A VISUAL ATTENTION BASED VGG19 NETWORK FOR FACIAL EXPRESSION RECOGNITION

Nguyen Thi Thanh Tam*, **Nguyen Thi Tinh**†
* Posts and Telecommunications Institute of Technology
† School of Information and Communication technology - Thai Nguyen university

Abstract **-** Facial emotion recognition (FER) is meaningful for human-machine interaction such as clinical practice, playing games, and behavioral description. FER has been an active area of research over the past few decades, and it is still challenging due to the high intra-class variation, the heterogeneity of human faces, and variations in images such as different facial poses and various lighting conditions. Recently, deep learning models have shown great potential for FER. Besides, the visual attention technique has helped deep learning networks improve. In this paper, we present a visual attention-based VGG19 network for FER. The proposed outperforms the state-of-the-art methods slightly on the FER2013 dataset.

Keywords **-** Facial expression recognition, Deep learning, VGGnet, Attention.

I. INTRODUCTION

FER using artificial intelligence is one of the most popular research areas nowadays. It has attracted the research community significantly because it has a wide range of practical applications such as mental diseases diagnosis and human social/physiological interaction detection, human-machine interaction, and games. FER has many challenges, such as the high intra-class variation, the heterogeneity of human faces, and variations in images because of different facial poses and various lighting conditions.

Many studies have been carried on FER. The traditional methods have used handcrafted features with shallow learning ([1], [2]). Meanwhile, due to the increase of the amount of collected dataset and chip processing abilities (GPU) and well-designed network architecture, studies in various fields, including FER, have begun to transfer to deep learning methods. Deep learning has achieved state-of-the-art recognition accuracy ([3], [4], [5], [6]). Among deep neural network architectures, VGGnet [4] is one of the best backbones that archives state-of-the-art recognition accuracy on the FER 2013 dataset [7].

Besides, visual attention helps the performances of deep neural networks increase ([8], [9], [10]). The visual attention model lets the neural network focus its attention on the crucial part of the image where it can get most of the information while paying less attention elsewhere.

Inspired by the success of VGGnet and visual attention model, we proposed a visual attention based VGG19 network for FER. Our proposed network is an end-to-end model that does not require an additional classifier for the classification purpose at the end. Furthermore, the attention module enable to capture interesting features of facial expression during the training process in order to get better discriminate between classes. Our main contributions in this paper are:

- We propose a deep learning model which is a combination of the VGG19 and the attention module.
- We evaluate our proposed model on a public dataset FER 2013. The experimental results shows that our model outperforms the state-of-the-art methods slightly on the FER 2013 datatset.

The remaining of this paper is organized as follows: Section II briefly reviews the related works in FER. Section III describes the proposed method. In Section IV, we show and analyze experimental results. Conclusion and future works are presented in Section V.

II. RELATED WORKS

The traditional methods have used handcrafted features with shallow learning ([1], [2]). In this section, we focus on reviewing the related works in FER based on deep learning and some work related to visual attention.

A. Deep facial expression recognition

Deep learning has achieved state-of-the-art performance for a variety of applications as well as FER

Contact author: Nguyen Thi Thanh Tam
Email: nttam@ptit.edu.vn
Manuscript Received: 10/2021, revised: 11/2021, accepted: 11/2021.

