Differential Convolution Feature Guided Deep Multi-Scale Multiple Instance Learning for Aerial Scene Classification

Beichen Zhou, Jingjun Yi, Qi Bi

School of Remote Sensing and Information Engineering, Wuhan University, China

ICASSP 2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

Introduction

Although deep learning approaches especially convolutional neural networks (CNNs) have boosted the performance of scene classification significantly, aerial scene classification remains to be challenging mainly due to some different characteristics between aerial images and ground images.

--(a) *Largely varied object sizes.*

--(b) *Arbitrary position and orientation.*

(a) (b)

Method

To stress the key local feature response while representing discriminative convolutional features from a variety of scales, we propose a deep multi-scale multiple instance learning (DMSMIL) framework, which consists of 3 parts.

Differential Dilated Convolutional Features. Our backbone is the widely-utilized VGGNet-16 in the aerial image community. Firstly, we utilize 256-channels, 3x3 window size, different dilated rate (r) convolution operators $\{d_i(\cdot)\}$ to extract multi-scale dilated convolutional features. We adopt four scales (i.e., $i = 1,2,3,4$) with the dilated rate $r = 1,3,5,7$ respectively.

Then, the convolutional feature from each scale is implemented $l-1$ norm (denoted as $\|\cdot\|_1$) with the convolutional feature from its adjacent scale. Also, the original features X is regarded as the 0^{th} scale. The four differential dilated convolutional features X_1, X_2, X_3, X_4 are calculated as shown on the right.

$$X_1 = \|d_1(X) - X\|_1,$$
$$X_2 = \|d_2(X) - d_1(X)\|_1,$$
$$X_3 = \|d_3(X) - d_2(X)\|_1,$$
$$X_4 = \|d_4(X) - d_3(X)\|_1.$$

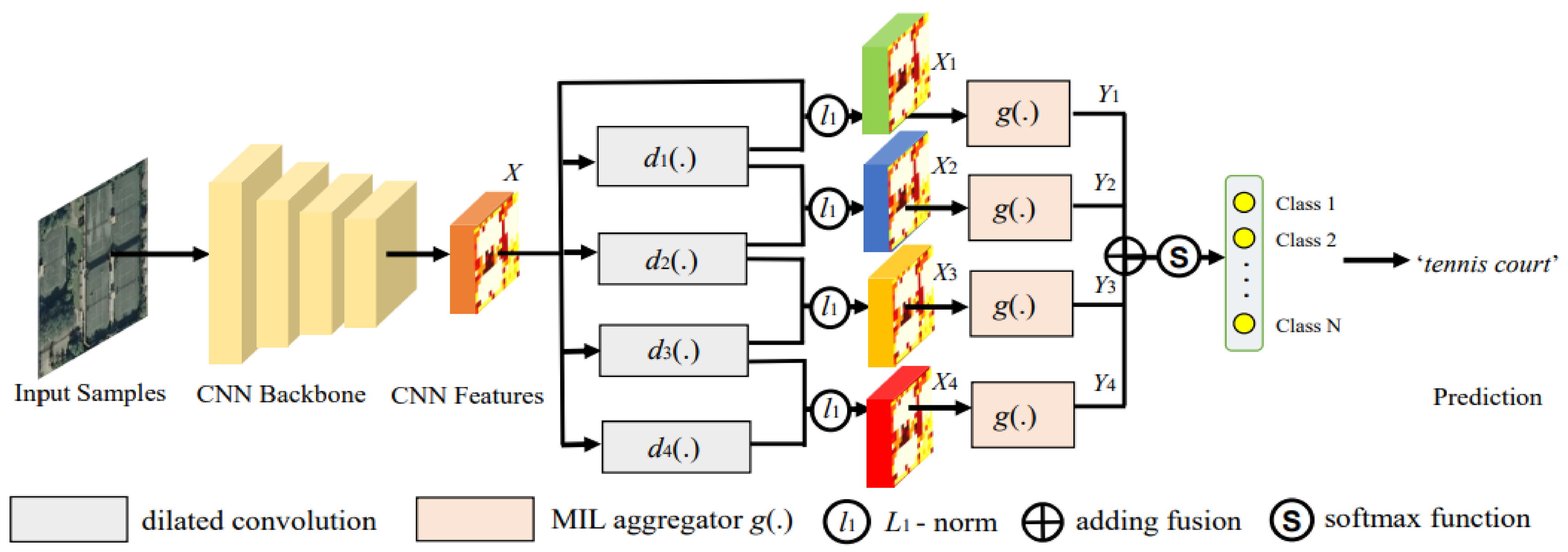


Fig. Demonstration of our proposed deep multi-scale multiple instance learning (DMSMIL) framework.

