DROPBAND: A CONVOLUTIONAL NEURAL NETWORK WITH DATA AUGMENTATION FOR SCENE CLASSIFICATION OF VHR SATELLITE IMAGES

Naisen Yang[a,b], Hong Tang[a,b], Hongquan Sun[c], Xin Yang[b]

[a] The Key Laboratory of Environmental Change and Natural Disaster,Beijing Normal University, China - yns@mail.bnu.edu.cn
[b] The State Key Laboratory of Earth Surface Processes and Resource Ecology,Beijing Normal University, China- (tanghong, yangxin)@bnu.edu.cn
[c] China Institute of Water Resources and Hydropower Research, China - sunhq@iwhr.com

KEY WORDS: dropband, dropout, data augmentation, neural networks, model combination

ABSTRACT:

Data augmentation is a common method that can prevent the overfitting of classification tasks in deep neural networks. This paper presents another kind of data augmentation method called DropBand that is useful for remote sensing image classification. Data augmentation is usually used along two dimensions of the image plane. This method executes this operation in the third dimension formed by all the spectral bands of an input image. With dropping a band of images out, the error rate of deep neural networks can be reduced. This method can also be viewed as a peculiar version of deterministic Dropout. The normal Dropout does not work well when it is applied to input channels of neural networks. To release this issue, dropping a band of input by schedule is employed. Moreover, model synthesis plays a key role in this procedure. To exclude the influence of increasing parameters, extra comparison groups are set up. The final experimental result shows that deep neural networks indeed benefit from the method of DropBand. This method improves the state-of-the-art on the latest SAT-4 and SAT-6 benchmarks.

1. INTRODUCTION

Data augmentation is widely used for preventing overfitting in a diverse range of machine learning technologies. It is also employed for training Deep Neural Networks (DNN) (Krizhevsky et al., 2012, Simonyan and Zisserman, 2014, Szegedy et al., 2015). Data augmentation enlarges the training datasets without touching the architectures of neural networks. By using label-preserving transformations, training dataset covers more regions of the input space.

In many previous works, the main forms of data augmentation, such as cropping and flipping, are performed along two dimensions of the image plane (Howard, 2013, Szegedy et al., 2015, Ciresan et al., 2012). Other forms of data augmentation include color casting (Wu et al., 2015) and intensity altering (Krizhevsky et al., 2012). But it is not thorough enough for remote sensing images. This paper proposes a new form of data augmentation that executes cropping along the third dimension formed by all the spectral bands of an input image. Below this method is referred as DropBand. We focus on convolutional neural networks (CNN) (Fukushima, 1980, LeCun et al., 1998) and show that deep neural networks trained with a subset of input channels can also achieve comparable accuracy.

Unsurprisingly, combining all these individual models trained on different subset of input bands can reduce error rate of models. Furthermore, we set up extra comparison groups which consist of models trained on all available bands with different random initialization. We observed that DropBand groups always surpass the corresponding comparison groups.

This paper is organized as follows. Section 2 describes the architecture of our method. The experimental setup and results are presented in section 3 and 4 respectively. In the end, section 5 gives a conclusion.

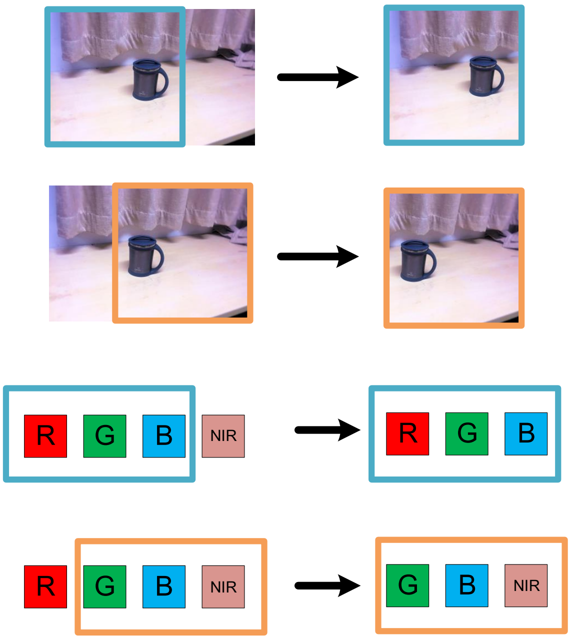

Figure 1. The concept of DropBand is an extension of cropping out a patch of the original input image for data augmentation. This method selects a subset of the input bands – four bands in this figure: red (R), green (G), blue (B) and Near Infrared (NIR).

2. METHODS

In this section, we give a concise definition of DropBand ,and discuss the connection between this method and other related methods.

