CNN and Deep learning with MNIST

Student ID: u5953029
Student Name: Zhuoqun Li

Abstract

As is mentioned in the assignment 1, the bp algorithm has many disadvantages. It makes the learning process of machines troubles. Thus, it can't be used in deep learning neural network.
In the assignment 2, I am trying to use deep learning(CNN) to let the machine automatically learn good features without the manual selection process. I try to deal with a new dataset, mnist, which is a classical dataset for CNN and deep learning.

Introduction

Suppose we have a system S, which has n layers (S1,...Sn) whose input is I and whose output is O, which is visually represented as: I => S1 => S2 =>..... => Sn => O, if the output O is equal to the input I, that is, the input I has no information loss after this system change (Oh, Daniel said that this is not possible. Inequality), suppose that processing a information to obtain b, and then b processing to obtain c, then it can be proved: mutual information of a and c will not exceed the mutual information of a and b. This shows that information processing will not increase information, most of the processing will Loss of information. Of course, if it is worthless to lose information that is useless, it remains unchanged, which means that input I goes through every layer of Si without any loss of information, ie, at any level of Si, it This is another representation of the original information (ie input I). Now back to our theme Deep Learning, we need to learn features automatically. Suppose we have a bunch of input I (like a bunch of images or text). Suppose we have designed a system S (with n layers). We adjust the parameters in the system. So that its output is still input I, then we can automatically get a series of hierarchical features of the input I, namely S1, ..., Sn.

For deep learning, the idea is to stack multiple layers, that is, the output of this layer as the input to the next layer. In this way, it is possible to hierarchically express the input information.

In addition, the front is assuming that the output is strictly equal to the input. This limit is too strict. We can relax this limit slightly. For example, we only need to make the difference between input and output as small as possible. This relaxation will lead to another type of different Deep. Learning method. The above is the basic idea of Deep Learning.

In 2006, Professor Geoffrey Hinton and his student Ruslan Salakhutdinov of the University of Toronto, Canada, and his student Ruslan Salakhutdinov published an article in Science that opened the wave of deep learning in the academic and industrial world. This article has two main viewpoints: 1) The multi-hidden layer artificial neural network has excellent feature learning capabilities, and the learned features have a more characterization of the data, which is conducive to visualization or classification; 2) deep neural networks Difficulties in training can be effectively overcome by layer-wise pre-training. In this article, layer-by-layer initialization is achieved through unsupervised learning.

Currently, most of the learning methods such as classification and regression are shallow structure algorithms. The limitation is that the ability to represent complex functions is limited in the case of finite samples and computational units, and the generalization ability of complex classification problems is restricted. Deep learning can learn a deep nonlinear network structure, realize the approximation of complex functions, represent the distributed representation of input data, and demonstrate a strong ability to learn the essential characteristics of data sets from a few sample sets. (The advantage of multi-layer is that you can express complex functions with fewer parameters.)

The essence of deep learning is to learn more useful features by building a machine learning model with a lot of hidden layers and massive training data, so as to ultimately improve the accuracy of classification or prediction. Therefore, "deep model" is the means, and "characteristic learning" is the purpose. Different from traditional shallow learning, the difference in deep learning is that: 1) The depth of the model structure is emphasized, usually 5, 6 or even 10 layers of hidden layer nodes; 2) The importance of feature learning is clearly highlighted. That is to say, through layer-by-layer feature transformation, the feature representation of the sample in the original space is transformed into a new feature space, thereby making classification or prediction easier. Compared with the method of constructing features of artificial rules, using big data to learn features makes it possible to describe the rich internal information of the data.

Method

If you train all layers at the same time, the time complexity will be too high; if you train one layer at a time, the deviation will be passed layer by layer. This will face the opposite problem of supervised learning above, and it will seriously under-fit because the depth of the network has too many neurons and parameters.

It is divided into two steps, one is to train one layer of network at a time, and the other is to tune, so that the high-level representation r generated by the original representation x upward and the x' generated by the high-level representation r are as consistent as possible. the way is:

1) First build a single layer of neurons layer by layer, so that each time you train a single-layer network.

2) After all layers have been trained, Hinton uses the wake-sleep algorithm for tuning.

Turn the weights of the layers except the topmost layer into bidirectional, so that the top layer is still a single-layer neural network, and other layers become the graph model. The upward weight is used for "cognitive" and the downward weight is used for "generating." Then use the Wake-Sleep algorithm to adjust all the weights. The consensus between cognition and generation is to ensure that the generated top-level representation can restore the underlying nodes as correctly as possible. For example, if a node at the top level represents the face, then the image of all faces should activate the node, and the resulting downward-looking image should be able to appear as a general face image. Wake-Sleep algorithm is divided into wake and sleep.

1) Wake phase: The cognitive process generates an abstract representation of each layer (node state) through external features and upward weights (cognitive weights), and uses gradient descent to modify the downlink weight between layers (generate weights). That is, "If the reality is different from what I have imagined, changing my weight makes my imagination something like

this."