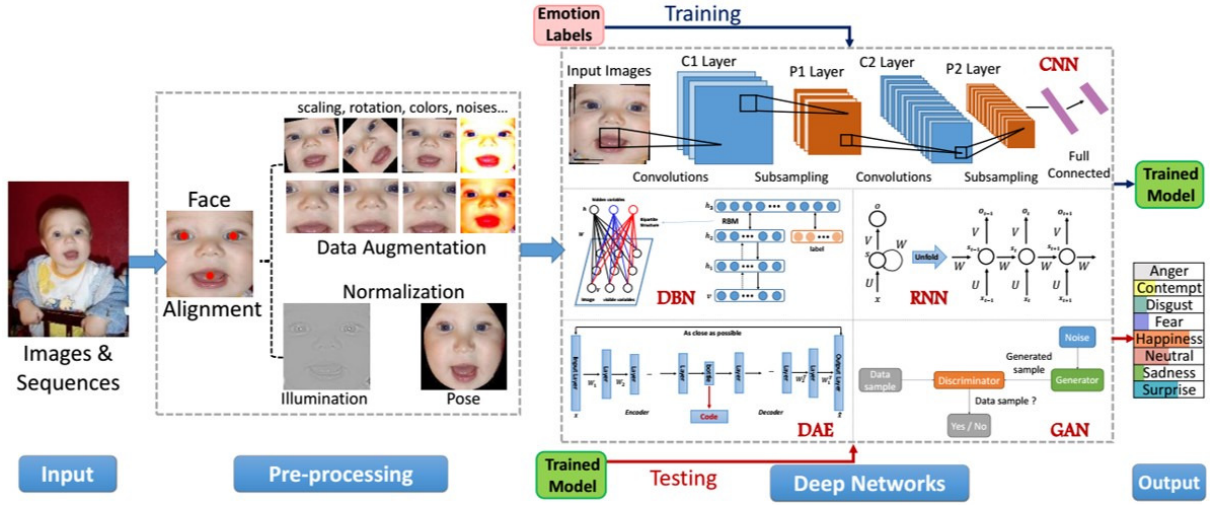


Fig. 1. The general pipeline of deep facial expression recognition systems [11]

[11]. Deep learning enables to capture of high-level features based on hierarchical architectures of multiple nonlinear transformations and representations. The traditional architectures of these deep neural networks are shown in Fig.1.

In 2002, Fasel Beat [12] indicate that the CNN is scale variations and robust to changes of face location. In 2015, Sun Bo et al. [13] extracted MSDF, LBP-TOP, HOG, LPQ-TOP, and acoustic features for emotion recognition. For each video frame, they extract MSDF, DCNN and RCNN features to represent the static facial expression. They use linear SVM for these features, then a fusion network at the decision level to combine all the extracted features.

Siqueira Henrique et al. [14] present experimental results on Ensembles with Shared Representations (ESRs) based on convolutional networks. The results show that ESRs reduce the remaining residual generalization error on the AffectNet and FER+ datasets and reach human-level performance on facial expression recognition in the wild using emotion and affect concepts.

Adrian Vulpe-Grigoraşi *et al.* [15] presented a method of optimizing the hyperparameters in order to increase the accuracy of a CNN model for FER. Their best model was trained and evaluated using the FER2013 database, obtaining an accuracy of 72.16%. While Christopher Pramerdorfer *et al.* overcame one existing bottleneck by employing modern deep CNNs leads to an improvement in FER2013 performance, obtaining an accuracy of 72.4%.

In 2021, Yousif Khaireddin *et al.* trained a VGGnet model on the FER-2013 dataset and got the state-of-the-art result with an accuracy of 73.28%.

B. Visual attention

When seeing a complex visual scene, humans do not tend to look at it in its entirety at once. Instead, they focus on a subset of the visual content to speed up the visual analysis process. Inspired by this phenomenon, the visual attention mechanism has become a hot topic in computer vision and deep learning. It is widely used in many computer vision tasks such as object segmentation, object recognition, human action recognition. Many CNNs and RNNs have achieved much better performance in various computer vision tasks than previous traditional methods. Recent progress in deep learning saw a close relation between deep learning and visual attention mechanism [16]. YifanCai *et al.* [17] presented a multi-task neural network called Temporal SonoEyeNet (TSEN) with a primary task to describe the visual navigation process of sonographers by learning to generate visual attention maps of ultrasound images around standard biometry planes of the fetal abdomen, head (trans-ventricular plane) and femur.

