Dynamic Routing Between Capsules

Avinash Kommineni
University at Buffalo
akommine@buffalo.edu

Abstract

This project is the understanding and implementation of the algorithm mentioned by Geoffrey E. Hinton, the godfather of deep learning in his paper called 'Dynamic Routing Between Capsules'.

1. Introduction

The conventional algorithms being used for image classification such as the Convolutional Neural Networks (CNNs) use translated replicas of learned feature detectors and this allows them to translate knowledge about good weight values acquired at one position in an image to other positions. CNNs have proved to be really good at this job of image recognition but they have got an intrinsic short coming since their inception such as rotation if all the trained images are in a particular rotation, pooling gives some translational invariance in much deeper layers but only in a crude way whereas the human brain must achieve translational invariance in a much better way. Human vision ignores irrelevant details by using a carefully determined sequence of fixation points to ensure that only a tiny fraction of the optic array is ever processed at the highest resolution.

Geoffrey E. Hinton coined the term capsule in 2011 paper 'Transforming Auto-encoders' as local pool of neurons that perform some quite complicated internal computations on their inputs and then encapsulate the results of these computations into a small vector of highly informative outputs. A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or object part. We use the length of the activity vector to represent the probability that the entity exists and its orientation to represent the instantiation parameters. Active capsules at one level make predictions, via transformation matrices, for the instantiation parameters of higher-level capsules.

This algorithm though in its budding state has created waves across the deep learning community for its unique way of thought and its potential growth. By this project I would have the opportunity of studying and understanding closely what are the recent advancements of the research community.

2. Related Work

This project is completely based of the recent publication, 'Dynamic Routing Between Capsules'[1] though the idea of capsules has been for a while and other state-of-the-art performing algorithms based on CNN from several communities like Alexnet, VGGNet, GoogleNet, ResNet, DenseNet are all aimed for bigger problems with higher dimensions.

3. Algorithm

Since Human vision ignores irrelevant details by using a carefully determined sequence of fixation points, in this paper we will assume that a single fixation gives us much more than just a single identified object and its properties. We will assume that our multi-layer visual system creates something like a parse tree on each fixation, and we will ignore the issue of how these single-fixation parse trees are coordinated over multiple fixations. Each layer will be divided into many small groups of neurons called capsules and each node in the parse tree will correspond to an active capsule. Using an iterative routing process, each active capsule will choose a capsule in the layer above to be its parent in the tree.

The activities of the neurons within an active capsule represent the various properties of a particular entity that is present in the image. These properties can include many different types of instantiation parameter such as pose (position, size, orientation), deformation, velocity, albedo, hue, texture etc. One very special property is the existence of the instantiated entity in the image. An obvious way to represent existence is by using a separate logistic unit whose output is the probability that the entity exists. The paper explores an interesting alternative which is to use the overall length of the vector of instantiation parameters to represent the existence of the entity and to force the orientation of the vector to represent the properties of the entity. It has also been made sure that the length of the vector output of a capsule cannot exceed 1 by applying a non-linearity that leaves the orientation of the vector unchanged but scales down its magnitude.

The fact that the output of a capsule is a vector makes it possible to use a powerful dynamic routing mechanism to ensure that the output of the capsule gets sent to an appropriate parent in the layer above. Initially, the output is routed to all possible parents but is scaled down by coupling coefficients that sum to 1. For each possible parent, the capsule computes a prediction vector by multiplying its own output by a weight matrix. If this prediction vector has a large scalar product with the output of a possible parent, there is top-down feedback which has the effect of increasing the coupling coefficient for that parent and decreasing it for other parents. This increases the contribution that the capsule makes to that parent thus further increasing the scalar product of the capsules prediction with the parents output.
Convolutional neural networks (CNNs) use translated replicas of learned feature detectors and this allows them to translate knowledge about good weight values acquired at one position in an image to other positions. This has proven extremely helpful in image interpretation. Even though we are replacing the scalar-output feature detectors of CNNs with vector-output capsules and max-pooling with routing-by-agreement, we would still like to replicate learned knowledge across space, so we make all but the last layer of capsules be convolutional. As with CNNs, we make higher-level capsules cover larger regions of the image, but unlike max-pooling we do not throw away information about the precise position of the entity within the region. For low level capsules, location information is place-coded by which capsule is active. As we ascend the hierarchy more and more of the positional information is rate-coded in the real-valued components of the output vector of a capsule. This shift from place-coding to rate-coding combined with the fact that higher-level capsules represent more complex entities with more degrees of freedom suggests that the dimensionality of capsules should increase as we ascend the hierarchy. In CNNs, the max-pooling allows neurons in one layer to ignore all but the most active feature detector in a local pool in the layer below. So a new type of routing-by-agreement mechanism: A lower-level capsule prefers to send its output to higher level capsules whose activity vectors have a big scalar product with the prediction coming from the lower-level capsule.

4. Task breakdown

1. Implement the algorithm in tensorflow.
2. Fine-tune the parameters and hyperparameters.
3. Compare the results with other algorithms currently in place.

5. Software or Libraries

I intend to use the Tensorlfow for this project in python.

6. Timline

Complete the implementation by the end of December and tune the parameters from then on. It takes a good amount of time to train as it is computatioinally very expensive.

References

[1] Sara Sabour, Nicholas Frosst, Geoffrey E. Hinton *Dynamic Routing Between Capsules*. Computer Vision and Pattern Recognition, 2017.