CenterAttentionFaceNet: A improved network with the CBAM attention mechanism

Thu Hien Nguyen[1#], Danh Vu Nguyen[1#], Trang Phung[2*]

[1]Thai Nguyen University of Education, Thai Nguyen, Vietnam
[2]Thai Nguyen University, Thai Nguyen, Vietnam
 # denotes the first two authors contributed equally
** is the corresponding author*

---***---

Abstract - *Convolutional Neural Network (CNN), one of common the Deep Learning models, is becoming more and more advanced, becoming the most widely used solution for most computer vision-related applications including facial recognition. Due to their high accuracy and practicality, facial recognition models play a key role in most real-world scenarios. However, the training process of these models is time-consuming and expensive. Therefore, designing a lightweight model with low computational cost and memory requirements is one of the most practical solutions for face recognition. CenterFaceNet is one of the popular lightweight networks to address the facial detection problem. In this paper, we proposed the combination of CenterFaceNet and attention modules to enhance performance while keeping the simplicity of lightweight architecture. Specifically, we propose to utilize CBAM attention that includes the Channel Attention Module and Spatial Attention Module after each block of the CenterFaceNet backbone. The test results of our proposed model on the WIDER FACE dataset show superiority to the original CenterFace model and state-of-the-art methods.*

Keywords: *Face Detection, Attention, Deep Learning, MobileNet, CBAM.*

1.INTRODUCTION

Face detection is an important area of object detection in computer vision [1], which has wide applications in areas such as security [2], recognition [3], image processing [4], video classification[5], etc. The goal of this process is to find and locate faces in an image. In recent years, the development of face detection algorithms has made significant progress, thanks to the development of deep learning models and the development of neural networks e.g., CNN - Convolutional Neural Networks) [6]. The previous face detection methods have inherited the model based on the common object detection framework. The results have shown that the combination with deep learning has significantly increased the performance and accuracy of the model. However, the problem of face located prediction seems inaccuracy due to many possible results in an image. In addition, high inference time cost and large model is also very challenging. In this paper, we export a simple and effective face detection and alignment model architecture based on CenterFace [7], which is lightweight but extremely powerful. CenterFace's network architecture is depicted in Figure 1.

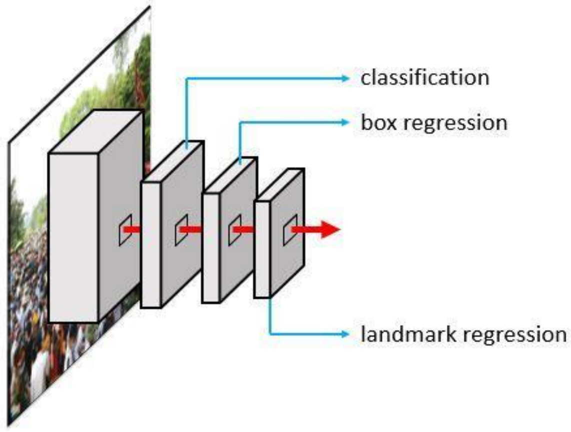

Figure 1: Overall of the CenterFaceNet architecture

2. RELATED WORK

2.1. Previous Methods

Continuing with related work, there are also several studies on learning multitasking [8, 9] for face detection. This approach involves the use of multiple monitoring labels to improve the accuracy of each task by using correlations between tasks. Detecting and aligning faces in a simultaneous model is widely used because the task of aligning and rearranging facial landmarks is done using the process of key feature extraction (backbone) [10] of a neural network, providing better features for the face classification task with information from face points. Similarly, RCNN significantly improved the detection performance by adding a branch to predict the faces of the subjects.

State-of-the-art studies on face detection were performed using cascaded CNN methods [6], using anchor points [11-13] and multitasking learning. Although each method has its own advantages and disadvantages, recent advances have shown that the anchor-based methods and its phases [14, 15] have made significant advances in both accuracy and efficiency. These methods densely sample face locations and scales on feature maps and use natural anchor points [11] or single points representing faces for regression, thus simple Simplify the training process and significantly reduce the training time.

