

# Multi-label Aerial Image Classification using A Bidirectional Class-wise Attention Network

1<sup>st</sup> Yuansheng Hua

Remote Sensing Technology Institute (IMF)  
German Aerospace Center (DLR)  
Wessling, Germany

Signal Processing in Earth Observation (SiPEO)  
Technische Universität München (TUM)  
München, Germany  
yuansheng.hua@dlr.de

2<sup>nd</sup> Lichao Mou

Remote Sensing Technology Institute (IMF)  
German Aerospace Center (DLR)  
Wessling, Germany

Signal Processing in Earth Observation (SiPEO)  
Technische Universität München (TUM)  
München, Germany  
lichao.mou@dlr.de

3<sup>rd</sup> Xiao Xiang Zhu

Remote Sensing Technology Institute (IMF)  
German Aerospace Center (DLR)  
Wessling, Germany

Signal Processing in Earth Observation (SiPEO)  
Technische Universität München (TUM)  
München, Germany  
xiaoxiang.zhu@dlr.de

**Abstract**—Multi-label aerial image classification is of great significance in remote sensing community, and many researches have been conducted over the past few years. However, one common limitation shared by existing methods is that the co-occurrence relationship of various classes, so called class dependency, is underexplored and leads to an inconsiderate decision. In this paper, we propose a novel end-to-end network, namely class-wise attention-based convolutional and bidirectional LSTM network (CA-Conv-BiLSTM), for this task. The proposed network consists of three indispensable components: 1) a feature extraction module, 2) a class attention learning layer, and 3) a bidirectional LSTM-based sub-network. Experimental results on UCM multi-label dataset and DFC15 multi-label dataset validate the effectiveness of our model quantitatively and qualitatively.

**Index Terms**—multi-label classification, high resolution aerial image, Convolutional Neural Network (CNN), class attention learning, Bidirectional Long Short-Term Memory (BiLSTM), class dependency.

## I. INTRODUCTION

With the booming of remote sensing techniques in the recent years, a huge volume of high resolution aerial imagery is now accessible and benefits a wide range of real-world applications. As a fundamental bridge between aerial images and these applications, image classification has obtained wide attentions, and many researches have been conducted recently [1], [5]. However, most existing studies assume that each image belongs to only one label (e.g., scene-level labels in Fig. 1), while in reality, an image is usually associated with multiple labels. Furthermore, numerous researches, i.e., semantic segmentation [2], [3] and object detection [4], have emerged recently, but unfortunately, the acquisition of ground truths for these studies are extremely labor- and time-consuming. Compared to these expensive labels, image-level labels (cf.

This work is jointly supported by the China Scholarship Council, the Helmholtz Association under the framework of the Young Investigators Group SiPEO (VH-NG-1018, [www.sipeco.bgu.tum.de](http://www.sipeco.bgu.tum.de)), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. ERC-2016-StG-714087, Acronym: So2Sat). In addition, the authors would like to thank the National Center for Airborne Laser Mapping and the Hyperspectral Image Analysis Laboratory at the University of Houston for acquiring and providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

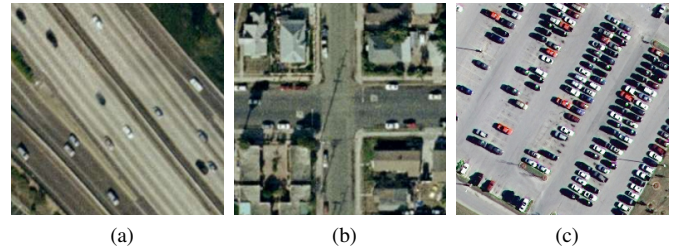


Fig. 1: Example high resolution aerial images with their scene labels and multiple *object* labels. Common label pairs are **highlighted**. (a) Free way: *bare soil*, *car*, *grass*, ***pavement***, and *tree*. (b) Intersection: *building*, *car*, *grass*, ***pavement***, and *tree*. (c) Parking lot: *car* and ***pavement***.

multiple object-level labels in Fig. 1) are at a fair low cost and readily accessible. Without a doubt, multi-label classification is arising and attracting an increasing attention.

Current aerial image multi-label classification methods [6], [7] consider such problem as a regression issue, where models are trained to fit a binary sequence, and each digit indicates the existence of its corresponding class. Besides, [8] formulates multi-label classification into several single-label classification tasks. Notably, one common assumption of these studies is that classes are independent of each other, and classifiers predict the existence of each category independently.

