**Applications of Information Theory in Neural Network**

Group Name: 安知鸿鹄之志

Author: 刘安 and 任志远

**Abstract**: Neural Network is hot and has been talked for a long time.But in fact, many people don’t why Neural Network can do something such well.So this article will show you how to understand Neural Network using a insight of Information Theory, and using this way we will also try to improve and rebuild our model, which not only has a better learning speed than traditional BP but also keep the capability of the network.

1. Understand network from information theory

A.network is similar to ‘encoder-decoder’

We use X to denote ‘Input Layer’ and Y to denote ‘True Labels’.And we have already known that neural network is designed for finding the function relationship between X and Y.

So, X can be treated as the high entropy distribution of Y.High entropy means that X has more information except Y.

For example, ‘the girl is nice!’ includes ‘positive’ information.But as the same time,it also includes other information like ‘the object is a girl’, ‘the sentence is present tense’ and so on.

Now imaging that each hidden layer can be described as a vector ‘H’,so the whole hidden layers can be denoted as H\_0, H\_1, H\_2 ......H\_(n-1).

Each layer is only determined by the prior layer,so in fact, the whole hidden layers is a markovi model.And it is showed in Fig.1.

So essentially each layer constructs a different piece of information in a different abstract form.

In the Markov expression of the neural network, each layer becomes part of the information which in the information theory is often viewed as a continuous refinement of the relevant information.

