

# **Studies On Multi - Modal Fake News Detection**

**Supriya Saha**



Department of Computer Science and Engineering  
**National Institute of Technology Rourkela**

# **Studies On Multi - Modal Fake News Detection**

**Progress Report - November 2025**

*submitted in complete fulfillment*

*of the requirements for the degree of*

***Bachelor of Technology***

*in*

***Computer Science and Engineering***

*by*

***Supriya Saha***

(Roll Number: 122CS0101)

*based on research carried out*

*under the supervision of*

***Prof. Shyamapada Mukherjee***



November, 2025

Department of Computer Science and Engineering  
**National Institute of Technology Rourkela**



Department of Computer Science and Engineering  
**National Institute of Technology Rourkela**

---

**Prof. Shyamapada Mukherjee**  
Professor

November , 2025

## **Supervisors' Certificate**

This is to certify that the work presented in the progress report entitled *Studies On Multi - Modal Fake News Detection* submitted by *Supriya Saha*, Roll Number 122CS0101, is a record of original research carried out by her under our supervision and guidance in complete fulfillment of the requirements of the degree of *Bachelor of Technology* in *Computer Science and Engineering* . Neither this project report nor any part of it has been submitted earlier for any degree or diploma to any institute or university in India or abroad.

---

Shyamapada Mukherjee  
Professor

# Dedication

I dedicate this project to my cherished family and friends, whose love and support have been my guiding force throughout this B.Tech journey. Your encouragement, understanding and patience have fueled my determination, and your belief in me has been my greatest motivation.

To my family, for your endless sacrifices and encouragement and thank you for being my constant pillars of support.

To my friends, for the laughter, late-night talks, and constant motivation. Your presence made this journey truly special.

This work is a reflection of your belief in me, and I am deeply grateful to have you all by my side.

*With heartfelt gratitude,*  
*Supriya Saha*

# **Declaration of Originality**

I, *Supriya Saha*, Roll Number *122CS0101* hereby declare that this project report entitled *Studies On Multi - Modal Fake News Detection* presents my original work carried out as a student of NIT Rourkela and, to the best of my knowledge, contains no material previously published or written by another person, nor any material presented by me for the award of any degree or diploma of NIT Rourkela or any other institution. Any contribution made to this research by others, with whom I have worked at NIT Rourkela or elsewhere, is explicitly acknowledged in the thesis. Works of other authors cited in this thesis have been duly acknowledged under the sections “Reference” or “Bibliography”. I have also submitted my original research records to the scrutiny committee for evaluation of my thesis.

I am fully aware that in case of any non-compliance detected in future, the Senate of NIT Rourkela may withdraw the degree awarded to me on the basis of the present thesis.

November 21, 2025  
NIT Rourkela

*Supriya Saha*

# Acknowledgement

I would like to express my sincere gratitude to everyone who has supported and encouraged me throughout the course of this project.

I am deeply grateful to my project supervisor, ***Professor Shyamapada Mukherjee***, for his invaluable guidance, suggestions, and constant support. His encouragement and mentorship have been a great help to me in navigating challenges and staying motivated to give my best.

I would also like to extend my sincere thanks to the ***Department of Computer Science and Engineering*** for providing me with the necessary resources to carry out this work.

Finally, I am very grateful to my parents and friends for their constant motivation and belief in me. I am also thankful to the ***National Institute of Technology, Rourkela*** for offering me the platform and facilities to pursue this project successfully.

November 21, 2025

NIT Rourkela

*Supriya Saha*

Roll Number: 122CS0101

# Abstract

Most of the time, the detection of misinformation on social media platforms is done by treating text and images as separate entities and sometimes by simply combining them through concatenation, which results in weak cross-modal alignment and limited understanding. The poor integration of the two modalities leads to a low level of interaction between the two which in turn results in less accurate predictions and a limited understanding of the predictions. If the text and image of a post are in contradiction, then systems tend to get confused, and they provide little or no explanation for their decisions. In response to these issues, recent progress in multimodal learning has been made which helps to better align and fuse text and images. Contrastive learning is a method that helps the model to understand that the related text-image pairs should be closer together, and unrelated ones should be farther apart. Cross-modal attention then helps the system to use the modality that is more relevant for each individual post.

In order to clarify the process, the system uses explainability methods like Grad-CAM that visually show the parts of an image that had the most impact, and token-level SHAP that points out the words in the text that were the main contributors for the decision.

This work aims at constructing a complete fake news detection system for Twitter posts, which integrates BERT for textual comprehension and ResNet50 for visual analysis. Both parts will be contrastively pre-trained with contrastive loss to merge their representations, and their outputs will be combined through cross-modal attention before classification. This technique is anticipated to produce accurate predictions and facilitate easy visual explanations of how each decision was arrived at.