Furthermore, there have been studies on the use of attentional mechanisms to enhance detection and recognition [16, 17]. The attention mechanism allows models to selectively focus on certain parts of the input, which can help improve accuracy and reduce misinformation. For example, the S3FD method [11] used the attention module to emphasize facial areas and suppress non-facial areas. Another study introduced a new mechanism that uses spatially variable Gaussian [18] filters to selectively enhance features in the facial region.

In general, face detection has been studied extensively and various approaches have been developed and developed over time. Although some methods are more suitable for specific situations or applications, recent advances have shown that anchor point-based methods and single-stage methods are often more efficient and accurate for face detection tasks. Attention mechanisms have also shown potential in improving detection performance and reducing false positive levels.

2.2. Our work

Different from mentioned face detection methods that focus on using anchor points and multitasking learning, in this paper, we mine the advantage of attention modules which have been demonstrated to bring better performance on various computer vision problems such as [16, 17]. Specifically, we introduce a compact module to exploit attention mechanisms. In the CBAM module [19] exploit all spatial and channel attention to enhance the network performance.

The CBAM is inserted after the convolutional network layer and before the output prediction layer. In addition, to solve the problem of aspect ratio mismatch in the training data, automatic techniques are also proposed to scale the images on the training and test data. We found that combining this module with the model is highly effective in face detection. The detail of the experiment result has been provided in Section 4 of this work.

3. METHODOLOGY

In this section, we present our proposed framework based on the combination of CenterFaceNet architecture with spatio-temporal attention module i.e., CBAM to increase the performance of the model.

3.1. Proposed method

MobileNetV3-Large. Introduced in [20] this model is the “Large” version that is targeted at the respective high resource use cases. This architecture consists of “bneck” blocks, in which, each block contains various convolution layers followed by a BatchNormalize layer and an activation layer e.g., hard-swish or ReLU (see Figure 2).

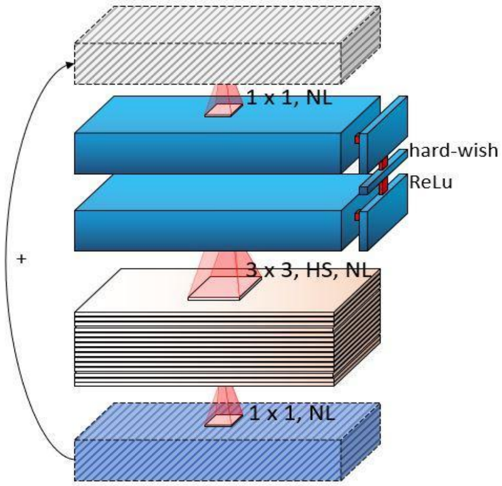

Figure 2: MobileNet V3 Architecture

The CenterFace model uses a backbone network namely MobileNet V3 [20] to extract features from the input image, the output of MobileNet is then combined with the head layers to perform tasks. This architecture consists of stacked blocks, each block has various convolution layers and nonlinear activation functions applied between convolution layers. The detail of MobileNet V3 [20] architecture is shown in Table 1.

Table 1: The detail of MobileNet V3 architecture. SE denotes Squeeze – Exciten [21]. NL denotes the type of nonlinearity used. In which, HS is H-Swish [20] and Re means ReLu. NBN [22] denotes No Batch Normalization. S presents the stride value.

