Wide Contextual Residual Network with Active Learning for Remote Sensing Image Classification

Conference Paper · July 2018

DOI: 10.1109/IGARSS.2018.8517855

5 authors, including:

Shengjie Liu
University of Southern California
13 PUBLICATIONS **180** CITATIONS
SEE PROFILE

Haowen Luo
The Chinese University of Hong Kong
5 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Ying Tu
Tsinghua University
16 PUBLICATIONS **186** CITATIONS
SEE PROFILE

Jun Li
Qingdao University of Science and Technology
207 PUBLICATIONS **9,142** CITATIONS
SEE PROFILE

WIDE CONTEXTUAL RESIDUAL NETWORK WITH ACTIVE LEARNING FOR REMOTE SENSING IMAGE CLASSIFICATION

Shengjie Liu, Haowen Luo, Ying Tu, Zhi He, Jun Li

Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation,
School of Geography and Planning, Sun Yat-sen University, Guangzhou, 510275, China.

ABSTRACT

In this paper, we propose a wide contextual residual network (WCRN) with active learning (AL) for remote sensing image (RSI) classification. Although ResNets have achieved great success in various applications (e.g. RSI classification), its performance is limited by the requirement of abundant labeled samples. As it is very difficult and expensive to obtain class labels in real world, we integrate the proposed WCRN with AL to improve its generalization by using the most informative training samples. Specifically, we first design a wide contextual residual network for RSI classification. We then integrate it with AL to achieve good machine generalization with limited number of training sampling. Experimental results on the University of Pavia and Flevoland datasets demonstrate that the proposed WCRN with AL can significantly reduce the needs of samples.

Index Terms— Residual networks; active learning; remote sensing; classification; hyperspectral image; SAR

1. INTRODUCTION

Deep learning, which can be considered as the extension of traditional artificial neural network, has been widely used for remote sensing image (RSI) classification [1]. Typical deep learning methods include the deep neural networks (DNNs), convolutional neural networks (CNNs) [2] and residual networks (ResNets) [3]-[6]. Specifically, ResNets, extended from CNNs, utilize skipped connections to facilitate the propagation of gradients, shown to be robust with very deep architecture [4]. However, its application to RSI classification remains a challenge due to the limited availability of training samples [1].

Active learning (AL), which aims at finding the most informative training set, can be used to minimize the number of required labeled data with a relative good machine generalization [7]. Rather than choosing the training set randomly, AL selects the training data actively based on a certain criterion such as the mutual information (MI) [8], the breaking ties (BT) [9], the modified BT (MBT) [7], etc. By using the most informative samples, the deep networks can achieve fast convergence [10]. In [10], active learning and transfer learning were integrated with AlexNet for biomedical image analysis, showing that the cost of sampling can be cut by at least half. In [11], AL with semi-supervised learning was used for SAR image recognition, demonstrating good effectiveness of AL for convolutional networks. In [1], a new AL algorithm, namely the weighted incremental dictionary learning (WI-DL) was integrated with a non-convolutional networks for HSI classification, showing that active learning with non-convolutional networks achieved higher accuracy with fewer training samples for HSI images.

In this paper, we propose the WCRN with AL for RSI classification, where a new network named WCRN is designed for RSI classification and then integrated with AL to improve the machine generalization. The remainder of the paper is organized as follows. Section II presents the methodology. Experimental results are shown in Section III, which illustrates the effectiveness of the proposed method. Finally, we draw some conclusions in Section IV.

2. METHODOLOGY

In this section, we will present the proposed method for RSI classification.

2.1. Wide Contextual Residual Network (WCRN)

The architecture of the proposed WCRN is shown in Fig. 1. Inspired by the works in [5], [6], [12], [13], we design the proposed WCRN with a multi-scale convolutional layer and one residual unit.

In the proposed WCRN, the number of kernels in a convolutional layer is significantly larger than that in the traditional CNN or ResNets. This is because RSIs contain many spectral bands or feature channels, e.g., Hyperspectral images venereally with hundreds of bands, a wide network therefore would better preserve the spectral/feature information. Specifically, the number of kernels per convolutional layer is 256 in WCRN, while the number of kernels in traditional CNN or ResNets ranges from 8 to 128 [3], [5], [6] .

Moreover, for the convolutional layer, we adopt a multi-scale filter bank that locally convolves the input image with 1×1 and 3×3 convolutional kernels to extract the spectral/feature correlations and spatial correlations, respectively.

For the residual unit, we adopt the newly developed residual unit in [5], as shown in Fig. 2. As shown in Fig. 2, we use batch normalization (BN) [12] to ensure appropriate inputs before rectified linear unit (ReLU). The BN layer, which normalizes each scalar feature independently, can regularize and speed up the training process. For a layer with d-dimensional input $\mathbf{x} = (x^{(1)}, ..., x^{(d)})$, each dimension can be normalized as follows,

$$\mathrm{BN}(\widehat{x}^{(k)}) = \frac{x^{(k)} - \mathbf{E}[x^{(k)}]}{\sqrt{\mathbf{Var}[x^{(k)}]}}, \tag{1}$$

for $k = 1, \ldots, d$, where the expectation, $\mathbf{E}[\cdot]$, and variance, $\mathbf{Var}[\cdot]$, are computed over the training data set.

This work was supported by National Natural Science Foundation of China under Grants 61771496 and 41501368, National Key Research and Development Program of China under Grant 2017YFB0502900, Guangdong Provincial Natural Science Foundation under Grant 2016A030313254, and the Fundamental Research Funds for the Central Universities under Grant No. 16lgpy04.

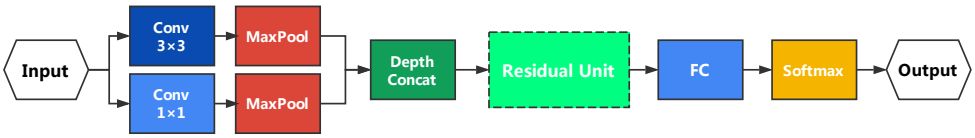

Fig. 1. The proposed WCRN.

As shown in [4], the key idea of ResNets is to add a shortcut connection every two layers,

$$\mathbf{x}_{l+1} = f(\mathbf{x}_l + \mathcal{F}(\mathbf{x}_l, \mathcal{W}_l)), \tag{2}$$

where $\mathbf{x}_l$ and $\mathbf{x}_{l+1}$ are input and output of the l-th unit, $\mathcal{W}_l$ is a set of weights (and biases) associated with the l-th unit, and $\mathcal{F}$ is a residual function and f is the activation function ReLU.

In this study, we apply the improved residual unit [5] to enhance the idea of shortcut by identity mappings, which is given by,

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \mathcal{F}(\hat{f}(\mathbf{x}_l), \mathcal{W}_l), \tag{3}$$

where $\hat{f}$ is an activation that only affects the $\mathcal{F}$ path (the non-skip part) of the unit.

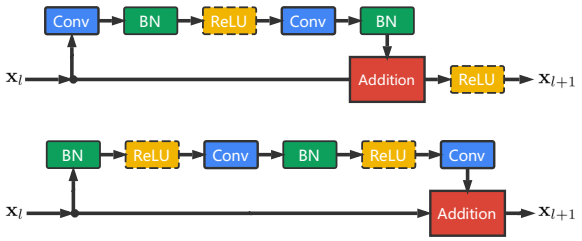

Fig. 2. The two residual units introduced in [4] (Top) and [5] (Bottom). In this work, we adopt the bottom one for our method.

2.2. Active Learning

The basic idea of AL is to iteratively enlarge the training set by requesting an expert to label new samples from the unlabeled set. A relevant question is what samples in the unlabeled set are informative and should be chosen for training. Base on the posterior probabilities produced by the proposed networks, we adopt four different sampling schemes for selection,

- Random selection (RS), where the new samples are randomly selected from the candidate set.
- Mutual information (MI)-based criterion, which aims at finding the samples maximizing the MI between the ResNet model and the class labels.
- Breaking ties (BT), which aims at finding the samples minimizing the distance between the first two most probable classes.
- modified BT (MBT), which aims at finding samples maximizing the probability of the large class for each individual class.