2) sleep stage: the generation process, through the top-level representation (concept learned when awakening) and the downward weight, to generate the underlying state, while modifying the upward weight between layers. That is, "if the dream scene is not a corresponding concept in my mind, changing my cognitive weight makes this scene seem to me the concept."

Details as follows:

1) Use non-supervised learning from the bottom up (that is, start from the bottom, layer by layer to top level training):

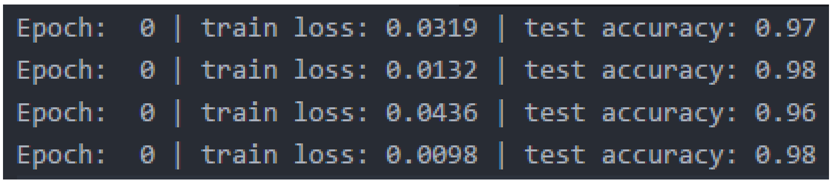
Using non-calibrated data (with calibration data also available) to stratify parameters at each level, this step can be seen as an unsupervised training process, which is the most different from the traditional neural network (this process can be seen as a feature learning process)

Specifically, the first layer is trained first with no calibration data, and the first layer parameters are learned first (this layer can be seen as a hidden layer of a three-layer neural network that minimizes the difference between output and input). The limitation of capacity and the sparsity constraint make the obtained model able to learn the structure of the data itself, so as to obtain features that are more capable of expressing than the input; after learning to obtain the n-1th layer, the output of the n-1 layer is taken as the first. The n-layer input trains the n-th layer, from which each layer's parameters are obtained.

2) Top-down supervised learning (that is, training with tagged data, error propagation from the top, fine-tuning the network):

Based on the parameters obtained in the first step to further fine-tune the parameters of the entire multi-layer model, this step is a supervised training process; the first step is similar to the neural network's random initialization initial value process, since the first step of DL is not random Initialization, but obtained by learning the structure of the input data, so that the initial value is closer to the global optimum, so that better results can be achieved; so the deep learning effect is largely due to the first step of the feature learning process.

Result



```
Epoch:  0 | train loss: 0.0319 | test accuracy: 0.97
Epoch:  0 | train loss: 0.0132 | test accuracy: 0.98
Epoch:  0 | train loss: 0.0436 | test accuracy: 0.96
Epoch:  0 | train loss: 0.0098 | test accuracy: 0.98
```

Conclusion

Deep learning is a new field in machine learning research. Its motivation lies in building and simulating the neural network of the human brain for analytical learning. It imitates the mechanism of the human brain to interpret data such as images, sounds, and texts. Deep learning is a kind of unsupervised learning.

The concept of deep learning stems from the study of artificial neural networks. A multilayer sensor with multiple hidden layers is a deep learning structure. Deep learning creates more abstract high-level representation attribute categories or features by combining low-level features to discover distributed representations of data.

Deep learning itself is a branch of machine learning. Simple can be understood as the development of neural networks. About two or three decades ago, neural network was once a particularly hot direction in the ML field, but it has since slowly faded out. The reasons include the following aspects:

1) It is easier to overfit, the parameters are more difficult to tune, and many tricks are needed;
2) The training speed is slower, and the effect is not better than other methods in the case where the level is relatively small (less than or equal to 3);

So for about 20 years in the middle, the neural network was little noticed. This time is basically the world of SVM and boosting algorithms. However, an infatuated old Mr. Hinton, he persisted, and finally (and others Bengio, Yann.lecun, etc.) put together a practical deep learning framework.

There are many differences between Deep learning and traditional neural networks.

The difference between the two is that deep learning adopts a similar hierarchical structure of neural networks. The system consists of a multi-layer network consisting of input layer, hidden layer (multi-layer), and output layer. Only the adjacent layer nodes have connections, the same layer. And cross-layer nodes are not connected to each other, each layer can be seen as a logistic regression model; this hierarchical structure is closer to the structure of the human brain.

In order to overcome the problems in neural network training, DL adopts a very different training mechanism from neural networks. In the traditional neural network, the method of back propagation is adopted. In simple terms, an iterative algorithm is used to train the entire network, the initial value is set at random, the output of the current network is calculated, and then the difference between the current output and the label is used. Change the parameters of the previous layers until convergence (the whole is a gradient descent method). Deep learning as a whole is a layer-wise training mechanism. The reason for this is because, if the mechanism of back propagation is used, for a deep network (above 7 layers), the residual spread to the frontmost layer has become too small, with the so-called gradient diffusio.

Convolutional neural network CNN is mainly used to identify displacement, scaling, and other forms of distortion-invariant two-dimensional graphics. Since CNN's feature detection layer learns through training data, when CNN is used, explicit feature extraction is avoided, and learning is implicitly performed from training data; in addition, the weights of neurons on the same feature mapping plane are used. The same, so the network can learn in parallel, which is also a great advantage of the convolutional network relative to the network where the neurons are connected to each other. Convolutional neural networks have unique advantages in terms of local recognition of their weights and image processing. Their layout is closer to the actual biological neural network. Weight sharing reduces the complexity of the network, especially multidimensional. The feature that the input vector image can be directly input to the network avoids the complexity of data reconstruction during feature extraction and classification.

