RS of Building High-level Features Using Large Scale Unsupervised Learning

Tianyang Liu, Su Pu, Yicheng Wang

link: https://docs.google.com/document/d/1T_X0Rq5FFaLZTduvytjW0VOTQia46yEhLkMi1bGPolQ/edit

This article is a big step in the area of artificial intelligence. The feature learning principle of GoogleBrain, is to use images without labels to learn those high-level features of human faces, as well as cat faces, therefore detectors are generated. This article uses big data to construct a local sparse-connected self-coding network of nine layers: the whole model has almost one billion connections, and data consists of ten million of 200 by 200 images. The authors use model parallelization and the asynchronous scheme, to train this model in three days using 16k cores. The final result exhibits that, we can accurately train human face detectors in situations that unlabeled images contain human faces versus not. Previously, people use feature learning methods including: RBM, auto encoders, sparse coding, and K-means. These methods, anyway, could only learn low-level features like edge features and clusters, this article learns high-level features and is more practical.

The algorithm principle of this is: sparse coding can generate accept areas through training on unlabeled data, but sparse coding has low structure, and therefore could only obtain low-level features. This article constructs a sparse deep self-coding network with three primary features: local accept area, pooling, and local comparison normalization. Local accept area is to process large images, while l2 pooling and comparison normalization is to maintain the form with regard to local transformation. Deep self-coding has three identical steps: local filtering, local pooling, and local comparison normalization, nine layers in total. The best feature of this network is the local connections of the nodes. The l1 accepts all image channels as inputs, of the dimension 18 by 18 as the accept area; l2 has just one channel. The l1 outputs line filtering response, while l2 outputs square root of the square sum, i.e. l2 pooling. Different from CNN, there is not weight sharing. The network nodes reach 10 billion but still a lot less than actual human visual system nodes by a million level.

The learning and optimization still have some specials to say. In learning, we use topographic ICA optimization, the first item coding important information as data, and the second item allocating identical features in pooling. In optimization, when training they use model parallelization to distribute local weights to different machines. Research group develops one framework called Disbelief to accelerate training which uses asynchronous SGD. This work is tested using 37k images, containing labeled human face images as well as some random images downloaded from ImageNet. And should test if this network can detect human faces among random images. As a result, the best neuron of this network receives a 81.7% accuracy in the case of human face detection, while random guess has 64.8% accuracy. Single network achieves 71%, while the best line classifiers get 74%.

Three strong points are:

1. Improve the previous approaches

Early approaches have the short back such as sparse coding, the architecture are shallow and typically capture low-level concepts (e.g., edge "Gabor" filters), but this paper goes beyond such simple features and captures complex invariances.

2. The special structure that stimulate the brain

The style of stacking a series of uniform modules, switching between selectivity and tolerance layers is actually the same architecture employed by the brain. In order to stimulate the brain architecture, the local receptive fields are not convolutional, the parameters and weights are not shared across different locations in the image. It is easier to learn more invariances other than translation invariances.

3. The scale is very large

The scale of the network is perhaps one of the larges know networks nowadays. It has 1 billion trainable parameters. However, the neurons in our visual cortex is still 10^6 bigger than the network.

Three weak points are:

1. The paper presents the accuracy of 15.8% when applied the network to ImageNet, compared with other supervised learning methods, this result is definitely unacceptable.
2. Of course, compared with previous method in unsupervised learning methods, which has the best accuracy of 9.3%, it's a great improvement.
3. The training time is really long, it cost the cluster of 1000 machine with 16 cpu core in each machine 3 whole days to finish.

Brainstorming for future work

The prospect of this article is to reduce the cost of training. We see that the whole training process costs 16k cores to finish but still cost up to three days, which is so much computational resources.

Reference:

[1] Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks[C]//Advances in neural information processing systems. 2012: 1223-1231.

[2] Coates A, Lee H, Ng A Y. An analysis of single-layer networks in unsupervised feature learning[J]. Ann Arbor, 2010, 1001(48109): 2.

[3] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 248-255.