However, this is violent and not accord with real life. As illustrated in Fig. 1, although images obtained in diverse scenes are assigned with multiple different labels, there are still common classes, e.g., car and pavement, coexisting in each image. This is because in the real-life world, some classes have strong correlation, for example, cars are often driven or parked on pavements. To this end, we propose a novel end-to-end network architecture, class attention-based convolutional and bidirectional LSTM network (CA-Conv-BiLSTM), which integrates feature extraction and high-order class dependency exploitation together for multi-label classification.

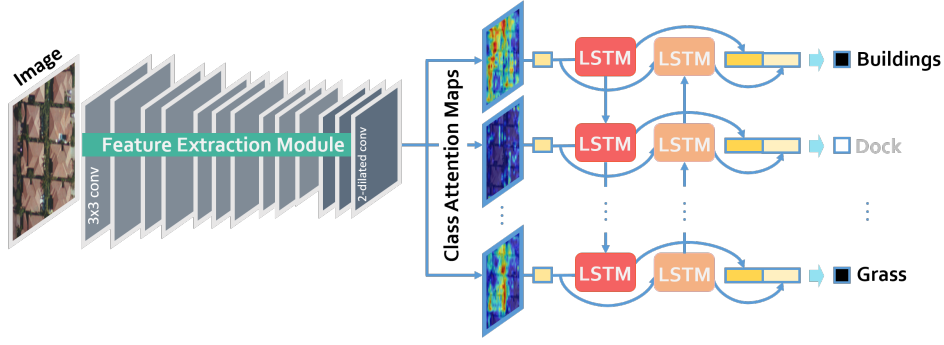


Fig. 2: The architecture of the proposed CA-Conv-BiLSTM for the multi-label classification of aerial images.

## II. NETWORK ARCHITECTURE

The proposed CA-Conv-BiLSTM, as illustrated in Fig. 2, is composed of three components: a feature extraction module, a class attention learning layer, and a Bidirectional LSTM-based recurrent sub-network.

### A. Dense High-level Feature Extraction

To learn efficient feature representations of input images, the feature extraction module takes a conventional CNN (e.g., VGG-16 [9]) as a prototype, which consists of 5 convolutional blocks (as illustrated in the left of Fig. 2). The receptive field of all convolutional filters is  $3 \times 3$ , and the convolution stride is 1 pixel. The spatial padding of each convolutional layer is set as 1 pixel. Among these convolutional blocks, max-pooling layers are interleaved, and the size of pooling windows is  $2 \times 2$  pixels. The pooling stride is 2 pixels, which halves feature maps in width and length.

Although features directly learned from a conventional CNN (e.g., VGG-16) are proved to be high-level and semantic, their spatial resolution is significantly reduced, which is not favorable for generating high-dimensional class-specific features in the subsequent class attention learning layer. To address this, max-pooling layers following the last two convolutional blocks are discarded in our model, and atrous convolutional filters with dilation rate 2 are employed in the last convolutional block for preserving original receptive fields.

Moreover, it is worth nothing that other popular CNN architectures can be taken as prototypes of the feature extraction module, and thus, we extend researches to GoogLeNet [10] and ResNet [11] for a comprehensive evaluation of CA-Conv-BiLSTM.

### B. Class Attention Learning Layer

Although Features extracted from pre-trained CNNs are high-level and can be directly fed into a fully connected layer for generating multi-label predictions, it is infeasible to learn high-order probabilistic dependencies by recurrently feeding it with identical features. Therefore, we propose a class attention learning layer to explore features with respect to each category, and the proposed layer, illustrated in the middle of Fig. 2, consists of the following two stages: 1) generating class attention maps via a  $1 \times 1$  convolutional layer with stride 1,

and 2) vectorizing each class attention map to obtain class-specific features. Formally, given feature maps  $\mathbf{X}$ , extracted from the feature extraction module, with a size of  $W \times W \times K$ , and let  $w_l$  represent the  $l$ -th convolutional filter in the class attention learning layer. The attention map  $M_l$  for class  $l$  can be obtained with the following formula:

$$M_l = \mathbf{X} * w_l, \quad (1)$$

where  $l$  ranges from 1 to the number of classes. Besides,  $*$  represents convolution operation. Given that the size of convolutional filters is  $1 \times 1$ , and the stride is 1, Eq. 1 can be further modified as:

