EXPLAINABLE ARTIFICIAL INTELLIGENCE IN DECODING HUMAN EMOTIONS THROUGH VISION TRANSFORMERS

Bojan Gutić[1],
[0000-0002-1085-4718]
Timea Bezdan[1],
[0000-0001-6938-6974]
Hojjatollah Farahani[2],
[0000-0002-9799-7008]
Peter Watson[3],
[0000-0002-9436-0693]
Marina Marjanović[1]
[0000-0002-9928-6269]

[1]Singidunum University,
Belgrade, Serbia

[2]Tarbiat Modares University,
Tehran, Iran

[2]MRC Cognition and Brain Sciences Unit,
University of Cambridge, Cambridge,
United Kingdom

Correspondence:
Petar Stevović

e-mail:
petar.stevovic.23@singimail.rs

Abstract:
In artificial intelligence and psychological research, understanding how artificial intelligence interprets human emotions through facial expressions is challenging, thus emotion recognition became a crucial task in many computer vision applications, and in the development of emotionally intelligent artificial intelligence systems. Although there have been significant advancements in the field, with many deep learning models achieving high accuracy, there is still a gap in developing models that are both highly accurate and explainable. This is particularly true in aligning with human psychological processes for emotion recognition. Addressing this issue, our research explores the capabilities of vision transformers, specifically focusing on how these models might mimic human attention to key facial features, such as the eyes and mouth, in the context of emotion recognition. The experiments are conducted on the well-known KDEF dataset. In the proposed model, the attention maps are analyzed, aiming to uncover whether artificial intelligence can replicate human-like processing in interpreting emotions. The findings reveal that the model's attention aligns with the psychologically significant facial regions, suggesting a level of human-like processing. Additionally, the model's performance is proven by a notable test accuracy of 95%. This research makes a significant contribution to the body of knowledge in both artificial intelligence and psychological domains by demonstrating the potential of vision transformers in accurately interpreting human emotions through facial expressions.

Keywords:
Artificial Intelligence, Computer Vision, Vision Transformers, Emotion Recognition, Explainable AI.

INTRODUCTION

Emotion recognition (ER) is a key area of study in artificial intelligence (AI). ER aims to identify human emotions from various data sources such as facial expressions, voice, body movements, and physiological signals. Accurate emotion recognition has high importance in numerous fields including healthcare, customer service, social media, and others. Despite advances, the task of understanding complex emotional states from facial expressions remains challenging, and traditional methods often struggle to recognize small human expressions within the images [1], [2], which highlights the need for more sophisticated approaches.

DOI: 10.15308/Sinteza-2024-168-174

In the era of artificial intelligence (AI), where algorithm utilization has significant influence, the critical need to understand and interpret their decision-making processes in different tasks, highlights the importance of explainable AI (XAI) [3]. Among the diverse applications of AI across various domains, emotion recognition stands out with significant potential to enhance human-computer interaction [4]. The introduction of Vision Transformers (ViT) revolutionized the field of computer vision [5]. Transformer model architectures, initially used in natural language processing (NLP), have been successfully adapted for visual tasks, and their adoption has disrupted traditional convolutional neural networks (CNNs) by taking advantages of self-attention mechanisms [6]. ViTs not only achieve superior performance compared to CNNs, but also require significantly fewer computational resources. Despite these advancements, the increased complexity and abstraction in models like ViTs present new challenges in understanding how decisions are made, this necessitates the integration of XAI techniques, which aim to make AI systems transparent and understandable to human users and to create a bridge between the black-box models and human understanding.

This study explores the potential of ViTs in recognizing human emotions. By conducting experiments using the ViT model on the Karolinska Directed Emotional Faces (KDEF) dataset [7], this paper provides insights into the effectiveness of ViTs for emotion recognition and their advantage in processing visual data associated with human facial expressions and getting insight into the model's decision by explainable AI.

The rest of the paper is organized as follows: Section 2 describes the methodology; Section 3 discusses the experimental setup and the dataset used and presents the experimental results along with their analysis. The conclusion summarizes our findings.

