POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Attention mechanism and Visual Transformer models for Facial Emotion Recognition in natural settings

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Giacomo Da Re

Advisor: Prof. Andrea Bonarini

Academic year: 2022-2023

1. Introduction

Emotions serve as foundation of human communication, transcending language and culture with their inherent spontaneity. In today's rapidly evolving world, where artificial intelligence and robotics are becoming integral to our daily experiences, there is a growing need for machines that can accurately interpret and respond to human emotions, particularly those conveyed through facial expressions.
Current intelligent systems have begun to incorporate facial analysis, analyzing traits beyond emotions, such as age, gender and ethnicity. However, the real potential of emotion recognition lies in its ability to facilitate seamless interactions in various fields, from health care to entertainment, making it a key element in the advancement of human-computer interaction.
Despite advances in the field, reliably recognizing emotions in uncontrolled environments poses a tough and unsurprising challenge. Traditional computer vision techniques struggle to adapt to the myriad expressions and environmental conditions encountered in the real world. This study aims to address these challenges by employing the attention mechanism and, more specifically, its use in computer vision through the Visual Transformer, an innovative approach that allows focusing on certain parts of the image (in our case, the face) to enable more accurate classification of emotion and its nuances.
The research explores various basic models, culminating in the development of a two-stream solution that analyzes images in two distinct ways: one that extracts general features of the input and another that focuses on fixed points of the face for more stable and accurate classification. This approach promises to create machines that can not only recognize but also react appropriately to human emotions in a natural and intuitive way, paving the way for more intelligent and consistent human-machine interaction in the future.

2. State of the art

2.1. Emotion Model

The emotional model adopted in this work is discrete in nature, anchoring itself on the foundations of Ekman's Discrete Model [4]. Ekman's framework identifies six core emotions — anger, disgust, fear, happiness, sadness, and surprise — as universally experienced, each linked to distinct facial expressions and innate reactions. For the purposes of this study, and to accommodate the requirements of the implemented technology, the model has been expanded to in-

clude 'neutral' and 'contempt' as additional basic emotional states, thereby tailoring the universal model to the nuanced demands of emotion recognition in computational applications.

2.2. Vision Transformer (ViT)

The attention mechanism, first introduced by Vaswani et al. in [9], has significantly impacted deep learning, enabling models to focus on various parts of input data based on relevance. This mechanism calculates a weighted sum of values, determined by attention scores as reported in Formula 1, enhancing the model's ability to selectively prioritize information. This approach has been adapted to Computer Vision (CV) through Vision Transformers (ViTs) [3]. Its architecture is presented in Figure 1.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

ViTs differ from Convolutional Neural Networks (CNNs) mainly in their approach to processing images. Rather than working directly on pixels, ViTs divide an image into fixed-size patches, linearly embedding these into vectors. The sequence of vectors, analogous to a sentence in NLP, is processed through a Transformer encoder. This method allows ViTs to catch the global context of images, offering an enhanced understanding compared to the local feature emphasis of CNNs.

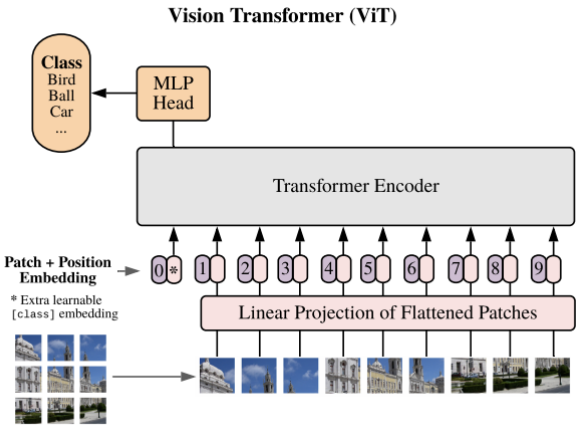

Figure 1: Vision transformer architecture proposed by Dosovitskiy et al. in [3].

One of the main advantages of ViTs is their global context awareness, unlike CNNs that concentrate on local features. ViTs scale efficiently with larger datasets and lack a fixed architectural pattern, offering flexibility for various tasks. They have shown promise in multiple CV tasks [3].