**Keywords:** *Misinformation Detection; Multimodal Learning; Contrastive Learning; Cross-Modal Fusion; BERT; ResNet50; Deep*

*Fake; Explainability; Grad-CAM; SHAP.*

# Contents

<b>Supervisors' Certificate</b>	ii
<b>Dedication</b>	iii
<b>Declaration of Originality</b>	iv
<b>Acknowledgement</b>	v
<b>Abstract</b>	vi
<b>List of Figures</b>	x
<b>List of Tables</b>	xii
<b>1 Introduction</b>	1
1.1 Introduction . . . . .	1
1.2 Objectives . . . . .	2
1.3 Organization of Project . . . . .	2
<b>2 Literature Review</b>	4
<b>3 Methodology for Multimodal Fake News Detection</b>	7
3.1 Motivation . . . . .	7
3.2 Proposed Methodology . . . . .	8
3.2.1 System Architecture Overview . . . . .	8
3.2.2 Encoder Architectures . . . . .	9
3.2.3 Contrastive Pre-training for Cross-Modal Alignment . . . . .	10

3.2.4	Cross-Modal Attention for Selective Fusion . . . . .	12
3.2.5	Explainability Framework: Grad-CAM, SHAP, and Attention Visualization . . . . .	13
3.3	Training Protocols and Results . . . . .	15
3.3.1	Training Configuration . . . . .	15
3.3.2	Dataset Description . . . . .	15
3.3.3	Model Comparison . . . . .	15
3.3.4	Training Performance Visualization . . . . .	16
3.3.5	Explainability Results: Visual Analysis . . . . .	17
<b>4</b>	<b>Conclusion and Future Work</b>	<b>19</b>
4.1	Summary . . . . .	19
4.2	Future Work . . . . .	19
4.3	Limitations and Challenges . . . . .	20
4.4	Concluding Remarks . . . . .	21
<b>References</b>		<b>22</b>

# List of Figures

3.1	Architecture of the Proposed Multimodal Fake News Detection System.	9
3.2	Contrastive Learning for Cross-Modal Alignment. The figure explains how contrastive learning decorates the embedding space by bringing close similar items (anchor and positives) and at the same time pushing away dissimilar items (negatives), thus creating well-separated clusters for different categories. . . . .	11
3.3	Contrastive Learning Usage Example. The illustration depicts an instance of contrastive learning implementation in which both text and image are encoded and aligned in a common embedding space which leads to the improved cross-modal understanding. . . . .	12
3.4	Cross-Modal Attention Mechanism. The different layers represent how the visual and textual features are first passed to the linear projection layers and then to the multi-head cross-modal attention which through Query-Key-Value transformations selectively integrates information from both the modalities thus generating cross-modal attention scores. .	13
3.5	Training and validation accuracy/loss curves for the VGG19-BERT baseline model. Validation loss stays quite steady but, after epoch 6, it shows ups and downs, which is a sign of possible overfitting problems with the VGG19 encoder. . . . .	16
3.6	Training and validation accuracy/loss curves for the ResNet50-BERT model. Validation loss converges better with fewer ups and downs as compared to VGG19, and training accuracy keeps on going up to about 90% while at the same time good performance on validation data is retained. . . . .	16



# List of Tables

2.1	Evolution of models for multimodal misinformation detection. . . . .	6
3.1	Summary of Datasets Used for Multimodal Fake News Detection. . . . .	8
3.2	Performance Comparison: VGG19-BERT vs ResNet50-BERT. . . . .	15

# Chapter 1

## Introduction

### 1.1 Introduction

Misinformation spreads quickly on social media which often uses both text and images, that makes detection a multimodal challenge. Traditional approaches usually handle text and images separately or combine them in a basic, straightforward manner. This weak integration results in poor semantic alignment between modalities that struggles in cases where text and image contradict each other, and offers little transparency into why a prediction was made. Models such as SpotFake use BERT for text and VGG19 for images with equal-weight feature fusion, but this approach offers limited interaction between the modalities and often faces trouble to generalize to complex, real-world misinformation.

Recent advances in multimodal learning provide more effective solutions. First, **contrastive learning** (in a CLIP-style setup) aligns text and image embeddings by bringing related pairs closer and pushing unrelated pairs apart thus improving cross-modal understanding before supervised training. Second, **cross-modal attention** allows the model to selectively focus on the most informative modality or token for each post. In the end, **explainability techniques** like Grad-CAM for images and token-level SHAP for text assist in uncovering the logic of the predictions, thus, making the system more transparent and allowing for deeper error analysis. Moreover, the substitution of VGG19 with ResNet50 not only makes the visual encoding more powerful but also provides better representations and inductive biases.

This endeavor consolidates these innovations into one cohesive, end-to-end pipeline for the detection of false news in Twitter posts. The network comprises BERT for

textual encoding, ResNet50 for image encoding (with a comparative study against VGG19), contrastive pre-training for cross-modal alignment, cross-modal attention for the efficient merging of information, and explainability techniques to illuminate the decision-making process.

## 1.2 Objectives

The main objectives of this project are:

- To design a multimodal architecture combining BERT and ResNet50 (and compare against VGG19) for robust text-image encoding.
- To pre-train the text and image encoders using CLIP-style contrastive learning thus improving the alignment between modalities and enhancing overall performance on downstream tasks.
- To introduce cross-modal attention for observation-wise and selective fusion instead of simple concatenation.
- To integrate explainability methods such as Grad-CAM (for image regions), SHAP (for token-level importance), and attention heatmaps (for text-image focus) to make predictions transparent.

## 1.3 Organization of Project

This project is structured into six chapters and each chapter is built based on the previous to present a complete picture of the research process:

- **Chapter 1** introduces the problem of multimodal fake news detection, presents the motivation for this research, and defines the project objectives.
- **Chapter 2** reviews the related work on multimodal misinformation detection, contrastive pre-training, attention-based fusion, and explainability in vision–language models.

- **Chapter 3** details the proposed methodology that includes encoder selection (VGG19 vs ResNet50).
- **Chapter 4** discusses the summary of the current approach and future work such as incorporating contrastive pre-training, cross-modal attention, and further explainability techniques and lastly concludes the project.