Input	Operator	exp size	Output	SE	NL	s
$224^2\times3$	con2d	-	16	-	-	2
$112^2\times16$	bneck, 3×3	16	16	-	Re	1
$112^2\times16$	bneck, 3×3	64	24	-	Re	2
$56^2\times24$	bneck, 3×3	72	24	-	Re	1
$56^2\times24$	bneck, 5×5	72	40	SE	Re	2
$28^2\times40$	bneck, 5×5	120	40	SE	Re	1
$28^2\times40$	bneck, 5×5	120	40	SE	Re	1
$28^2\times40$	bneck, 3×3	240	80	-	HS	2
$14^2\times80$	bneck, 3×3	200	80	-	HS	1
$14^2\times80$	bneck, 3×3	184	80	-	HS	1
$14^2\times80$	bneck, 3×3	184	80	-	HS	1
$14^2\times80$	bneck, 3×3	480	112	SE	HS	1
$14^2\times112$	bneck, 3×3	672	112	SE	HS	1
$14^2\times112$	bneck, 5×5	672	160	SE	HS	2
$7^2\times160$	bneck, 5×5	672	160	SE	HS	1
$7^2\times160$	bneck, 5×5	960	160	SE	HS	1
$7^2\times160$	conv2d, NBN	-	24	-	HS	1
$1^2\times160$	conv2d, NBN	-	960	-	HS	1
$1^2\times960$	conv3d, NBN	-	320	-	HS	1
$1^2\times320$	conv4d, NBN	-	24	-	HS	1

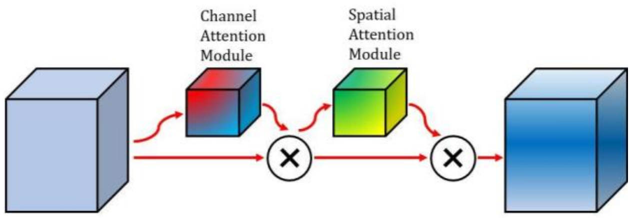

Figure 3: Overview of CBAM module

The CBAM module illustrated in Figure 3, consists of two main mechanisms that play an important role in image feature enhancement: the Channel Attention Module and the Spatial Attention Module.

Channel Attention Module. Generates a channel attention map by exploiting the channel relationships of features. This mechanism focuses on enhancing the input data of the layers in the network. It does this by marking important channels in the input data and maximizing high-quality features. This is done through the use of marked important communication channels to convey information to the next layers and ignore the unimportant channels.
Channel Attention Module is illustrated in Figure 4. Specifically, the input of this module is passed to two different Pooling layers including Avg and Max to get two feature vectors. These vectors are then passed to MLP (Multi-Layer Perceptron) [23] layers with the number of neurons reduced to produce a channel importance vector. Finally, the two vectors of the two channels are concatenated and passed through a Sigmoid layer to normalize the value, the resulting vector is then used to refine the input critical information channels and minimize the input channels not important to produce a more precise input. To summarize, the Channel Attention Module is illustrated as follows:

$$M_c(F)=\sigma(MLP(AvgP(F))+MLP(MaxP(F))) \quad (1)$$

where *σ denotes the sigmoid layer.*

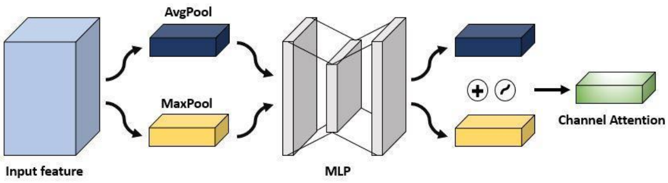

Figure 4: Channel Attention Module

Spatial Attention Module. This module aims to create a spatial attention map using the spatial relationships of features to focus on important parts of the image. To determine the spatial attention (as shown in Figure 5), we first apply the mean (AvgP) and the maximum value (MaxP) along the channel axis and concatenate them to create an efficient feature descriptor. The application of pooling layers is proven to be effective in highlighting regions of information. Then these two vectors are merged together and passed into a convolution layer with a kernel size of 7×7. Finally, the results are normalized and passed through the activation function (sigmoid) to generate a vector containing the values 0 and 1, representing the importance of each position on the image. To summarize, the Spatial Attention Module is illustrated as follows:

$$M_s(F)=\sigma(f^{7\times7}([AvgP(F);MaxP(F)])) \quad (2)$$

where σ denotes the sigmoid layer and $f^{7\times7}$ means the convolution layer with a kernel size of 7×7.