The classification of streams is almost always based on statistical features, which means that certain features must be extracted before resolving. However, explicit feature extraction is not easy and it is not always reliable in some application problems. Convolutional neural network, which avoids explicit feature sampling, implicitly learns from training data. This makes convolutional neural networks distinctly different from other classifiers based on neural networks. It fuses feature extraction functions into multi-layer perceptrons through structural reorganization and weight reduction. It can directly process grayscale images and can be used directly to process image-based classifications.

The convolutional network has the following advantages over the general neural network in image processing: a) The input image and the topology of the network can be well matched; b) Feature extraction and pattern classification are performed simultaneously and simultaneously in the training; c) Weights Sharing can reduce the training parameters of the network, making the neural network structure simpler and more adaptable.

The close relationship between such interlayer connections and airspace information in CNNs makes it suitable for image processing and understanding. Moreover, it also shows superior performance in automatically extracting the salient features of the image. In some examples, the Gabor filter has been used in an initial preprocessing step to simulate the human visual system's response to visual stimuli. In most of the current work, researchers have applied CNNs to a variety of machine learning problems, including face recognition, document analysis, and language detection. In order to achieve the purpose of finding the coherence between frames in a video, CNNs currently train through a temporal coherence, but this is not unique to CNNs.

1) Deep learning summary

Deep learning is an algorithm for multilayered (complex) expressions that automatically learn the potential (implicit) distribution of the data to be modeled. In other words, the deep learning algorithm automatically extracts low-level or high-level features needed for classification. A high-level feature means that the feature can be hierarchically dependent on other features. For example, for machine vision, a deep learning algorithm learns from the original image to obtain a low-level representation of it, such as edge detectors, wavelet filters, etc. , and then establish expressions based on these low-level expressions, such as linear or nonlinear combinations of these low-level expressions, then repeat this process, and finally obtain a high-level expression.

Deep learning can obtain features that better represent data. At the same time, because the model has many levels, parameters, and sufficient capacity, the model has the ability to represent large-scale data. Therefore, this feature is not obvious for images and speech (manual design and many Problems with no direct physical meaning can achieve better results on large-scale training data. In addition, from the perspective of pattern recognition features and classifiers, the deep learning framework combines features and classifiers into a framework to use data to learn features, reducing the enormous workload of manual design features in use (this is currently the industry Engineers pay the most efforts. Therefore, not only can the effect be better, but also there are many conveniences for use. Therefore, it is a set of frameworks that are worth paying attention to. Everyone who does ML should pay attention to it.

Of course, deep learning itself is not perfect, nor is it a tool to solve any ML problem in the world, and it should not be magnified to an omnipotent degree.

2) Deep learning future

Deep learning still has a lot of work to study. The current focus is still to learn from the field of machine learning some methods that can be used in deep learning, especially in the field of dimensionality reduction. For example, at present, a job is sparse coding. Compressive sensing theory reduces dimensionality of high-dimensional data, so that very few elements of the vector can accurately represent the original high-dimensional signal. Another example is semi-supervised pop learning. By measuring the similarity of training samples, this similarity of high-dimensional data is projected into low-dimensional space. Another inspiring direction is evolutionary programming approaches, which can be conceptually adaptive learning and change the core architecture by minimizing engineering energy.

Deep learning also has many core issues that need to be resolved:

(1) For a particular framework, how many dimensions of input it can perform better (if it is an image, it may be millions of dimensions)?

(2) Which architecture is effective for capturing short-term or long-term time dependence?

(3) How to fuse multiple perceptual information for a given deep learning architecture?

(4) What is the correct mechanism to enhance a given deep learning architecture to improve its robustness and invariance to distortion and data loss?

(5) Are there other more effective and theoretically based deep model learning algorithms in the model?

Exploring the new feature extraction model is worth studying in depth. In addition, an effective parallel training algorithm is also worth studying. Currently, the minimal-batch-based stochastic gradient optimization algorithm is difficult to perform parallel training on multiple computers. The usual approach is to use a graphics processing unit to speed up the learning process. However, a single machine GPU is not suitable for large-scale data recognition or similar task data sets. In terms of deep learning application development, how to rationally and fully utilize deep learning to enhance the performance of traditional learning algorithms is still the focus of research in all areas.

Reference

1. F. Giannini, V. Laveglia, A. Rossi, D. Zanca, A. Zugarini(2017). Neural Networks for Beginners A fast implementation in Matlab, Torch, TensorFlow
2. Gregory Cohen, Saeed Afshar, Jonathan Tapson & Andre van Schaik. EMNIST: an extension of MNIST to handwritten letters
3. Patrice Y. Simard, Dave Steinkraus, John C. Platt(2003). Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis