MULTI-LABEL REMOTE SENSING IMAGE CLASSIFICATION WITH DEFORMABLE CONVOLUTIONS AND GRAPH NEURAL NETWORKS

Yingyu Diao, Jingzhou Chen, Yuntao Qian

College of Computer Science, Zhejiang University, Hangzhou, China

IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium | 978-1-7281-6374-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IGARSS39084.2020.9324530

ABSTRACT

Multi-label remote sensing image classification is a significant yet difficult task due to intra-class variations and label dependencies among land-cover classes. In this paper, we propose a novel multi-label classification model based on deformable convolutions and graph neural networks. Specifically, we first use deformable convolutions to learn image features with geometric transformation invariance and adaptive receptive field. Then we adopt attention mechanism to extract label-related image features. After that, a directed graph is constructed to model the label dependencies, and the label-related features are fused through graph propagation mechanisms. Experiments on UC-Merced and DOTA data sets demonstrate its effectiveness.

Index Terms— Multi-label classification, Deformable convolution, Graph neural networks

1. INTRODUCTION

In the traditional single-label task, each image is associated with a unique semantic label. However, a single label may be insufficient for annotating the remote sensing scenes with complex semantic information. Multi-label remote sensing image classification, which aims to extract elements of interest (e.g., buildings, ships) and generate multiple labels, has become crucial for understanding remote sensing images.

Compared to the single-label task, the multi-label classification is much more difficult due to the overwhelming size of output space [1]. One feasible solution is using the prior information of label dependencies. Extensive research has been devoted to capture label dependencies for natural scene images. In some recent works, Chen et al. leveraged the graph structure to explore the label dependencies and utilized graph convolutional networks (GCNs) to propagate information between multiple labels [2]. Similarly, in [3] a semantic-specific graph representation learning model (SSGRL) was proposed, which incorporates category semantics to guide learning semantic-specific features and explore their interactions to facilitate multi-label classification. Few attempts have been made in the context of multi-label classification for remote sensing images. In [4], a multi-attention mechanism combined with convolutional neural network (CNN) and recurrent neural network (RNN) was presented, in which the joint occurrence of multiple land-cover classes is considered and the attention-based local descriptors is learned. In [5] a label relation inference module is designed to take advantage of pairwise label relations to infer multiple labels. However both of methods do not explicitly model the label dependencies via prior knowledge. In this paper the complex dependencies between class labels will be clearly represented by the graph structure and their interaction will be learned in a data-driven way.

Moreover, another key challenge of remote sensing image classification is how to address huge intra-class variations in the scale, orientation, and shape. CNNs can extract rich and discriminative features with hierarchically, locally and shared filtering, but CNNs are inherently limited to model large and unknown transformations due to its fixed geometric structures. Deformable convolutional networks [6] adds 2D offsets to the regular grid sampling locations in the standard convolution, so that dense spatial transformation can be learned for sophisticated vision tasks. We will use deformable convolutions for extracting the features of remote sensing images.

In summary, this paper propose a novel multi-label classification model based on deformable convolutional networks and graph neural networks (GNNs) [7], abbreviated as DCN-GNN. This method can learn the remote sensing image features with geometric transformation invariance and adaptive receptive field by deformal convolution, construct a graph to explicitly model label dependencies in remote sensing images, and utilize GNNs to explore the semantic interaction among land-cover labels.

2. METHOD

Multi-label classification aims to predict multiple labels corresponding to a given remote sensing image. We define $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M\}$ as a remote sensing archive that contains M images. Suppose $\mathcal{L} = \{l_1, l_2, \cdots, l_N\}$ is a finite label set that consists of N labels. We assume that each image $\mathbf{x}$ in the archive $\mathcal{X}$ is associated with a binary label vector $\mathbf{y} \in \{0, 1\}^N$ and each element of $\mathbf{y}$ sequentially denotes

This work was supported by the National Key Research and Development Program of China 2018YFB0505000, and the 2030 National Key AI Program of China 2018AAA0100500.

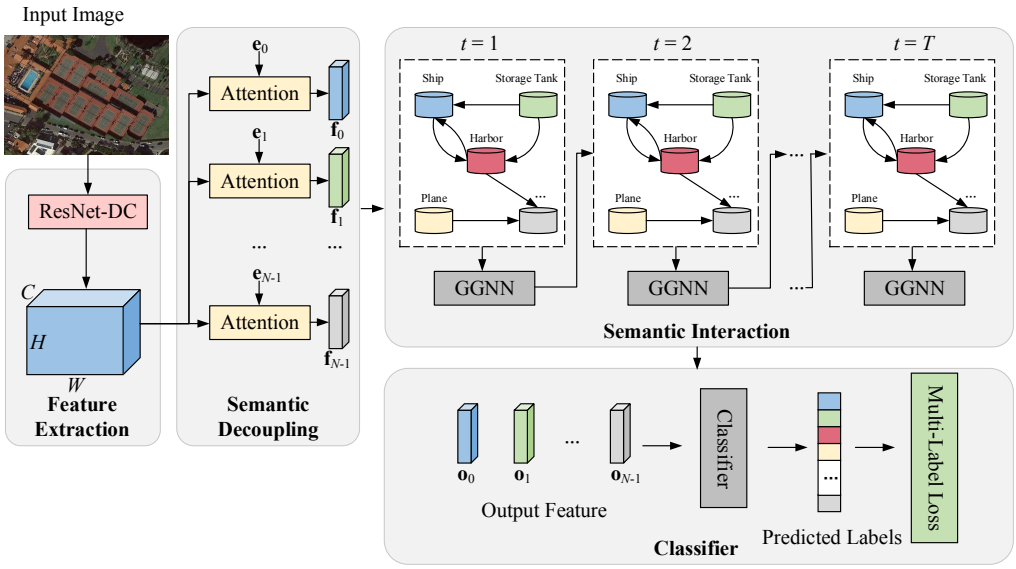


Fig. 1. The framework of DCN-GNN

wether the n-th label $l_n \in \mathcal{L}$ appears in this image or not. Based on the training samples, we want to train a classifier that maps new remote sensing images $\mathbf{x}^*$ to the multiple labels $\mathbf{y}^*$. The DCN-GNN framework mainly consists of four modules: 1) feature extraction; 2) semantic decoupling; 3) semantic interaction and 4) classification. Fig. 1 illustrates the overall framework of DCN-GNN.

2.1. Feature Extraction

The feature extraction module aims to learn image features with geometric transformation invariance and adaptive receptive field. We replace the standard convolutional layers in CNNs with deformable convolutions. Distinguished from standard convolutions, deformable convolutions add 2D offsets to the spatial sampling locations and model various geometric transformations. These offsets are learned from the input feature maps by standard convolutional layers. Only with small amount of additional parameters and computations, deformable convolutions can result in noticeable performance gains. Compared with data augmentation technique, its has better generalization to new tasks possessing unknown geometric transformations.

2.2. Semantic Decoupling

Once the image features are obtained, the semantic decoupling module extracts label-related features via attention mechanism. In this module, the weights are generated for different locations in original feature maps, then semantic features that focus on the related regions of a specific class label are learned. For an input image $\mathbf{x}$, the feature extraction module outputs its feature maps $\mathbf{f}^{\mathbf{x}} \in \mathbb{R}^{H\times W\times C}$, where H, W, C are the height, width and channels respectively. For each label l, a d_s-dimensional semantic representation $\mathbf{e}_l$ is calculated as $\mathbf{e}_l = f_g(l)$. f_g is the pretrained model for obtaining word representations such as GloVe, and $\mathbf{e}_l$ is the word embedding of label l. Then for the location (h, w) in feature maps, the corresponding C-dimensional feature vector $\mathbf{f}^{\mathbf{x}}_{hw}$ and the label embedding $\mathbf{e}_l$ are fused for each label l as follows:

$$\tilde{\mathbf{f}}^{\mathbf{x}}_{l,hw} = \mathbf{P}^T \left(\tanh\left(\left(\mathbf{U}^T\mathbf{f}^{\mathbf{x}}_{hw}\right) \odot \left(\mathbf{V}^T\mathbf{e}_l\right)\right)\right) + \mathbf{b} \quad (1)$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, $\mathbf{U} \in \mathbb{R}^{C\times d_1}$, $\mathbf{V} \in \mathbb{R}^{d_s\times d_1}$, $\mathbf{P} \in \mathbb{R}^{d_1\times d_2}$, $\mathbf{b} \in \mathbb{R}^{d_2}$ are the parameters to be learned, $\odot$ is the element-wise multiplication operation, and d_1, d_2 are the dimension of joint embeddings and output features respectively. Afterwards attention coefficients $\boldsymbol{\alpha}$ are generated

$$\boldsymbol{\alpha}_{l,hw} = \mathrm{softmax}\left(f_a\left(\tilde{\mathbf{f}}^{\mathbf{x}}_{l,hw}\right)\right) \quad (2)$$

where f_a is a fully-connected layer and a softmax function is used for coefficients normalization. The attention coefficients indicate the importance of each location for each class label. Finally, we sum over the original feature maps weighted by corresponding attention coefficients and obtain the label-related feature vector:

$$\mathbf{f}_l = \sum_{h,w} \boldsymbol{\alpha}_{l,hw}\mathbf{f}^{\mathbf{x}}_{hw} \quad (3)$$

2.3. Semantic Interaction

In this module, we first construct a directed graph $\mathcal{G} = (V, A)$. The node set $V = \{v_0, v_1, \cdots, v_{N-1}\}$ represents the label-related features, i.e., each class label l is associated with a

node $v_l \in V$. The edges indicate dependencies between pairwise labels, represented by adjacent matrix A with the size of $N \times N$. Similar to the graph construction method used in [2], we define such dependencies in the form of prior label co-occurrence patterns within the data set. The weight of an edge between vertices v_l and $v_{l'}$ is $a_{ll'}$ in A, which refers to the conditional probability of label l' if label l appears. Then we utilize gated GNN (GGNNs) [7] to propagate information over the graph and generate contextual semantic features. At time step t, v_l has a hidden state $\mathbf{h}_l^t$. When $t = 0$, the hidden state is initialized with the label-related feature vector $\mathbf{f}_l$. In GGNNs, each node in graph aggregates information $\mathbf{a}_l^t$ from their adjacent nodes at time step t

$$\mathbf{a}_l^t = \left[\sum_{l'} \left(a_{ll'}\right) h_{l'}^{t-1} + \sum_{l'} \left(a_{l'l}\right) h_{l'}^{t-1} \right] \tag{4}$$

Based on the aggregated information $\mathbf{a}_l^t$ and the hidden state $\mathbf{h}_l^{t-1}$ at previous time step $t-1$, each node updates the hidden state over time step, formulated as:

$$\begin{aligned} \mathbf{z}_l^t &= \sigma\left(\mathbf{W}^z \mathbf{a}_l^t + \mathbf{U}^z \mathbf{h}_l^{t-1}\right) \\ \mathbf{r}_l^t &= \sigma\left(\mathbf{W}^r \mathbf{a}_l^t + \mathbf{U}^r \mathbf{h}_l^{t-1}\right) \\ \tilde{\mathbf{h}}_l^t &= \tanh\left(\mathbf{W} \mathbf{a}_l^t + \mathbf{U}\left(\mathbf{r}_l^t \odot \mathbf{h}_l^{t-1}\right)\right) \\ \mathbf{h}_l^t &= \left(1 - \mathbf{z}_l^t\right) \odot \mathbf{h}_l^{t-1} + \mathbf{z}_l^t \odot \tilde{\mathbf{h}}_l^t \end{aligned} \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{W}^z$, $\mathbf{U}^z$, $\mathbf{W}^r$, $\mathbf{U}^r$, $\mathbf{W}$, $\mathbf{U}$ are the parameters to be learned. The update process is executed by T times.

2.4. Classification

In this module, the initial and final hidden state are fused and sent to the multi-label classifier. Specifically, this module can be formulated as $\mathbf{o}_l = f_o\left(\mathbf{h}_l^T, \mathbf{h}_l^0\right)$ and $s_l = f_c\left(\mathbf{o}_l\right)$, in which f_o is a fully-connected layer, $\mathbf{o}_l$ is the output features corresponding to label l, f_c is a multi-label classifier and s_l is the confidence of label l. Here we adopt common multi-label cross entropy as the loss function

$$L = \sum_{l=0}^{N-1} y_l \log\left(s_l\right) + \left(1 - y_l\right) \log\left(1 - s_l\right), \tag{6}$$

where y_l is the ground truth of label l for the training image. All four modules can be end-to-end trained by back propagation algorithm.

3. EXPERIMENTS

For the proposed DCN-GNN framework, ResNet-101 [8] is used as the feature extraction backbone, and deformable convolutions are applied in all 3×3 conv layers in last three stages. In semantic decoupling module, d_s, d_1 and d_2 are set to 300, 1024, 1024. In the semantic interaction module, the dimension of the hidden state is 2048 and the number of steps of GGNN $T = 3$. During training, we use SGD as the optimizer with momentums of 0.9 and weight decay of 10^{-4}. The learning rate is initialized as 0.001 and divided by 10 for every 30 epochs. Totally we train the model for 100 epochs.

Two data sets, i.e., UC-Merced [9] and DOTA [10], are used for experiments. UC-Merced contains 2100 images of the size 256×256, and these images are categorized into 17 classes. DOTA consists of 2806 images of the size of about 4000×4000, and these images are annotated with 15 classes. For UC-Merced, each input image is resized to 512×512 and random cropped into 448×448. For DOTA, each input image is resized to 864×864 and random cropped into 800×800. We adopt the mean average precision (mAP), and overall precision, recall, F1-measure (OP, OR, OF1) and per-class precision, recall, F1-measure (CP, CR, CF1) for performance evaluation.

The quantitative results are respectively presented in Table 1 and Table 2 for two datasets. Six methods are compared: 1) ResNet50 [8]; 2) ResNet101 [8]; 3) ML-GCN [2]; 4) SSGRL [3]; 5) ResNet-DC that is ResNet101 with deformable convolutions; 6) our DCN-GNN. As is shown in tables, the proposed DCN-GNN outperforms other methods on two data sets in most cases. During the inference stage, our model runs at 116 fps for UC-Merced, and 44 fps for DOTA on Nvidia Geforce GTX 1080ti.

We further visualize the effects of deformable convolutions and attention mechanisms in our model respectively. Fig. 2 shows the sample locations (red points) in last three stacked deformable convolutions corresponding to the activation units (green points) on the top feature. It demonstrates that deformable convolutions can obtain adaptive receptive field for different scales and shapes of land-cover targets. Fig. 3 shows the attention weights of different labels for two remote sensing images. For example, Fig. 3(b) and Fig. 3(c) are the generated attention maps for tennis court and large vehicle. Similarly, Fig. 3(e) and Fig. 3(f) are attention maps that associated with small vehicle and plane respectively.

4. CONCLUSION

In this paper, we propose a multi-label remote sensing image classification framework based on deformable convolutions and GNNs, in which deformable convolutions are introduced to model various geometric transformations and GNNs are adopted to capture contextual semantic features. Experimental results on UC-Merced and DOTA demonstrate the effectiveness of the proposed method.

5. REFERENCES

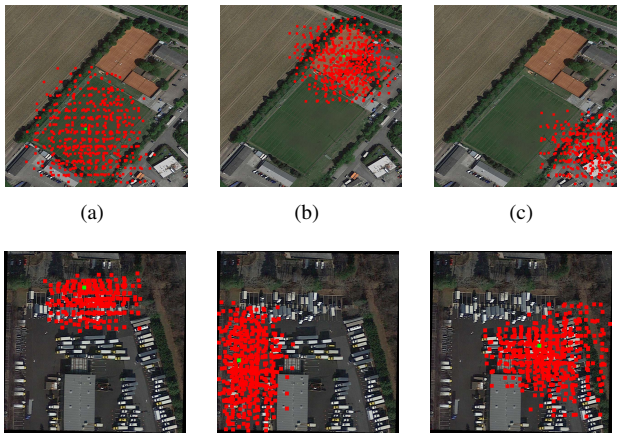
[1] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowledge Data Eng.*,