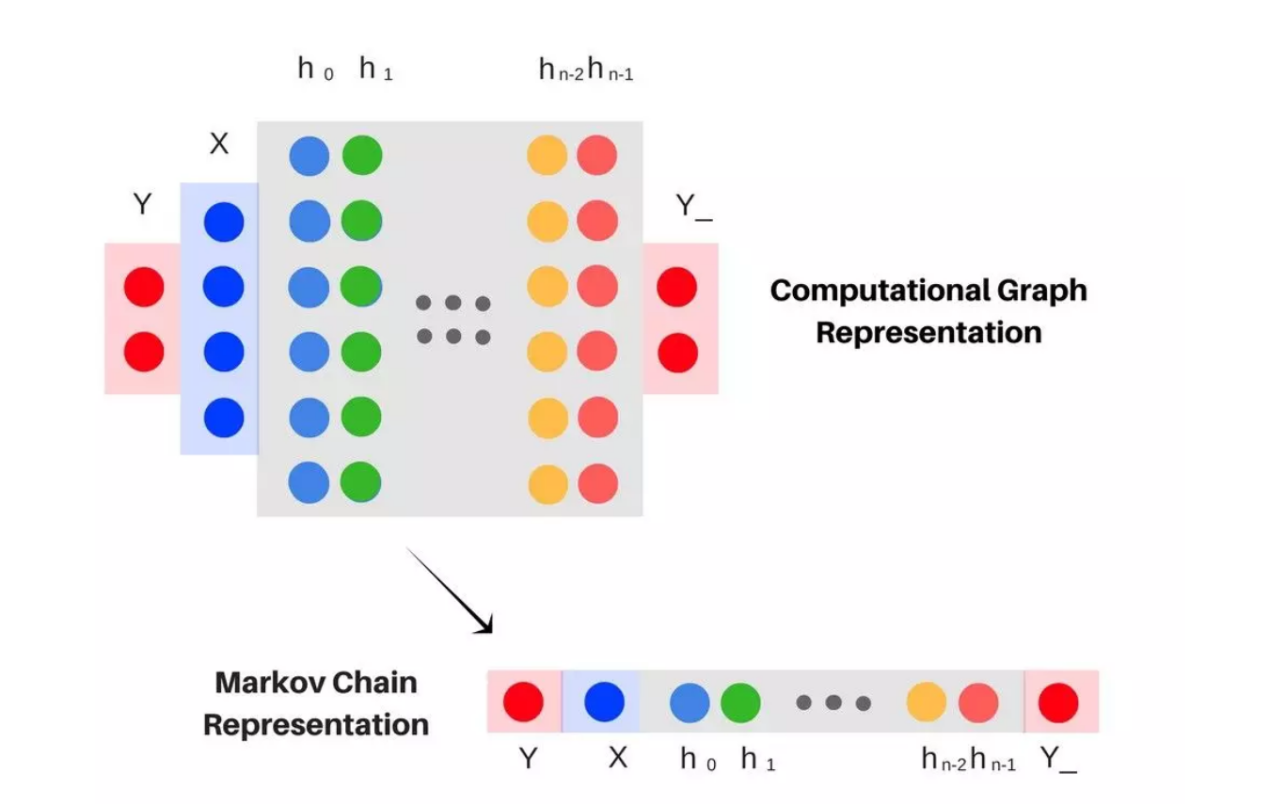


Fig.1 A formal visual neural network of Markov chains

But from a new perspective ‘encoder-decoder’,we can take the view that X is encoded and Y\_ is the result of the decoder as Fig.2 shows.

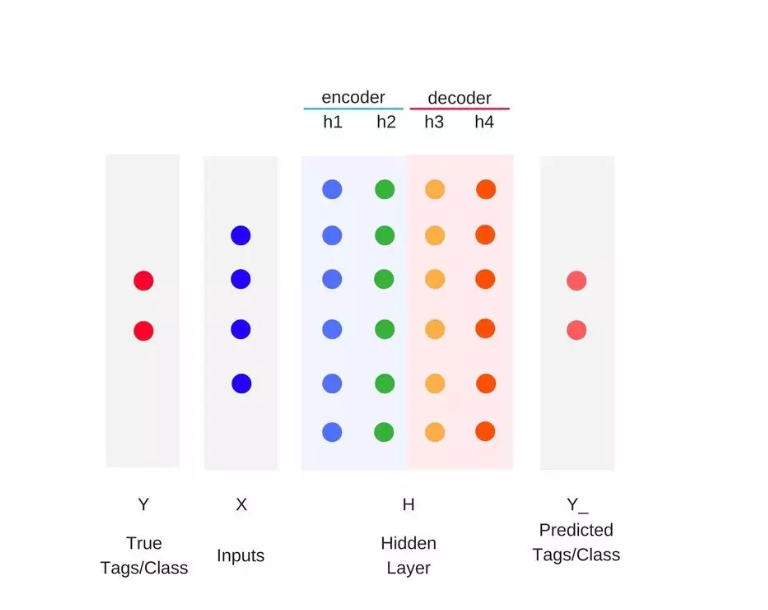


Fig.2 Hidden Layer can be taken as encoder and decoder

Then, for enough hidden layers: the complexity of

neural network sampling is determined by the mutual information encoded by the last hidden layer, and the accuracy is determined by the mutual information decoded by the last hidden layer.

B.Visualization of training stage

We compute I(H,X) and I(H,Y) ,H is the whole hidden layers we refer to above.

