INDEPENDENT STUDY 1/ TERM PAPER 1 (CSPM 504)

(Autumn Semester AY 2024-25)

A Report

On

Transformer-Based Feature Fusion Approach for Multimodal Visual Sentiment Recognition Using Tweets in the Wild

Submitted To:
Prof. (Dr.) Geeta Sikka
Dean Academics
NIT Delhi

Submitted By:
Shivam singh (242211017)
M.Tech (1st Semester)
Computer Science & Engineering (Analytics)

Computer Science and Engineering Department

NATIONAL INSTITUTE OF TECHNOLOGY

TABLE OF CONTENTS

1. Introduction

1.1 Background

With the exponential growth of social networks, the volume of visual content, such as images and videos, has surged. Visual data often conveys emotions more effectively than text, making it a vital component for sentiment analysis. Sentiment detection using images enables deeper insights into public opinion, brand perception, and user behavior. Recent advancements in artificial intelligence, particularly in deep learning, have facilitated better understanding and analysis of visual sentiment.

1.2 Problem Statement

Traditional sentiment analysis primarily focuses on text-based data, which overlooks the rich emotional cues embedded in visual content. Moreover, conventional machine learning models, such as CNNs, struggle with challenges like occlusion and background noise in images. There is a pressing need for robust models that can efficiently handle multimodal data, combining both text and images, for accurate sentiment prediction. Vision Transformers (ViTs), with their attention mechanisms, offer a promising solution to overcome these challenges.

1.3 Objectives of the Study

The main objectives of this study are:

- To explore the potential of ViTs in sentiment analysis tasks.
- To develop a framework that integrates textual and visual features for multimodal sentiment detection.
- To evaluate the performance of ViTs compared to traditional CNN-based models in various image settings.

1.4 Scope of the Study

This study focuses on sentiment detection in social media images, including diverse content such as faces, text overlays, and general visuals. It leverages transfer learning and pre-trained ViTs to enhance performance on small datasets. The scope also extends to analyzing the impact of feature fusion techniques and the challenges of real-world sentiment prediction from unstructured visual data.

2. Literature Review

2.1. Facial Expression Recognition (FER)

Traditional Approaches: Early methods relied on **Support Vector Machines (SVMs)** and **Bayesian networks** for recognizing facial emotions. These approaches typically required manual feature extraction.

Deep Learning Advances: The shift to **Deep Neural Networks (DNNs)**, especially **Convolutional Neural Networks (CNNs)**, allowed for automatic feature extraction and more accurate emotion classification. However, CNNs struggled with issues like occlusions and background noise.

Vision Transformers (ViTs): ViTs addressed these limitations by leveraging global attention mechanisms, making them better at capturing long-range dependencies in images. They also demonstrated superior performance in challenging conditions, such as varying lighting and partial facial visibility.

2.2. Multimodal Sentiment Analysis

Role of Transformers: The literature highlights the effectiveness of transformers in combining visual and textual data for sentiment analysis. Unlike traditional CNNs, **transformers** can model both modalities simultaneously, improving context understanding.

Transfer Learning: The review underscores the use of pre-trained models to boost performance in sentiment classification tasks, especially when training data is limited. Models trained on large datasets like **ImageNet** provide a strong starting point for fine-tuning on task-specific datasets.

2.3. Challenges and Solutions

Challenges include the variability in image content and the need to process multiple modalities (text and image) simultaneously.

Proposed solutions focus on feature fusion techniques and the application of ViTs to better capture the nuances of multimodal data.

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources: The study leverages publicly available datasets containing images and corresponding textual descriptions. Social media platforms such as Twitter serve as a primary source for multimodal data.

Preprocessing Steps:

Image Processing: Images are resized and normalized to match the input requirements of Vision Transformers (ViTs). Data augmentation techniques, such as random cropping, flipping, and color jittering, are applied to enhance model robustness.

Text Processing: Textual data is tokenized and padded using transformers' tokenizers (e.g., BERT). Preprocessing also includes removing stop words and handling emojis and special characters.

