IMPROVING FACIAL ATTRACTIVENESS PREDICTION VIA CO-ATTENTION LEARNING

Shengjie Shi[1], Fei Gao[1,2,*], Member, IEEE, Xuantong Meng[1], Xingxin Xu[1], Jingjie Zhu[1]

[1] Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China.

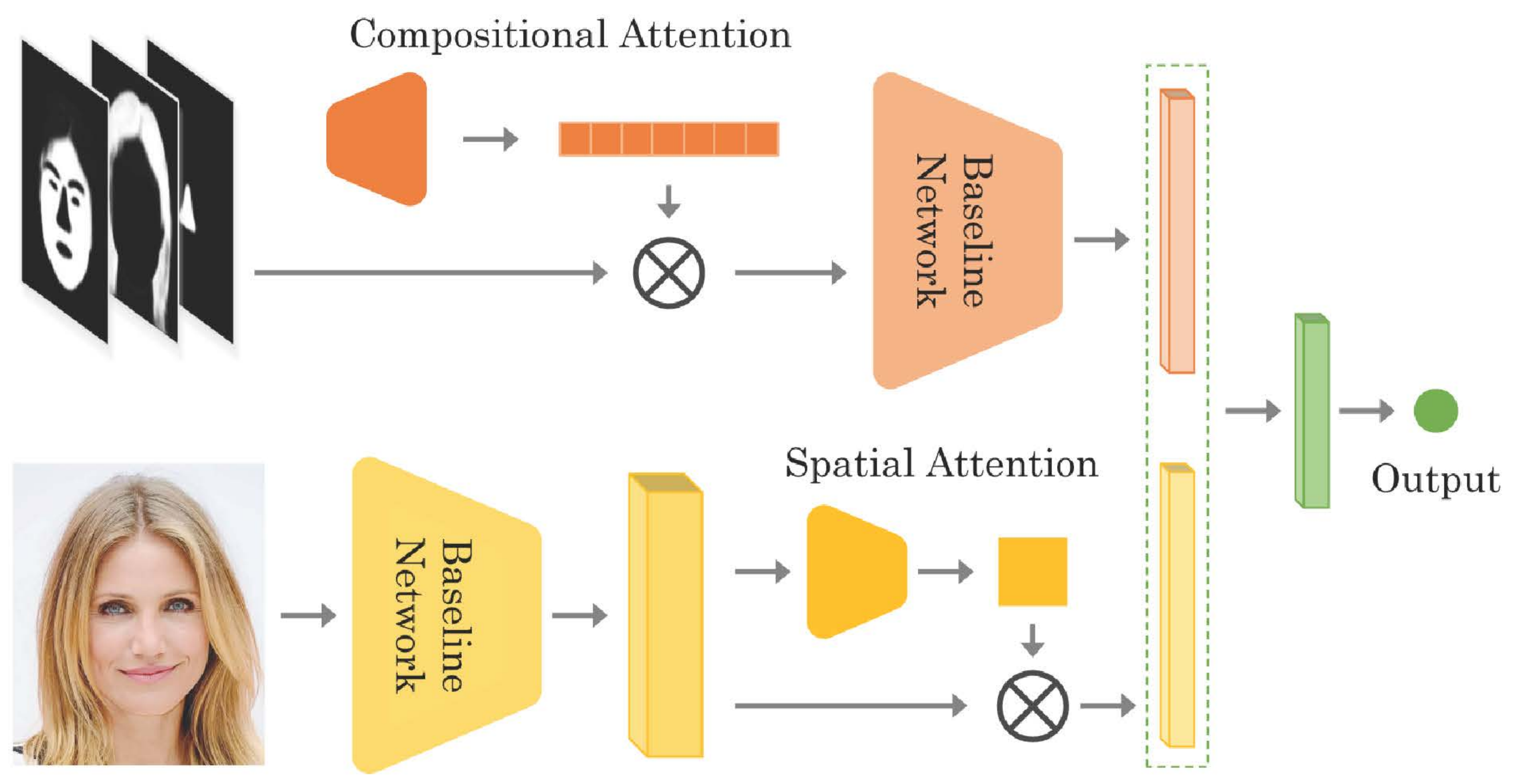
[2] State Key Laboratory of Integrated Services Networks, Xidian University, Xi′an 710071, China.

Abstract

Facial attractiveness prediction has drawn considerable attention from image processing community. Despite the substantial progress achieved by existing works, various challenges remain.

- One is the lack of accurate representation for facial composition, which is essential for attractiveness evaluation. In this paper, we propose to use pixel-wise labelling masks as the meta information of facial composition, and input them into a network for learning high-level semantic representations.
- The other challenge is to define to what degree different local properties contribute to facial attractiveness. To tackle this challenge, we employ a co-attention learning mechanism to concurrently characterize the significance of different regions and that of distinct facial components.
- We conduct experiments on the SCUT-FBP5500 and CelebA datasets. Results show that our co-attention learning mechanism significantly improves the facial attractiveness prediction accuracy. Besides, our method consistently produces appealing results and outperforms previous advanced approaches.

Proposed



Objective

1. We formulate the former as a binary classification problem, and use Binary Cross-Entropy (BCE) loss in the learning process.
2. We formulate score prediction as a regression task and use the L2 loss for training the network.

