Stacked Capsule Autoencoders

Abraham Jose

1 Summary

The author proposes an intuitive and novel architecture for learning representation using capsule auto encoders to encode the objects in image and the geometric relationship between the objects w.r.t viewer. This novel approach is robust to the viewpoint changes in the image and it helps in creating the model that can be used to reduce the data augmentation required for the model. There are 2 encoders which is Constellation Autoencoder(CCAE) and Part Capsule Autoencoder(PCAE) which encodes special feature, poses and object viewer relationship. The decoder is a Object Capsule Encoder which reassembles the parts and increasing the likelihood from that of the original image.

2 Good points

The strongest point of this paper is that the this stacked autoencoder is capable of performing well in unsupervised object classification as this model learns the best learning representation for every image and it does not rely on the amount of the mutual information. This means the model can work with less augmentation or requires less number of data to perform really well. Wiring an inductive bias towards geometric shapes is another key contribution.

3 Weak points

The concern I have about the architecture is the scalability of the model for the real world tasks w.r.t. inference time and complexity of the geometric relationships that it can encode and successfully decode to reassemble. If we introduce deeper hierarchy to learn relationship in the images with complex objects(objects can be pose invariant like a sphere, it could be occluded etc..) the performance and the inference time required might be hindrance to make it viable for deployment in real world.

4 Questions

How can we create input-dependent shape functions that can be used for the training? CNN features are inductive biased and we cannot use them to learn the input-dependent features(which is less frequent in the rest of the images).
How much time it takes to infer an image(training and testing)?

5 Ideas

Incorporating SIFT like features which are scale independent and input-dependent can be used as an input to Constellation Autoencoder. Or the features derived from SIFT like features.