3. EXPERIMENTAL RESULTS

In this section, two real RSIs are used to evaluate the performance of the proposed WCRN with AL. The first one is a hyperspectral image collected by the ROSIS optical sensor over the urban area of the University of Pavia, Italy, on July 8, 2002. The spatial resolution is 1.3 m/pixel, and the number of spectral bands in the acquired image is 103 (with a spectral range from 0.43 to 0.86 μm). Fig. 4 (a) shows a false color composite of the image, while Fig. 4 (b) shows a ground truth map, which contains 42776 samples and 9 ground truth classes of interest. The other data used here is the AirSAR L-band PolSAR dataset, obtained by National Aeronautics and Space Administration/Jet Propulsion Laboratory (NASA/JPL) in 1989 over the Flevoland site in the Netherlands. Pauli composite and the ground truth are displayed in Fig. 6 (a) and (b), respectively. The Flevoland image, with a size of 375 $\times$ 512, contains 54276 samples in the reference data and 11 classes.

Before describing the obtained results, we introduce the experimental settings in this work.

- The proposed WCRN is implemented by using Keras with TensorFlow backend. All convolutional layers are initialized with a zero-mean Gaussian distribution with standard deviation of 0.01. In addition, Adadelta optimizer [14] is used to speed up the training process.
- To avoid overfitting, training samples are augmented four times by mirroring across the horizontal, vertical, and diagonal axes [6]. Furthermore, in order to increase the model stability, in the end of AL, models of the final nine epoches are used to predict a group of results, while the final results are generated by majority voting.

3.1. University of Pavia Dataset

In the first experiment, we analyze the impact of the number of residual units and kernels in the proposed WCRN. Table 1 shows the results by using different number of residual units and kernels. It can be observed that, the results obtained by using one residual unit with more kernel are better. This is because, as aforementioned, a wide network better preserves the spectral/feature information due to the fact that there are 103 bands in the considered dataset. Furthermore with one residual unit, the parameters involved in the learning are much less that the other cases. Since the training information is limited, less parameters would lighten the computational burden. Therefore, in the following experiments, we empirically set the number of units and kernels as 1 and 256, respectively.

In the second experiment, we compare the proposed WCRN with a newly developed contextual CNN approach [6]. We also adopt the majority voting strategy of multiple models to this contextual CNN approach, for more robust performance. Table 2 shows the obtained overall accuracies (OAs) along with the standard deviation, from 10 independent runs, on the University of Pavia dataset. Two difference scenarios are considered in this experiment. A first case, without AL, uses 1800 training samples (200 per class, same as

Table 1. The OAs, along with the standard deviations, obtained by the proposed WCRN by using different number of residual units and kernels.

No. of Residual Units	No. of Kernels	OA
1	32	97.69 ± 0.21 %
1	64	98.07 ± 0.20 %
1	128	98.28 ± 0.25 %
1	256	$\mathbf{98.36 \pm 0.26}$ **%**
2	32	97.88 ± 0.36 %
2	64	98.13 ± 0.37 %
2	128	98.24 ± 0.20 %
2	256	98.27 ± 0.72 %

that in [6]) randomly selected from the reference data. In the other case, a total of 600 samples with 510 were actively selected by using different strategies. Notice that, for the proposed approach, we set a batch size as 20. The total epoches are 200 to make sure that the network can be well trained. As can be observed from Table 1, without AL, the proposed approach achieved the best results in comparison with the contextual CNN. Furthermore, by adopting AL, the networks significantly reduces the requirement of the training samples.

Table 2. The obtained overall accuracies(OAs, averaging from 10 independent runs), along with standard deviations for the University of Pavia dataset). For the results obtained with AL, 90 samples (10 per class) were used as the initial training set, and 510 samples (10 per iteration) were actively selected by using different strategies. The best results are given in bold.

Method	OA	Samples
Contextual CNN [6]	95.97 ± 0.46 %	1800
Contextual CNN (vote)	97.75 ± 0.24 %	1800
WCRN	$\mathbf{98.36 \pm 0.26}$ **%**	1800
Contextual CNN (vote)	94.06 ± 1.19 %	600
WCRN (RS)	96.22 ± 0.38 %	600
WCRN (MI)	99.41 ± 0.09 %	600
WCRN (BT)	$\mathbf{99.43 \pm 0.08}$ **%**	600
WCRN (MBT)	99.31 ± 0.08 %	600

