Depthwise Separable Convolution for Facial Landmarks Detection

Savina Colaco and Dong Seog Han*
School of Electronics and Electrical Engineering, Kyungpook National University,

Daegu, Republic of Korea

savinacolaco@knu.ac.kr, dshan@knu.ac.kr*

Abstract

Facial landmarks detection has been used as important information for problems such as head pose estimation, facial emotion expression, face recognition, and face modelling. The facial keypoints are point information on the face such as eye corners, eye centre, mouth corners, nose, jawline, eyebrow points and so on. In this paper, depthwise separable convolution is used to predict vital keypoints on the face which is trained with public datasets with additional data. The face is detected with a widely used face detector. The predicted keypoints are mapped on a face detected to detect keypoints in real-time. The model is evaluated with wing loss and adaptive wing loss.

Ⅰ. Introduction

Facial landmark detection is also known as a facial alignment problem, is one of the challenging problems in the field of computer vision [1]. The landmarks can be used in applications such as face recognition, facial emotion recognition, self-driving cars and so on. With improved landmark detection, facial information can solve various facial alignment problems. Building a system with Convolutional neural networks (CNN) has been widely popular since it outperforms the traditional approach in speed and accuracy. Using CNN with deep structure, landmarks can be predicted and detected simultaneously. It can extract a high level of features needed for the prediction. In this paper, the deep learning approach depthwise separable convolution is used to predict the facial keypoints and mapped with the detected face in real-time.

Ⅱ. Experiment

The model trained with 300W [2] dataset which consists of XM2VTS, AFW, HELEN, LFPW and IBUG with additional data adding up to 112K grayscale images with 68 (x, y) coordinates. The input size scaled to 112x112 resolution. The model is implemented with Keras framework with epoch at 300 and batch size of 100. It uses Adam optimizer with learning rate fixed to 10^{-3} throughout the training. The model is evaluated with wing [4] and adaptive wing loss [5] as described by equations 1 and 2 respectively.

$$wing(x) = \begin{cases} \omega \ln\left(1 + \frac{|x|}{\varepsilon}\right) & \text{if } |x| < \omega \\ |x| - C & \text{otherwise} \end{cases} \quad (1)$$

where ω is a non-negative constant which sets the range of the nonlinear part to $(-\omega, \omega)$, ε limits the curvature of the nonlinear region and $C = \omega - \omega \ln(1 + \omega / \varepsilon)$ is a constant that smoothly links linear and nonlinear parts. The parameters are set to $\omega = 10$ and $\varepsilon = 2$.

$$Awing(y, \hat{y}) = \begin{cases} \omega \ln\left(1 + \left|\frac{y - \hat{y}}{\varepsilon}\right|^{\alpha - y}\right) & \text{if } |(y - \hat{y})| < \theta \\ A|y - \hat{y}| - C & \text{otherwise} \end{cases} \quad (2)$$

where y and $\hat{y}$ are ground truth and predicted values, respectively. Unlike wing loss ω as the threshold, Θ is the new variable threshold to switch between linear and non-linear part. The ω, Θ, ε, and α are positive values. $A = \omega(1/(1+(\Theta/\varepsilon)^{(\alpha - y)}))(\alpha - y)((\Theta/\varepsilon)^{(\alpha - y - 1)})(1/\varepsilon)$ and $C = (\Theta A - \omega \ln(1+(\Theta/\varepsilon)^{(\alpha - y)}))$ and are used to make smooth and continuous loss function at $|y - \hat{y}| = \Theta$. Similar settings from the paper (11) are used such as $\omega = 14$, $\Theta = 0.5$, $\varepsilon = 1$, and $\alpha = 2.1$.