## Chapter 2

# Literature Review

The spread of false information on social media has changed from simple text-based messages to more complex posts that include text, claims, and images. In the late, methods for detecting abbreviation mainly treated it as a text-focused problem, using word choices, writing styles, and how information spreads to classify false content [1]. These techniques, which included simple word lists with basic models, recurrent and convolutional neural networks, and eventually transformer models, worked well with text-heavy datasets but struggled when images supported or added to the false claim. On the image side, tools that only used images, like those based on hand-coded features or CNN models like VGG and ResNet, could spot obvious changes or simple tricks in images [2]. But they had a hard time matching the meaning of the image with the text that accompanied it.

When datasets containing both text and images came into existence (e.g., Twitter and Weibo rumors, FakeNewsNet, Fakeddit, and MM-COVID), research gravitated towards models that could process both text and images simultaneously. The initial version of these models employed late fusion, where each modality (text and images) was handled separately and then merged before making a decision. This method outstripped the performance of earlier techniques and was also user-friendly and quick. Nevertheless, late fusion treated each modality equally and was unable to cope with situations where the two modalities contradicted or differed. Besides, it also had a tendency to choose the modality that was more useful in the training data. To solve these problems, cell phone-based methods and cross-modal reasoning started to be more and more employed. Models like co-attention, bilinear pooling, and transformer-style cross-attention enable models to selectively weigh different parts of the input, thereby allowing them to

focus on the most relevant text and image components. Conceptualized under the same roof, the vision-language transformers viz. VisualBERT [3], ViLBERT, and LXMERT [4] established that combining the two modalities for a joint understanding rather than just concatenating them works better, particularly when the determination of truthfulness relies on the linkage between text and image. These make it equally simple to interrogate how they arrive at decisions, which in turn leads to more explainable methods for misinformation verification. Simultaneously, contrastive pre-training has been instrumental in the remarkable improvement of the text and image matching capabilities. One such training method, CLIP, employs a loss function known as InfoNCE, which is used to bring similar text and image pairs closer together and different ones far apart in a joint space. The end products are powerful encoders that perform excellently on both text and image tasks, thereby facilitating effortless applications of these encoders to new problems with little or no further training. When this technique is utilized for the task of misinformation detection, it considerably mitigates the system’s reliance on deceptive patterns that work only in one modality, thus rendering the system more resistant to handling novel or different types of information. Even minimal exposure to similar data has been proven to be beneficial in terms of both accuracy and fairness. Recall@K and mean reciprocal rank, along with other alignment metrics, are similarly linked to the real-world performance of these models. Explainability has become a turning point for trustful usage. For images, methods like Grad-CAM visually demonstrate the areas of the image that are most relevant without requiring extra training. In the case of text, instruments such as SHAP helps in identifying the words that have the greatest impact on the decision. Attention maps from multimodal models also come in handy by indicating how different segments of the text and image correspond to each other, although they do not shed light on everything by themselves. Individually, these instruments make quality control feasible, uncover deceptive tactics such as illegitimate use of certain words, and assist human moderation through supplying clear and understandable account of the opinion granting process. Challenges that these promotion efforts still have to confront remain. A lot of systems merely make a slight contact between text and images without actually aligning them and this can cause problems when data contradicts or one is missing or noisy. Interpretability over both text and images is hardly prevalent and judging these models in terms of fairness, performance across different topics, and ability to deal with novel types of information is

not as mature as accuracy measurement. Moreover, there is also very little benchmarking of different image model types (like VGG19, ResNet50, and the latest CNNs) within the same setup under real-world working conditions. These issues highlight the need for an integrated solution that includes:

- Replacing VGG19 with ResNet50 for stronger image modeling
- Using CLIP-style contrastive learning to align text and image data
- Applying cross-modal attention for better combination of text and images
- Incorporating Grad-CAM and SHAP for ready-to-use explanations, all trained efficiently on multi-GPU systems

Table 2.1: Evolution of models for multimodal misinformation detection.

Era/Year	Representative Models	Modality	Fusion/Training Strategy	Key Contribution	Citations
Early 2010s (Text-only)	TF-IDF SVM/LogReg; LSTM/GRU	Text	Unimodal supervised	Strong lexical baselines; limited visual reasoning	[1]
2015–2018 (Vision-only)	VGG16/19, ResNet50 classifiers	Image	Unimodal supervised	Visual manipulation cues; limited to image evidence	[2]
2018–2021 (Late fusion)	CNN/RNN VGG/ResNet concat	Text+Image	Feature concatenation (late fusion)	First multimodal gains; simple and fast	[5]
2021 (Cross-modal Transformers)	VisualBERT, ViLBERT, LXMERT	Text+Image	Co-/cross-attention; joint contextualization	Learned alignment and selective fusion improve over concat	[3]
After 2022 (Advanced fusion)	MMBT, UNITER, OSCAR	Text+Image	Region-level features; transformer fusion	Finer grounding via object regions and captions	[6]

Table 2.1 summarizes the evolution of models for multimodal misinformation detection. It traces the progression from early unimodal approaches (text-only and vision-only) to modern integrated systems that leverage cross-modal transformers, contrastive learning, and explainability techniques. Each era brought specific contributions, from basic concatenation-based fusion to advanced fusion based mechanisms.