Yang Liu *et al.* [18] designed four visual attention blocks and embedded them in the existing CNNs model to give a mineral image classification models. Luca Cultrera *et al.* [19] proposed an learning-based agent using an attention model. The attention model enables us to understand what part of the image is important and better perform in a standard benchmark using the CARLA driving simulator.

Aleksandar Vakanski *et al.* [20] proposed an approach for integrating visual saliency into a deep learning model for breast tumor segmentation in ultrasound images. Visual saliency refers to image maps containing regions that are more likely to attract radiologists' visual attention. Their approach introduces attention blocks into a U-Net architecture and learns feature representations that prioritize spatial regions with high

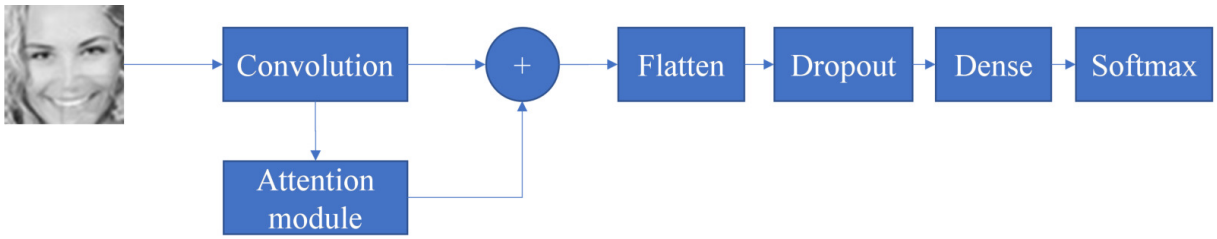


Fig. 2. Overall diagram of the proposed network

saliency levels. The experimental results showed that the salient attention layers made accuracy for tumor segmentation increase.

In 2019, Shervin Minaee *et al.* [21] proposed a deep learning approach based on an attentional convolutional network that can focus on important parts of the face and achieved significant improvement over previous models. However, their achievement is still limited. Despite the excellent performance of VGGnet and visual attention, there is still much room for improvement.

TABLE I
INFORMATIONS OF THE PROPOSED MODEL

Layer (type)	Output shape
VGG19 (Model)	7 x 7 x 512
Lamda layer (Avg pooling)	7 x 7 x 1
Lamda layer (Max pooling)	7 x 7 x 1
Concat layer	7 x 7 x 2
Conv layer	7 x 7 x 1
Concat layer	7 x 7 x 513
Flatten	25137
Dropout	25137
Dense	4096
Dense	7

III. VISUAL ATTENTION BASED VGG19 NETWORK FOR FACIAL EXPRESSION RECOGNITION

Our proposed network is based on VGG19, a pre-trained deep learning model, and an attention module. VGG19 was proposed in 2014 by the Visual Geometry Group (VGG) [4] which was the runner-up of the 2014 ILSVRC.

The VGG model has convolutional layers using the ReLU activation function. Following the activation function is a single max-pooling layer and several fully connected layers using the ReLU activation function. The final layer of the architecture is a Softmax layer for classification. The pre-trained model can classify images into 1000 object categories: mouse, keyboard, pencil, and many animals (Figure 3). Therefore, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224 x 224. We can use pre-trained weights trained on more than a million images from the ImageNet database [22].

Table I and Table II shows the informations of the proposed and VGG19 models, respectively.

In our work, a transfer learning technique is applied to the pre-trained weights of ImageNet. The proposed network consists of four main building blocks: convolution, visual attention, FC-layers, and softmax classifier. Fig. 2 shows the overall diagram of the proposed network.

Convolution module: In our method, we use the convolution module of the VGG19 model using filters with a small receptive field: 3x3 (which is the smallest size to capture the notion of left/right, up/down, center). The convolution stride is 1 pixel; the spatial padding of convolutional layer input is 1 pixel. The scale-invariant convolution module captures the interesting clues of the image. The features from other layers (higher or lower) are more general nor more specific.

