Reading Assignments

Bingzhang Zhu

December 15, 2021

1 ImageNet Classification with Deep Convolutional Neural Networks

1.1 Paper Summary

The paper trained a large and deep convolutional neural network to classify high-resolution images and it achieved the new state-of-the-art. What is special with this model is that:

(1) it has millions of parameters and 650,000 neurons, which is huge compared with normal machine learning models.

(2) it consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. And this framework works best and dropping any layer would result in a worse performance.

(3) To make training faster, the non-saturating neurons and a very efficient GPU implementation of the convolution operation were used.

(4) To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective.

(5) It trained the nextwork on the raw RGB values of the pixels.

(6) It introduces the feature vectors in the last hidden layer to help explain the model.

1.2 Suggestions/Extensions

This paper introduced the deep neural networks to the public image classification use. There are a lot of good ideas in this paper that could be used for the image classification, like using RGB values to represent image data, using SGD to optimize the model, using the deep and wide network for a complex task and using multiple GPU to train the model and thus save training time.

2 Deep Residual Learning for Image Recognition

2.1 Paper Summary

The depth of the network is critical to model performance. However, as the depth deepens, the network will encounter the famous "gradient disappearance" and "gradient explosion" problems during training. They prevent convergence from the beginning of training. To some extent, these problems can be solved by normalized initialization and intermediate normalization, which makes dozens of layers The network of is able to converge on the stochastic gradient descent of backpropagation.

However, as the number of network layers deepens, a problem called "degradation" for training accuracy is exposed: as the network depth increases, the accuracy on the training set reaches saturation and then rapidly declines. Obviously, this phenomenon is not caused by over-fitting, because over-fitting will make the accuracy on the training set extremely high. If you add more layers to a model with appropriate depth, it will bring higher training error.

In order to solve the degradation problem, this paper proposed a network called Deep residual learning framework in this paper. In the structure of this framework, each stacked layer fits residual mapping instead of directly fitting the underlying mapping expected by the entire building block. This paper also provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth.

2.2 Suggestions/Extensions

This work could be a backbone in the deep learning network for different tasks of computer vision, because it is easier to optimize, and can gain accuracy from considerably increased depth. For example, it can be used for more complex image classification task. Besides, it could work as the baseline for network improvements.

3 Explaining and Harnessing Adversarial Examples

3.1 Paper Summary

Early attempts focused on nonlinearity and overfitting to explain the existence of adversarial examples - inputs formed by applying small but intentionally worst-case perturbations to examples from the datase.

However, this article proposes a linear hypothesis different from the previous papers, which is simpler, and can also explain why softmax regression is vulnerable to adversarial examples. At the same time, the paper proposes a simple method for generating adversarial samples - Fast Gradient Sign Method (FGSM). The core idea is to add disturbances in the opposite direction of the gradient to increase the distance between the adversarial example and the original sample. After generating, these samples could be used for adversarial training, which could result in regularization, which helps reducing the test set error of a maxout network on the MNIST dataset.

In general, this article mainly describes three aspects of adversarial samples: 1. Existence, 2. Attack method, 3. Defense method.

3.2 Suggestions/Extensions

FGSM is a good way to generate adversarial examples for adversarial training. Therefore, when we want to do the adversarial training to protect machine learning models against adversarial examples, we could use the FGSM method, where the gradient can be computed efficiently using backpropagation. In this way, the model could be robust against perturbations in the data.

4 GloVe: Global Vectors for Word Representation

4.1 Paper Summary

There are problems with the existing two mainstream word vector representations: global matrix factorization and local context window. The former effectively uses statistical information, but

performs poorly on the word analogy task; the latter, like the Skip Gram models of word2vec performs well on the vocabulary analogy task, but its training focus on the vocabulary around the central word only, and does not make good use of the statistical information of the entire document.

Therefore, this paper proposed a new global log-bilinear regression model that combines the advantages of the two major model families. This model efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition.

4.2 Suggestions/Extensions

GloVe, like word2vec, is one of the word representation methods. It can be used in any NLP tasks like Named Entity Recognition and Sentiment analysis, as the word embedding. Compared to word2vec, GloVe allows for parallel implementation, which means that it's easier to train over more data.

5 A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning

5.1 Paper Summary

Many systems lack a unified framework to perform deep semantic tasks. These systems have three major shortcomings: 1) They obtain shallow semantics because they are all linear. 2) In order to make the linear classifier perform well, they constructed many features manually. 3) They are separated from other tasks, leading to propagation errors.

To solve this problem, this paper proposed a method using a single convolutional neural network training jointly on multiple tasks using weight-sharing. In this way, given a sentence, the model could outputs a host of language processing predictions, such as part-of-speech tags, chunks, named entity tags, semantic roles. All of these tasks are integrated into a single system which is trained jointly. All the tasks except the language model are supervised tasks with labeled training data. The language model is trained in an unsupervised fashion on the entire Wikipedia website. Training this task jointly with the other tasks comprises a novel form of semi-supervised learning.

This paper also show how both multitask learning and semi-supervised learning improve the generalization of the shared tasks, resulting in state-of-the-art performance.

5.2 Suggestions/Extensions

The novel form of semi-supervised learning proposed in this paper and the jointly multi-task training could be adopted in many ways. For example, we could uses this method to pre-trained BERT. Besides, it could be adopt in other fields like learning from different classification dataset to get the videos representation.