2. METHODOLOGY

ViT represents a novel application of transformer architectures, which were traditionally used NLP [8]. In classical transformers, the inputs are tokens (words) that are transformed into embeddings. For images, ViT adapts this approach by using fixed-size patches and transforms images into sequences of these patches. Unlike traditional CNNs [9] that process the entire image, ViT treats images as sequences of smaller patches, enabling the model to capture relationships between different parts of an image.

These patches are flattened into one-dimensional vectors and then transformed through a linear projection, which acts as a simple neural network layer. This produces the embeddings. Additionally, after the linear projection, positional information is added to each patch embedding to indicate its original location within the input image; this is known as positional embedding. The sequence of embeddings is then passed into the transformer encoder. Within the encoder, the embeddings undergo multiple layers of processing, including multi-head self-attention and feed-forward neural networks. The output of the transformer encoder can then be pooled and passed through a multilayer perceptron (MLP), which is used to predict the class of the image [10].

Our study uses ViT model for emotion recognition. The ViT model begins with an embedding layer, where image patches are extracted using a convolutional operation. The embedding uses convolutional layer with 768 filters and a filter size of 16x16 pixels, striding over the image to produce flattened image tokens. These tokens are then projected into a 768-dimensional space. The encoder consists of 12 transformer layers. Each layer comprises a self-attention mechanism, an intermediate layer, an output layer, and layer normalization. Self-attention mechanism allows the model to weigh the importance of different patches when processing an image. The intermediate layer is a dense layer, which expands the dimensionality from 768 to 3072 with a GELU activation function [11], enhancing the model's ability to learn complex features. The output layer, another linear transformation maps the dimensions back from 3072 to 768, followed by normalization to stabilize the learning process. Each transformer layer is preceded and followed by layer normalization, ensuring that activations are normalized, which helps in stabilizing the learning. The model architecture is presented in Listing 1.

```
<ViTModel(
  (embeddings): ViTEmbeddings(
    (patch_embeddings): ViTPatchEmbeddings(
      (projection): Conv2d(3, 768, kernel_size=(16, 16), stride=(16, 16))
    )
  )
  (encoder): ViTEncoder(
    (layer): ModuleList(
      (0-11): 12 x ViTLayer(
        (attention): ViTAttention(
          (attention): ViTSelfAttention(
            (query): Linear(in_features=768, out_features=768, bias=True)
            (key): Linear(in_features=768, out_features=768, bias=True)
            (value): Linear(in_features=768, out_features=768, bias=True)
          )
          (output): ViTSelfOutput(
            (dense): Linear(in_features=768, out_features=768, bias=True)
          )
        )
        (intermediate): ViTIntermediate(
          (dense): Linear(in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation()
        )
        (output): ViTOutput(
          (dense): Linear(in_features=3072, out_features=768, bias=True)        )
        (layernorm_before): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      )
    )
  )
  (layernorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
)>
```

Listing 1. Model Architecture.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The experiment in this study is designed to explore the potential of ViTs in ER. This paper demonstrates how these deep learning models can effectively mimic the way humans focus on facial features like the eyes and mouth, essential to processing emotions. The experiments were conducted using PyTorch [12] with dual NVIDIA T4 GPUs.

3.1. EXPERIMENTAL SETUP AND DATASET

The ViT model utilized in the experiments is the 'google/vit-base-patch16-224-in21k' [13] model with 16x16 patch size for 224x224 input images, pre-trained on the ImageNet-21k [14] dataset. To adapt the model to emotion recognition, we fine-tuned the pre-trained ViT model on KDEF [7] dataset. The parameter configuration for the experimental study is described in Table 1.

Table 1. Parameter configuration.

Parameter	Value
Batch size	16
Learning rate	1e-4
Wurmup ratio	0.1
Weight decay	0.01
Epochs	50

Additionally, the learning rate scheduled is used to decrease the learning rate linearly during the training process.