2.1 DropBand

The DropBand method simply selects the subsets of input bands as training data. As shown on the top of Figure 2, DNN is trained on the dataset enlarged by DropBand. At test time, the scores of final predictions are a combination of predictions of each subset of input bands. Actually, this form of combination can be viewed as a combination of a number of networks that share same weights (on the bottom of Figure 2). For the reason that the model trained by DropBand can only capture the relevance among the subsets of bands, a base net trained with whole bands is added to the final model.

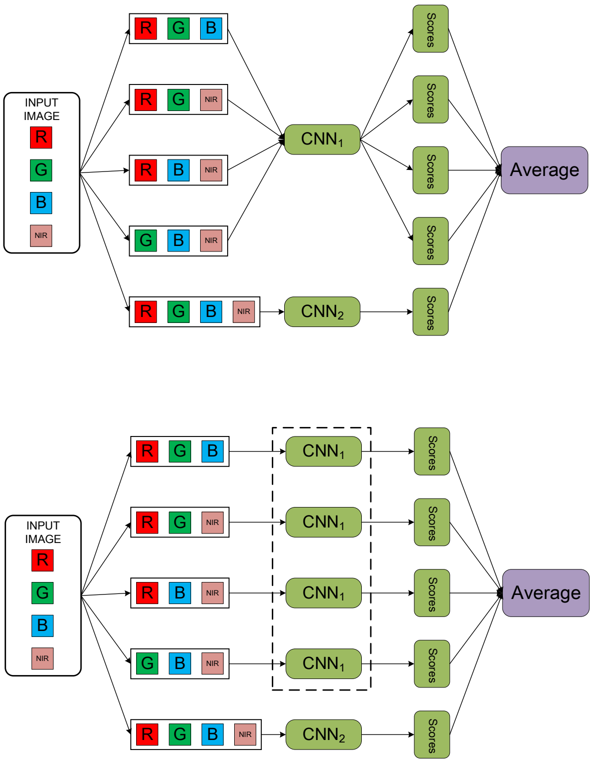

Figure 2. Architecture of the shared DropBand. The model on the bottom is an equivalence of the model on the top.

Unfortunately, by comparing with the base net, feeding the enlarged datasets to the input directly would lead to the changing of hyperparameters of neural networks. For the consideration of parameter tuning, we also train individual models on each maximum proper subset of whole input bands. This version of DropBand is illustrated in Figure 3. To distinguish the two forms of the DropBand method, we refer the model with weight sharing as the shared DropBand (on the top of Figure 2), and name the model in Figure 3 DropBand.

2.2 From A Data Augmentation Perspective

Augmenting data by domain knowledge improves the generalization of models (Krizhevsky et al., 2012, Szegedy et al., 2015). Transformations of data augmentation applied to samples preserve their original labels. This is the most common method to enlarge the dataset with little extra effort. In image classification tasks, the most common forms of data augmentation are flipping, cropping, and rotation. These operations are performed along two dimensions of the image plane.

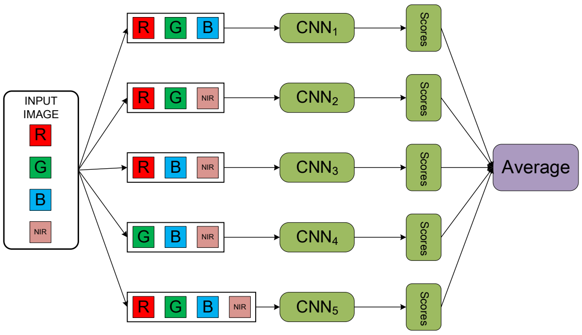

Figure 3. Architecture of the DropBand.

Altering intensities of RGB channels is also adopted widely (Krizhevsky et al., 2012). But for remote sensing images, it is not thorough enough. The most significant difference between the images of everyday objects and remote sensing images is that the latter have a wide range of available bands. These bands form the third dimension, so we can crop part of them along this dimension to augment datasets. Cropping along the dimension of bands amounts to set intensities of some bands to zero. In other words, this means that we have to abandon other bands at the same time. Therefore we call this form of data augmentation DropBand, and it has a deep relationship with Dropout (Srivastava et al., 2014) (this will be explained in section 2.3).

2.3 From A Dropout Perspective

By far, the most simple regularization method for training deep neural networks is Dropout (Srivastava et al., 2014). Dropout is an efficient method of model combination with weight sharing. It randomly drops a portion of the hidden units out of the neural network during training time. This amounts to training a single thinned network. All the hidden units work together during testing. Dropout significantly reduces overfitting and prevents the hidden units from co-adapting.