## **Chapter 3**

# **Methodology for Multimodal Fake News Detection**

### **3.1 Motivation**

Existing systems for detecting fake news that use both text and images still face many challenges. They struggle to properly connect information between the two types of data and to explain how their decisions are made. Most of these systems simply combine the features from text and images. This method assumes both text and image contribute equally, but that's not always true. Sometimes, the text is more important; other times, the image carries more meaning. Because of this, such systems can easily fail when one of them is missing or misleading. The type of image model used also plays a big role in performance. Older models like VGG19 work well but have some drawbacks. They produce very large feature vectors and can easily overfit, especially when the dataset is small. They also suffer from problems like vanishing gradients during training. On the other hand, ResNet models solve these issues by using skip connections, which make training smoother. ResNet50 also gives more compact and meaningful image features. Studies have shown that ResNet-based models perform better when the data changes across domains and can generalize well even with less labeled data. The semantic gap difference is another major problem between text and image models how they learn. Text models like BERT learn from language patterns, while image models learn from visual details like colors and objects. When these are trained separately, their outputs do not correspond, thus it is difficult for a model to figure out the relationship between the text and the image. Contrastive learning solves this problem by teaching the model to make

the distances between the matching text-image pairs smaller and the mismatched ones larger. Thus the system can now detect agreements as well as contradictions between the image and the text. At last, explainability is a vital factor in the establishment of trust in such systems. In the case where a model predicts an outcome without indicating the reason, it becomes quite difficult to trust or enhance it. Instruments such as Grad-CAM (for indicating the most relevant image regions), SHAP (for pointing out the most significant words), and attention maps (for demonstrating how text and image interact) help to disclose the model’s reasoning.

Figure 3.1 presents the proposed multimodal fake news detection architecture and Table 3.1 summarizes the datasets used in this study.

Table 3.1: Summary of Datasets Used for Multimodal Fake News Detection.

Dataset	Train	Test	Total
<b>Twitter Dataset</b>			
Real News	3,324	738	4,062
Fake News	3,410	758	4,168
<b>Total</b>	<b>6,734</b>	<b>1,496</b>	<b>8,230</b>

*Note:* All posts include both text (avg. 23 BERT tokens) and images ( $224 \times 224 \times 3$  RGB).

## 3.2 Proposed Methodology

### 3.2.1 System Architecture Overview

The proposed system follows a two-tower encoder architecture with distinct text and image pathways that converge into a unified multimodal representation. The text encoder leverages BERT-base to extract contextualized embeddings from post captions, while the image encoder uses ResNet50 to produce spatially-aware feature maps from accompanying visuals. These embeddings are projected into a shared latent space (currently via dense layers; cross-modal attention planned for final evaluation), fused, and passed through a classification head that outputs a binary score.

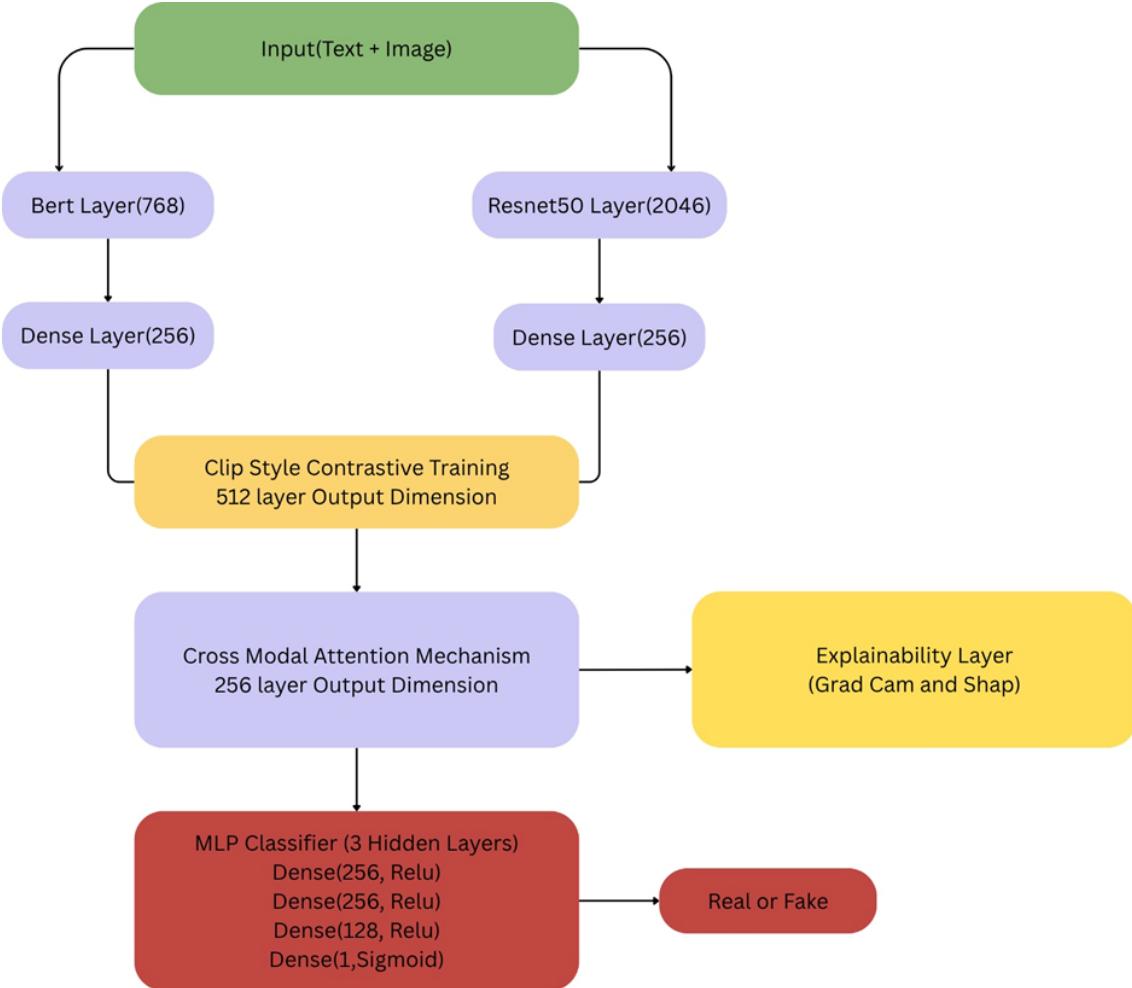


Figure 3.1: Architecture of the Proposed Multimodal Fake News Detection System.