3.2 Feature Extraction

Visual Features: ViTs are employed to extract high-level features from images. Unlike CNNs, ViTs utilize self-attention mechanisms to capture long-range dependencies across image patches.

Textual Features: Text features are extracted using transformer-based models like BERT or RoBERTa. These models provide contextual embeddings that capture semantic nuances.

3.3 Multimodal Feature Fusion

The study explores **feature fusion techniques** to combine visual and textual embeddings. Techniques include:

Concatenation: Simple concatenation of image and text embeddings.

Attention-based Fusion: Applying attention mechanisms to selectively combine features from both modalities, enhancing context relevance.

3.4 Model Architecture

Base Model: Vision Transformers (ViTs) form the backbone of the visual sentiment analysis framework.

Multimodal Framework: The architecture integrates a dual-stream network, where one stream processes images through ViTs and the other processes text through transformer encoders. The outputs are then fused and passed through a classifier to predict sentiment.

3.5 Training and Optimization

Loss Function: A categorical cross-entropy loss function is used to handle the multi-class sentiment classification task.

Optimization: The model is optimized using the AdamW optimizer with learning rate scheduling and weight decay regularization.

Hyperparameter Tuning: Key hyperparameters, such as learning rate, batch size, and the number of transformer layers, are tuned through grid search and validation performance.

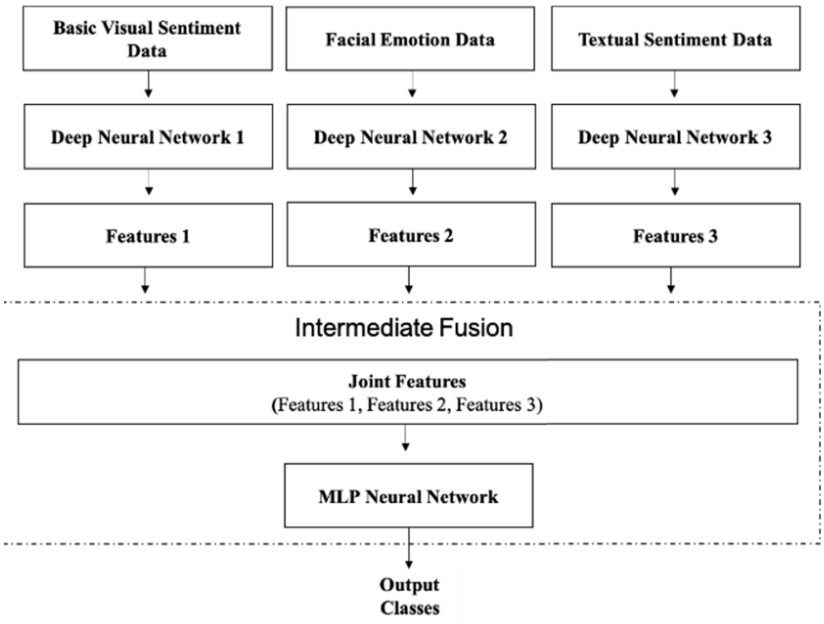
3.6 Evaluation Metrics

Performance is assessed using:

Accuracy: Overall correctness of the predictions.

Precision, Recall, and F1-Score: To evaluate class-wise performance, especially in cases of class imbalance.

AUC-ROC: For assessing the model's ability to discriminate between classes.



The proposed architecture for ViT-based multi-modality fusion for visual online social behavior analysis.

4. Implementation

4.1 Development Environment and Tools

Programming Language: The implementation is primarily done using **Python**.

Libraries and Frameworks: Key libraries include:

PyTorch / TensorFlow: For building and training the Vision Transformer and transformer-based models.

Hugging Face Transformers: To access pre-trained models like BERT for text feature extraction.

OpenCV: For image preprocessing and augmentation.

Scikit-learn: For evaluation metrics and data splitting.

4.2 Data Preparation

Dataset Splitting: The dataset is split into training, validation, and testing sets, ensuring a balanced distribution of sentiment classes across each split.