In the third experiment, we evaluated the proposed WCRN with the AL scheme. Fig. 3 presents the obtained OAs as a function of the number of training samples in the AL scheme of the proposed WCRN, by using 90 samples in total (10 per class) as the initial training set. In this experiment, for the AL, 10 samples were actively selected per iteration. It can be observed that, by adopting AL, the performance is greatly boosted as the number of training sample increases. Furthermore, for the considered AL strategies, BT and MBT achieved better results, in comparison with MI, when the number of training samples are small. This is expected due to the fact that the training samples selected by BT and MBT are with more diversity than those by MI. Finally, as the number of training samples increases to a relative medium size, i.e., around 480 in this experiment, all AL strategies achieved very similar and robust performance with respect to the OA, which are much better than that obtained by RS. This is expected, as the number of samples increases by AL, the sample uncertainty decreases.

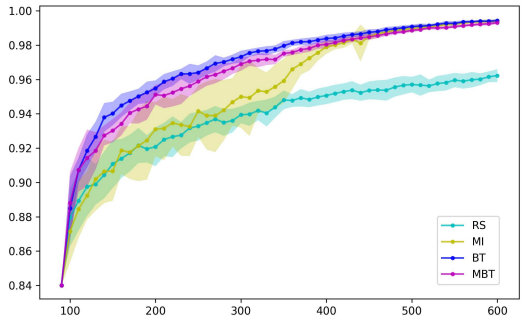

Fig. 3. The OAs, along with the standard deviations, as a function of the number of training samples in the AL scheme of the proposed WCRN over the University of Pavia dataset, where 90 samples in total (10 per class) were used as the initial training set, and 10 samples were actively selected per iteration.

Finally, for illustrative purposes, Fig. 4 (c) and (d) show the classification maps by RS and BT, respectively, where 400 training samples were used for training, with 310 samples were actively selected by the AL strategies. It can be observed that the results obtained by BT are remarkable, which are better than that obtained by MI.

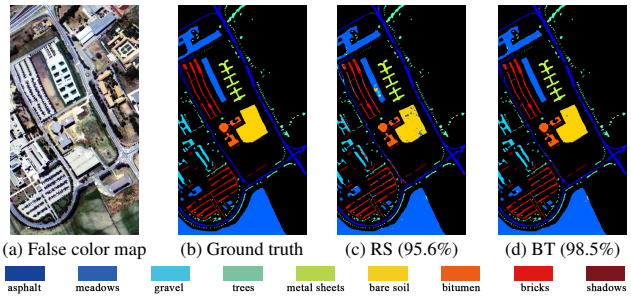

Fig. 4. The classification maps, along with the OAs, of the University of Pavia dataset, where 400 samples were used for training, with 310 were actively selected by the AL strategy.

3.2. The Flevoland Dataset

For the Flevoland SAR dataset, in the first experiment, the proposed WCRN is compared with contextual CNN [6]. It should be noted here, similar to the pervious experiments, we use the contextual CNN with a voting strategy, as it can produce better and more robust results. Table 3 shows the obtained OAs along side with standard deviation from 10 independent runs. It can be seen that the proposed WCRN achieved better performance than the competitor. Furthermore, by including the AL strategy, the proposed method can obtain a very good classification accuracy.

In the second experiment, we evaluate the proposed WCRN with AL as a function of number of training samples, by using 110 samples (10 per class) as the initial training set and 10 samples actively selected per iteration. Fig 5 presents the obtained OAs, along with the standard deviations, as a function of the number of training samples in the AL scheme of the proposed approach. Similar observations can be obtained as to the experiments of the University of Pavia dataset. First of all, it can be seen that, with the AL strategies, the

results obtained are much better than that of RS. For instance, with 620 samples (510 were actively selected), the results obtained by BT, MBT and MI are about 98.5%, which is 4% higher than that of RS, which is about 94.5%. Furthermore, when the number of training samples are small, BT and MBT achieved much better accuracies than MI. As the number of training samples reaches around 500, all AL methods converges to similar performance. This is, again, a similar observation to the pervious dataset. As aforementioned, the sample uncertainty significantly decreases with the increase of number of samples by the AL strategies, which aim at finding the most informative samples.

Finally, for illustrative purposes, Fig. 6 (c) and (d) show the obtained classification maps by using 600 training samples, with 110 ones as the initial set and 490 actively selected, respectively.

