Multi-task Relative Attributes Prediction by Incorporating Local Context and Global Style Information Features

Yuhang He
www.heyuhang.com
Long Chen
http://www.carlib.net
Jianda Chen
http://www.carlib.net

School of Data and Computer Science
Sun Yat-sen University
Guangzhou, P.R China

Relative attribute represents the correlation degree of one attribute between an image pair. Fine-grained or appearance insensitive relative attribute prediction still remains as a challenging task. To address this challenge, we propose a multi-task trainable deep neural networks by incorporating an object's both local context and global style information to infer the relative attribute. In particular, we leverage convolutional neural networks (CNNs) to extract feature, followed by a ranking network to score the image pair. In CNNs, we treat features arising from intermediate convolution layers and full connection layers in CNNs as local context and global style information, respectively. Our intuition is that local context corresponds to bottom-to-top localised visual difference and global style information records high-level global subtle difference from a top-to-bottom scope between an image pair. We concatenate them together to escalate overall performance of multi-task relative attribute prediction. Finally, experimental results on 5 publicly available datasets demonstrate that our proposed approach outperforms several other state of the art methods and further achieves comparable results when comparing to very deep networks, like 152-ResNet and inception-v3.

0.1 Feature Learning

We propose to learning discriminative feature by incorporating both local context and global style information feature. Local context stores object's local and obvious feature, while global style information stores more abstract and high-level feature. We achieve this by extracting final full connection layer feature and intermediate layer feature, and further concatenate them together to form the final feature vector [1] (see fig.2 for framework pipeline):

$$\psi^i = \psi^i_{fc} + \psi^i_{local} + \psi^i_{global} \qquad (1)$$

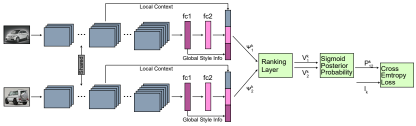

Figure 1: Framework pipeline: we feed the image pair to two CNNs with the same network architecture and shared parameters. Features learned by CNNs intermediate layers and final several full connection layers are concatenated together to form the final feature. The feature pair is further fed to the ranking layer to score each attribute.

0.2 Relative Attribute Prediction via Ranking

The image feature extracted above is mapped to a real value (or a real value vector) through a matrix W and a bias term b: $v = W \cdot \psi + b$. Then We calculate the posterior probability for each relative attribute and squash it between [0-1] via a sigmoid function[2],

$$P^k_{1,2} = \frac{1}{1 + e^{-(v^i_1 - v^i_2)}} \qquad (2)$$

Finally, we utilise cross entropy loss to rank each relative attribute,

$$\mathcal{L}_i = \sum_{k=1}^{K} l^i_k \log(P^k_{1,2}) - (1 - l^i_k) \log(-P^k_{1,2}) \qquad (3)$$

0.3 Experiment

see the paper for detailed experiment discussion.

[1] W. Choi F. Yang and Y. Lin. Exploit all layers: fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proc. CVPR*, 2016.

[2] E. Adeli-Mosabbed Y. Souri, E. Noury. Deep relative attributes. *arXiv preprint arXiv:1512.04103*, 2015.