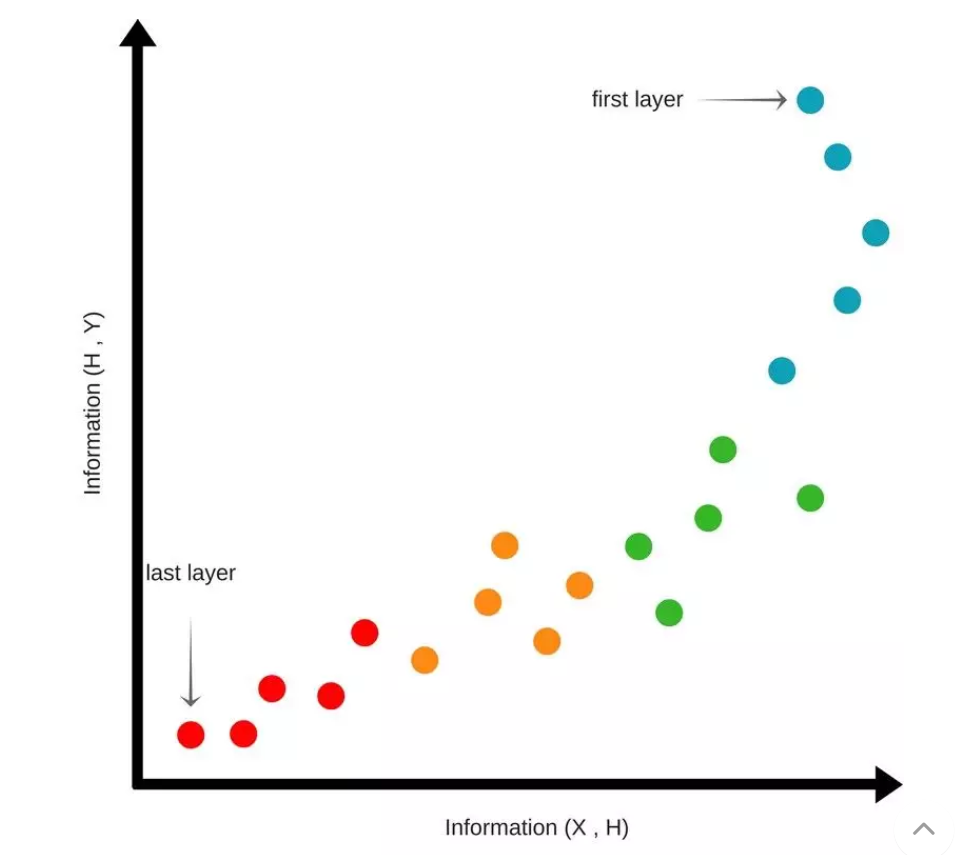


Fig.3 Begin status

Begin status(Fig.3):

At the beginning, We randomly initialize the weights of the network.H don’t know anything about X and Y.So their mutual information is lower as the layer becomes deeper.

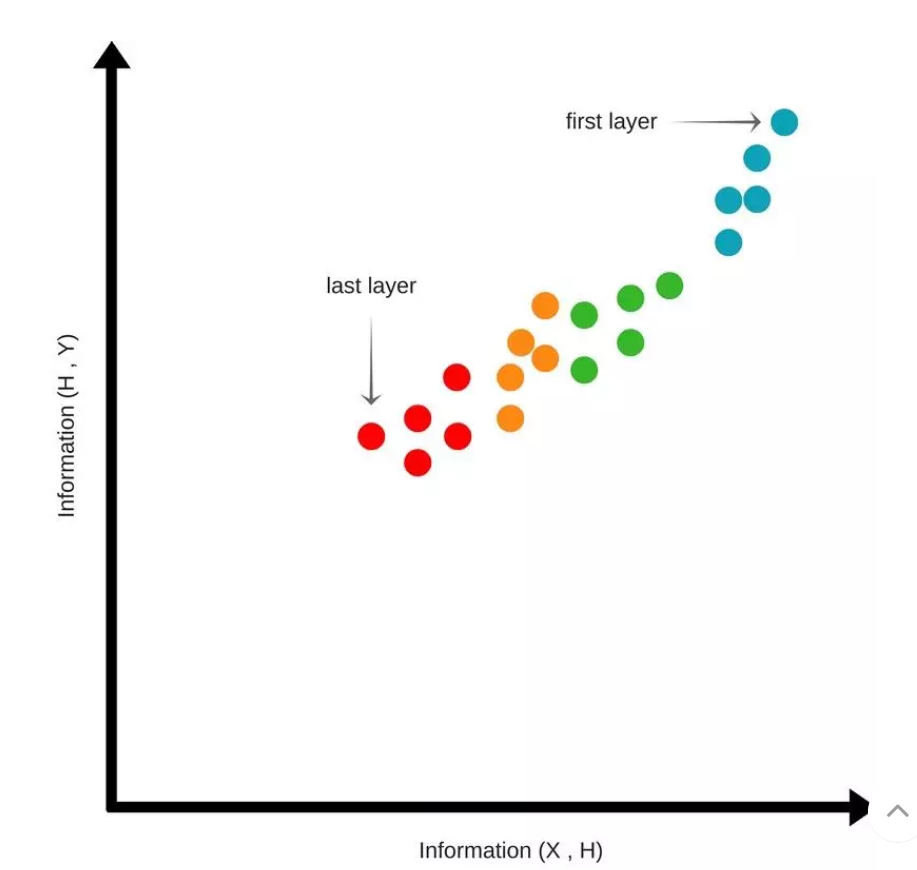


Fig.4 Compression status

Compression status(Fig.4):

As the training is going on,the point ‘last layer’ is moving up and right,which means mutual information between H and X(Y) is increasing.