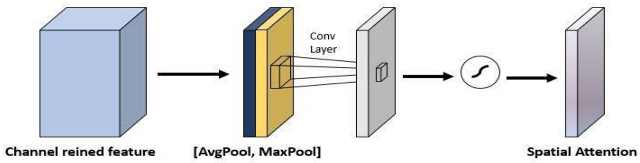

Figure 5: Spatial Attention Module

3.2. CenterAttentionFaceNet

In this part, we present the CenterAttentionFaceNet by combining the original CenterFaceNet with the CBAM modules including Spatial Attention Module and Channel Attention Module. Specifically, CBAM is placed after each "bneck" block of Mobilenet V3 (see Table 2). This aims to analyze spatial and input channels and generate better local features. The output of each "bneck" block has feature maps of higher importance and reduces the repetition of unnecessary features. Then, these features after being processed will be transferred to the last three layers to define and localize the face i.e., classification layer, box regression layer, and landmark regression layer.

Table 2: The architecture of MobileNet V3 with CBAM attention modules

Input	Operator	exp size	Output	SE	NL	s
224^2×3	conv2d	-	16	-	-	2
112^2×16	bneck, 3×3	16	16	-	Re	1
	CBAM	-	-	-	-	-
112^2×16	bneck, 3×3	64	24	-	Re	2
	CBAM	-	-	-	-	-
56^2×24	bneck, 3×3	72	24	-	Re	1
	CBAM	-	-	-	-	-
56^2×24	bneck, 5×5	72	40	SE	Re	2
	CBAM	-	-	-	-	-
28^2×40	bneck, 5×5	120	40	SE	Re	1
	CBAM	-	-	-	-	-
28^2×40	bneck, 5×5	120	40	SE	Re	1
	CBAM	-	-	-	-	-
28^2×40	bneck, 3×3	240	80	-	HS	2
	CBAM	-	-	-	-	-
14^2×80	bneck, 3×3	200	80	-	HS	1
	CABM	-	-	-	-	-
14^2×80	bneck, 3×3	184	80	-	HS	1
	CBAM	-	-	-	-	-
14^2×80	bneck, 3×3	184	80	-	HS	1

	CBAM	-	-	-	-	-
$14^2 \times 80$	bneck, 3×3	480	112	SE	HS	1
	CBAM	-	-	-	-	-
$14^2 \times 112$	bneck, 3×3	672	112	SE	HS	1
	CBAM	-	-	-	-	-
$14^2 \times 112$	bneck, 5×5	672	160	SE	HS	2
	CBAM	-	-	-	-	-
$7^2 \times 160$	bneck, 5×5	672	160	SE	HS	1
	CBAM	-	-	-	-	-
$7^2 \times 160$	bneck, 5×5	960	160	SE	HS	1
	CBAM	-	-	-	-	-
$7^2 \times 160$	conv2d, CBN	-	24	-	HS	1
$1^2 \times 160$	conv2d, CBN	-	960	-	HS	1
$1^2 \times 960$	conv3d, NBN	-	320	-	HS	1
$1^2 \times 320$	conv4d, NBN	-	24	-	HS	1

4. EXPERIMENT

4.1. Dataset

WIDER FACE DATASET. We use the WIDER FACE dataset to train the CenterAttentionFaceNet model. This is a widely used dataset in the field of facial recognition. This dataset contains more than 32,000 images with more than 50,000 faces marked with landmarks. The WIDER FACE dataset is divided into three parts including a Training set (40%), a validation set (10%), and a test set (50%). We report the results of the test in Tables 3 and 4, respectively. Model evaluation at the three levels that the model achieves is Easy, Medium, and Hard.

4.2. Implementation Detail

We train our proposed model with the ADAM optimizer and learning rate $\eta = 0.002$ and drop it by x10 after the validation loss saturates. We set $\beta 1 = 0.5$ and $\beta 2 = 0.99$. We set the mini-batch size to 32 images.