2.3. POSTER++

The POSTER++ design, proposed by Mao et al. [7], integrates a facial landmark detector and an image-bearing structure. A key aspect of POSTER++ is the window-based cross-attention mechanism designed for efficient linear computation. In this system, image features are divided into non-overlapping windows aligned with landmark features for efficient integration. This leads to the development of the window-based multi-headed cross-attention mechanism (W-MCSA). These features are then combined using a pre-trained Vision Transformer. This process uses attention mechanisms to handle long-range dependencies at various scales.

3. Method and materials

3.1. AffectNet

The AffectNet dataset [8], serving as the foundation for this work, categorizes emotions into eleven discrete categories, including Neutral, Happiness, Sadness, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face. However, this study focuses on the first eight of these categories. A key feature of AffectNet is its scale and diversity, encompassing a wide range of emotions across different ethnic backgrounds and age groups, making it one of the largest manually annotated databases for facial expression recognition in in-the-wild settings.

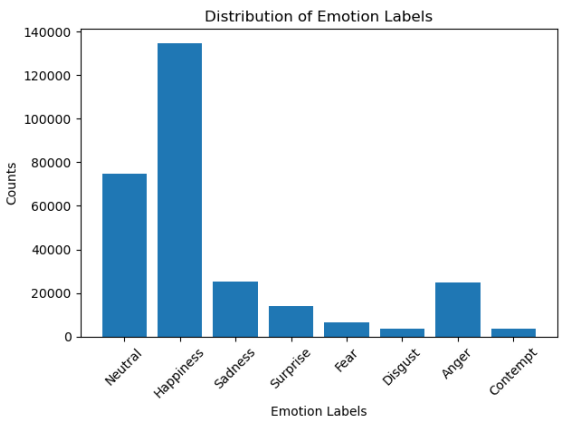

Figure 2: Class distribution in AffectNet.

An analysis of the dataset reveals a significant imbalance in the number of images available for each emotion, as shown in Figure 2. To address this imbalance and enhance the model's learning process, the study employs weighted random sampling and extensive data augmentation techniques to generalize the input data.

3.2. Proposed Solutions

3.2.1 ViT

The initial experiment we performed adopted a basic Vision Transformer (ViT) model without pre-training, which resulted in subpar accuracy. This prompted for subsequent experiments with both basic and large ViT models, as detailed in Table 3, with the inclusion of pre-training on ImageNet1K to evaluate the potential advantages of transfer learning.

Model	Layers	Hidden size	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M

Figure 3: Comparison of ViT-Base and ViT-Large model configurations as described in [3].

The culmination of these efforts resulted in a significant improvement in model performance. However, despite the increased capabilities offered by pre-training, the results stabilized, with the maximum accuracy reaching only 0.3865 on validation set. The model reaches full accuracy saturation around the 12th epoch, as shown in Figure 4.

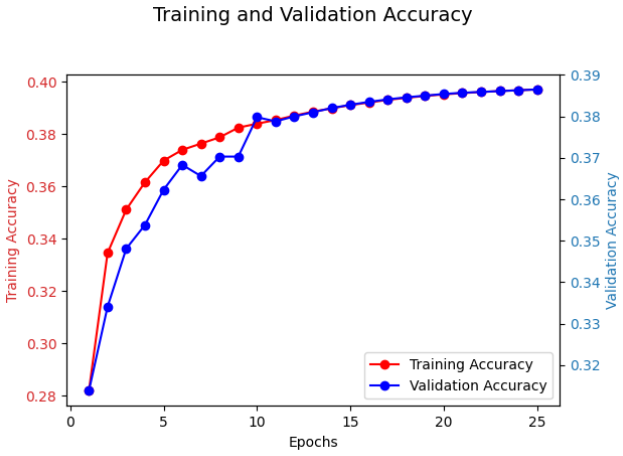

Figure 4: Training and Validation Accuracy over 25 epochs for the ViT-Large model with pre-training on ImageNet1k.

This result indicates that although the pre-trained ViT models (particularly the large ones) possess an initial advantage, achieving substantial and consistent performance in facial emotion recognition remains is still a long way off.