The original Dropout is a stochastic technique that brings noises to the units. DropBand can be regarded as a deterministic version of the original Dropout. It drops some of the input channels consistently instead of at random. This deterministic manner can ameliorate the drawback of misconvergence for the case of dropping the input channels out. When applying Dropout to the input channels with some probabilities of retaining an unit, the validation accuracy of neural network does not follow the increasing training accuracy and oscillates dramatically. This is shown in Figure 4. After using the DropBand, this phenomenon does not appear any more.

2.4 From A Feature Bagging Perspective

Feature bagging (Sutton et al., 2006) is to ameliorate the weight undertraining of conditional random fields (CRFs) (Lafferty et al., 2001). Weight undertraining is caused by which a few strong features dominate the result of classifiers. They build a collection of feature bags. Each feature bag consists of overlapping subset of input features. Different models are trained on the corresponding feature bags. Finally, the synthetic model is obtained by averaging the individual CRFs.

DropBand can also be interpreted as a kind of feature bagging if the input bands are seen as a kind of features. The proposed paper (Sutton et al., 2006) have shown the validity of feature bagging

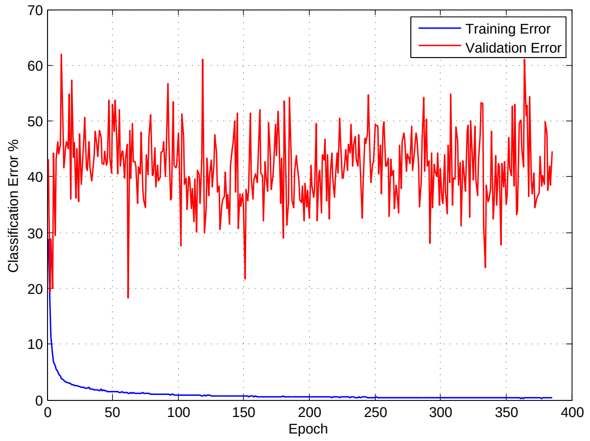

Figure 4. Effect of applying Dropout to the input channels directly. It is obtained by adding Dropout to the input channels with $p = 0.75$ on the SAT-6 dataset. Details of this model is described in section 3.

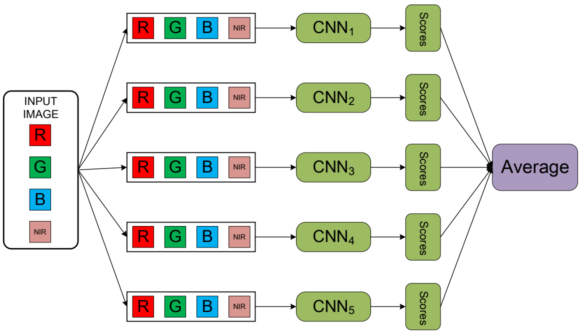

Figure 5. Architecture of comparison group. It consists of five base networks.

by comparing the synthetic model with a single CRF. Furthermore, we set the comparison groups. These are also the synthetic models consisting of the individual models. But these individual models of comparison groups are trained on all the input bands with different initialization (as shown in Figure 5).

3. EXPERIMENTS

This section presents the datasets and the configuration of the CNNs used in the experiments.

3.1 Datasets

3.1.1 SAT-6 The SAT-6 dataset, released in 2015 (Basu et al., 2015), consists of 405,000 image patches of size 28 × 28 selected from the National Agriculture Imagery Program (NAIP). It has 6 landcover classes : barren land, trees, grassland, roads, buidings and water bodies. One fifth of them are testing dataset (81,000 images) and the remains are training dataset (324,000 images). Each image contains 4 bands – red (R), green (G), blue (B) and Near Infrared (NIR).

3.1.2 SAT-4 The SAT-4 dataset (Basu et al., 2015) is similar to SAT-6. It has 500,000 image patches in total and contains four classes – barren land, trees, grassland and a class composed of all the other landcover classes. One fifth of the SAT-4 dataset are testing dataset (100,000 images) and the remains are training dataset (400,000 images). Each image also contains 4 bands – red (R), green (G), blue (B) and Near Infrared (NIR).

3.2 Experimental Setup

The DNN architecture used in this paper is VGG-like (Simonyan and Zisserman, 2014) (as shown in Figure 6). The convolutional kernel is 3 × 3 without zero-padding. The nonlinear activation function uses rectified linear unit (ReLU). Dropout is used for the purpose of preventing the overfitting (Srivastava et al., 2014).