4.3. Benmark Results

The experiment results on val set of the WIDER FACE dataset have shown in Table 3. All methods have evaluated by the SIO (Single Inference on the Original) metrics. Specifically, our proposed approach achieves the 94.3% (Easy), 93.6% (Medium) and 88.4% (Hard) that ouperforms state-of-the-art methods such as FaceBoxes [24], FaceBoxes3.2× [25], RetinaFace-mnet [26], LFFD-v1 [25], LFFD-v2, CenterFace [7] by a margin 0.8% - 14.5% on Easy set, 1.2% - 17% on Medium set and 0.9% - 48.9% on Hard set.

Table 3: The results on val set of the WIDER FACE dataset

Method	Easy Set	Medium Set	Hard Set
FaceBoxes [24]	0.840	0.766	0.395
FaceBoxes3.2× [25]	0.798	0.802	0.715
RetinaFace-mnet [26]	0.896	0.871	0.681
LFFD-v1 [25]	0.910	0.881	0.780
LFFD-v2	0.837	0.835	0.729
CenterFace [7]	0.935	0.924	0.875
CenterAttentionFaceNet (Ours)	**0.943**	**0.936**	**0.884**

The same with experiment on val set, for the test set, we also utilize SIO metric and compare our proposed approach to state-of-the-art methods in Table 4. The experiment results have shown that our CenterAttentionFaceNet outperforms all recent proposed methods such as FaceBoxes [24], FaceBoxes3.2×[25], RetinaFace-mnet[26], LFFD-v1[25], LFFD-v2, CenterFace [7]. Specifically, our network is better than CenterFace [3] with 1.7% on Easy set, 1.5% on Medium set, and 1.9% on Hard set. Comparing to RetinaFace [26], our framework achieve better performance on all sets by a margin of 5.3%, 6.5% and 21.1%, respectively.

Table 4: The results on test set of the WIDER FACE dataset

Method	Easy Set	Medium Set	Hard Set
FaceBoxes [24]	0.839	0.763	0.396
FaceBoxes3.2× [25]	0.791	0.794	0.715
RetinaFace-mnet [26]	0.896	0.871	0.681
LFFD-v1 [25]	0.910	0.881	0.780
LFFD-v2	0.837	0.835	0.729
CenterFace [7]	0.932	0.921	0.873
CenterAttentionFaceNet (Ours)	**0.949**	**0.936**	**0.892**

5. CONCLUSIONS

In this paper, we have presented a method to improve the performance of the original CenterFace model. Specifically, we propose to utilize the CBAM attention modules after each convolution block in the CenterFace backbone. CBAM is used to enhance the features after each block of the CNN network to create feature maps of higher importance and reduce the repetition of unnecessary features. From there, it helps to increase the accuracy of the classification classes and reduce the error in predicting the position of the face. Experiment results have shown that CenterAttentionFaceNet works well compared to state-of-the-art methods. Moreover, due to still keeping the lightweight architecture e.g., MobileNet V3, our framework is easy to deploy on embedded and mobile devices that have speed and memory limits. In the future, we will continue to improve and combine them with several other methods to enhance the performance and improve the accuracy of the model even more. On the other hand, we will apply the model to other problems in the field of computer vision and image processing.

REFERENCES

[1] Voulodimos A, Doulamis N, Doulamis A, et al. Deep Learning for Computer Vision: A Brief Review. Comput Intell Neurosci. 2018;2018.

[2] Duc, Q. V., Phung, T., Nguyen, M., Nguyen, B. Y., & Nguyen, T. H. (2022). Self-knowledge Distillation: An Efficient Approach for Falling Detection. In Artificial Intelligence in Data and Big Data Processing: Proceedings of ICABDE 2021 (pp. 369-380). Cham: Springer International Publishing

[3] Vu, D. Q., Le, N. T., & Wang, J. C. (2022, August). (2+ 1) D Distilled ShuffleNet: A Lightweight Unsupervised Distillation Network for Human Action Recognition. In 2022 26th International Conference on Pattern Recognition (ICPR) (pp. 3197-3203). IEEE.