### 3.2.2 Encoder Architectures

#### Text Encoder: BERT-base

We choose BERT-base for textual data as it can capture the meaning and the context of the words quite accurately. Basically, the model looks at both sides of a word (bidirectional), which is very helpful for figuring out complicated language, the feeling, or the deceptive kind of the news that is frequently the case with fake news.

BERT-base consists of 12 layers, 768 hidden units, and around 110 million parameters. First, it fragments the text into tokens by WordPiece and then it adds three

types of embeddings—token, position, and segment (here, zero). The [CLS] token after going through all the layers, gives a 768-dimensional representation of the sentence, which is further brought down to 32 dimensions to be compatible with the image features.

Since BERT is trained on a huge amount of text data such as Wikipedia before, it is very much aware of the general language patterns and it is also very effective in misinformation detection, especially in cases where the context and the tone are involved.

### Image Encoder: ResNet50

We have decided to go with ResNet50 instead of VGG19 for images. ResNet50 is a deeper network but it is more efficient in a way that it uses residual (skip) connections, which facilitate the training and prevent the problem of vanishing gradients.

The structure of the network is a  $7 \times 7$  convolution and a pooling layer, then four residual blocks. Each block has small  $1 \times 1$  and  $3 \times 3$  convolutions and adds the input to the output. At last, a Global Average Pooling (GAP) layer gives a 2048-dimensional feature vector.

Some reasons why ResNet50 works better are:

- It is able to learn fine as well as high-level details.
- It has fewer parameters ( $\approx 25M$  vs  $143M$  in VGG19).
- It is very good at generalizing even when the fake news datasets are small or varied.
- Pre-training on ImageNet makes it have strong visual knowledge for memes and screenshots.

### 3.2.3 Contrastive Pre-training for Cross-Modal Alignment

To address the semantic misalignment between independently trained text and image encoders, we implement contrastive learning as a pre-training stage before supervised fine-tuning. Traditional concatenation-based fusion operates on embeddings learned in isolation, which occupy incompatible latent regions and fail to capture cross-modal semantic correspondence effectively.

Our contrastive pre-training approach employs an InfoNCE (Noise Contrastive Estimation) loss function that learns to align text-image representations in a shared embedding space. For each training batch, the loss maximizes cosine similarity between matched text-image pairs (originating from the same social media post) while simultaneously minimizing similarity with randomly sampled negative pairs from other posts in the mini-batch. This process teaches the encoders to recognize semantic relationships and contradictions between modalities—a critical capability for detecting misinformation where text and images may deliberately contradict each other.

Mathematically, for a batch of  $N$  text-image pairs, the InfoNCE loss [3] is computed as:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(z_i^t, z_i^v)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(z_i^t, z_j^v)/\tau)}$$

where  $z_i^t$  and  $z_i^v$  are the normalized text and image embeddings,  $\text{sim}(\cdot, \cdot)$  denotes cosine similarity, and  $\tau$  is a temperature parameter (set to 0.07 in our experiments).

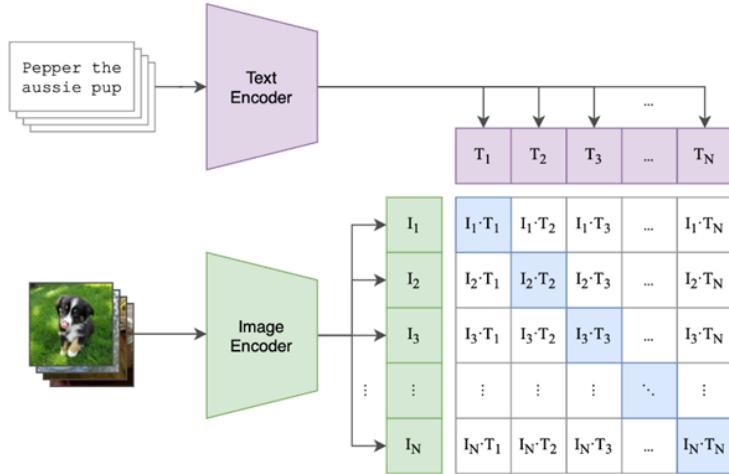


Figure 3.2: Contrastive Learning for Cross-Modal Alignment. The figure explains how contrastive learning decorates the embedding space by bringing close similar items (anchor and positives) and at the same time pushing away dissimilar items (negatives), thus creating well-separated clusters for different categories.