The KDEF dataset is developed by the Department of Clinical Neuroscience at Karolinska Institute. It consists of facial expressions captured from multiple angles. The dataset includes images of 70 individuals (35 male and 35 female) representing seven different emotional states: happiness, sadness, anger, fear, surprise, disgust, and a neutral. Samples from each emotional states are depicted in Figure 1.

Figure 1. Sample images from KDEF dataset.

The dataset was divided into training (70%), validation (15%), and testing (15%) sets. We implemented a series of preprocessing steps, such as image cropping, resizing, and image normalization to optimize the dataset for the ViT model. The training was conducted over 50 epochs, similarly as in [15] where the authors used the same dataset and conducted experiments with other deep learning techniques, to evaluate the model, we focused on key metrics, such as accuracy, precision, recall, and f1-score, then the obtained results are compared to the state-of-the-art methods presented in [15].

3.2. EXPERIMENTAL RESULTS AND DISCUSSION

This subsection describes the results of the conducted experiments using the ViT model on the KDEF dataset. The performance of the ViT model in terms of accuracy, precision, recall, and f1-score compared against the established state-of-the-art methods, the results of VGG [16] and EFL-LCNN is taken from [15]. The comparative analysis is presented in Table 2.

Table 2. Comparative analysis.

Metric	VGG16	EFL-LCNN	ViT
Accuracy	91%	93%	95%
Precision		92%	95%
Recall		92%	95%
F1-score		93%	95%

The overall accuracy of the model is 0.95, that highlights the potential of ViTs in ER tasks. The macro and weighted averages for precision, recall, and f1-score all align at 0.95, indicating consistent performance across different emotions.

In the experiment, the ViT model demonstrated high performance in emotion recognition. For anger, it showed a strong ability with a precision of 0.92 and recall of 0.94. The model excelled in recognizing disgust, achieving both precision and recall at 0.96. Fear presented a slight challenge for the model, with a lower precision of 0.88, which might point to difficulties in differentiating it from similar expressions; however, the high recall of 0.92 indicated a sensitivity to this emotion. Exceptional performance was observed in detecting happiness, with 1.0 precision and recall. The model was also effective in identifying neutral expressions, crucial for differentiating emotional from non-emotional states, achieving 0.98 in both precision and recall. The recognition of sadness, with a lower recall of 0.86, suggested some challenges in identifying this emotion. Finally, the model showed a balanced capability in recognizing surprise, with both precision and recall at 0.95. The confusion matrix of the emotional states is depicted in Figure 3.

Figure 2 illustrates the training and validation loss over the course of epochs, as well as the validation accuracy. Decreasing trend in both, training and validation loss, indicates to effective learning of the model and shows that the model generalizes to the validation data, the upward trend in the validation accuracy indicates to the improvement of the performance of the model on unseen data.

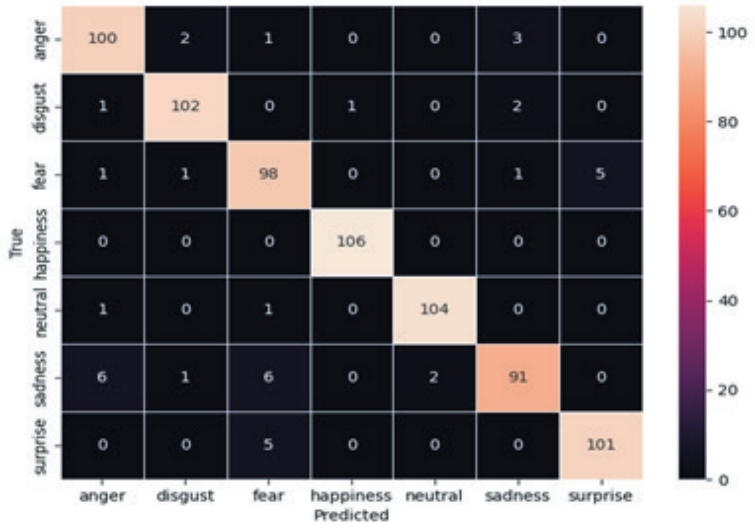

Figure 2. Confusion Matrix.

