Two-stream contextualized CNN for fine-grained image classification

Jiang Liu[1], **Chenqiang Gao**[1,*], **Deyu Meng**[2], **Wangmeng Zuo**[3]

[1]Chongqing University of Posts and Telecommunications, Chongqing, China
[2]Xi'an Jiaotong University, Xi'an, China
[3]Harbin Institute of Technology, Harbin, China
liujiang4work@outlook.com, gaocq@cqupt.edu.cn, dymeng@mail.xjtu.edu.cn, wmzuo@hit.edu.cn

Abstract

Human's cognition system prompts that context information provides potentially powerful clue while recognizing objects. However, for fine-grained image classification, the contribution of context may vary over different images, and sometimes the context even confuses the classification result. To alleviate this problem, in our work, we develop a novel approach, two-stream contextualized Convolutional Neural Network, which provides a simple but efficient context-content joint classification model under deep learning framework. The network merely requires the raw image and a coarse segmentation as input to extract both content and context features without need of human interaction. Moreover, our network adopts a weighted fusion scheme to combine the content and the context classifiers, while a subnetwork is introduced to adaptively determine the weight for each image. According to our experiments on public datasets, our approach achieves considerable high recognition accuracy without any tedious human's involvements, as compared with the state-of-the-art approaches.

Introduction

We notice that when people discriminate different images, they will have different attentions between content and context information. Inspired from this cognition mechanism and combining it with the recent deep learning method(Simonyan and Zisserman 2014a), we propose a novel two-stream contextualized Convolutional Neural Network(CNN), to address fine-grained level classification problem. Our contribution mainly has the following three aspects:

- The content and context information are finely compensated with each other. Thus, for a given image, we employ a two-stream CNN architecture to simultaneously utilize both of them.
- The fusion weight for compensating the content and context classifiers can be optimally adapted in an unified optimization framework, which finely accords with the cognition process of humans.
- Our method achieves the state-of-the-art performance on public datasets so far without any human interactions.

Methodology

Overview

Our two-stream contextualized CNN network utilizes two parallel nets: *content net* and *context net*, which are modified and fine-tuned from the convolutional layers of vgg-16 net(Simonyan and Zisserman 2014b), to capture both object and background features. The input of our two-stream network requires an image coupling with its coarse image segmentation to separate content and context regions. We name them as the content image and context image in our work, respectively. We get the content image from the raw image by calculating the bounding box of the content part given the segmentation result. The context counterpart is obtained by filling the content region with the pixels from equivalent position of the mean image calculated across the training set. Two parallel nets, the content net and the context net are employed to handle the corresponding images and extract feature maps. We also conduct a Spatial Pyramid Pooling(SPP)(He et al. 2014) operation after the last convolutional layer in order to obtain fixed length features without distorting and cropping the input image. Finally, we design a optimal fusion model to automatically learn the fusion weight from content and context features and output the final recognition result.

Classifier and optimal fusion model

Our optimal fusion model is illustrated in Figure 1. Given the *i*-th input image x_i, off-the-shelf segmentation result $M(x_i)$, we can obtain corresponding context feature $f^c(x_i)$ and content feature $f^o(x_i)$. We then formalize three independent networks: W_o-Net as the content classifier, W_c-Net as the context classifier and W_q-Net as the fusion weight regression model to construct the optimal fusion framework. Both the W_o-Net and W_c-Net are fully connected to $f^o(x_i)$ and $f^c(x_i)$ to output probabilistic classification results φ_{ij}^o and φ_{ij}^c, respectively. The W_q-Net is constructed with a fully-connected layer followed by a logistic regression model to output the fusion weight q_i ranging from 0 to 1. Finally, we define the probabilistic classification result using a fusion