$$M_l(p, q) = \sum_{k=1}^K w_{l,k} \mathbf{X}_k(p, q), \quad (2)$$

where  $p, q = 1, 2, \dots, W$ , and  $M_l(p, q)$  and  $\mathbf{X}_k(p, q)$  indicate activations of the class attention map  $M_l$  and the  $k$ -th channel of  $\mathbf{X}$  at a spatial location  $(p, q)$ , respectively.  $w_{l,k}$  is the  $k$ -th channel of  $w_l$ . The modified formula highlights that a class attention map  $M_l$  is intrinsically a linear combination of all channels in  $\mathbf{X}$ , and  $w_{l,k}$  depicts the importance of the  $k$ -th channel of  $\mathbf{X}$  for class  $l$ . Therefore,  $M_l(p, q)$  with a strong activation suggests that the region is highly relevant to class  $l$ , and vice versa. With this design, the proposed class attention learning layer is capable of tracking distinctive attention of the network when predicting different classes, and extracted class attention maps are abundant in discriminative class-specific semantic information. Subsequently, class attention maps  $M_l$  are transformed into class-wise feature vectors  $v_l$  of  $W^2$  dimensions by vectorization.

### C. Class Dependency Learning via a BiLSTM-based Sub-network

Instead of fully connecting class attention maps to each hidden unit in the following layer, we seek to model class dependencies with an LSTM-based RNN, which has shown great performance in processing long sequences [15]. Specifically, we construct class-wise connections between class attention maps and their corresponding hidden units, i.e., corresponding time steps in an LSTM layer in our network. In this way, features fed into different units are retained to be class-specific

TABLE I: Quantitative Results on UCM Multilabel Dataset

Model	M. $F_1$ (%)	M. $F_2$ (%)	P <sub>e</sub> (%)	R <sub>e</sub> (%)
VGGNet	78.54	80.17	79.06	82.30
CA-VGG-LSTM	79.57	80.75	80.64	82.47
<b>CA-VGG-BiLSTM</b>	<b>79.78</b>	<b>81.69</b>	79.33	83.99
GoogLeNet	80.68	82.32	80.51	84.27
CA-GoogLeNet-LSTM	81.78	<b>85.16</b>	78.52	88.60
<b>CA-GoogLeNet-BiLSTM</b>	<b>81.82</b>	84.41	79.91	87.06
ResNet-50	79.68	80.58	80.86	81.95
CA-ResNet-LSTM	81.36	83.66	79.90	86.14
<b>CA-ResNet-BiLSTM</b>	<b>81.47</b>	<b>85.27</b>	77.94	89.02

M.  $F_1$  and M.  $F_2$  indicates the mean  $F_1$  and  $F_2$  score.  
P<sub>e</sub> and R<sub>e</sub> indicate example-based mean precision and recall.

TABLE II: Quantitative Results on DFC15 Multilabel Dataset

Model	M. $F_1$ (%)	M. $F_2$ (%)	P <sub>e</sub> (%)	R <sub>e</sub> (%)
VGGNet	73.86	74.09	76.16	74.95
CA-VGG-LSTM	75.46	75.85	77.95	76.95
<b>CA-VGG-BiLSTM</b>	<b>76.25</b>	<b>76.93</b>	78.27	78.30
GoogLeNet	74.99	73.41	81.01	73.01
CA-GoogLeNet-LSTM	75.67	<b>75.46</b>	79.08	76.12
<b>CA-GoogLeNet-BiLSTM</b>	<b>78.25</b>	76.80	83.97	76.52
ResNet-50	78.10	76.21	84.89	75.64
CA-ResNet-LSTM	78.78	76.65	85.66	75.84
<b>CA-ResNet-BiLSTM</b>	<b>83.65</b>	<b>80.61</b>	91.93	79.12

discriminative and significantly contribute to exploitation of the dynamic class dependency in the subsequent LSTM layer.

However, taking into account that the class dependency is bidirectional, a single-directional LSTM-based RNN is insufficient to draw a comprehensive picture of inter-class relevance. Therefore, a bidirectional LSTM-based RNN, composed of two identical recurrent streams but with reversed directions, is introduced in our model, and the hidden units are updated based on signals from not only their preceding states but also subsequent ones.

In order to practically adapt a bidirectional LSTM-based RNN to modeling the class dependency, we set the number of time steps in our bidirectional LSTM-based sub-network equivalent to that of classes under the assumption that distinct classes are predicted at respective time steps.

