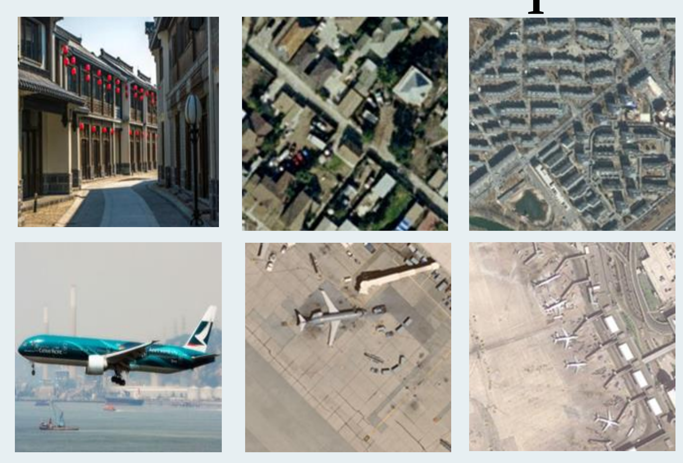
Multiple Instance Dense Connected ConvNet for Aerial Image Scene Classification

Qi Bi, Kun Qin*, Zhili Li, , Han Zhang, Kai Xu
School of Remote Sensing and Information Engineering, Wuhan University, China

Introduction

Compared with ground scenes, aerial scenes are quite challenging because of



Varied object distribution.

Complicated spatial arrangement.

Strong background information.

Current ConvNets tend to preserve global features, while recent studies point out the following solutions for aerial scene classification.

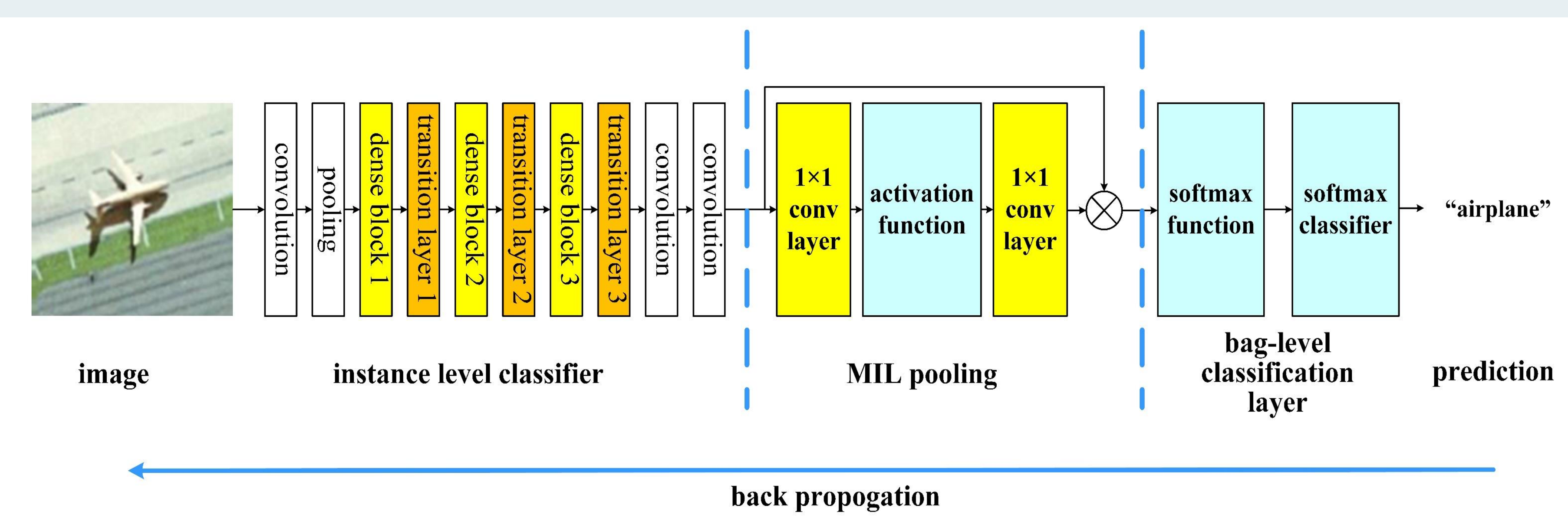
Enhancing local semantic representation.

Preserving more shallower features.

Method

We propose a multiple instance dense connected ConvNet (MIDC-Net) for aerial scene classification.

Local patches: instances; image scenes: bags



Instance-level classifier. We introduce a simplified dense connection structure as our backbone.

(1)More orderly multi-level feature representation.

The forth dense block is removed.

(2)Much fewer convolutional layers.

Only three composite functions in each dense block.

(3)More channels to highlight local features.

MIL pooling. We propose a trainable MIL pooling operator based on spatial attention mechanism.

(1)Selecting instances relevant to the scene label via assigning higher weights.

$$a_{ij} = softmax({w_2}^T tanh(W_1 F_{ij}^T + b))$$

(2)Calculating a bag-level probability distribution.

$$g(\{p_{ij}\}) = \sum_i \sum_j a_{ij} p_{ij}$$

Bag-level classification layer. We utilize cross-entropy loss function to optimize the entire framework. It is under the direct supervision of bag labels.