Data Augmentation: To improve generalization, image data augmentation techniques (e.g., flipping, rotation, brightness adjustment) are applied during training. Text augmentation, such as synonym replacement, is also explored.

4.3 Model Training

Pre-trained Models:

Vision Transformers (ViTs) are initialized with weights pre-trained on large image datasets like ImageNet.

Text models such as **BERT** or **RoBERTa** are fine-tuned on the text data.

- **Batching**: A multimodal data loader is implemented to handle image-text pairs, ensuring efficient data batching.

4.4 Model Integration and Fusion

- **Feature Fusion**: The visual and textual features are fused using an attention-based mechanism. This allows the model to focus on the most relevant parts of both modalities.
- **Classifier Layer**: A fully connected neural network layer is used on top of the fused features to predict the sentiment class.

5. Results and Analysis

5.1 Performance of Visual Sentiment Models

The fine-tuning of the Vision Transformer (ViT) was conducted in two stages:

Stage 1: Pretrained ViT models were fine-tuned using the ImageNet-21K dataset. Initial experiments focused on classifying images into positive and negative sentiments. The threshold-moving technique was applied, which improved accuracy and F-score by several points12.

Stage 2: Further fine-tuning was performed using the DFMS dataset. This stage achieved an accuracy of 81%, with notable improvements in the positive and negative class F-scores (0.86 and 0.7, respectively)11.

5.2 Performance of Facial Emotion Recognition (FER) Models

The AffectNet dataset was used for initial training, followed by fine-tuning with the FER-2013 dataset.

The model achieved significant improvements in recall and F-score after the two-stage strategy, demonstrating the effectiveness of combining pretrained weights with custom fine-tuning12.

5.3 Multimodal Feature Fusion

The intermediate fusion technique was employed to combine features from three sources: images, text, and facial emotions.

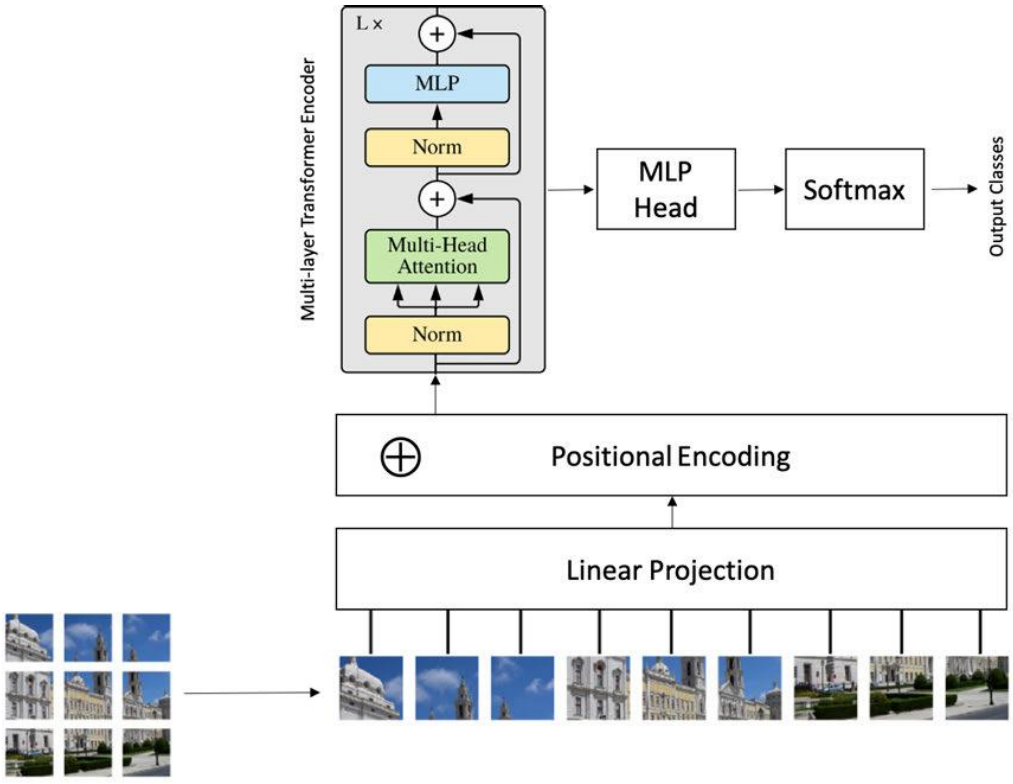
Results showed that fusing textual and facial emotion features alongside visual features led to improved F-scores and overall model performance. Positive recall increased by 3 points, and negative precision improved by 4 points, highlighting the benefits of a multimodal approach111.