### III. EXPERIMENTS AND DISCUSSION

#### A. Data Description

The first experimental dataset, UCM multi-label dataset [12], is reproduced from UCM dataset [13] by reassigning them with multiple object labels. UCM multi-label dataset consists of 2100 aerial images, which is collected by cropping manually from aerial ortho imagery provided by the United States Geological Survey (USGS) National Map. The size of them is  $256 \times 256$  pixels, and the spatial resolution is one foot. The total number of object classes is 17: airplane, sand, pavement, building, car, chaparral, court, tree, dock, tank, water, grass, mobile home, ship, bare soil, sea, and field. To train and test our network on UCM multi-label dataset, we select 80% of sample images evenly from each scene category for training and the rest as the test set.

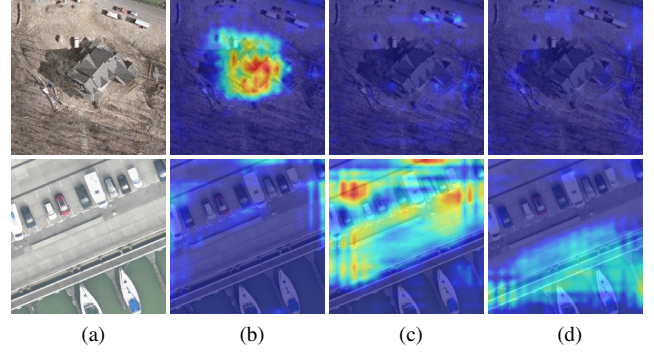


Fig. 3: Example class attention maps of (a) images in UCM (top row) and DFC15 (bottom row) multi-label dataset with respect to (b) building, (c) car, and (d) water. Red indicates strong activations, while blue represents non-activations. Besides, normalization is performed based on each row.

The second experiment dataset, DFC15 multi-label dataset, is built based on a semantic segmentation dataset, DFC15 [14], which was published and first used in 2015 IEEE GRSS Data Fusion Contest. DFC15 multi-label dataset contains 3342 images of  $600 \times 600$  pixels, and the spatial resolution of them is 5 cm. All images are assigned with multiple object labels according to labels of each pixel: impervious, water, clutter, vegetation, building, tree, boat, and car. To conduct evaluation, 80% of images are randomly selected as the training set, while the others are utilized to test our network.

#### B. Training Details

The proposed CA-Conv-BiLSTM is initialized with separate strategies with respect to three dominant components: 1) the feature extraction module is initialized with CNNs pre-trained on ImageNet dataset, 2) convolutional filters in the class attention learning layer is initialized with a Glorot uniform initializer, and 3) all weights in the bidirectional 2048-d LSTM layer are randomly initialized in the range of  $[-0.1, 0.1]$  with a uniform distribution. Notably, weights in the feature extraction module is trainable and fine tuned during the training phase of our network.




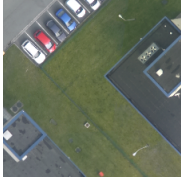
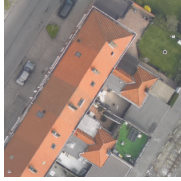

Regarding the optimizer, we chose Nestrov Adam, claimed to converge faster than stochastic gradient descent (SGD), and set parameters of the optimizer as recommended:  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 1e - 08$ . The learning rate is set as  $1e - 04$ , and decayed by 0.1 when the validation accuracy is saturated. The loss of the network is simply defined as mean squared error. We implement the network on TensorFlow and train it on one NVIDIA Tesla P100 16GB GPU for 100 epochs. The size of training batch is 32 as a trade-off between GPU memory capacity and training speed. To avoid overfitting, we stop training procedure when the loss fails to decrease in five epochs.

#### C. Discussion of the Results

To evaluate the performance of CA-Conv-BiLSTM for multi-label classification of high resolution aerial imagery, we



TABLE III: Example Predictions on UCM and DFC15 Multi-label Dataset

Images						
Ground Truths	building, car, pavement, and tree	building, court, pavement, grass, and tree	car, pavement, mobile-home, and tree	impervious, vegetation, car, and building	impervious, vegetation, building, clutter, and car	water, vegetation, tree
Predicted Labels	building, car, pavement, and tree	building, court, pavement, grass, and tree	car, pavement, mobile-home, tree, and <b>grass</b>	impervious, vegetation, car, and building	impervious, vegetation, building, clutter, and car	<b>impervious</b> , <b>water</b> , <b>tree</b> , and vegetation

Red predictions indicate false positives, while blue predictions are false negatives.

calculate  $F_1$  and  $F_2$  score as follows:

$$F_\beta = (1 + \beta^2) \frac{p_e r_e}{\beta^2 p_e + r_e}, \quad \beta = 1, 2, \quad (3)$$

where  $p_e$  is the example-based precision of predicted multiple labels, and  $r_e$  indicates the example-based recall. They are computed by:

$$p_e = \frac{TP_e}{TP_e + FP_e}, \quad r_e = \frac{TP_e}{TP_e + FN_e}, \quad (4)$$

where  $TP_e$ ,  $FP_e$ , and  $FN_e$  indicate numbers of true positives, false positives, and false negatives in an example (i.e., an image with multiple object labels in our case), respectively. Then, the average of  $F$  scores of each example is formed to assess the overall accuracy of multi-label classification tasks.