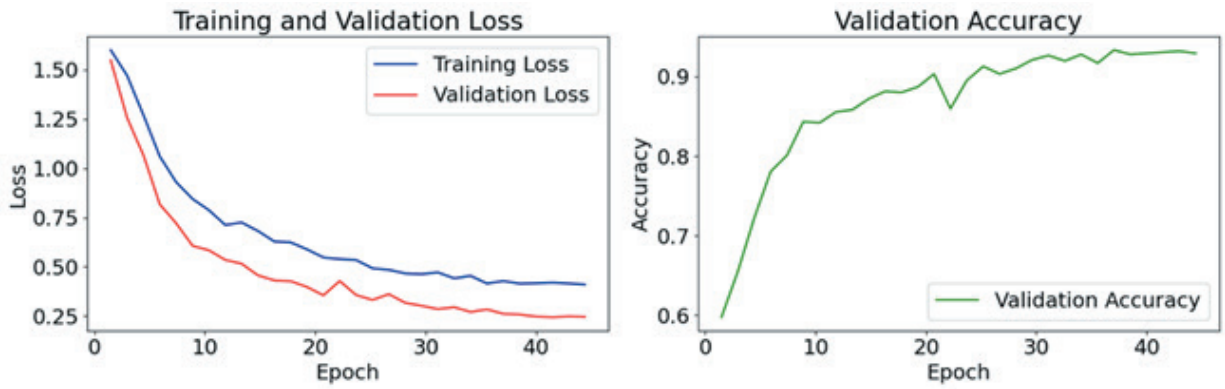

Figure 3. The first plot (on the left) shows both the training loss and the validation loss across epochs. The second plot (on the right) displays the validation accuracy over epochs.

We conducted an in-depth analysis of the attention maps generated by the ViT. This analysis was crucial in understanding how the model focuses on specific facial features, such as the eyes and mouth, essential for emotion recognition. Samples of the attention maps are depicted in Figure 4. The attention maps serve as indicators of the model's focus areas when identifying different emotions, providing a layer of explainability, which is crucial for undersendinf the decisions and actions taken by ViT model. The following samples illustrate the attention given by the model to each emotional state: i) For fear, the model concentrates on regions typically involved in expressing fear, including the eyebrows, eyes, and mouth. ii) In the case of anger, the heat map reveals the model's primary focus around the eyebrow and eye regions, with notable attention on the mouth area as well. These regions are aligned with key facial muscles as identified by the Facial Action Coding System (FACS) [17] for the expression of anger. iii) With surprise, highlighting areas around the eye suggests that the model views wide-open eyes as important feature of the emotion surprise, which aligns with FACS. iv) Regarding sadness, intensified coloration around the cheek and nose might signal the model's association of these areas with sadness. FACS typically points to the eyebrows and mouth for cues of sadness. v) For neutral expressions, the model's attention on the eyes could indicate their use as a reference point to confirm the absence of emotional expression. vi) As for happiness, the heat map indicates that the model is detecting a smile as important sign of happiness, which is in line with the universal interpretation of smiles as an indicator of happiness.

