MMIM: An Interpretable Regularization Method for Neural Networks (Student Abstract)

Nan Xie, Yuexian Hou*
College of Intelligence and Computing, Tianjin University, Tianjin, China
{xienan6, yxhou}@tju.edu.cn

Abstract

In deep learning models, most of network architectures are designed artificially and empirically. Although adding new structures such as convolution kernels is widely used, there are few methods to design new structures and mathematical tools to evaluate feature representation capabilities of new structures. Inspired by ensemble learning, we propose an interpretable regularization method named **M**inimize **M**utual **I**nformation **M**ethod(MMIM), which minimize the generalization error by minimizing the mutual information of hidden neurons and provides ideas for designing new structures. The experimental results also verify the effectiveness of our proposed MMIM.

Introduction

Deep learning has become the mainstream framework for various tasks. It increases the number of network layers or designs of new network structures on the basis of classic DNN, CNN and other networks to improve the performance of the model. Convolutional neural networks, from LeNet to ResNet, increase the number of layers from 5 to 100, and add convolution kernels, dropout and other structures. Although the result of such complex networks are usually better, the network is a black box. On the one hand, it is difficult to understand what features neural networks learn when the number of layers increases. The work in this area includes visual interpretability mainly involved in visualization of hidden units, etc. On the other hand, it lacks methods to design a new structure and mathematical tools to evaluate feature representation capabilities of networks.

In response to the second problem above, Zhang et al. propose a generalization error upper bound of neural networks from an information-theoretic perspective. They introduce a new definition of redundancy to describe the diversity of hidden units by mutual information and only use two mutual information in the upper bound as a regularizer. However, this does not sufficiently minimize the entire upper bound, only a part of it. We propose a two-step regularization method named **M**inimize **M**utual **I**nformation **M**ethod(MMIM) that minimizes the entire upper bound instead of a part. Our experiments minimize the mutual information by *Mutual Information Neural Estimation* (Belghazi et al. 2018) to update network parameters. In addition, inspired by ensemble learning, we interpret why MMIM performs better.

The main idea of ensemble learning is to construct a strong learner which has better generalization performance through multiple weak learners. Weak learners are also called base learners. Obviously, base learners learn basic features, and strong learners learn overall features. For example, when using a convolutional neural network for image recognition tasks, different convolution kernels can be regarded as different base learners, and they may learn texture and contour of pictures. The strong learner, the final output of the neural network, learns what the picture is, which may be a dog or a cat.

We use mutual information to measure the independence between hidden layer neurons. Mutual information can help us explain our method from the perspective of ensemble learning. If we regard the neurons in a hidden layer of a neural network as a group of base learners, we hope that the more independent the neurons, the better. Therefore they learn different feature instead of learning same things. Obviously, the smaller the mutual information of random variables, the more independent they are.

Method

First, let's introduce Zhang's work. Let $P(\cdot)$ denote either a probability mass function (PMF), or a probability density function(PDF), depending on the random variable having either discrete or continuous support. The symbol $\mathbb{E}_X(\cdot)$ denotes expectation of the random object within the brackets with respect to the subscript random variable X.

For a stochastic variable X, its entropy is defined as

$$H(X) = -\mathbb{E}_X(\log P(X)) \tag{1}$$

The multivariate mutual information is defined as

$$I(X_1, X_2, \ldots, X_n) = \mathbb{E}_{(X_1, X_2, \ldots, X_n)}\left(\log \frac{P(X_1, \ldots, X_n)}{P(X_1), \ldots, P(X_n)}\right) \tag{2}$$

g is the error on the new sample. Zhang et al. derived a generalization error upper bound of the neural network

$$g \leq \sqrt{2\sigma^2}\sqrt{\begin{array}{c} I(h_1(X), \ldots, h_m(X)) - I(h_1(X), \ldots, h_m(X) \mid Y) \\ -\sum_{i=1}^{n} I(Y, h_i(X)) + H(Y) + \frac{H(W) - H(S_y, W \mid \hat{h}(S_x))}{n} \end{array}} \tag{3}$$

*Corresponding author:Yuexian Hou (yxhou@tju.edu.cn).

Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the label space. $X \in \mathcal{X}$. $Y \in \mathcal{Y}$. The layer before the output layer is called the sub-high-level layer. The output of neurons in the sub-high-level layer is $h_i(X)$ and there are *m* hidden units. σ is a subgaussian constant of loss function. *n* is the size of dataset *S*. *W* are the neural network parameters from the sub-high-level layer to the output layer. $S_x = \{X_1, X_2, \ldots, X_n\}, S_y = \{Y_1, Y_2, \ldots, Y_n\}$.$\hat{h} = (h_1, h_2, \ldots, h_m)$.