For a fair validation of CA-Conv-BiLSTM, we decompose the evaluation into two components: we compare 1) CA-Conv-LSTM with standard CNNs to validate the effectiveness of employing LSTM-based recurrent sub-network, and 2) CA-Conv-BiLSTM with CA-Conv-LSTM for further assess the significance of the bidirectional structure. Table I and II exhibit results on UCM and DFC15 multi-label datasets, respectively. The increments on both datasets demonstrate the effectiveness and robustness of our CA-Conv-BiLSTM for high resolution aerial image multi-label classification. Besides, Table III exhibits several example predictions in both datasets.

In addition to validate classification capabilities of the network by computing the mean  $F_1$  and  $F_2$  score, we further explore the effectiveness of class-specific features learned from the proposed class attention learning layer by feature visualization. It is observed that class attention maps highlight discriminative areas for different categories and exhibit almost no activations w.r.t. absent classes (as shown in Fig. 3).

#### IV. CONCLUSION

In this paper, we propose a novel network, CA-Conv-BiLSTM, for the multi-label classification of high resolution aerial imagery. The proposed network is composed of three indispensable elements: 1) a feature extraction module, 2) a class attention learning layer, and 3) a bidirectional LSTM-based sub-network. We evaluate our network on two datasets,

UCM multi-label dataset and DFC15 multi-label dataset, and experimental results validate the effectiveness of our model from both quantitative and qualitative respects. On one hand, the mean  $F_1$  and  $F_2$  score are increased compared to other competitors. On the other hand, visualized class attention maps demonstrate that they are class-specific and discriminative. Looking into the future, the application of our network can be extended to weakly supervised object localization.

#### REFERENCES

- [1] X. X. Zhu, D. Tuia, L. Mou, S. Xia, L. Zhang, F. Xu, F. Fraundorfer, Deep learning in remote sensing: A comprehensive review and list of resources, IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, 8–36, 2017.
- [2] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, CVPR, 2015.
- [3] L. Mou, X. X. Zhu, Vehicle Instance Segmentation from Aerial Image and Video Using a Multi-Task Learning Residual Fully Convolutional Network, IEEE Trans. Geosci. Remote Sens., no. 99, 1–13, 2018.
- [4] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, NIPS, 2015.
- [5] Y. Hua, L. Mou, X. X. Zhu, LAHNet: A convolutional neural network fusing low-and high-level features for aerial scene classification, IGARSS, 2018.
- [6] A. Zeggada, F. Melgani, Y. Bazi, A deep learning approach to UAV image multilabeling, IEEE Geosci. Remote Sens. Lett. vol. 14, no. 5, pp. 694–698, 2017.
- [7] S. Koda, A. Zeggada, F. Melgani, R. Nishii, Spatial and structured SVM for multilabel image classification, IEEE Trans. Geosci. Remote Sens., pp. 1–13, 2018.
- [8] K. Karalas, G. Tsagkatakis, M. Zervakis, P. Tsakalides, Land classification using remotely sensed data: Going multilabel, IEEE Trans. Geosci. Remote Sens., vol. 54 no. 6, pp. 3548–3563, 2016.
- [9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large scale image recognition, arXiv:1409.1556.
- [10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, CVPR, 2015.
- [11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CVPR, 2016.
- [12] B. Chaudhuri, B. Demir, S. Chaudhuri, L. Bruzzone, Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method, IEEE Trans. Geosci. Remote Sens. vol. 56, no. 2, pp. 1144–1158, 2018.
- [13] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, SIGSPATIAL ACM, 2010.
- [14] 2015 IEEE GRSS data fusion contest, <http://www.grss-ieee.org/community/technical-committees/data-fusion>, online.
- [15] L. Mou, L. Bruzzone, X. X. Zhu, Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery, IEEE Trans. on Geoscience and Remote Sensing, 2019, 57(2) 924–935.