Neural Networks 104 (2018) 60–67

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Biased Dropout and Crossmap Dropout: Learning towards effective Dropout regularization in convolutional neural network

Alvin Poernomo, Dae-Ki Kang *

Department of Ubiquitous IT, Dongseo University, Busan, South Korea

ARTICLE INFO

Article history:
Received 23 July 2017
Received in revised form 14 November 2017
Accepted 22 March 2018
Available online 9 April 2018

Keywords:
Dropout
Regularization
Convolutional neural network

ABSTRACT

Training a deep neural network with a large number of parameters often leads to overfitting problem. Recently, Dropout has been introduced as a simple, yet effective regularization approach to combat overfitting in such models. Although Dropout has shown remarkable results on many deep neural network cases, its actual effect on CNN has not been thoroughly explored. Moreover, training a Dropout model will significantly increase the training time as it takes longer time to converge than a non-Dropout model with the same architecture. To deal with these issues, we address Biased Dropout and Crossmap Dropout, two novel approaches of Dropout extension based on the behavior of hidden units in CNN model. Biased Dropout divides the hidden units in a certain layer into two groups based on their magnitude and applies different Dropout rate to each group appropriately. Hidden units with higher activation value, which give more contributions to the network final performance, will be retained by a lower Dropout rate, while units with lower activation value will be exposed to a higher Dropout rate to compensate the previous part. The second approach is Crossmap Dropout, which is an extension of the regular Dropout in convolution layer. Each feature map in a convolution layer has a strong correlation between each other, particularly in every identical pixel location in each feature map. Crossmap Dropout tries to maintain this important correlation yet at the same time break the correlation between each adjacent pixel with respect to all feature maps by applying the same Dropout mask to all feature maps, so that all pixels or units in equivalent positions in each feature map will be either dropped or active during training. Our experiment with various benchmark datasets shows that our approaches provide better generalization than the regular Dropout. Moreover, our Biased Dropout takes faster time to converge during training phase, suggesting that assigning noise appropriately in hidden units can lead to an effective regularization.

1. Introduction

Nowadays, deep learning model has been evolving rapidly and showing remarkable results in various areas, including image recognition (Krizhevsky, Sutskever, & Hinton, 2012), speech recognition (Dahl, Yu, Deng, & Acero, 2012), language modeling (Arisoy, Sainath, Kingsbury, & Ramabhadran, 2012), information retrieval (Huang et al., 2013), etc. Its outstanding ability to learn very complex, both linear and non-linear relationships directly from the data makes deep neural network perfectly suitable to perform intelligent tasks similar to those performed by human brain. Recently, Convolutional neural network (CNN), one of the popular deep learning models which originally introduced for computer vision task, has constantly shown promising results in large areas and applications. With the advancement of GPU-based parallel computing and the availability of large public datasets on the Internet, such as Imagenet, CNN marks its superiority over the conventional computer vision approaches in ILSVRC2012 competition. A typical CNN architecture consists of several layers of alternating convolution and pooling layers at lower stages, which at higher stage, those layers are usually combined with fully connected layers, which correspond to a traditional MLP, to perform the classification task.

Deep models, including CNN with a large number of parameters has shown to be very powerful, but overfitting becomes a crucial problem as a big network with millions of parameters can be easily overfit, especially when provided with insufficient number of training data. This will lead to a poor prediction performance since the network fails to generalize the unseen or testing data. Correspondingly, many methods have been developed for regularizing deep neural networks, including early stopping of training as soon as its performance on validation data becomes degraded. Other methods, like introducing weight penalties and soft weight sharing (Nowlan & Hinton, 1992) have also been proven to improve the generalization performance.

* Corresponding author.
E-mail address: dkkang@dongseo.ac.kr (D.-K. Kang).

https://doi.org/10.1016/j.neunet.2018.03.016
0893-6080/

Recently, Dropout regularization approach has been introduced and considered as the most effective way to reduce overfitting of deep neural network (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012). Inspired by combining and averaging the results of many models to improve the performance of neural network, Dropout network literally tries to combine exponentially many different neural network architectures which share the same weights by randomly omitting a set of hidden units at each training case. Extensive research work have shown that Dropout significantly reduces overfitting and improves generalization performance (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Since then, there have been several extensions to improve the Dropout performance, including DropConnect (Wan, Zeiler, Zhang, Cun, & Fergus, 2013). It was proposed as a generalization of Dropout, which introduces sparsity within the weights W rather than the output vectors of a layer.