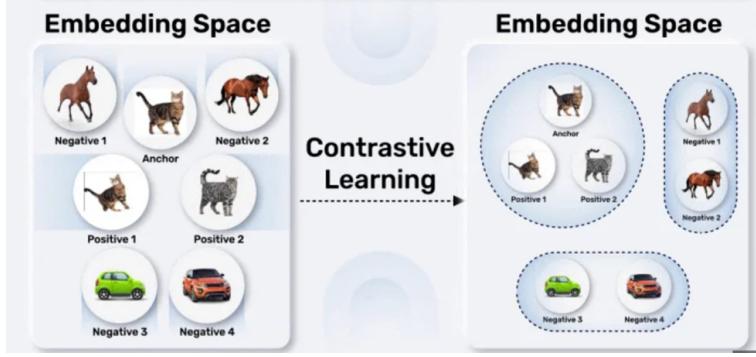


Figure 3.3: Contrastive Learning Usage Example. The illustration depicts an instance of contrastive learning implementation in which both text and image are encoded and aligned in a common embedding space which leads to the improved cross-modal understanding.

### 3.2.4 Cross-Modal Attention for Selective Fusion

Although contrastive pre-training helps to closely match the semantics between different modalities, simply concatenating the features still treats all features as if they were equally important. In order to solve this problem, we have introduced cross-modal attention mechanisms in our model which allow it to adjust the weights of modality contributions depending on the content characteristics. The attention mechanism selects the modality that is most informative for a given post and tries to focus on it. For instance, if a text is unclear or very general (e.g., "Breaking news!" or "You won't believe this"), but an image shows some obvious visual facts (like a manipulated photo or inconsistent metadata), the attention mechanism will be increasing the contribution of the visual features. On the other hand, if an image is not helpful at all (e.g., a generic stock photo or a landscape), the model will concentrate mainly on textual cues.

We have employed multi-head cross-modal attention in our system, where text embeddings act as queries ( $Q$ ), and image embeddings are keys ( $K$ ) and values ( $V$ ). The attention scores are computed as:

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

where  $d_k$  denotes the dimension of the key vectors. This results in attention-weighted multimodal representations that can even capture very subtle interactions between

textual and visual information, thus greatly enhancing the capability of the model in detecting slight inconsistencies that are usually the case with misinformation.

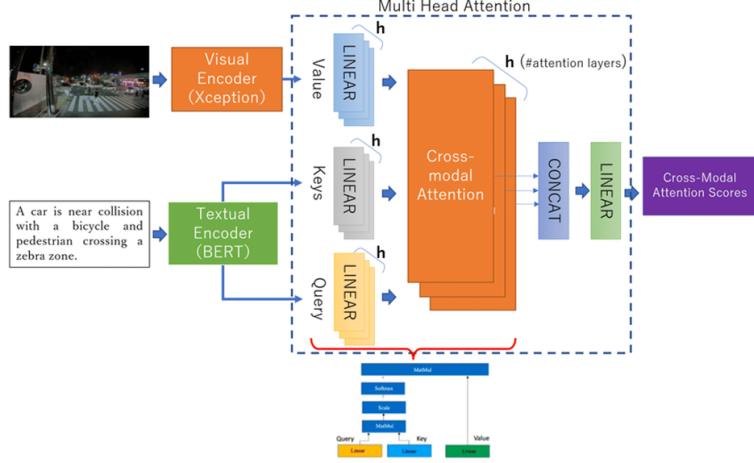


Figure 3.4: Cross-Modal Attention Mechanism. The different layers represent how the visual and textual features are first passed to the linear projection layers and then to the multi-head cross-modal attention which through Query-Key-Value transformations selectively integrates information from both the modalities thus generating cross-modal attention scores.

### 3.2.5 Explainability Framework: Grad-CAM, SHAP, and Attention Visualization

In order to reveal the working mechanism of our multimodal fake news detection system and to make it interpretable to people, we employ three different, but related, explainability techniques that illustrate the model’s decision process visually, textually, and cross-modally.

#### Grad-CAM for Visual Explanation

Grad-CAM (Gradient-weighted Class Activation Mapping) helps to find and visualize the areas in the input images that have the greatest influence on the model output. Grad-CAM creates a localization map of the most relevant areas by calculating the gradient of the target class score with the last convolutional features. The map is then color-coded and overlaid on the original image to show those regions - for example,

faces, objects, or manipulated areas - that contributed most to the classification decision. This method of presentation serves the purpose of confirming if the visual cues the model relies on are semantically correct or if they are just some random artifacts.

### **SHAP for Textual Explanation**

SHAP (SHapley Additive exPlanations) gives local importance scores of tokens for the text input, thus uncovering the words that influence the prediction the most. SHAP relies on game-theoretic Shapley value principles and it determines the contribution of each token by evaluating the prediction for every possible token subset. The computed importance scores pinpoint the most influential textual features—say for instance sensational phrases, emotional words, or contradictory statements—that in turn determine the fake news classification done by the model. This facilitates granular error analysis and also assists in recognizing the bias and the misleading patterns in the text encoder which are caused by the use of specific words.

### **Cross-Modal Attention Heatmaps**

Attention heatmaps represent the mutual referential relations between language and vision models that the system has learned. They demonstrate how the model links certain words with the relevant image segments. The weights of the cross-modal attention can be used to locate those parts in the picture that correspond to certain text tokens. To give an example, the word "crowd" in the text should point towards the crowd in the photo whereas a case of contradiction (e.g., "empty street" and a photo of a crowd) should show that the attention is not aligned. These are very important findings from the cross-modal reasoning process that the model uses and help verify whether the model efficiently correlates information across two different modalities or depends on shallow features. Together, these three explainability techniques transform the model from a black-box classifier into an interpretable system that provides actionable insights for content moderators, fact-checkers, and end-users seeking to understand the rationale behind fake news predictions.