3.2.2 Swin ViT

Swin Transformer [5] differs from other vision transformers because of its characteristic "shifted window" mechanism, which, compared with the traditional Vision Transformer (ViT) model, allows Swin Transformer to handle visual data with greater skill, being highly efficient and scalable:
As explained in [5], Ω(MSA) represents the computational complexity of a global Multi-head Self-Attention (MSA) module, which is quadratic with respect to the total number of patches hw. Ω(W-MSA) denotes the complexity of a window-based MSA module, which is linear with respect to hw when the window size M is fixed.

$$\begin{aligned} \Omega(\text{MSA}) &= 4hwC^2 + 2(hw)^2C, \\ \Omega(\text{W-MSA}) &= 4hwC^2 + 2M^2hwC, \end{aligned} \tag{2}$$

By generating hierarchical feature maps that fit various image scales and resolutions, Swin Transformer holds promise for high-resolution applications. It is precisely because of this last feature that Swin Transformer was thought of for the FER task that this paper deals with. It would allow more attention to be placed on parts of the face and microexpressions that classical ViT could not have captured.
Consider Figure 5. Disregarding predictions, however, even with pretraining on the ImageNet1K dataset, the Swin Transformer model shows a peak accuracy of about 0.3416 at the twelfth epoch, slightly worse than the performance of the large-scale ViT models in the previous experiment.

3.2.3 Two Stream

The proposed architecture, illustrated in Figure 6, shows a dual flow approach applicable to FER tasks.
The Two Stream model integrates facial landmark detection with direct image analysis through two dedicated paths. By using MobileFaceNet [2] for the detection of reference points,

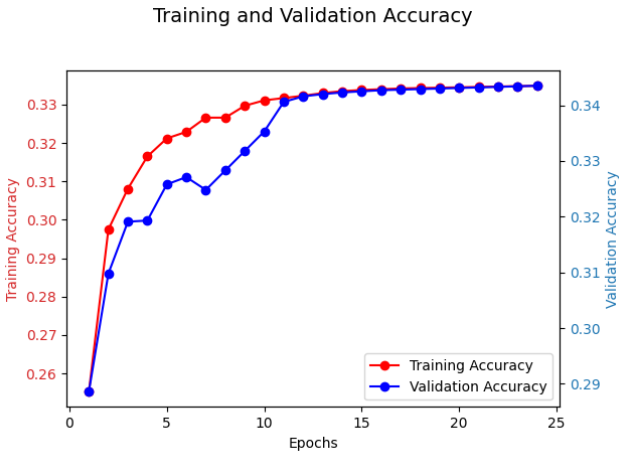

Figure 5: Training and Validation Accuracy over 25 epochs for the Swin Transformer model with pre-training on ImageNet1k.

the precise identification of key facial points is ensured. This phase is crucial to sharpen the focus on the essential areas to interpret facial expressions.

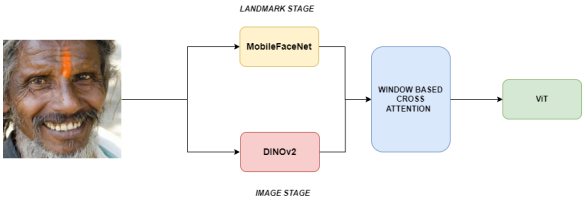

Figure 6: Two stream architecture.

To complement this, the DINOv2 network [1] replaces traditional convolutional backbones, bringing the power of attention mechanisms to the forefront. Trained across various Vision Transformer (ViT) configurations, the models harness the extensive LVD-142M dataset [1], comprising 142 million diverse images, to refine their pre-training effectiveness and bolster their performance in computer vision tasks. Notably, the self-supervised learning methodology of DINOv2 has shown promise in capturing robust and versatile visual features from a broad image spectrum without reliance on labeled data. The learning capability of DINOv2 is particularly advantageous, offering a robust foundation for machine learning and computer vision advancements without the constraints of annotated datasets.
At the confluence of these two streams, window-based cross-focus modules combine the distinct insights of each path. This convergence is directed towards a Vision Transformer (Vit) phase, which combines the different features in a unified and informed output, potentially setting a new standard for facial recognition and emotion detection systems.
The first experiment with the Two Stream Model without the Landmark Stage showed strong volatility in validation accuracy, peaking at 0.5501. This instability, as shown in the figures, suggested a model that could struggle with consistent performance in different real-world applications.
The complete Two Stream Model was then enhanced with pre-training on DINOv2, utilizing the CelebA dataset [6] known for its extensive collection of celebrity faces. Contrary to expectations, this additional pre-training did not yield the anticipated improvement in performance. This outcome challenges the widely held belief that pre-training invariably confers benefits. With a peak validation accuracy of 0.5554, the results exhibited considerable volatility. This observation raises questions about the direct applicability of transfer learning for facial expression recognition tasks, suggesting that more targeted pre-training strategies or a more suitable pre-training dataset may be required.