Table 1. Network architecture. Each line describes a sequence of 1 or more identical layers, repeated n times. All layers in the same sequence have the same number c of output channels. (This table follows [16])

Input	Layer	c	n
MobileNetV2 (baseline network)			
$224^2 \times 3$	Conv	32	1
$112^2 \times 32$	bottleneck	16	1
$112^2 \times 16$	bottleneck	24	2
$56^2 \times 24$	bottleneck	32	3
$28^2 \times 32$	bottleneck	64	4
$14^2 \times 64$	bottleneck	96	3
$14^2 \times 96$	bottleneck	160	3
$7^2 \times 160$	bottleneck	320	1
$7^2 \times 320$	Conv 1×1	1280	1
$7^2 \times 1280$	avgpool 7×7	-	1
spatial attention module			
$7^2 \times 1280$	Conv	1280	1
$7^2 \times 1280$	Tanh	–	1
$7^2 \times 1280$	Conv	1	1
$7^2 \times 1$	Softmax	–	1
compositional attention module			
1×1	FC	7	1
1×7	Softmax	–	1

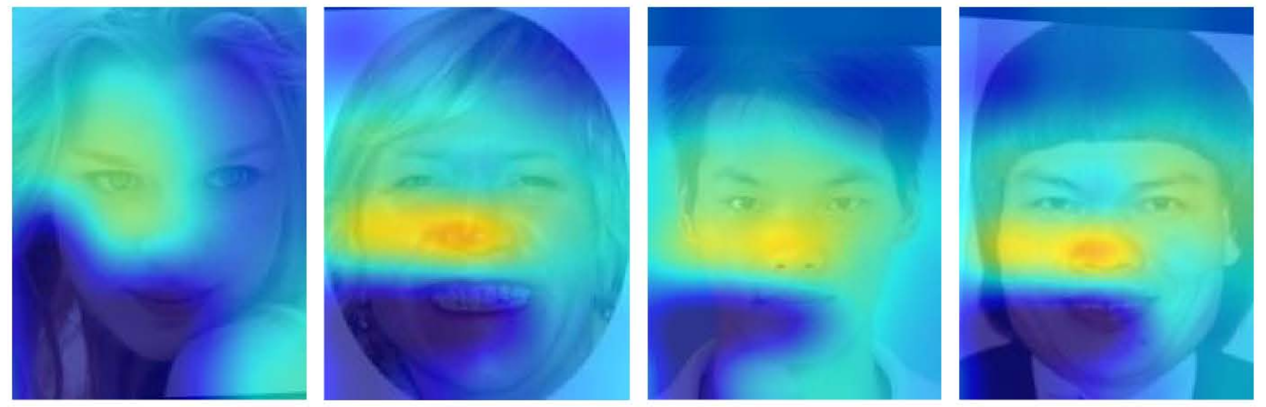
Co-attention Mechanisms

Spatial Attention

Let $\mathbf{A}^{(s)} = \{a_{i,j}^{(s)}\}_{i,j=1}^{7}$ denotes the learned spatial attention. $\mathbf{A}^{(s)}$ is used to integrate local activation vectors by:

$$\mathbf{x}_a = \sum_{i=1}^{7}\sum_{j=1}^{7} a_{i,j}^{(s)} \mathbf{X}_{i,j}. \tag{1}$$

$\mathbf{x}_a \in \mathbb{R}^{1\times 1280}$ is concatenated with the output of the composition branch, and then used for attractiveness prediction.



(3.80, 4.00) (2.13, 1.93) (2.25, 2.50) (1.65, 1.35)

Compositional Attention

We denote the compositional attention vector by:

$$\mathbf{a}^{(c)} = (a_1^{(c)}, a_2^{(c)}, ..., a_7^{(c)}), \text{with } \sum_{i=1}^{7} a_i^{(c)} = 1. \tag{2}$$

$a_i^{(c)}$ measures the correlation between the i^{th} component and facial attractiveness. Afterwards, $\mathbf{a}^{(c)}$ is used to integrate the pix-wise labelling masks by:

$$\mathbf{M}_a = \sum_{i=1}^{7} a_i^{(c)} \mathbf{M}^{(i)}. \tag{3}$$

$\mathbf{M}_a$ is input into a network for learning high-level representation of facial composition, and finally used in attractiveness prediction.

Experimental Results

Table 2. Results of ablation study.

Model Variants	CelebA Acc.(%)	SCUT-FBP5500 PLCC	SCUT-FBP5500 SRCC
image	83.4	0.920	0.909
image+spat.att.	84.4	0.925	0.914
masks	84.1	0.806	0.785
masks+comp.att.	84.1	0.835	0.813
full	**85.2**	**0.926**	**0.916**

Table 3. Performance on the SCUT-FBP5500 dataset.

Method	PLCC	SRCC	MAE	RMSE
LBP+GR [13]	0.674	–	0.391	0.509
Gabor+SVR [13]	0.807	–	0.401	0.518
AlexNet [13]	0.863	–	0.265	0.348
ResNet-18 [13]	0.890	–	0.242	0.317
ResNeXt-50 [13]	0.900	–	0.229	0.302
Ours	**0.926**	**0.916**	**0.202**	**0.266**

Table 4. Performance on the CelebA dataset.

Method	Publication	Acc.(%)
PANDA [22]	CVPR'14	81.0
Liu *et.al* [18]	ICCV'15	81.0
MOON [23]	ECCV'16	81.7
Ding *et.al* [24]	Arxiv'17	82.9
Cao *et.al* [21]	CVPR'18	84.4
Ours	ICASSP'19	**85.6**

• References

- L. Liang, L. Lin, L. Jin, D. Xie, and M. Li, "SCUTFBP5500: A diverse benchmark dataset for multiparadigm facial beauty prediction," 2018.
- M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," arXiv:1801.04381v3, 2018.
- Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in Proc. IEEE Int. Conf. Comput. Vis., Dec 2015, pp. 3730–3738.
- Y. Y. Fan, S. Liu, B. Li, Z. Guo, A. Samal, J. Wan, and S. Z. Li, "Label distribution-based facial attractiveness computation by deep residual learning," IEEE Trans. Multimedia, vol. 20, no. 8, pp. 2196–2208, 2018.