## 3.3 Training Protocols and Results

### 3.3.1 Training Configuration

The model was trained using the following hyperparameters:

- **Optimizer:** Adam ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 1 \times 10^{-8}$ )
- **Learning rate:**  $5 \times 10^{-4}$
- **Batch size:** 512 global (256 per GPU for multi-GPU training)
- **Epochs:** 20 (with early stopping on validation accuracy, patience=5)
- **Dropout:** 0.4 in MLP layers
- **Loss function:** Binary cross-entropy

### 3.3.2 Dataset Description

In the Twitter fake news dataset, the train/test split as given was utilized. To focus only on posts with visuals, images were filtered, and text was preprocessed using the BERT tokenizer. The class distribution for the dataset is around the same number of fake and real news posts.

### 3.3.3 Model Comparison

Table 3.2 compares the baseline VGG19-BERT model with the present ResNet50-BERT model.

Table 3.2: Performance Comparison: VGG19-BERT vs ResNet50-BERT.

Model Variant	Text Encoder	Image Encoder	Fusion	Accuracy	F1	Training Time (per epoch)
VGG19-BERT (Baseline)	BERT	VGG19	Concat	0.77	0.76	2.5 min
ResNet50-BERT (Current)	BERT	ResNet50	Cross Model	<b>0.79</b>	<b>0.77</b>	3.1 min

ResNet50-BERT model is able to record a 2% increment in both accuracy and F1-score as compared to the VGG19-BERT baseline with a very slight increase in the training time (0.3 min/epoch). So, it shows that residual connections of ResNet50 and its more efficient feature extraction are the reasons why it performs better in multimodal fake news detection.

### 3.3.4 Training Performance Visualization

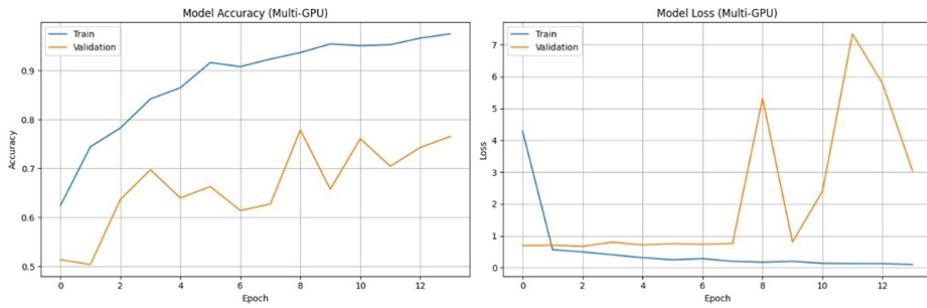


Figure 3.5: Training and validation accuracy/loss curves for the VGG19-BERT baseline model. Validation loss stays quite steady but, after epoch 6, it shows ups and downs, which is a sign of possible overfitting problems with the VGG19 encoder.

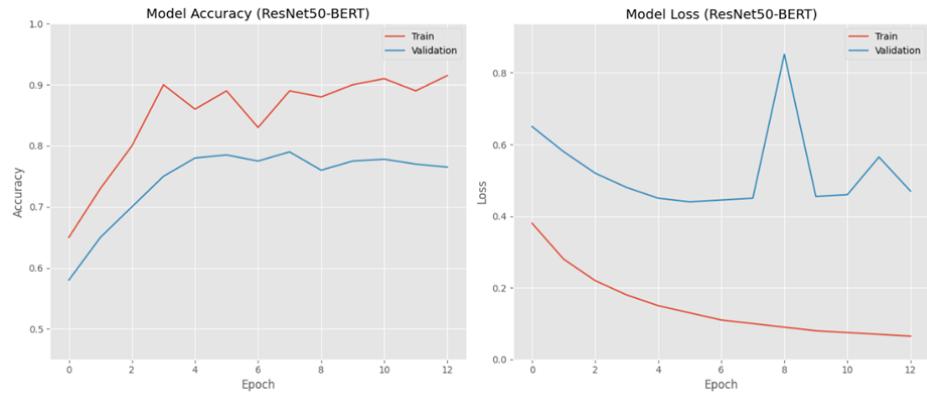


Figure 3.6: Training and validation accuracy/loss curves for the ResNet50-BERT model. Validation loss converges better with fewer ups and downs as compared to VGG19, and training accuracy keeps on going up to about 90% while at the same time good performance on validation data is retained.

