LABEL RELATION INFERENCE FOR MULTI-LABEL AERIAL IMAGE CLASSIFICATION

Yuansheng Hua[1,2], *Lichao Mou*[1,2], *Xiao Xiang Zhu*[1,2]

[1]Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

Multi-label aerial image classification is a challenging visual task and obtaining increasing attention recently. Most of the existing methods resort to training independent classifier for each label, while underlying label correlations are not fully exploited while making predictions. To this end, we propose an innovative inference network, which takes advantage of pairwise label relations to infer multiple object labels of a high-resolution aerial image. Specifically, we first employ a feature extraction module to extract high-level feature representations of an aerial image, and then, feed them into a relational inference module to predict the presence of each object label. We evaluate our network on the UCM multi-label dataset and experiment with various popular convolutional neural networks (CNNs) as the backbone of the feature extraction module. Experimental results demonstrate that the proposed network behaves superiorly in comparison with other existing methods.

Index Terms— label relation, relational inference network, multi-label classification, CNN

1. INTRODUCTION

Aerial image classification is a fundamental visual mission, which aims at assigning images with various semantic categories. 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 [1]. With this intention, multi-label classification is now arising and obtaining increasing attention due to that 1) it provides a comprehensive picture of objects present in an aerial image, and 2) the acquisition of image-level labels (cf. multiple object-level labels in Fig. 1) is at a fair low cost. Along with such benefits, challenges have come up inevitably. On the one hand, it is difficult to extract high-level features from high-resolution images owing to its complex spatial structure. Conventional hand-crafted features and mid-level semantic models suffer from the poor performance of capturing holistic semantic features, which leads to an unsatisfactory classification ability. On the other hand, underlying correlations between dependent labels (cf., car and pavement in Fig. 1) are required to be explored for an efficient prediction of multiple object labels. However, the recently proposed multi-label classification methods [2, 3] assumed that labels are independent and employed a set of binary classifiers [2] or a regression model [3] to infer the existence of each label separately. Although [4] assumes that labels are correlated and employs a bidirectional Long short-term memory (LSTM) network to model such underlying label dependencies for multi-label classification, this method relies on a chain propagation structure and the correlation between each label is loose, especially those at both ends. Therefore, an efficient approach to fully explore a compact and synthetic correlation among all labels is essential for a high-performance multi-label classification model.

(a) (b) (c)

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

Recently, a relational reasoning network [5] has been proposed for visual question answering, where all inter-entity relations are computed for reasoning the object's properties. Experimental results demonstrate that it has even surpassed human performance in certain tasks. Later, [6] proposes a temporal relation network to exploit multi-scale temporal relations between frames to infer activities in a video. In [7], the authors propose an object relation module, which allows modeling relationships among sets of objects, for object detection tasks. Our work is motivated by the success of these works, but we focus on modeling potential label correlations. In this paper, we propose a novel label-wise relation (LR) network for capturing pairwise label relations for predicting the presence of each object label.

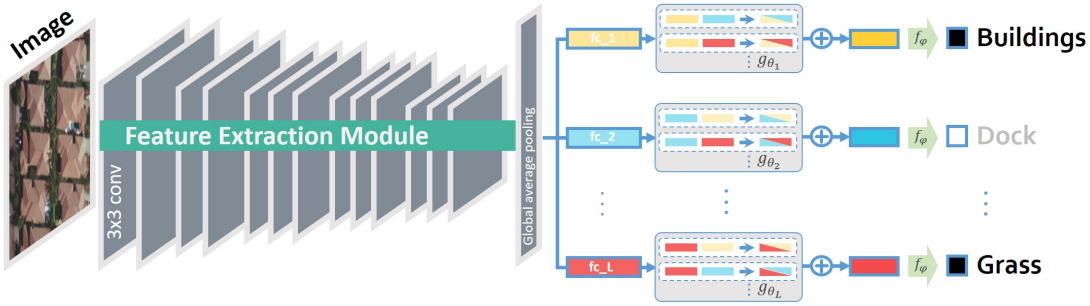

Fig. 2: The architecture of label relation network

2. NETWORK ARCHITECTURE

The proposed network consists of two indispensable components: 1) the label-wise feature extraction module and 2) the label relational inference module. The former module is designed for learning discriminative high-level features with respect to all object labels, while the latter module focuses on exploiting all pairwise relations among these features to predict the presence of each object. Details of these two modules are introduced in the following Section 2.1 and 2.2, respectively.

2.1. Label-wise feature extraction module

Learning efficient feature representations of input images is extremely crucial for the image classification task. Conventional methods mainly rely on hand-craft features, while extracted features are either low-level or mid-level and contain little semantics. To tackle this, a modern popular trend is now arising, which aims at employing a CNN architecture to automatically extract high-level features and then feeding them to trainable classifiers. Many recent studies [8] have achieved great progress in a wide range of classification tasks, such as image classification and semantic segmentation.

