EmoticDetR : Emotion Detection Framework Leveraging Contextual and Facial Cues with Transformer Architecture

2020230057 LeeSeohyun

1. Introduction

1.1. Motivation

Recent advances in emotion recognition have highlighted the need for models that can understand emotions in a context-rich environment. Traditional approaches, focusing solely on facial expressions or body language, often fail to capture the full spectrum of human emotions, especially in dynamic or crowded scenes. This project is motivated by the challenge of integrating contextual information to enhance the accuracy and reliability of emotion detection.

1.2. Problem Definition

This paper address the problem of emotion recognition in complex environments, where the context, including the surrounding environment and body language, plays a crucial role. The challenge lies in effectively fusing these multimodal inputs to accurately interpret emotional states.

1.3. Contribution

The main contribution is the development of an emotion recognition model that combines the strengths of transformer-based architectures with context-aware learning. The model uniquely integrates features from body postures, facial expressions, and scene context, employing an enhanced transformer mechanism with Low-Rank Adaptation (LoRA) layers for improved interpretability and performance.

2. Methods

3. Significance and Novelty

3.1. Significance

The EmoticDetModel represents an advancement in emotion recognition, offering enhanced capabilities for understanding complex emotions in real-world scenarios. Its hybrid architecture, combining ResNet-18 for individual facial feature processing and the encoder part of DETR for contextual analysis, allows for efficient and comprehensive emotion detection. The integration of scenario data through BERT embeddings prior to Transformer processing adds a depth to the model's understanding of context, further enhancing its emotion recognition capabilities.

3.2. Novelty

The novelty of the EmoticDetModel lies in several key aspects:

A hybrid approach combining ResNet-18 with DETR's encoder, enabling the model to efficiently capture both intricate facial features and the broader context within images.

The integration of scenario data processed by BERT with image features before Transformer processing, enriching the model's input for a more nuanced understanding of emotional context.

The model's output is concatenated with CLIP features, post transformer, creating a comprehensive representation of emotional cues.

3.3. Main Challenges and Solutions

Developing an advanced emotion detection system using multimodal data presented several significant challenges.

Challenge 1: Efficient Integration of Multimodal Data.

Solution: Implemented a data processing pipeline using BERT for text and a custom loader for diverse data integration.

Challenge 2: Balancing Computational Complexity.

Solution: Utilized only DETR's encoder and incorporated LoRA in Transformer for efficient feature extraction.

Challenge 3: Optimized Feature Extraction from Facial Expressions.

Solution: Developed a custom decoder focused on facial expressions for accurate emotion detection.

Challenge 4: Comprehensive Scenario Processing.

Solution: Employed BLIP and GPT-3 for scenario generation, combining with DETR objects and CLIP features for rich contextual understanding.

Challenge 5: Enhancing Feature Representation Post-Transformer.

Solution: Concatenated Transformer output with CLIP features for a comprehensive representation of emotions.

3.4. Main Figure

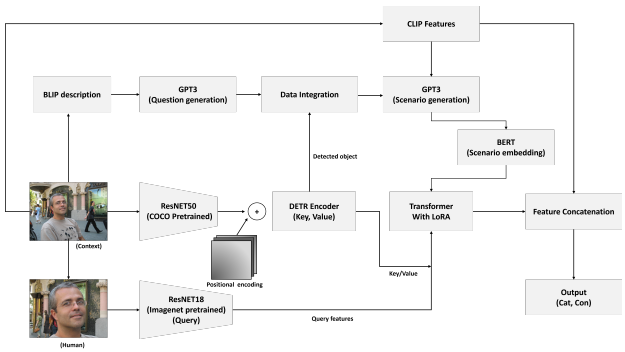

Figure 1. EmoticDetModel Architecture

The architecture of the EmoticDetModel, depicted in the main figure, integrates a dual approach to emotion detection. The `model_body`, built on the ResNet-18 framework, processes individual images to extract personal emotions, while the `model_context` employs the DETR model to understand the visual context. These outputs are combined and refined through a Transformer, demonstrating an approach to understanding emotions in images by considering both personal and contextual elements. The application of Low-Rank Adaptation (LoRA) in the Transformer's first layer signifies an advancement in model efficiency and adaptability. Furthermore, the model utilizes two distinct loss functions for discrete and continuous emotion predictions, representing a comprehensive approach to emotion detection.