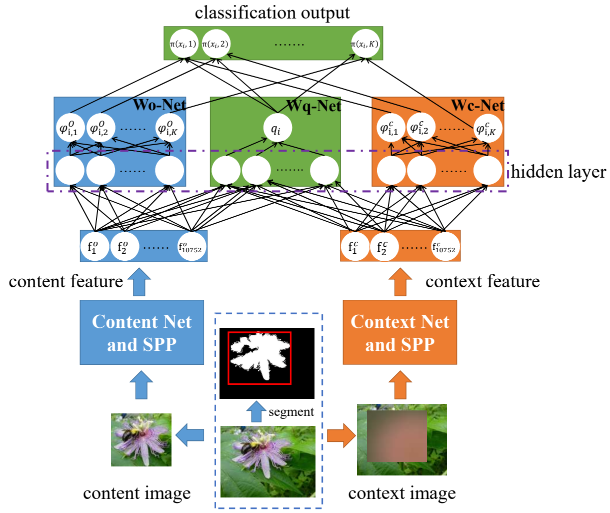

Figure 1: The structure of our two-stream network and the optimal fusion model: content and context net are designed to caputure content and context feature simultaneously; W_o-Net and W_c-Net are two classifiers for content and context features; W_q-Net is intended to output the fusion weight.

function $\pi(x_i, j)$ as following:

$$\pi(x_i,j) = \frac{1}{1+\exp(-(q_i\varphi^c_{ij} + (1-q_i)\varphi^o_{ij}))}, \qquad (1)$$

where j is the index of each dimension of the classification result vector. The hidden layer parameters of W_o-Net, W_c-Net and W_q-Net could be easily learned using an alternative Stochastic Gradient Descent(SGD) method by updating one particular network and fix the rests at one time.

Experiment

We test our two-stream contextualized CNN framework on three popular datasets: Oxford Flower 102(Flower102)(Nilsback and Zisserman 2008), Caltech-UCSD Birds 200-2010(CUB2010)(Welinder et al. 2010) and Caltech-UCSD Birds 200-2011(CUB2011)(Wah et al. 2011) using their corresponding evaluation metrics. The well-established BiCoS segmentation approach(Chai, Lempitsky, and Zisserman 2011) is employed for both Flower102 and CUB2010 images. As to the CUB2011, we directly employ the ground truth segmentation result. We select five baseline methods using either hand-crafted or CNN features which do not rely on part annotation, denoted as: template-based(Chen et al. 2012), codebook-based(Khan et al. 2011), segmentation-based(Chai, Lempitsky, and Zisserman 2011), CNN-1(Razavian et al. 2014) and CNN-2(Azizpour et al. 2015). Finally, our results are listed in Table 1. We can observe that our method outperforms the baselines obviously and the optimal fusion strategy further improves the performances of traditional content/context fusion methods.

Method	Accuracy		
	Flower102	CUB2010	CUB2011
codebook-based	73.3	22.4	-
templated-based	82.6	19.2	-
segmentation-based	80.0	23.3	-
CNN-1	86.8	-	65.0
CNN-2	91.3	-	67.1
Ours(Early Fusion)	93.7	31.0	66.7
Ours(Late Fusion)	93.8	39.2	73.3
Ours(Optimal Fusion)	**94.5**	**41.8**	**76.9**

Table 1: The performance comparision of our approaches with several baselines. Our method based on content and context CNN features with the optimal fusion model exhibits significant better results than traditional appraoches and fusion strategies.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No.61571071, 61102131).

References

Azizpour, H.; Razavian, A.; Sullivan, J.; Maki, A.; and Carlsson, S. 2015. From generic to specific deep representations for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 36–45.

Chai, Y.; Lempitsky, V.; and Zisserman, A. 2011. Bicos: A bi-level co-segmentation method for image classification.

Chen, Q.; Song, Z.; Hua, Y.; Huang, Z.; and Yan, S. 2012. Hierarchical matching with side information for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 3426–3433. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Computer Vision–ECCV 2014*. Springer. 346–361.

Khan, F. S.; Weijer, J.; Bagdanov, A. D.; and Vanrell, M. 2011. Portmanteau vocabularies for multi-cue image representation. In *Advances in neural information processing systems*, 1323–1331.

Nilsback, M.-E., and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*.

Razavian, A. S.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, 512–519. IEEE.

Simonyan, K., and Zisserman, A. 2014a. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, 568–576.

Simonyan, K., and Zisserman, A. 2014b. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.