Table 1. Experimental results on UC-Merced

Methods	OP	OR	OF1	CP	CR	CF1	mAP
ResNet50	0.8857	0.9101	0.8978	**0.9420**	0.9367	0.9393	0.9716
ResNet101	0.8862	0.9329	0.9096	0.9218	0.9457	0.9338	0.9723
ML-GCN	0.9099	0.9175	0.9137	0.9238	0.9335	0.9287	0.9769
SSGRL	0.9078	0.9244	0.9161	0.9206	0.9325	0.9265	0.9820
ResNet-DC	0.8848	**0.9370**	0.9109	0.9314	0.9463	0.9389	0.9741
DCN-GNN	**0.9381**	0.9162	**0.9272**	0.9311	**0.9497**	**0.9404**	**0.9876**

Table 2. Experimental results on DOTA

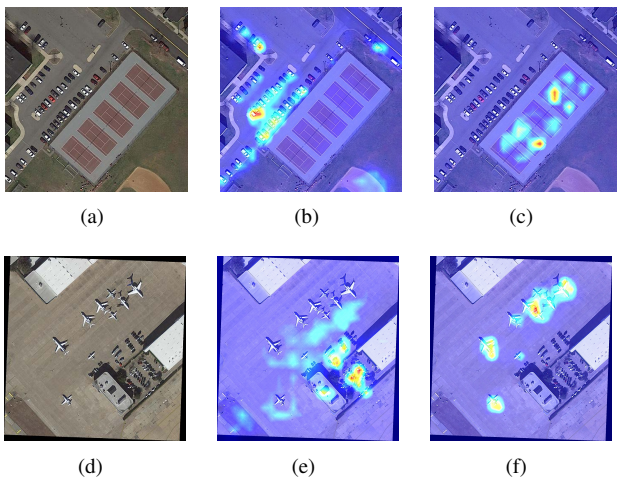
Methods	OP	OR	OF1	CP	CR	CF1	mAP
ResNet50	0.7996	0.6583	0.7221	0.7774	0.5471	0.6422	0.7195
ResNet101	0.7554	0.6351	0.6901	0.7308	0.5658	0.6378	0.7269
ML-GCN	0.7836	0.6524	0.7120	0.7497	0.5905	0.6606	0.7478
SSGRL	**0.8134**	**0.7072**	**0.7566**	0.7677	0.5875	0.6657	0.7544
ResNet-DC	0.7623	0.6479	0.7004	0.7438	0.5626	0.6406	0.7467
DCN-GNN	0.8026	0.6288	0.7051	**0.7777**	**0.6457**	**0.7056**	**0.7713**



(a) (b) (c)

(d) (e) (f)

Fig. 2. Deformation convolutions on two images.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Attention maps on two images

vol. 26, no. 8, pp. 1819–1837, Aug 2014.

[2] T. Chen, M. Xu, et al., "Learning semantic-specific graph representation for multi-label image recognition," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, October 2019, pp. 522–531.

[3] Z. Chen, X. Wei, et al., "Multi-label image recognition with graph convolutional networks," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.(CVPR)*, 2019.

[4] G. Sumbul and B. Demir, "A novel multi-attention driven system for multi-label remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.(IGARSS)*, July 2019, pp. 5726–5729.

[5] Y. Hua, L. Mou, and X. X. Zhu, "Label relation inference for multi-label aerial image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.(IGARSS)*, July 2019, pp. 5244–5247.

[6] J. Dai, H. Qi, et al., "Deformable convolutional networks," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Oct 2017, pp. 764–773.

[7] Y. Li, D. Tarlow, et al., "Gated graph sequence neural networks," *arXiv preprint arXiv:1511.05493*, 2015.

[8] K. He, X. Zhang, et al., "Deep residual learning for image recognition," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.(CVPR)*, June 2016, pp. 770–778.

[9] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb 2013.

[10] G. Xia, X. Bai, et al., "Dota: A large-scale dataset for object detection in aerial images," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.(CVPR)*, June 2018, pp. 3974–3983.