This means the information in X is compressed to the part information of X which is related to Y.

Extension status(Fig.5):

After the Compression status,point starts to move up and left.This means network tends to lose some information from X ,but keep the information of Y;

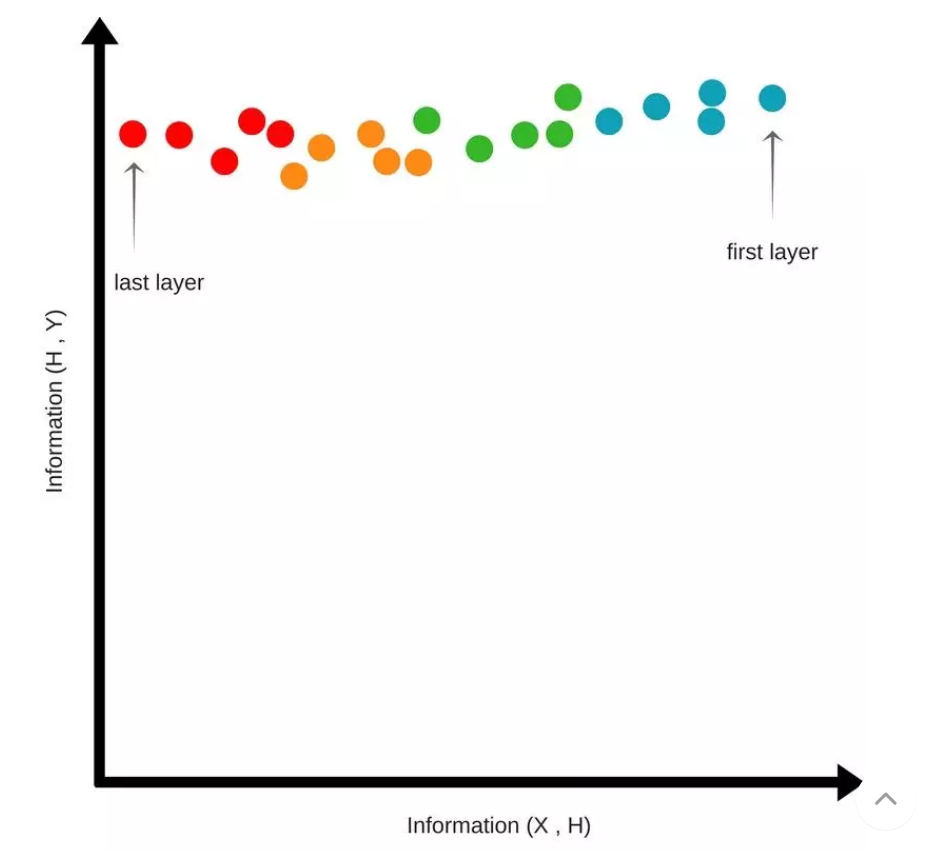


Fig.5 Extension status

1. information bottleneck

Information bottleneck can be cut into two steps.

1. Fitting; 2. Compression

Fitting means remember all the information.

Compression means forget all the unrelated information.

The left part is knowledge.

In fact, this is how people learn.What’s more,it is how neural network do.

Using the information bottleneck, we can easily answer some questions in neural network.

(1)The speed to train model have a tendency of converging very fast at the beginning and becoming slower at the end,because the time used to compress information increases exponentially at the later stage.

(2)In the case of without sufficient samples,DNN tends to have better generalization ability than expected because of the compression I(X; T) reduces the number of bits per compression by half, and the number of samples which is

required to maintain the same generalization error is

also halved.