Multi-scale Multiple Instance Learning Module. Each differential dilated convolutional feature $X_i (i = 1,2,3,4)$ is fed into a deep MIL module f_i, which converts X_i into a set of bag-level probability distribution Y_i.

$$Y_i = f_i(X_i).$$

-- *Converting the convolutional features X into an instance-level feature representation X' by a 1×1 convolutional layer.*

$$Y_i = g(X').$$

--*Averaging all the instance feature responses in each channel to get the corresponding bag-level feature response.*

$$Y_{i,k} = \frac{\sum_{w=1}^{W} \sum_{h=1}^{H} X'_{w,h,k}}{W \times H}$$

Semantic Prediction Fusion. The final bag probability distribution Y is the adding fusion of Y_i from each scale. This process can be presented as

$$Y = softmax(\sum_{i=1}^{4} Y_i)$$

Results

Comparison with state-of-the-art methods. We conduct our experiments on UCM, AID and NWPU datasets. Three widely-used aerial image scene classification benchmarks. It can be seen that our approach outperforms all the SOTA approaches and the corresponding baseline models in five out of six experiments with an obvious improvement. In AID 50% experiment, our approach only performs a little bit worse than the D-CNN.

Method	UCM		AID		NWPU	
	50%	80%	20%	50%	10%	20%
AlexNet [7]	93.98±0.67	95.02±0.81	86.86±0.47	89.53±0.31	76.69±0.21	79.85±0.13
VGGNet-16 [7]	94.14±0.69	95.21±1.20	86.59±0.29	89.64±0.36	76.47±0.18	79.79±0.15
GoogLeNet [7]	92.70±0.60	94.31±0.89	83.44±0.40	86.39±0.55	76.19±0.38	78.48±0.26
SPP-Net [5]	94.77±0.46	96.67±0.94	87.44±0.45	91.45±0.38	82.13±0.30	84.64±0.23
MIDC-Net [1]	95.41±0.40	97.40±0.48	88.51±0.41	92.95±0.17	86.12±0.29	87.99±0.18
TEX-Net [18]	94.22±0.50	95.31±0.69	87.32±0.37	90.00±0.33	—	—
D-CNN [19]	—	98.93±0.10	90.82±0.16	**96.89±0.10**	89.22±0.50	91.89±0.22
MSCP [6]	—	98.36±0.58	91.52±0.21	94.42±0.17	85.33±0.17	88.93±0.14
FV [8]	—	98.57±0.34	—	—	—	—
ARCNet [15]	96.81±0.14	99.12±0.40	88.75±0.40	93.10±0.55	—	—
DMSMIL (ours)	**99.09±0.36**	**99.45±0.32**	**93.98±0.17**	95.65±0.22	**91.93±0.16**	**93.05±0.14**

For specific evaluation protocols, parameter settings and development environment, please refer to our paper.

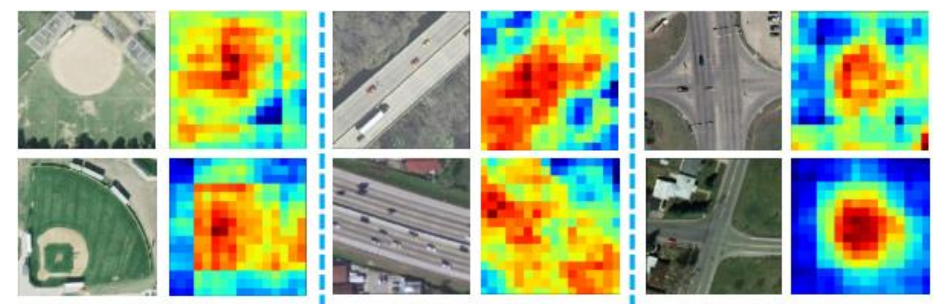
Ablation Study. We performed ablation experiments on the NWPU dataset for our DMSMIL framework and the results are shown below.

Our approach (VGG+DDC+MSMIL) achieves the best performance, and using it leads to an obvious performance boost compared with using single-scale deep MIL (VGG+DDC+SMIL)

Visualized Samples. The figure offers some visualized feature response maps when processed by our DMSMIL framework, which comes from the adding fusion and normalization of X_1, X_2, X_3 and X_4.

It indicates that our framework could provide discriminative feature representation and could be transferred to the task of aerial image object detection and segmentation.

	50%	80%
VGG	76.69±0.21	79.85±0.13
VGG+DDC	85.36±0.18	88.73±0.17
VGG+SMIL	87.23±0.16	89.42±0.19
VGG+DDC+SMIL	89.02±0.15	91.35±0.18
VGG+DDC+MSMIL (ours)	**91.93±0.16**	**93.05±0.14**



(a) baseball diamond (b) freeway (c) intersection

Conclusion

-- Our DMSMIL framework is for aerial scene recognition, taking the wide range of object sizes and the complicated object distribution into account.

-- Our framework includes a differential convolutional feature extractor, a multi-scale multi-instance learning module and a semantic prediction fusion module. Experiments on three datasets validate the effectiveness of our proposed approach.

Key references

[1] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in Int. Conf. Mach. Learn., 2018.

[2] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected convnet for aerial scene classification," in IEEE Trans Image Process., 2020, vol. 29, pp. 4911–4926.

[3] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," in Pattern Recognit., 2018, vol. 74, pp. 15–24.

Contact: q_bi@whu.edu.cn , zbc350715695@gmail.com