Recurrent Transformer Networks for Remote Sensing Scene Categorisation

Zan Chen[1]
zanchen2@gmail.com
Shidong Wang[2]
shidong.wang@uea.ac.uk
Xingsong Hou[1]
houxs@xjtu.edu.cn
Ling Shao[3]
ling.shao@ieee.org

[1] Xi'an Jiaotong University
Xi'an, China
[2] University of East Anglia
Norwich, UK
[3] Inception Institute of Artificial Intelligence
Abu Dhabi, United Arab Emirates

Remote sensing scene categorisation is a task to distinguish the basic level scene images in accordance with the contents of the subordinate level feature representations. This gives rise to a significant semantic gap between subordinate level features and the basic level scene contents. In this work, we propose recurrent transformer networks (RTN) to mitigate the above problem. RTN incorporates learning transformation-invariant regions with transformer based attention mechanism, thus reducing the semantic gap efficiently. It can learn the canonical appearance for the most relevant regions based on the subordinate level contents of the remote sensing scene images. The predictions of both transformation parameters and classification score are derived from the bilinear CNN pooling regression. The whole network is differentiable and can be learned end-by-end by only acquiring the basic level labels. As shown in Figure 1, our RTN framework composes of three major parts which are recurrent warp operation, bilinear pooling operation, and intra-scale loss L_{intra} and inter-scale loss L_{inter}.

The original STN [2] includes multiple independent streams, and each stream learns its own spatial transformation independently, which neglects the latent relationship of each stream. To address these disadvantages, we extract the relevant multi-scale region-based feature representations progressively. Specifically, our warp operation runs in a recurrent manner, which can be denoted as

$$I^{(s)} = f_{warp}(\theta^{(s)}\tau^{(s)}, I^{(s-1)}), \tag{1}$$

where $I^{(s)}$ is the s-th scale image (e.g., $I^{(0)}$ is the raw image), θ^s is the transformation parameters computed by the localisation function $\theta^s = f_{loc}^{(s)}(I^{(s-1)})$, and $\tau^{(s)}$ is the target coordinates of the regular grid in the output image. Each warp operation f_{warp} has the similar progress to the original STN. The warp operation requires applying a sampling kernel on the input image $I^{(s-1)}$, to produce the value at a particular pixel in the finer scale image $I^{(s)}$. With repeatedly calling the warp operation, the network can progressively yield multi-scale discriminative regions.

We merge intra-scale loss for each stream and inter-scale loss for neighbouring streams to optimise the network. The final loss is defined as

$$L = \sum_{s=0}^{S} L_{intra}^{(s)} + \alpha \sum_{s=0}^{S-1} L_{inter}^{(s)}, \tag{2}$$

where α is a hyper-parameter to adjust the total loss and learn the latent relationship between the neighbouring scales. Suppose $P^{(s)}$ and P^* as the predicted label vector from a specific scale and the ground truth label respectively, then the intra-scale loss $L_{intra}^{(s)}$ can be written as

$$L_{intra}^{(s)} = -\sum_{k=1}^{n} P_k^* \log P_k^{(s)}, \tag{3}$$

where n is the number of the classification. To ensure the streams learning in a mutual reinforcement way, we impose inter-scale loss for the adjoining scales and define it as

$$\begin{aligned} L_{inter}^{(s)} &= \max(0, \sum_{k=1}^{n} P_k^* (\log P_k^{(s)} - \log P_k^{(s+1)}) - \text{margin}) \\ &= \max(0, L_{intra}^{(s+1)} - L_{intra}^{(s)} - \text{margin}), \end{aligned} \tag{4}$$

which enforces $L_{intra}^{(s+1)} < L_{intra}^{(s)} + \text{margin}$ during the training phase. In such way, each scale can refer to the adjoining scales to progressively learn sub-region feature representations. With gradually attending at the finer scale, the extracted features are able to decrease the semantic gap by degrees and boost the performance of the proposed architecture on the RSSC datasets.

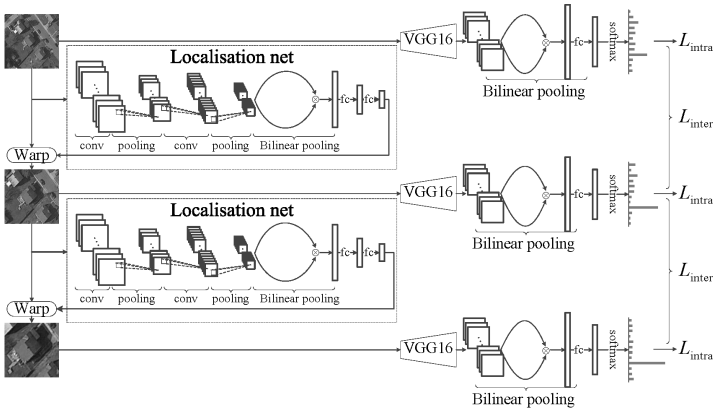

Figure 1: The architecture of recurrent transformer networks (RTN). Given an input image, the localisation network will learn to predict the transformer parameters. With recurrently applying the warp operation, the network can progressively attend to the discriminative regions and produce multi-scale relevant sub-images.

We conduct experiments on three publicly available remote sensing image datasets, including NWPU-RESISC45, UC Merced, and AID. As shown in Table 1, we can obvious that CNN-based feature approaches have a much better performance on predicting the categories for RSSC tasks compared with the hand-craft feature approaches. The best accuracies on RSSC datasets are made by the recently proposed D-CNN method [1], which takes the metrics learning as the regularisation term. Compared to the state-of-the-art results of D-CNN, our RTN achieves improvements to the categorisation accuracies on all experiment datasets.

Table 1: Comparison results of our RTN to baselines and previous work. We experiment the accuracy in different training ratios on three public RSSC datasets.

	Method	NWPU-RESISC45		AID	UC-Merced
		10%	20%	20%	80%
Handcraft Feature	SPM+SIFT	27.83	32.96	38.43	60.02
	LLC+SIFT	38.81	40.03	58.06	72.55
	BoVW+SIFT	41.72	44.97	62.49	75.52
Deep Feature	GoogLeNet+SVM	82.57	86.02	87.51	96.82
	VGG16+SVM	87.15	90.36	89.33	97.14
	D-CNN with GoogLeNet [1]	86.89	90.49	88.79	97.07
	D-CNN with VGG16 [1]	89.22	91.89	90.82	98.93
	RTN with VGG16 (ours)	**89.90**	**92.71**	**92.44**	**98.96**

[1] Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 2018.

[2] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.