[4] Menon, S., Damian, A., Hu, S., Ravi, N., & Rudin, C. (2020). Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 2437-2445).

[5] Vu, D. Q., & Wang, J. C. (2021, December). A novel self-knowledge distillation approach with siamese representation learning for action recognition. In 2021 International Conference on Visual Communications and Image Processing (VCIP) (pp. 1-5). IEEE.

[6] Li H, Lin Z, Shen X, et al. A convolutional neural network cascade for face detection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2015;07-12-June:5325–5334.

[7] Xu Y, Yan W, Yang G, et al. CenterFace: Joint Face Detection and Alignment Using Face as Point. Sci Program. 2020;2020.

[8] Newell A, Huang Z, Deng J. Associative embedding: End-to-end learning for joint detection and grouping. Adv Neural Inf Process Syst. 2017;2017-Decem:2278–2288.

[9] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Process Lett. 2016;23:1499–1503.

[10] Gao SH, Cheng MM, Zhao K, et al. Res2Net: A New Multi-Scale Backbone Architecture. IEEE Trans Pattern Anal Mach Intell. 2021;43:652–662.

[11] Zhang S, Zhu X, Lei Z, et al. Single Shot Scale-invariant Face Detector Shifeng. :192–201.

[12] Shaifee MJ, Chywl B, Li F, et al. Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video. J Comput Vis Imaging Syst. 2017;3.

[13] Zhao Y, Han R, Rao Y. A new feature pyramid network for object detection. Proc - 2019 Int Conf Virtual Real Intell Syst ICVRIS 2019. 2019;428–431.

[14] Chi C, Zhang S, Xing J, et al. Selective refinement network for high performance face detection. 33rd AAAI Conf Artif Intell AAAI 2019, 31st Innov Appl Artif Intell Conf IAAI 2019 9th AAAI Symp Educ Adv Artif Intell EAAI 2019. 2019;8231–8238.

[15] Pathak BK, Srivastava S. Integrated ANN-HMH Approach for Nonlinear Time-Cost Tradeoff Problem. Int J Comput Intell Syst. 2014;7:456–471.

[16] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[17] Phung, T., Nguyen, V. T., Ma, T. H. T., & Duc, Q. V. (2022). A (2+ 1) D attention convolutional neural network for video prediction. In Artificial Intelligence in Data and Big Data Processing: Proceedings of ICABDE 2021 (pp. 395-406). Cham: Springer International Publishing.

[18] Zong B, Song Q, Min MR, et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. 6th Int Conf Learn Represent ICLR 2018 - Conf Track Proc. 2018;1–19.

[19] Woo S, Park J, Lee J, et al. Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.pdf. Eur. Conf. Comput. Vis. 2018.

[20] Howard A, Wang W, Chu G, et al. Searching for MobileNetV3 Accuracy vs MADDs vs model size. Int. Conf. Comput. Vis. 2019. p. 1314–1324.

[21] Hu J. Squeeze-and-Excitation_Networks_CVPR_2018_paper.pdf. Cvpr [Internet]. 2018;7132–7141. Available from: http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

[22] Garbin C, Zhu X, Marques O. Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimed Tools Appl. 2020;79:12777–12815.

[23] Ramchoun H, Amine M, Idrissi J, et al. Multilayer Perceptron: Architecture Optimization and Training. Int J Interact Multimed Artif Intell. 2016;4:26.

[24] Zhang S, Zhu X, Lei Z, et al. FaceBoxes: A CPU real-time face detector with high accuracy. IEEE Int Jt Conf Biometrics, IJCB 2017. 2018;2018-Janua:1–9.

[25] He Y, Xu D, Wu L, et al. LFFD: A Light and Fast Face Detector for Edge Devices. 2019; Available from: http://arxiv.org/abs/1904.10633.

[26] Li Q, Guo N, Ye X, et al. Video Face Recognition System: RetinaFace-Mnet-Faster and Secondary Search. Adv Intell Syst Comput. 2021;1364 AISC:625–636.