Following this trend, we employ three classical CNN architectures (i.e., VGGNet, GoogLeNet, and ResNet) to extract high-level features, which is further utilized to produce label-wise features. Specifically, for VGGNet, we feed outputs of *"block5_pool"* to a global average pooling layer for producing high-level feature representations of the input image. For GoogleNet and ResNet, outputs of their global average pooling layers are regarded as extracted semantic features.

However, these features are not label-specific, and thus can not be directly fed into the label relational inference module. Concerning this, label-wise fully connected layers are attached to learn label-wise semantic features. To be more specific, the output of aforementioned CNN architectures, e.g., a 512-D vector from VGGNet, is fed into separate fully connected layers (cf. *"fc1"*, *"fc2"*, and *"fcL"* in Fig 2). The number of fully connected layers is equivalent to that of all candidate object labels, denoted as L. Afterwards, L label-specific feature vectors are obtained, which are then fed into the subsequent label relational inference module.

2.2. Label Relational inference module

Inspired by recent successful relation networks, which has obtained huge attention for its great performance of inferring underlying relationships between feature pairs [5, 6], we propose a label relational inference module to exploit inherent label correlations to predict multiple object labels of an image.

In our label relational inference module, the presence of the object i is inferred according to its pairwise relations with all other objects. Formally, we define the synthetic label relation for the object i as a composite function with the following equation:

$$\mathrm{LR}(\boldsymbol{x}_i) = f_{\phi_i}(\sum_j g_{\theta_i}(\boldsymbol{x}_i, \boldsymbol{x}_j)), \tag{1}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote two label-wise feature vectors with respect to object label i and j, and j ranges from 1 to the number of labels, excluding i. The function g_{θ_i} is used to encode pairwise relations between labels i and j, and f_{ϕ_i} focuses on blending all these relation to form the synthetic label relation for object i, i.e., $\mathrm{LR}(\boldsymbol{x}_i)$. Here, we employ trainable multi-layer perceptrons (MLP) as g_{θ_s} and f_{ϕ_i}.

With this design, the network is capable of learning potential relations between object i and all other objects inherently. Compared to [4], where label dependencies are modeled in a propagation way, our proposed module explores a more direct and comprehensive label relation for all object labels. Afterward, $\mathrm{LR}(\boldsymbol{x}_i)$ is fed into a fully connected layer and activated by a sigmoid function to predict the presence of the object i.

3. EXPERIMENTS AND DISCUSSION

3.1. Data description

UCM multi-label dataset [9] is reproduced from UCM dataset [10] by reassigning them with multiple object labels. This

dataset consists of 2100 aerial images of 256×256 pixels, and the spatial resolution of each image is one foot. All images are collected by cropping manually from aerial ortho imagery provided by the United States Geological Survey (USGS) National Map, and each of them is assigned with one or more labels based on their primitive objects. The total number of newly defined 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.

3.2. Training details

The proposed LR network is initialized with its corresponding pre-trained CNN model and fine-tuned on the UCM multi-label dataset, where 80% of sample images are selected for training, and the rest for testing. Regarding the optimizer, we chose Adam with Nesterov momentum, 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 defined as the binary cross entropy.

We implement the network on TensorFlow and train it on one NVIDIA Tesla P100 16GB GPU for 100 epochs. The size of the 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. Concerning ground truths, multiple labels of an image are encoded into a multi-hot binary sequence, of which the length is equivalent to the number of all candidate labels. For each digit, 1 indicates the existence of its corresponding label, while 0 denotes the absent label.

3.3. Discussion of the result

To evaluate the performance of the proposed relation inference network for multi-label classification of high resolution aerial imagery, we calculate the example-based F_1 score as follows:

$$F_1 = (1+\beta^2)\frac{p_e r_e}{\beta^2 p_e + r_e}, \quad \beta = 1, \tag{2}$$

where p_e is the example-based precision of predicted multiple labels, and r_e indicates the example-based recall. Then, the average of F_1 scores of each example is formed to assess the overall accuracy of multi-label classification tasks. Besides, example- and label-based mean precision and mean recall are calculated to assess the performance from perspectives of the example and label, respectively.

For a comprehensive evaluation of our proposed network, which is denoted as LR-CNN in the following discussion, we compare with standard CNNs, and two relevant existing methods [2, 4]. Notably, considering standard CNNs are designed for single-label classification, we modify them by substituting last softmax layers with sigmoid layers to predict multi-hot binary sequences, where each digit indicates the probability of the presence of its corresponding label. In this way, each unit in the last fully connected layer can be considered as an independent classifier for predicting the presence of its corresponding object. To calculate evaluation metrics, we binarize outputs of all models with a threshold of 0.5 for producing binary sequences.