Visual attention module: When seeing a complex visual scene, humans do not look at the entire visual scene at once. Instead, people focus on a subset of the visual content to speed up the visual analysis process. Inspired by this observation, the visual attention mechanism has become an interesting research topic in computer vision and deep learning. Compared to previous traditional methods, many CNNs and RNNs have achieved much better performance in various computer vision and natural language processing tasks.

In the proposed network, the attention module captures the spatial relationship of visual clues in the facial images extracted from the 5th pooling layer of the VGG19. We rely on the spatial attention concept in [23], using both max pooling and average pooling on the input tensor. These max pooled 2D tensors and average pooled 2D tensors are then concatenated to each other. After that, a convolution of filter size of 7x7 using Sigmoid function.

Fully connected layers: We use three flatten, dropout, and dense layers to convert the concatenated features into one-dimensional features, Fig. 2.

Softmax: The final layer is the soft-max layer. We use this softmax layer to classify the features extracted from the FC-layers. The soft-max layer performs seven classes of the FER2013 dataset and thus contains seven channels (one for each class).

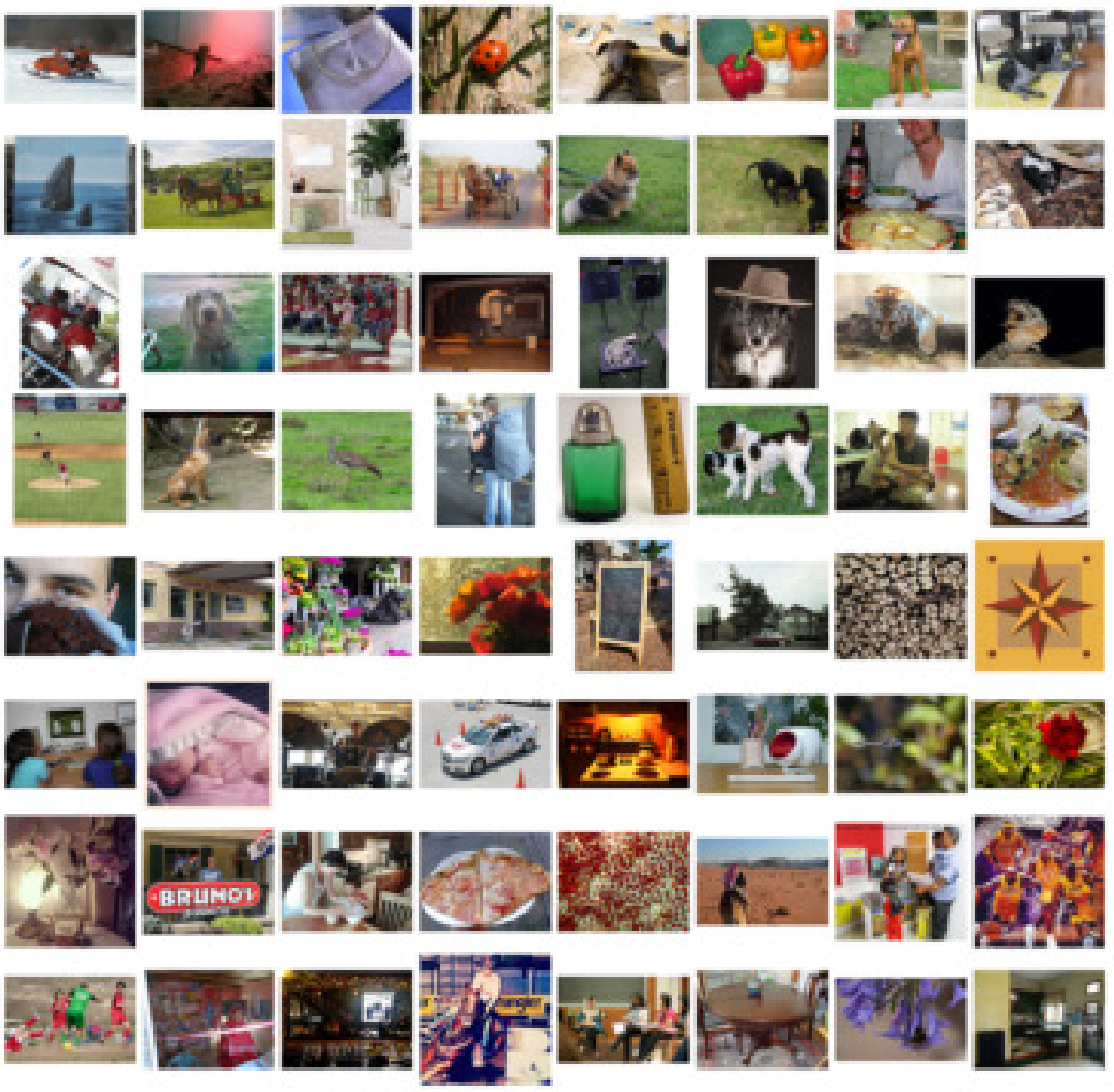


Fig. 3. Samples from ImageNet dataset

IV. EXPERIMENTS

We evaluate the proposed method on FER 2013 dataset [25] and compare it with the the-sate-of-the-art methods.

FER2013 is a commonly used benchmark dataset for facial expression recognition. It was first introduced at The 30th International Conference on Machine Learning (ICML) in Representation Learning [25]. This dataset contains 35,887 images of 48 × 48 resolution, most taken in wild settings. It is divided into the training set containing 28,709 images and the validation and test sets, including 3589 images each. The FER2013 dataset encompasses the difficult naturalistic conditions and challenges because this database has many variations in the images, including facial occlusion (mainly with a hand), partial faces, low-contrast images, and eyeglasses. There are seven categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Figure 4 shows sample images from the FER dataset. Images in the FER2013 dataset are stored in CSV format.

The experimental results is shown in Table III. The proposed method archives an accuracy of 73.35%, slightly higher than the state-of-the-art method [7]. Although using attention in VGG19 is a reasonable approach, the proposed model has only achieved nearly the same accuracy as the stat-of-the-art method. This result may be because the mechanism of the attention module and its location in the network architecture was not suitable.