(3)From low layer to high layer, The compression of upper limit can be achieved by using some tricks to change the upper limit of structural information carrying capacity. Reducing the number of neurons

(including Max pooling and dimensionality reduction,

as well as temporary drop out) and reducing the enumeration number of discrete data (including using nonlinear activation functions, normalization, and argmax),which reduced the generalization error in some extent.

(4)If it is a model with good prediction after training convergence,I(Y;Ti) is close to H(Y) at each layer, so Y can be accurately predicted no matter which layer is cut off and the eigenvalue T is re-entered to replace the original input X.This provides a window for analyzing the eigenvalue meaning of DNN layer by layer.

1. Build a model using information theory

Let's take the 2-classification problem as an example.

Firstly,we introduce some signs. ‘N’ denotes the number of training samples. denotes the number of the training samples of type 1.denotes the number of the training samples of type 2.We think of the nature of a neural network as a hyperplane.A hyperplane divides the training samples into 2 sets(+ and -).So there will be 4 results.

|  |  |
| --- | --- |
|  | Type 1 in + set |
|  | Type 1 in - set |
|  | Type 2 in + set |
|  | Type 2 in - set |

And they have the following relationship:

(1)

And our goal is to make G(Information gain) big as we can.

(2)

En means entropy.S is training samples.k can be + or -.

When we get the whole training samples.En(S) is certain.

So to make G larger,we should try to make the second term smaller.Let’s call the second term ‘E’.

(3)

For neural network, N+ is the number of neural units whose output is 1.So we get the following part:

(4)

The rest can be done in the same manner:

(5)

D\_j is jth training sample’s expected output.O\_j is jth training sample’s true output.x\_jk is jth training sample’s kth component.w\_lk is the weight between lth neural unit to kth neural unit in next layer.In this way,we cannot get an integer, but we can gain on that number.

So, here E in(3) become a function of weight in neural network .And neural network’s job is to find a set of weight to accomplish the task.According to this, we can get the learning rule.

(6)

(7)

(8)

So, weight can be modified by this rule.

(9)

We call this network is Entropy-Based Neural Network.If we compare it with BP, we will know its advantages.

(1)From Fig.6,we can know that the rate of identification and the rate of generalization of EBNN is better than those of BP.

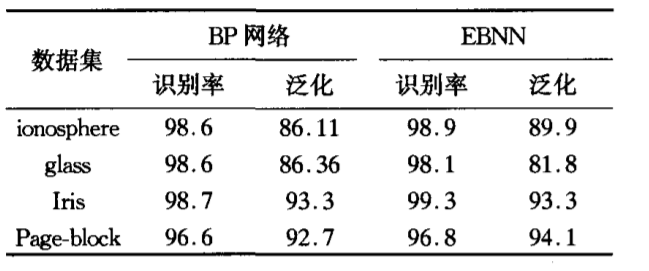


Fig.6 BP VS EBNN

(2)At learning speed,in BP network,the number of layer and the number of neural units in each layer is uncertain.

So,when constructing the network, project needs many times of learning to modify the network.But EBNN only needs 3 layers and it can be built in the base of data sets.

1. Conclusion

|  |  |
| --- | --- |
| Help understand neural network | Similar to ‘encoder-decoder’ |
| Visualization of training stage |
| information bottleneck |
| Help build neural network | Example: binary classification |

This article talks about how Information Theory help in neural network:

1. We can take the neural network as a structure like “encoder-decoder”.So we can use the theory related to “encoder-decoder”to analysis neural network.
2. Using mutual information, we can visualize the training stage, and we can see different appearance of different status.
3. Using information bottleneck, we can know why neural network can work so well:remember all the information and forget useless information.
4. Using information theory,we create a new network called EBNN, and its performance is better than BP.

**References**

1. <信息理论框架下的神经网络构建> 李爱军，罗四维，刘蕴辉，黄 华. (北京交通大学计算机与信息技术学院，北京100044)
2. <https://www.cnblogs.com/ws12345/p/8036710.html>
3. <https://www.jiqizhixin.com/articles/2018-04-11-2>
4. <https://blog.csdn.net/qq_35267530/article/details/78084462>