Zhang et al. propose a regularization method named redundancy decrease method(RDM) by mimimizing total loss

$$T_{\text{loss}} = E_{\text{loss}} + \lambda(I(h_1(X), \ldots, h_m(X)) - I(h_1(X), \ldots, h_m(X) \mid Y)) \tag{4}$$

where E_{loss} is the premier loss function of neural networks without any regularizer.

Note that they only used two terms in the Equation 3 and do not consider the influence of other terms. Our method takes each item into consideration. Since

$$\begin{aligned} H(W) - H\left(S_y, W \mid \hat{h}(S_x)\right) &\leq H(W) - H\left(W \mid \hat{h}(S_x)\right) \\ &= I\left(W, \hat{h}(S_x)\right) \end{aligned} \tag{5}$$

and the generalization error upper bound changes to

$$g \leq \sqrt{2\sigma^2}\sqrt{\begin{array}{c} I(h_1(X), \ldots, h_m(X)) - I(h_1(X), \ldots, h_m(X) \mid Y) \\ -\sum_{i=1}^{n} I(Y, h_i(X)) + H(Y) + \frac{I(W, \hat{h}(S_x))}{n} \end{array}} \tag{6}$$

We decompose terms of the second radical in g into C_1, C_2, C_3.

$$\begin{aligned} C_1 =& I(h_1(X), \ldots, h_m(X)) - I(h_1(X), \ldots, h_m(X) \mid Y) \\ & - \sum_{i=1}^{n} I(Y, h_i(X)) \\ C_2 =& H(Y) \\ C_3 =& \frac{I\left(W, \hat{h}(S_x)\right)}{n} \end{aligned} \tag{7}$$

Our method is as follows. C_2 is constant. So we minimize g by minimizing C_1 and C_3. C_1 are not related to W. C_3 is related to W. Therefore, firstly we minimize C_1. Then fix the network parameters before the sub-high-level layer, and minimize C_3, which minimize g. We call the above algorithm Mimimize Mutual Information Method(MMIM).

We use MINE to estimate mutual information in C_1 and C_3. MINE can be used to minimiaze mutual information and is linearly scalable in dimensionality as well as in sample size, trainable through back-prop, and strongly consistent.

Experimental Results

To verify the performance of MMIM, we test our method on fully connected neural networks and convolutional neural networks on the Fashion-MINST dataset.

We train two networks with different hyperparameter settings and apply MMIM to original network structures. In addition, we test the performance on networks by using C_4, C_5 and C_6 as a regularizer separately.

$$\begin{aligned} C_4 &= I(h_1(X), \ldots, h_m(X)) \\ C_5 &= -I(h_1(X), \ldots, h_m(X) \mid Y) \\ C_6 &= -\sum_{i=1}^{n} I(Y, h_i(X)) \end{aligned} \tag{8}$$

Methods	Accuracy	Methods	Accuracy
DNN	89.15	CNN	90.89
DNN with MMIM	**90.82**	CNN with MMIM	**92.39**
DNN with C_4	90.1	CNN with C_4	91.85
DNN with C_5	89.64	CNN with C_5	91.3
DNN with C_6	89.37	CNN with C_6	91.06

Table 1: Results of different methods on two networks.

We can see results from Table 1. We call fully connected networks as DNN in the table. From the table, we can see that MMIM or using C_4, C_5 or C_6 as a regularizer is effective. We think there are three reasons from the perspective of ensemble learning why MMIM is effective.

1. Regard neurons in the sub-high-level layer as base learners. The more independent the output of base learners, the better the generalization performance. According to C_4, the smaller the mutual information, the more independent the output of base learners.
2. The smaller C_5 means that the network learn the overall feature representation that more relevant to *Y*, which means the smaller generalization error upper bound.
3. The smaller C_6 is, the more relavant between the output of base learners and *Y*, which leads to smaller upper bound.

The first point ensures the diversity of base learners. The second point and the third point ensure that each base learner covers some samples, that is, learning ability. For example, convolution kernels can be regarded as some base learners that are independent of each other and have learning ability. In the future, we should design new network structures based on C_4, C_5 and C_6 with both diversity and learning ability.

Conclusions

In this paper, we propose a new interpretable regularization method MMIM with a better performance. Our results suggest that designing a new network structure should take into account both diversity and learning ability, and quantify feature representation capabilities of the network through our equations. In the future, we will focus on why and how network structures ensure their diversity and learning ability.

Acknowledgments

This work is funded in part by the National Key RD Program of China (2017YFE0111900), the National Natural Science Foundation of China (61876129) and the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 721321.

References

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062* .

Zhang, C.; Hou, Y.; Song, D.; Ge, L.; and Yao, Y. 2020. Redundancy of Hidden Layers in Deep Learning: An Information Perspective. *arXiv preprint arXiv:2009.09161* .