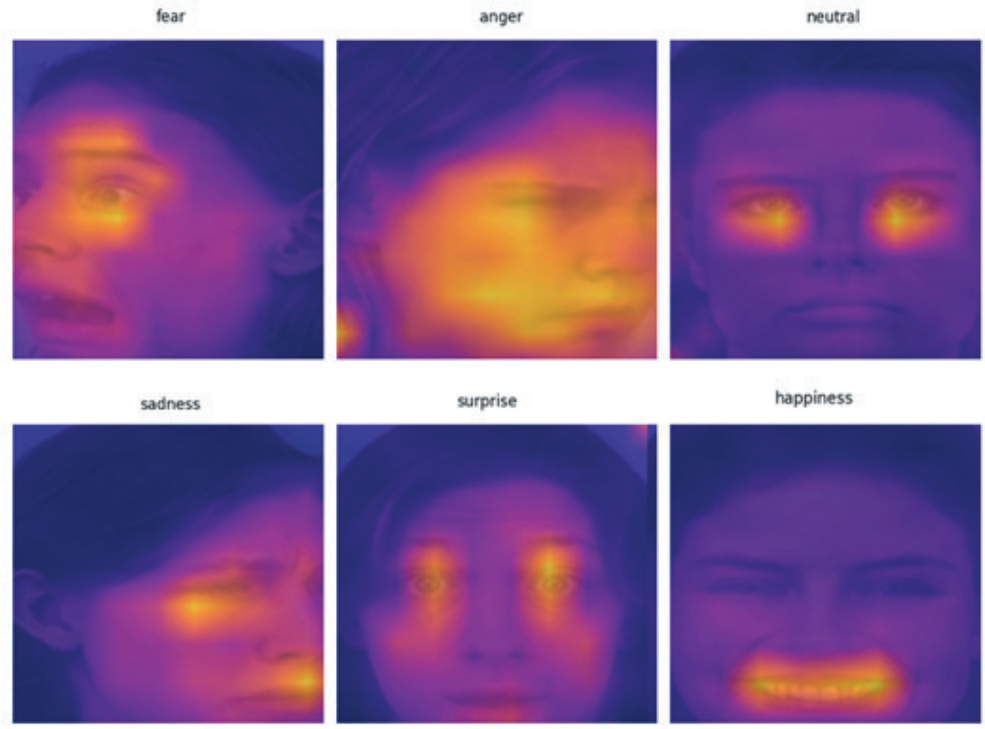

Figure 4. Heatmap visualization for the model's attention for each emotional state.

4. CONCLUSION

This work has taken a significant step forward in addressing the complex challenge of emotion recognition within the field of artificial intelligence and psychological research. By leveraging the advanced capabilities of vision transformers, we have moved closer to developing AI-based systems that not only achieve high accuracy but also offer a layer of explainability that human emotional cognition. The detailed analysis of attention maps within these models provided evidence that AI can mirror the human to focus on crucial facial features for ER. The alignment of the model's attention with significant facial regions, and high test accuracy of 95%, underscores the promise and robustness of vision transformers in recognizint emotions. Our findings not only augment the current understanding of the potential of AI in emotion recognition but also lay the groundwork for future innovations in creating empathetic and emotionally intelligent artificial systems that can interact with humans more naturally and intuitively. While the results are promising, future work could focus on further improving the model's sensitivity to emotions like sadness and optimization of the model to reduce more the gap between the training and validation loss.

5. REFERENCES

[1] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 443–449.

[2] J. Li et al., "Facial expression recognition with faster R-CNN," *Procedia Comput. Sci.*, vol. 107, pp. 135–140, 2017.

[3] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[4] R. Cowie et al., "Emotion recognition in human-computer interaction," *IEEE Signal Process*. Mag., vol. 18, no. 1, pp. 32–80, 2001.

[5] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv Prepr. ArXiv201011929*, 2020.

[6] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst*., vol. 30, 2017.

[7] "Karolinska directed emotional faces," *PsycTESTS Dataset*, vol. 91, p. 630, 1998.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv Prepr. ArXiv181004805*, 2018.

[9] T. Bezdan and N. B. Džakula, "Convolutional neural network layers and architectures," in *International scientific conference on information technology and data related research*, Singidunum University Belgrade, Serbia, 2019, pp. 445–451.

[10] N. Park and S. Kim, "How do vision transformers work?," *ArXiv Prepr. ArXiv220206709*, 2022.

[11] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *ArXiv Prepr. ArXiv160608415*, 2016.

[12] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[13] B. Wu et al., "Visual transformers: Token-based image representation and processing for computer vision," *ArXiv Prepr. ArXiv200603677*, 2020.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," presented at the 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[15] S. Bellamkonda and L. Settipalli, "EFL-LCNN: Enhanced face localization augmented light convolutional neural network for human emotion recognition," *Multimed. Tools Appl.*, vol. 83, no. 4, pp. 12089–12110, Jan. 2024, doi: 10.1007/s11042-023-15899-5.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv Prepr. ArXiv14091556*, 2014.

[17] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.