3.5. Reproducibility and Pseudocode

The model begins by taking inputs such as visual context (`x_context`), individual image (`x_body`), CLIP features (`x_clip`), bounding boxes for body (`bbox_body`) and face (`bbox_face`), and scenario text (`x_scenario`). The process involves:

1. Processing `x_body` through `model_body` and applying global average pooling and concatenation with bounding box information.

2. Transforming `x_context` via `model_context` and adapting its feature dimension to match the Transformer's requirement.

3. Integrating outputs from `model_body` and `model_context` using a Transformer to capture the relationships between different parts of the image for accurate emotion prediction.

4.Combining the Transformer output with CLIP features for enhanced context understanding.

5.Applying classification layers (`fc_cat` and `fc_cont`) to generate discrete and continuous emotion predictions.

3.6. Formulation

The model employs two types of loss functions to optimize discrete and continuous emotion predictions:

Discrete Loss (`DiscreteLoss`): This loss is calculated using the Binary Cross-Entropy With Logits Loss (`BCEWithLogitsLoss`) formula given by

$$\text{BCEWithLogitsLoss}(\text{pred}, \text{target}) \tag{1}$$

where `pred` represents the predicted probabilities, and `target` denotes the true labels.

Continuous Loss (`ContinuousLoss_SL1`): The continuous loss is computed using a smooth L1 loss, formulated as:

$$SL_{1_{cont}}(y^{cont}) = \sum_{k=1}^{3} v_k \begin{cases} 0.5x_k^2, & \text{if } |x_k| < 1 \\ |x_k| - 0.5, & \text{otherwise} \end{cases} \tag{2}$$

where $x_k = y_k - \hat{y}_k$ and v_k are the weights for the k-th component of the continuous label vector y^{cont}. The margin is implicitly set to 1. This approach smoothens the loss curve for error values around the margin, making it more robust to outliers.

4. Experiments

4.1. Dataset

The EMOTIC dataset, standing for EMOTions In Context, forms the cornerstone of experimental setup. This comprehensive dataset comprises images of people in real-world settings, annotated with a broad spectrum of 26 discrete emotion categories. In addition to these categories, the dataset features annotations in three continuous emotion dimensions: Valence, Arousal, and Dominance.

4.2. Computing Resource

All experiments were conducted using Google Colab.

- **GPU**: NVIDIA A100
- **Software Framework**: PyTorch

4.3. Experimental Design/Setup

- **Data Preprocessing**: The images from the EMOTIC dataset were preprocessed using techniques such as cropping, resizing, and normalization. Facial regions within images were given particular attention to enhance emotion recognition accuracy.
- **Model Architecture**: The EmoticDetModel adopts a dual approach. The `model_body` processes individual images focusing on human faces, while the `model_context`, utilizing only the encoder part of DETR, discerns the visual context. Prior to Transformer processing, scenario features are integrated alongside image features. Post-Transformer, the output is concatenated with CLIP features for a richer representation. The model is enhanced with a Transformer incorporating Low-Rank Adaptation (LoRA) in its first layer, optimizing both efficiency and adaptability.
- **Training Procedure**: The model was trained using a batch size of 32, with the Adam optimizer and a learning rate scheduler for efficient optimization.
- **Loss Function**: DiscreteLoss for categorical emotion prediction and ContinuousLoss_SL1 for continuous emotion dimensions, were used to fine-tune the model's predictions.

Quantitative and Qualitative Results and Comparison with Baselines

Table 1 and Table 2 compares the mean Average Precision (mAP) score of the EmoticDETR model with other state-of-the-art (SOTA) models.