Results

Comparison with state-of-the-art methods

UCM dataset

Method	Training ratio 50%	Training ratio 80%
PLSA(SIFT) [6]	67.55±1.11	71.38±1.77
BoVW(SIFT) [6]	73.48±1.39	75.52±2.13
LDA(SIFT) [6]	59.24±1.66	75.98±1.60
AlexNet [6]	93.98±0.67	95.02±0.81
VGGNet-16 [6]	94.14±0.69	95.21±1.20
GoogLeNet [6]	92.70±0.60	94.31±0.89
SPP with AlexNet [47]	94.77±0.46*	96.67±0.94
D-CNN with AlexNet [5]	—	96.67±0.10
TEX-Net with VGG [15]	94.22±0.50	95.31±0.69
Gated attention [54]	94.64±0.43*	96.12±0.42*
MIDC-Net (ours)	**94.93±0.51**	**97.00±0.49**

—: not reported, *: not reported & conducted by us

AID dataset

Method	Training ratio 20%	Training ratio 50%
PLSA(SIFT) [6]	56.24±0.58	63.07±1.77
BoVW(SIFT) [6]	62.49±0.53	68.37±0.40
LDA(SIFT) [6]	51.73±0.73	68.96±0.58
AlexNet [6]	86.86±0.47	89.53±0.31
VGGNet-16 [6]	86.59±0.29	89.64±0.36
GoogLeNet [6]	83.44±0.40	86.39±0.55
SPP with AlexNet [47]	87.44±0.45*	91.45±0.38*
D-CNN with AlexNet [5]	85.62±0.10	**94.47±0.12**
TEX-Net with VGG [15]	87.32±0.37	90.00±0.33
Gated attention [54]	87.63±0.44*	92.01±0.21*
MIDC-Net (ours)	**88.26±0.43**	92.53±0.18

—: not reported, *: not reported & conducted by us

NWPU dataset

Method	Training ratio 10%	Training ratio 20%
BoVW(SIFT) [9]	41.72±0.21	44.97±0.28
AlexNet [9]	76.69±0.21	79.85±0.13
VGGNet-16 [9]	76.47±0.18	79.79±0.15
GoogLeNet [9]	76.19±0.38	78.48±0.26
SPP with AlexNet [47]	82.13±0.30*	84.64±0.23*
D-CNN with AlexNet [5]	85.56±0.20	87.24±0.12
Gated attention [54]	84.94±0.22*	86.62±0.22*
MIDC-Net (ours)	**85.59±0.26**	**87.32±0.17**

—: not reported, *: not reported & conducted by us

	Parameters (in million)	Model size (in MByte)
AlexNet	60	434
VGG-VD-16	138	1024
GoogLeNet	6.8	91.1
MIDC-Net(ours)	**0.5**	**9.94**

Gated attention: method in [1]

Comparison of MIL pooling operators

	UCM 50%	UCM 80%	AID 20%	AID 50%	NWPU 10%	NWPU 20%
No MIL pooling	94.52±0.63	96.21±0.67	87.37±0.41	91.49±0.22	83.97±0.19	85.63±0.18
Mean_pooling [53], [55], [65]	94.82±0.54	96.41±0.44	87.87±0.37	92.19±0.24	84.94±0.18	86.37±0.18
Max_pooling [53], [55], [65]	93.81±0.49	95.91±0.55	86.41±0.39	91.21±0.27	82.88±0.22	85.23±0.21
Attention (ours)	**94.93±0.51**	**97.00±0.49**	**88.26±0.43**	**92.53±0.18**	**85.59±0.26**	**87.32±0.17**

Influence of simplified dense connection structure

	UCM 50%	UCM 80%	AID 20%	AID 50%	NWPU 10%	NWPU 20%
Dense4 [30]	93.75±0.55	95.81±0.55	85.85±0.43	91.92±0.21	83.91±0.27	85.93±0.19
#Dense4+#conv	94.16±0.44	96.22±0.53	87.41±0.51	91.99±0.19	84.98±0.29	86.08±0.20
#Dense4+conv (ours)	**94.93±0.51**	**97.00±0.49**	**88.26±0.43**	**92.53±0.18**	**85.59±0.26**	**87.32±0.17**

Dense4:original dense connection structure

Number of convolutional layers	UCM	AID	NWPU
1	95.01±0.62	90.07±0.58	84.79±0.20
2	95.91±0.63	91.87±0.25	85.97±0.22
3	**97.00±0.49**	**92.53±0.18**	**87.32±0.17**
4	96.14±0.58	92.06±0.26	86.82±0.23
5	95.90±0.50	91.92±0.22	86.05±0.21

Conclusion

(1) Our MIDC-Net outperforms many state-of-the-art methods with much fewer parameters. It offers an end-to-end solution for the combination of MIL and ConvNet under the direct supervision of bag labels.

(2) Our proposed attention based MIL pooling operator outperforms non-trainable operators such as mean or maximum pooling operator, and the recently proposed gated attention based MIL pooling operator.

(3) Simplified dense connection structure preserves features from different levels well and outperforms the original dense connection structure.

Key Reference

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

[2] X. Wang, Y. Yan, T. Peng, B. Xiang, and W. Liu, "Revisiting multiple instance neural networks," Pattern Recognit., vol. 74, pp. 15–24, 2016.

[3] H. Gao, L. Zhuang, L. Maaten, and K. Weinberger, "Densely connected convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*(CVPR), 2017.

Contact: q_bi@whu.edu.cn , 2009biqi@163.com