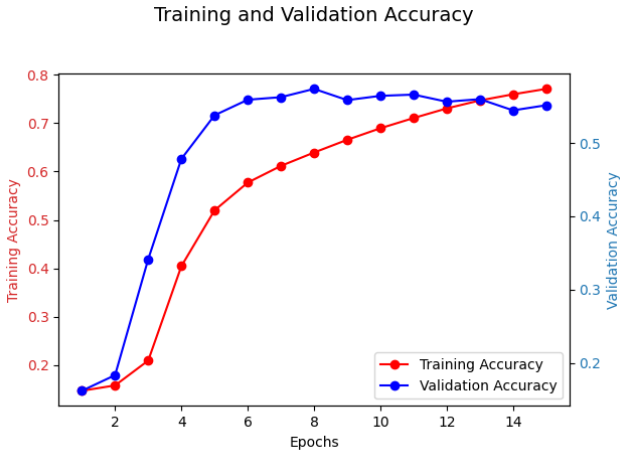

Figure 7: Training and validation accuracy trends for the Two Stream Model without pre-training over 15 epochs.

The Two Stream Model's full configuration, inclusive of the Landmark Stage but without any DINOv2 pre-training, presented the most favorable results among the series of experiments conducted. As depicted in Figure 7, this version of

the model achieved a peak validation accuracy of 0.5739, indicating a more robust performance compared to prior experiments. Figure 8 shows the confusion matrix of the performance of the model, useful to delineate the frequency of each emotion being predicted compared to the true labels. Diagonal cell's high value indicates that there are a lot of correct predictions for 'Happiness' in the matrix.
The model is struggling to differentiate similarities in expression patterns between 'Anger' and 'Disgust', which indicate a noticeable confusion between certain emotions. Despite an initial rapid improvement in training accuracy, which eventually plateaued, the model demonstrated a discernible improvement in generalizing to unseen data, although fluctuations in validation accuracy suggested there was still room for refinement to address overfitting concerns.

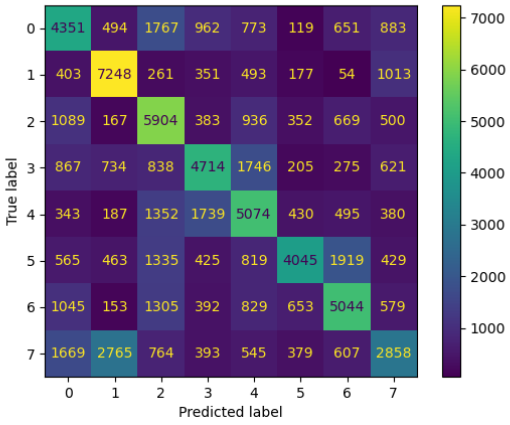

Figure 8: Confusion Matrix for the Two Stream Model without pre-training, showing the classification performance for emotions labeled from 0 to 7 corresponding to 'Neutral', 'Happiness', 'Sadness', 'Surprise', 'Fear', 'Disgust', 'Anger', and 'Contempt', respectively.

The Two Stream Model's evaluation highlights challenges in facial expression recognition (FER), particularly in generalization. Pre-training with DINO did not yield expected improvements, suggesting that FER may require more refined pre-training approaches or datasets. Incorporating the Landmark Stage proved crucial for stability, while the absence of pre-training led to potential overfitting. State-of-the-art models show that achieving high accuracy in FER remains challenging, indicating the need for continued innovation in model development and training methodologies.

4. Conclusions

This thesis advances the field of facial emotion recognition (FER) using neural networks, with a focus on the attention mechanism and the computer vision models related to it. It explores the effectiveness and challenges of various architecture, proposing the use of a two stream model to be expolited in FER tasks. Despite achieving good performance, the research highlights issues like overfitting and the need for better data diversity.
The Two Stream model, designed for a more general approach to image input, shows promise in capturing emotional nuances. The study emphasizes the potential of attention models in FER and suggests future research directions, including enhancing dataset quality, exploring more pre-training, and integrating different types of facial analysis.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[2] Cunjian Chen. PyTorch Face Landmark: A fast and accurate facial landmark detector, 2021. Open-source software available at *https://github.com/cunjia/pytorch_face_landmark*.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*, volume 19, pages 207–283, 1971.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and

Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[7] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang. Poster v2: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*, 2023.

[8] A. Mollahosseini, B. Hasani, and M.H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10:18–31, 2019.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, łLukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.