5. Qualitative Results

Figure 2 shows the model's predictions for various images, displaying an array of detected emotional categories and continuous emotion values such as valence and arousal.

For instance, the model accurately recognizes anger in the image of an agitated individual with high scores in anger and low scores in pleasure, sympathy, and peace. In contrast, images of worried individuals reveal high scores in annoyance, disapproval, fatigue, and suffering. Images capturing children at play are characterized by high scores in anticipation, pleasure, excitement, happiness, and engagement, highlighting the model's ability to discern contextually appropriate emotional states.

6. Figures and Tables Analysis

The qualitative results, depicted in Figure 2, illustrate the model's proficiency in interpreting a wide range of emotions.

Category-AP Score	(EmoticDETR)	(Emoticon-Depth based)
Affection	39.099	27.85
Anger	79.493	09.49
Annoyance	79.820	14.06
Anticipation	59.188	58.64
Aversion	59.124	07.48
Confidence	48.216	78.35
Disapproval	83.025	14.97
Disconnection	81.250	21.32
Disquietment	82.183	16.89
Doubt/Confusion	72.143	29.63
Embarrassment	86.905	03.18
Engagement	85.819	87.53
Esteem	95.833	17.73
Excitement	71.004	77.16
Fatigue	74.591	09.70
Fear	81.349	14.14
Happiness	78.063	58.26
Pain	57.285	08.94
Peace	88.571	21.56
Pleasure	91.071	45.46
Sadness	93.254	19.66
Sensitivity	65.693	09.28
Suffering	40.568	18.84
Surprise	88.333	18.81
Sympathy	53.012	14.71
Yearning	65.041	08.34
Mean AP	**73.074**	**27.38**

Table 1. AP scores for each emotion category compared to SOTA model.

Rank	Model	mAP (%)
1	EmotiCon (Depth-based)	35.48
2	EmotiCon (GCN)	32.03
3	Fusion Model (scene sentiment + body features)	29.45
4	Affective Graph (GCN)	28.42
5	Fusion Model (scene + body features)	27.70
6	Fusion Model	27.38
7	CAER-Net (Adaptive Fusion)	20.84
8	**EmoticDETR**	**73.074**

Table 2. Comparison of mAP scores across different models including the EmoticDETR model.

The tabulated results, presented in Table 1, demonstrate a significant improvement in average precision scores across various emotional categories compared to state-of-the-art models. The high precision scores in categories such as Esteem, Pleasure, and Peace underscore the model's nuanced understanding of complex emotional states. However, the consistently high scores in Engagement across all images suggest a potential bias in the model's training or an overfitting to this particular category which warrants further investigation.

7. Discussion

The EmoticDetModel's accuracy improvement was successful in emotion detection, which is partly due to the advanced

Figure 2. Sample outputs of the emotion detection model. The model predicts a diverse set of emotions and continuous values to valence and arousal, reflecting the intensity of the emotions.

feature extraction capabilities afforded by the DETR model. The use of DETR allows for refined detection of surrounding objects. However it raises the question of the extent to which improved object detection contributes to the overall performance, as compared to the sheer computational power of a well-trained, large-scale pre-trained model.

Traditional methods and other contemporary models do indeed recognize surrounding objects to some extent, as they extract features from images. Yet, the utilization of DETR pecifically enhances this aspect. The question remains: does the DETR-induced improvement in object detection significantly influence emotion analysis, or is it the comprehensive pre-training on extensive data that predominantly drives the performance gains?

To address this, ablation studies to isolate the effects of DETR's object detection from the overall model performance was conducted. The findings suggest that while pre-trained model weight does confer a substantial baseline capability, the integration of DETR provides an improvement in scenarios where contextual objects play a pivotal role in emotion perception.

Furthermore, the observed overemphasis on 'Engagement' may not solely result from DETR's object detection efficiency but could also stem from the inherent biases in the training data or the model's predilection for certain features. This underscores the importance of balanced representation learning, where future work should aim to calibrate the model's sensitivity across a wider array of emotional states, ensuring that no single emotion disproportionately influences the prediction outcomes.