There is some method tested on the FER-2013 dataset obtaining higher accuracy. However, these

TABLE II
INFORMATIONS OF THE VGG19 MODEL

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv4 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv4 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv4 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000

TABLE III
COMPARISON OF THE PROPOSED METHOD WITH OTHER METHODS ON THE FER 2013 DATASET.

Method	Accuracy (%)
DeepEmotion [21]	70.02
Inception [7]	71.6
CNN Hyperparameter Optimisation [15]	72.16
ResNet [24]	72.4
VGG [24]	72.7
VGGnet [7]	73.28
The proposed method (VGG19+Attention)	**73.35**

Fig. 4. Some example images from the FER2013 datatset

methods are trained on the FER-2013 dataset plus extra data. Therefore, we only compare the proposed method with models using only FER-2013.

V. CONCLUSION AND FUTURE WORKS

In this paper, we propose a visual attention-based VGG19 network for FER. The attention module helps

the performance of the proposed model improve slightly. We plan to research a more suitable attention mechanism in the future and evaluate the proposed method on other datasets.

ACCKNOWLEGDE

This research is funded by Posts and Telecommunications Institute of Technology under grant number 07-2021-HV-ĐPT-CN.

REFERENCES

[1] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[2] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

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

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on fer2013," *arXiv preprint arXiv:2105.03588*, 2021.

[8] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks," *arXiv preprint arXiv:1312.6082*, 2013.

[9] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[10] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[11] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[12] B. Fasel, "Head-pose invariant facial expression recognition using convolutional neural networks," in *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, 2002, pp. 529–534.

[13] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, and Q. Wei, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 497–502.

[14] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5800–5809.

[15] A. Vulpe-Grigoraşi and O. Grigore, "Convolutional neural network hyperparameters optimization for facial emotion recognition," in *2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*. IEEE, 2021, pp. 1–5.