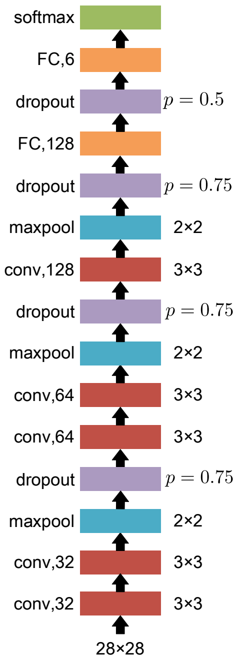

Figure 6. Basic CNN architecture for the SAT-6 dataset. It is also used for the SAT-4 dataset. For the reason that the SAT-4 dataset contains four classes, the last fully connected (FC) layer of the network for SAT-4 has four output channels.

All the network weights follow the Glorot initialization (Glorot and Bengio, 2010). We use the Adadelta optimization method (Zeiler, 2012) for training. Configuration of the Adadelta's parameters follows the recommendation in the proposed paper (Zeiler, 2012). Mini-batch size is 512 in both SAT-6 and SAT-4.

We split out one fifth of training dataset for validation in each experiment. After selecting hyperparameters, the validation dataset is not incorporated to the training dataset.

4. RESULTS AND DISCUSSION

SAT-6: Experimental results of DropBand and the shared DropBand on SAT-6 are summarized in Table 1. Due to the large number of training samples, the single base CNN, which achieves a good enough performance, has a testing error rate of 0.038%. The shared DropBand (shown on the bottom of Figure 2) yields a very low 0.010% error rate. It is identical with the error rate of the comparison group consisting of five base networks with different initializations (Figure 5). The DropBand achieves the best result of 0.006%. By comparing the DropBand with the comparison group, we can observe that they have nearly equal number of parameters. This means that the improvement of accuracy comes from DropBand rather than increasing number of models.

Method	**Test error %**
DeepSat (Basu et al., 2015)	6.084
GoogLeNet (Ma et al., 2016)	3.963
One Base CNN	0.032
Five Base CNNs (comparison group)	0.010
Shared DropBand	0.010
DropBand	**0.006**

Table 1. Comparison of different methods on SAT-6.

SAT-4: Table 2 shows results on the SAT-4 dataset. Comparing to the SAT-6 dataset, the SAT-4 dataset contains more samples but less classes. Therefore, the DropBand method obtains almost perfect result of 0.003%. It means that only 3 images are misclassified in testing dataset (100,000 images). Comparison group also reach an error rate of 0.004%, which means that 4 images are misclassified. For the model of the shared DropBand, underfitting occurs.

Taken together, our method, DropBand, improves the state-of-the-art on SAT-6 and SAT-4.

Method	**Test error %**
DeepSat (Basu et al., 2015)	2.054
GoogLeNet (Ma et al., 2016)	1.592
One Base CNN	0.018
Five Base CNNs (comparison group)	0.004
Shared DropBand	0.017
DropBand	**0.003**

Table 2. Comparison of different methods on SAT-4.

5. CONCLUSION

In this work, a new form of data augmentation, DropBand, is proposed. It has shown that this method can improve the performance of deep neural network significantly. Besides, results of the comparison groups also demonstrated that the effectiveness of DropBand does not come from the increasing number of models.

Abundant information of colors, which is underutilized by the deep neural networks, is contained in remote sensing images. It is the most notable characteristic different from other datasets. Therefore, an interesting area for future work is creating more refined structures of neural networks to make full use of multi-spectral information.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 41571334, in part by the Fundamental Research Funds for the Central Universities, and in part by IWHR Research&Development Support Program (JZ0145B592016).

REFERENCES

Basu, S., Ganguly, S., Mukhopadhyay, S., DiBiano, R., Karki, M. and Nemani, R., 2015. Deepsat: A learning framework for satellite imagery. In: *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '15, ACM, New York, NY, USA, pp. 37:1–37:10.

Ciresan, D., Meier, U. and Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, pp. 3642–3649.

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics* 36(4), pp. 193–202.

Glorot, X. and Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: *International conference on artificial intelligence and statistics*, pp. 249–256.

Howard, A. G., 2013. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: F. Pereira, C. Burges, L. Bottou and K. Weinberger (eds), *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., pp. 1097–1105.

Lafferty, J., McCallum, A. and Pereira, F., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the eighteenth international conference on machine learning, ICML*, Vol. 1, pp. 282–289.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11), pp. 2278–2324.

Ma, Z., Wang, Z., Liu, C. and Liu, X., 2016. Satellite imagery classification based on deep convolution network. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 10(6), pp. 1031–1035.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1), pp. 1929–1958.

Sutton, C., Sindelar, M. and McCallum, A., 2006. Reducing weight undertraining in structured discriminative learning. In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Association for Computational Linguistics, pp. 89–95.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.

Wu, R., Yan, S., Shan, Y., Dang, Q. and Sun, G., 2015. Deep image: Scaling up image recognition. *arXiv preprint arXiv:1501.02876*.

Zeiler, M. D., 2012. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.