Table 3. The obtained overall accuracies (OAs, averaging from 10 independent runs), along with standard deviations for the Flevoland dataset). For the results obtained with AL, 110 samples (10 per class) were used as the initial training set, and 510 samples (10 per iteration) were actively selected by using different strategies. The best results are given in bold.

Method	OA	Samples
Contextual CNN (vote)	93.38 ± 0.38 %	600
WCRN (RS)	94.05 ± 0.48 %	600
WCRN (MI)	97.68 ± 0.47 %	600
WCRN (BT)	$\mathbf{98.23 \pm 0.15}$ **%**	600
WCRN (MBT)	98.01 ± 0.17 %	600

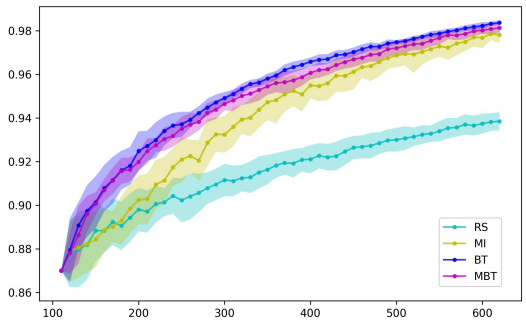

Fig. 5. The OAs, along with the standard deviations, as a function of the number of training samples in the AL scheme of the proposed WCRN over the Flevoland dataset, where 110 samples in total (10 per class) were used as the initial training set, and 10 samples were actively selected per iteration.

4. CONCLUSION

In this paper, we design a wide contextual residual network (WCRN) for the classification of remote sensing images (RSIs). Then, we introduce active learning into the proposed WCRN, aiming at reducing the necessarity of labeled training information. The advantages of the proposed approach are two folds. On the one hand, the proposed WCRN can extract and maintain the abundant spectral/feature information, as well as spatial information in the input RSIs, by taking advantage from the contextual convolutional layers with a large number of kernels. On the other hand, the integration of AL leads to good machine generalization with limited number of training samples by finding the most informative samples. Experimental results on two RSIs, including the University of Pavia hyperspectral dataset, and the Flevoland SAR dataset, demonstrate that the proposed method can significantly reduce the needs of training samples.

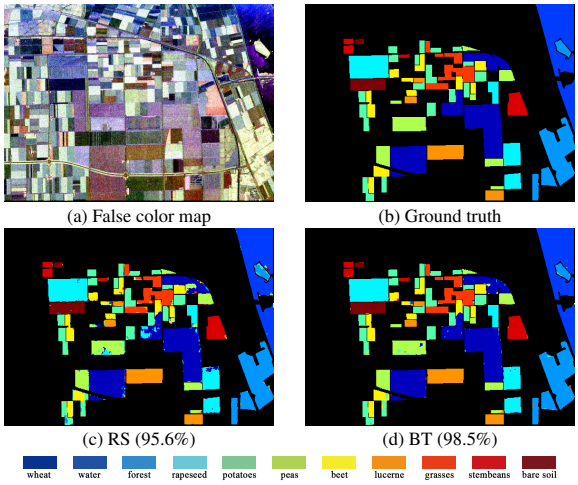

Fig. 6. The classification maps, along with the OAs, of the Flevoland dataset, where 600 samples were used for training, with 490 were actively selected by the AL strategy.

5. REFERENCES

[1] P. Liu, H. Zhang, and K. B. Eom, "Active deep learning for classification of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 712–724, 2017.

[2] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, 2016.

[3] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. PP, no. 99, pp. 1–12, 2017.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*. Springer, 2016, pp. 630–645.

[6] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, 2017.

[7] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new bayesian approach with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, 2011.

[8] D. MacKay, "Information-based objective functions for active data selection," *Neural comput.*, vol. 4, no. 4, pp. 590–604, 1992.

[9] T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins, "Active learning to recognize multiple types of plankton," *J. Mach. Learn. Res.*, vol. 6, pp. 589–613, 2005.

[10] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7340–7349.

[11] F. Gao, Z. Yue, J. Wang, J. Sun, E. Yang, and H. Zhou, "A novel active semisupervised convolutional neural network algorithm for SAR image recognition," *Comput. Intell. Neurosci.*, vol. 2017, 2017.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[13] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[14] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.