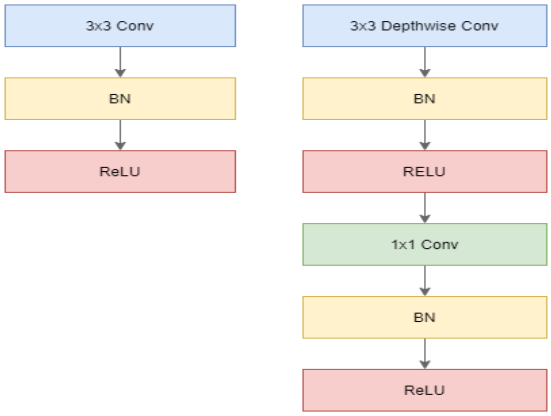

Figure 1: (Left) standard convolution layer and (right) MobileNet

The facial landmarks are predicted using mobileNet [6] model which is based on depthwise separable convolutions. The depthwise separable convolution is depthwise convolutions followed by pointwise convolution. In Fig. 1, the standard convolution layer is followed by batch normalization (BN) and rectified Linear units (ReLU). In MobileNet, the depthwise separable convolutions with depthwise and pointwise layers followed by batch normalization and ReLU. The depthwise separable convolution into two layers, where one layer for filtering and another for combining inputs whereas the standard convolution both filters and combines inputs into a new set of outputs in one step.

Single-shot detector (SSD) with ResNet as the backbone model is used to detect the user's face in the input image. In Figs. 2 and 4, the model accuracy with wing and adaptive wing loss functions are depicted in the plot with width multiplier 1 and 0.5, respectively.

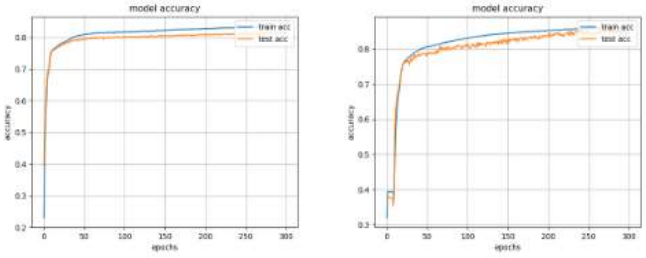

Figure 2: Model accuracy Loss functions with width multiplier=1

Figure 3: Facial keypoint detection with width multiplier = 1

In Figs. 3 and 5, the facial keypoints detection in real-time is demonstrated with width multiplier 1 and 0.5 respectively for wing and adaptive wing loss functions.

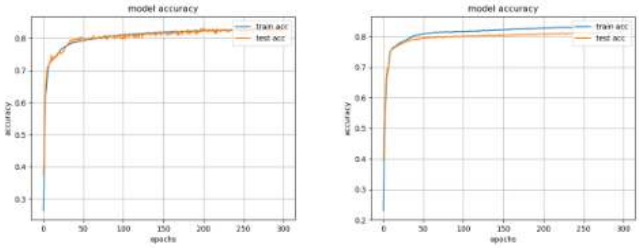

Figure 4: Model accuracy Loss functions with width multiplier=0.5

Figure 5: Facial keypoint detection with width multiplier = 0.5

Table 1: Model Accuracy of MobileNetV1 with wing loss and Adaptive wing loss

Model	Loss function	Width multiplier	Accuracy
MobileNet V1	Wing loss	α =1	84.4%
		α =0.5	82.5%
	Adaptive wing loss	α =1	84.9%
		α =0.5	81.5%

The model trained is 3.3M and 0.9M parameters in total for width multiplier 1 and 0.5, respectively. The model trained with adaptive wing loss has higher accuracy compared to wing loss. The model loss of adaptive wing loss is considerably lower than wing loss. The model sensitive to extreme head poses orientation and occlusion.

Ⅲ. Conclusion

In the paper, MobileNetV1 is used to predict the 68 facial keypoints. It is mapped with the detected face using an SSD detector with ResNet. The facial landmarks can be lost if the initial face detector failed to detect faces. Facial landmark detection is sensitive to extreme facial poses, occlusion and illumination conditions which can be improved.

ACKNOWLEDGMENT

This research was supported by the Ministry of Trade, Industry & Energy (MOTIE), Korea Institute for Advancement of Technology (KIAT) through 5G-based autonomous driving convergence technology demonstration platform task (task number: 1415169669).

References

[1] S. Shi, “Facial Keypoints Detection,” arXiv preprint arXiv:1710.05279, 2017.

[2] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-Wild challenge: The first facial landmark localization challenge,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops, Dec. 2013, pp. 397– 403

[3] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, “Wing loss for robust facial landmark localisation with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, 2018, pp. 2235– 2245.

[4] X. Wang, L. Bo, and L. Fuxin, “Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression,” in Proc. IEEE/CVF Int. Conf. Comput. Vis, Seoul, Korea, 2019, pp. 6971-6981.

[5] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *ArXiv,* vol. abs/1704.04861, 2017.