Table 1: Numerical Results on UCM Multi-label Dataset (%)

Methods	m.P_e	m.R_e	m.P_l	m.R_l	m.F_1
VGGNet [11]	79.1	82.3	**86.0**	80.2	78.5
VGG-RBFNN [2]	78.2	83.9	81.9	**82.6**	78.8
CA-VGG-BiLSTM [4]	79.3	**84.0**	85.3	76.5	79.8
LR-VGGNet	**84.9**	82.5	84.9	72.3	**82.1**
GoogLeNet [12]	80.5	84.3	87.5	80.9	80.7
GoogLeNet-RBFNN [2]	80.0	86.8	86.2	**84.9**	81.5
CA-GoogLe-BiLSTM[4]	79.9	**87.1**	86.3	84.4	81.8
LR-GoogLeNet	**83.4**	85.2	**89.8**	79.0	**83.0**
ResNet [13]	80.9	82.0	88.8	79.0	79.7
ResNet-RBFNN [2]	79.9	84.6	86.2	83.7	80.6
CA-Res-BiLSTM [4]	77.9	**89.0**	86.1	**84.3**	81.5
LR-ResNet	**87.1**	85.8	**90.0**	81.2	**85.3**

m.F_1 indicates the mean F_1 score.
m.P_e and m.R_e indicate mean example-based precision and recall.
m.P_l and m.R_l indicate mean label-based precision and recall.

Table 1 exhibits results on the UCM multi-label dataset, and it can be seen that compared to directly applying standard CNNs to multi-label classification, LR-CNN framework performs superiorly as expected due to taking all pairwise label correlations into consideration. LR-VGGNet increases the mean F_1 score by 3.6% with respect to VGGNet, while for LR-GoogLeNet, an increment of 2.3%, is obtained compared to GoogLeNet. Mostly enjoying this framework, LR-ResNet achieves the best mean F_1 score of 85.3% and an increment of 5.6% in comparison with other LR-CNN models and ResNet, respectively. Moreover, LR-ResNet shows an improvement of 3.8% of the mean F_1 score in comparison with CA-Res-BiLSTM, and compared to CA-CNN-BiLSTM architectures, LR-CNN obtains an increment of at least 1.2 in terms of the mean % F_1 score of 85.16%. To summarize, all comparisons demonstrate that the exploitation of compact pairwise label relation plays a key role in multi-label classification. Several example predictions are exhibited in Table 2.

Table 2: Example Predictions on UCM Multi-label Dataset

Images	Ground Truths	Predictions
	airplane, car, building, and pavement	airplane, car, building, and pavement
	bare soil, grass, building, court, tree, and pavement	bare soil, grass, building, court, car, tree, and pavement
	bare soil, building, car, and pavement	bare soil, building, car, tree, and pavement
	bare soil, car, and pavement	bare soil, car, and pavement
	bare soil, building, car, tree, and pavement	bare soil, building, car, court, grass, tree, and pavement
	car and pavement	car and pavement
	bare soil, building, car, court, grass, tree, and pavement	bare soil, building, car, court, grass, tree, and pavement

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

4. CONCLUSION AND OUTLOOK

In this paper, we proposed a novel RL-CNN network to exploit compact underlying label-wise relation for inferring multiple object labels of a high-resolution aerial image. Experiments are conducted on the UCM multi-label dataset and comparisons with relevant existing methods are performed for a comprehensive evaluation. The experimental results demonstrate that our RL-CNN network performs superiorly compared to other methods, and example predictions are exhibited to provide a distinct view of the performance of our model. Further work mainly comprises extraction of efficient label-wise features and modeling precise label relations for multi-label aerial image classification.

5. ACKNOWLEDGEMENTS

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.sipeo.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*).

6. REFERENCES

[1] Q. Tan, Y. Liu, X. Chen, and G. Yu, "Multi-label classification based on low rank representation for image annotation," *Remote Sensing*, vol. 9, no. 2, pp. 109, 2017.

[2] A. Zeggada, F. Melgani, and Y. Bazi, "A deep learning approach to UAV image multilabeling," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 694–698, 2017.

[3] S. Koda, A. Zeggada, F. Melgani, and R. Nishii, "Spatial and structured SVM for multilabel image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2018.

[4] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional lstm network for multi-label aerial image classification," *arXiv:1807.11245*, 2018.

[5] A. Santoro, D. Raposo, D. G.T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *NIPS*, 2017.

[6] B. Zhou, A. Andonian, and A. Torralba, "Temporal relational reasoning in videos," in *ECCV*, 2018.

[7] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *CVPR*, 2018.

[8] X. X. Zhu, D. Tuia, L. Mou, S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[9] B. Chaudhuri, B. Demir, Subhasis Chaudhuri, and Lorenzo Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2018.

[10] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, 2010.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.