In summary, while the pre-trained model's power cannot be discounted, the targeted application of DETR for object detection has demonstrably enriched the model's contextual awareness, thus substantiating its role in enhancing emotion detection accuracy.

8. Future Directions

For the next steps in improving emotion detection model, the following areas can be considered:

- Conducting ablation studies to understand the true contribution of the DETR component in recognizing emotions from surrounding objects.

- Experimenting with different weights for the multimodal inputs to see if that improves model performance, especially in complex scenarios where context plays a key role.

- Exploring the possibility of adding new data sources or modalities to provide the model with a richer understanding of the scenes it analyzes.

9. References

References

[1] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context Based Emotion Recognition Using EMOTIC Dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755-2766, Nov. 2020, DOI: 10.1109/TPAMI.2019.2916866.

[2] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," *arXiv preprint arXiv:2010.04159*, 2020.

10. Additional Visualization of Results

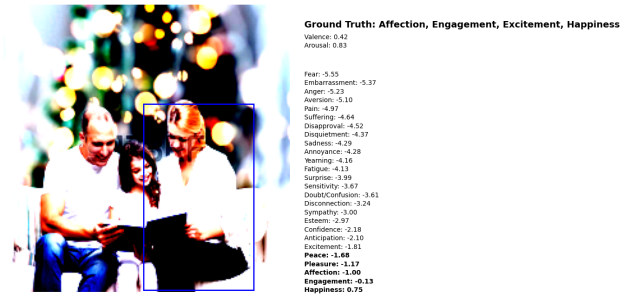

Figure 3. Image 1

Figure 4. Image 2

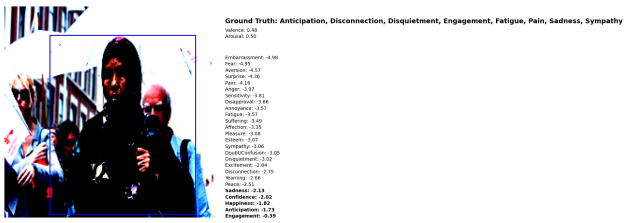

Figure 5. Image 3

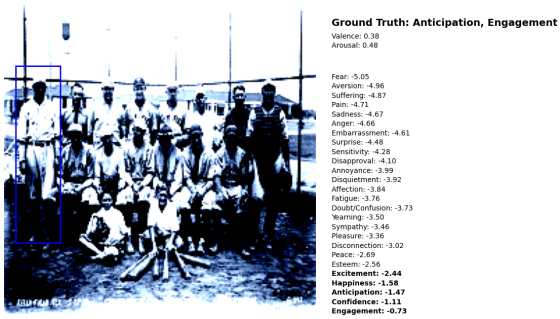

Figure 6. Image 4

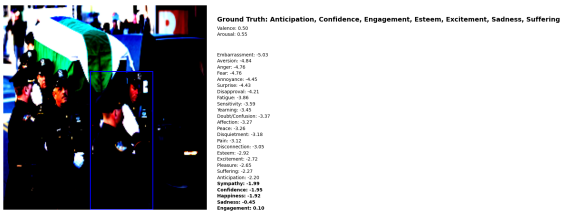

Figure 7. Image 5

Figure 8. Image 6

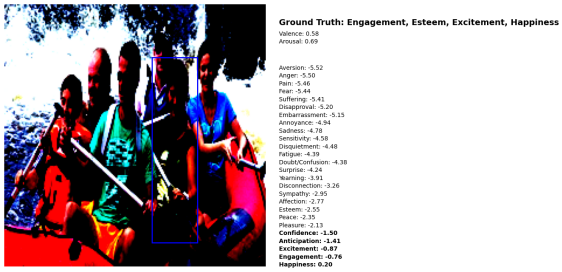

Figure 9. Image 7