[16] X. Liu and M. Milanova, "Visual attention in deep learning: a review," *Int Rob Auto J*, vol. 4, no. 3, pp. 154–155, 2018.

[17] Y. Cai, R. Droste, H. Sharma, P. Chatelain, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Spatio-temporal visual attention modelling of standard biometry plane-finding navigation," *Medical Image Analysis*, vol. 65, p. 101762, 2020.

[18] Y. Liu, Z. Zhang, X. Liu, W. Lei, and X. Xia, "Deep learning based mineral image classification combined with visual attention mechanism," *IEEE Access*, vol. 9, pp. 98 091–98 109, 2021.

[19] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. Del Bimbo, "Explaining autonomous driving by learning end-to-end visual attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 340–341.

[20] A. Vakanski, M. Xian, and P. E. Freer, "Attention-enriched deep learning model for breast tumor segmentation in ultrasound images," *Ultrasound in Medicine & Biology*, vol. 46, no. 10, pp. 2819–2833, 2020.

[21] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network. arxiv 2019," *arXiv preprint arXiv:1902.01019*.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

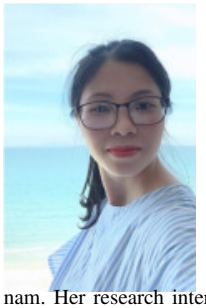
[24] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," *arXiv preprint arXiv:1612.02903*, 2016.

[25] I. Goodfellow, D. Erhan, P.-L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," 2013. [Online]. Available: http://arxiv.org/abs/1307.0414

MỘT PHƯƠNG PHÁP NHẬN DẠNG CẢM XÚC KHUÔN MẶT DỰA TRÊN MẠNG VGG19 VÀ VISUAL ATTENTION

Tóm tắt - Nhận dạng cảm xúc khuôn mặt (FER) có ý nghĩa đối với việc tương tác giữa người và máy như trong nhiều lĩnh vực khác nhau như y học, trò chơi điện tử. FER đã là chủ đề nghiên cứu được chú ý trong vài thập kỷ qua. Bài toán này vẫn còn nhiều thách thức do sự đa dạng trong một biểu cảm, sự không đồng nhất của khuôn mặt người và các biến thể trong hình ảnh như các tư thế khuôn mặt khác nhau và các điều kiện ánh sáng khác nhau. Gần đây, các mô hình học sâu đã cho thấy tiềm năng lớn đối với bài toán FER. Bên cạnh đó, kỹ thuật visual attention đã giúp cải thiện hiệu năng của mạng nơ ron học sâu. Trong bài báo này, chúng tôi trình bày một phương pháp nhận dạng cảm xúc khuôn mặt dựa trên mạng VGG19 và visual attention. Phương pháp đề xuất cho kết quả tốt hơn một chút so với các phương pháp khác trên tập dữ liệu FER2013.

Từ khoá - Nhận dạng cảm xúc khuôn mặt, Học sâu, VGGnet, Attention.



Nguyen Thi Thanh Tam received her bachelor's degree in Information Technology from Thai Nguyen University in 2004. In 2017, she received her degree of master of science in Information systems from the University of Engineering and Technology, Ha Noi National University. Tam has 11 years of teaching experience. Currently, she is a lecturer at the Faculty of Multimedia - The Posts and Telecommunications Institute of Technologies (PTIT), Hanoi, Vietnam. Her research interests include Machine Learning, Multimedia Application Development.

Nguyen Thi Tinh is an Information Technology lecturer at Thai Nguyen University of Information and Communication Technology (ICTU). She received her engineer's degree in Information Technology from the Faculty of Information Technology - Thai Nguyen University in 2008. Then, she received her Master of Science degree in Information Technology from Manuel S. Enverga University Foundation – Lucena City - the Philippines in 2010. Tinh's research interests are human activity recognition, artificial intelligence, and machine learning/deep learning.