In this paper, we address two major issues which can improve the effectiveness of Dropout, especially in CNN models. First, regular Dropout introduces noise within the hidden units by temporarily dropping a random set of units during each training case. Each unit has the same probability p of being omitted during training. Introducing the same proportion of noise to each hidden units could render the effectiveness of Dropout itself, since each unit has different saliency and contribution to the network final performance. Biased Dropout is introduced to address this issue by segmenting the hidden units into two groups, "important" and "less-important" group and assigning a different Dropout rate for each group. The main purpose of this approach is to retain the "important" group by giving a lower Dropout rate as this group contains units with high-activated values, thus considered as important, whose deletion will have a major impact on the network performance. Viewed another way, limiting the noise on the "important" part of the units will make the network learn and converge faster as the parameter update for this important part of the network is not as noisy as using the regular Dropout. The second issue is that Dropout is commonly used only in the fully connected layers and its effect on convolution layer has not been sufficiently investigated. In convolution layer, units are organized in such planes, called feature maps and each feature map share a strong correlation between each other since all of them are the output of the same input with different filters. Instead of omitting a random set of units in every feature map, Crossmap Dropout tries to break the correlation between each unit with respect to all corresponding feature maps by omitting a random set of units in one feature map and applying the same Dropout mask to all remaining maps. This will ensure that each unit or pixel in the same position of all feature maps, which resembles the same feature, is either dropped or active during the training phase. This paper is structured as follows. Section 2 describes relevant previous work. Section 3 describes the motivation and the details of our proposed approaches. In Section 4, we present our experimental results with different datasets and the comparison with other related approaches. Section 5 describes the analysis and discussions of our proposed approaches. Section 6 covers the summary of our work.

2. Related work

Dropout as a regularization approach for deep fully connected networks was first introduced by Hinton et al. (2012). The idea is similar to Bagging approach (Breiman, 1996) where a set of models are trained on different subsets of the same training data and the prediction results of each model are averaged in order to improve the performance. Instead of training many different architectures, which would take a lot of computation time, especially when dealing with very big architectures and be difficult to find the optimal hyperparameter values for each model, Dropout provides a way of approximately combining exponentially many different networks by randomly dropping units with a fixed probability p for each training case. Therefore, training a Dropout neural network with n hidden units can be seen as training a collection of 2^n different "thinned" models with extensive weight sharing. These thinned networks can be combined into a single network at the testing phase by retaining all units and scaling the weights by $(1 - p)$. It is mentioned in the extension of the work of Srivastava et al. (Srivastava et al., 2014) that Dropout gives additional performance improvement when it is applied to the convolution layers. Following the success of Dropout, several work have been done on improving the performance and efficiency of Dropout.

Goodfellow, Warde-Farley, Mirza, Courville, and Bengio (2013) proposed Maxout Network to facilitate optimization and accuracy improvement of Dropout. The maxout unit picks the maximum value within a group of linear pieces as a generalization of rectified activation function which is capable of approximating the arbitrary convex function. They reported state-of-the-art results at that time by combining a convolution maxout network and Dropout approach, including a 0.45% and 2.47% test error rate on MNIST and SVHN dataset respectively. Wan et al. (2013) also proposed DropConnect, an extension of regular Dropout by introducing noise in weight vector, instead of the hidden units. DropConnect introduces a dynamic weight sparsity in the model, modifying e.g. a fully connected layer into a sparsely connected layer with its connections are chosen randomly during training. Although DropConnect offers a marginally slower training time than the regular Dropout, its performances are proven effective on some cases, including a 1.12% test error rate on MNIST dataset using a normal MLP model and 1.94% test error rate on SVHN dataset using CNN model.

The idea to give different Dropout rate to each hidden unit has been proposed by Ba and Frey (2013). They proposed another extension of Dropout, named Adaptive Dropout or Standout, which suggested that finding an optimal Dropout rate for each node based on the results of the previous layers will increase the performance of Dropout. The reason is that there might be certain hidden units that can individually make confident predictions for the presence or absence of an important feature or combination of features. Regular Dropout will ignore this confidence and drop the units with the same p probability at all times. Therefore, in Standout approach, Dropout rates are adaptively trained for each hidden node instead of putting a consistent rate. This process is equivalent to learning a separate belief network of Dropout rates on top of the existing neural network. Adaptive Dropout approach has been proven effective on several cases, such as MNIST and NORB datasets, although the additional computation time of binary belief network becomes its drawback.

Another modification of regular Dropout was introduced by Tompson, Goroshin, Jain, LeCun, and Bregler (2015), named Spatial Dropout. It is designed for convolution layer, where a random subset of feature maps will be dropped during each training case, instead of a random subset of units. Since natural images exhibit strong spatial correlation and the feature map activations in the convolution layer are also strongly correlated, regular Dropout tends to fail to improve generalization performance in convolution layers. They have found out that Spatial Dropout implementation improves performance on small-training-set-size dataset, such as FLIC (Frames Labeled in Cinema) dataset over the regular Dropout.

3. Proposed approach

3.1. Biased Dropout

In the original Dropout approach, each hidden unit shares a constant probability p of being omitted during training phase, despite that each hidden unit has different contributions to the network