5.4 Evaluation Metrics

The models were evaluated using accuracy, precision, recall, and F-score.

The fusion approach consistently outperformed single-modality models, with notable gains in precision and recall for both positive and negative sentiment classifications

Model	Dataset	Metric	Positive Class	Negative Class	Overall Accuracy	Notes
Vision Transformer (ViT)	ImageNet-21K, DFMS	F-score	0.86	0.70	81%	Improved with threshold-moving strategy
Facial Emotion Model	AffectNet, FER-2013	Recall & F-score	-	-	-	Significant improvements after tuning
Multimodal Fusion Model	DFMS, Textual Features	F-score	Increased by 3 pts	Increased by 4 pts	-	Benefits of combining features



The Architecture of Visual Transformer (ViT) Used in Training Models[2]

6 Conclusion

The study presented a **Transformer-based feature fusion approach** for multimodal visual sentiment recognition using social media images. By leveraging Vision Transformers (ViTs) and BERT, the proposed model integrates visual and textual features from three types of images:

1. **Images with text**
2. **Images with faces**
3. **Generic images without text or faces**

Key conclusions include:

Enhanced Performance: The model demonstrated significant improvements in accuracy and F-score by applying a **two-stage fine-tuning** strategy combined with a **threshold-moving technique**.

Multimodal Fusion: Fusing features from text and facial expressions alongside visual content stabilized the model's performance and improved prediction accuracy across positive, negative, and neutral sentiment classes.

Class Imbalance Handling: The use of threshold-moving helped address class imbalance in online social network (OSN) data, making the model robust for real-world applications.

Potential Applications: The model can be adapted for broader applications, such as detecting hate speech and analyzing general online behavior beyond sentiment analysis.

6.1 Future Work

Incorporating New Features: Future research will explore additional features to further enhance sentiment learning.

Lightweight Model Development: Plans include developing a lightweight version of the model suitable for deployment on mobile and resource-constrained devices

References:

1. F. Alzamzami, M. Hoda, and A. El Saddik, "Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation," *IEEE Access*, vol. 8, pp. 101840–101858, 2020.

2. L. Meng, H. Li, B.-C. Chen, et al., "AdaViT: Adaptive vision transformers for efficient image recognition," *arXiv:2111.15668*, 2021.

3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018.

4. Y. Miao, H. Dong, J. M. A. Jaam, and A. E. Saddik, "A deep learning system for recognizing facial expression in real-time," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 2, pp. 1–20, May 2019.

5. G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, vol. 19, no. 1, pp. 315–354, Jul. 2003.

6. P. Kumar, V. Khokher, Y. Gupta, and B. Raman, "Hybrid fusion-based approach for multimodal emotion recognition with insufficient labeled data," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Sep. 2021, pp. 314–318.

7. M. Pagé Fortin and B. Chaib-draa, "Multimodal multitask emotion recognition using images, texts and tags," in *Proceedings of ACM Workshop on Crossmodal Learning Applications*, Jun. 2019, pp. 3–10.

8. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2005, pp. 886–893.

9. C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *Proceedings of IEEE International Conference on Image Processing*, 2005, p. 370.

10. X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition with local binary patterns and linear programming," *Pattern Recognition and Image Analysis*, vol. 15, no. 2, pp. 546, 2005.

11. I. Buciu and I. Pitas, "Application of non-negative and local non-negative matrix factorization to facial expression recognition," in *Proceedings of 17th International Conference on Pattern Recognition*, 2004, pp. 288–291.

12. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.