 02139
cqgz21@mit.edu

Abstract

A distinguishing feature of Capsule Neural Network (CapsNet) is that the input and output neurons of a Caps Layer are vectors. The orientation of a neuron vector encodes features of a particular entity, but the semantic meanings of such orientation are entangled and thus uninterpretable. Therefore, in this project, I seek to use a training procedure that enables CapsNet to learn disentangled and semantically interpretable latent representations. Demonstrating the feasibility of this concept is the primary goal, hence MNIST dataset is used and only two features (digit rotation and scale) are experimented.

In order to achieve this goal, a deliberate training procedure is implemented. The first part creates supporting training batches by creating Rotation Batches (only rotation transformation), Scale Batches (only scaling transformation) and Standard Batches (no transformation). The second part ensures that only one group of variable(s)/dimension(s) is allowed to learn at each step by imposing restrictions on neuron activations and gradients flow during the forward and backward stages, respectively.

Under the new method, almost all 10 digits unanimously choose the first and second dimensions of their encodings to explain rotation and scaling, respectively. Under the original method, the best explaining dimension for each feature is rather arbitrary. Moreover, regressing the true rotation and scaling against the best explaining dimension under the original methods has a mean R^2 of just 0.45, suggesting the lack of disentangled representations for rotation and scaling. Under the new methods, the mean R^2 for rotation and scaling are 0.63 and 0.83, respectively. The significant improvements show that the first and second dimensions have truly learned to explain the rotation and scaling of a given image.

It is worth mentioning that the 0.63 R^2 for rotation, although better than the old method counterpart, is not impressive. This could be largely attributed to the sub-optimal quality of the Standard Batches. The images in the Standard Batches are supposed to have no rotation and scaling so irrelevant dimensions in the latent representations will not be exposed to any variance in these two features during training. Nevertheless, MNIST does have substantial innate variance in rotation, leading to sub-optimal Standard Batches. This undermines the constraint on the first dimension to explain rotation alone and hence resulting in a less powerful rotation encoding. It also explains why the scale encoding achieves great performance and the rotation encoding does not, suggesting that rotation encoding could be as powerful if presented with better Standard Batches (ones with less variance in rotation).