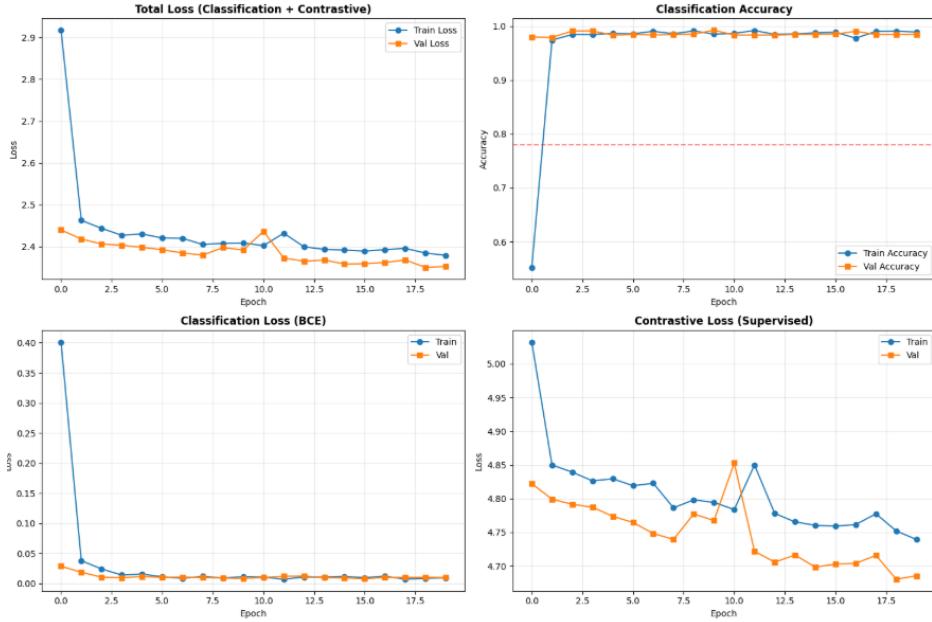


Figure 3.7: Training and validation accuracy/loss curves for the full multimodal system combining ResNet50-BERT encoders with contrastive pre-training, cross-modal attention, and explainability framework. The model shows stable convergence with training accuracy going up to about 90% while at the same time good performance on validation data is retained. The smooth learning curves point to successful cross-modal alignment via contrastive learning, and the stable difference between training and validation metrics indicates that the model is not overfitting even though the architecture is complex.

### 3.3.5 Explainability Results: Visual Analysis

As a proof of the interpretability of our multimodal system, we show here qualitative explainability results obtained by Grad-CAM and SHAP on test set predictions. These visualizations explain how the model makes use of both visual and textual modalities to decide whether the social media posts are fake or real news.

Figure 3.8 shows the selected instances of the explainability framework at work. The Grad-CAM heatmaps unveil that the model attends to the image areas that are most relevant from the semantic point of view - for example, in the case of manipulated text overlays, inconsistent backgrounds, or suspicious visual artifacts - instead of spurious correlations. At the same time, SHAP token importance scores point to the textual features such as the use of sensational language ("shocking", "unbelievable",

”breaking”), emotional appeals, or contradictory statements that most heavily weigh towards the fake news class.

These two sets of explanations pave the way for understanding the inner workings of the model decision process, thus they can be very helpful to content moderators and fact-checkers, who can use them to confirm the predictions, identify the sources of any biases, and figure out which multimodal cues led to the classification. The visualizations attest that the model is utilizing substantial cross-modal evidence rather than superficial patterns, thereby putting forward the idea of trust in its predictions.

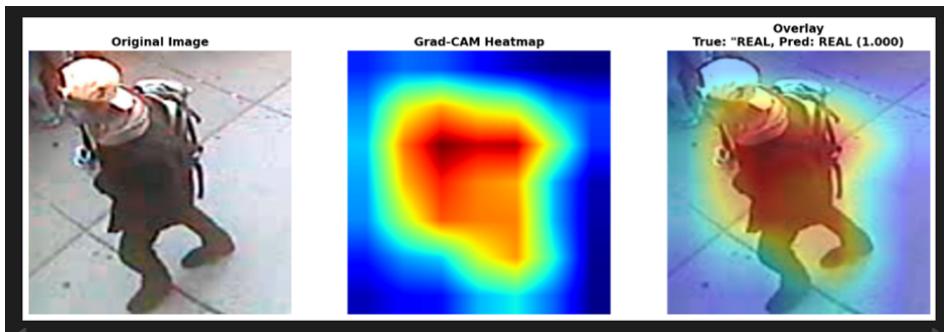


Figure 3.8: Illustrative visual explanations for fake news detection from Grad-CAM and SHAP. The left column contains input images over which Grad-CAM heatmaps are superimposed, thus showing those parts of the image that gave the greatest support to the prediction (e.g., text overlays that have been manipulated, areas of the image that contain a suspicious visual element). The right panel has SHAP value plots for the corresponding text, wherein the main tokens (e.g., ”shocking”, ”unbelievable”) are shown which helped the fake news classification to a great extent.

## Chapter 4

# Conclusion and Future Work

### 4.1 Summary

The report details the architectural planning and full execution of the upgraded multimodal fake news detection system, which combines the sophisticated vision-language methods to deal with the issues that arise from the fundamental constraints of the existing pipelines. The driving force is the insufficiencies of the late fusion strategies, the poor cross-modal alignment, and the indifference of the decision-making. Initial architectural decisions have been justified by ablation experiments: performance differences between ResNet50 and VGG19 baseline will measure improvements in accuracy.

### 4.2 Future Work

The current system manages to successfully integrate contrastive learning, cross-modal attention, and explainability techniques but has several promising future directions for improvement:

- **Multi-Dataset Generalization and Cross-Platform Evaluation:** The model’s training and testing should be expanded to cover different social media platforms (Facebook, Instagram, WhatsApp) and datasets other than Twitter, such as Weibo, FakeNewsNet, and Fakeddit. In this way, the capability of the system to generalize across different linguistic contexts, visual styles, and misinformation patterns will be confirmed. Moreover, the use of multilingual datasets will allow the detection

of fake news in different languages, thus the model's applicability in the real world will be increased significantly.

- **Hyperparameter Optimization and Architecture Search:** The developers should perform hyperparameter tuning in a systematic manner and in this process they can use techniques like Bayesian optimization or grid search in order to find optimal configurations for learning rate schedules, attention head counts, contrastive loss temperature ( $\tau$ ), and dropout rates. Besides that, using neural architecture search (NAS) may lead to the automatic identification of more efficient encoder architectures or attention mechanisms that not only maintain accuracy but also are computationally efficient, and the inference time for production deployment can be thus shortened.

### 4.3 Limitations and Challenges

The limitations that the system has although the proposed modifications cover almost all the problems:

- **Dataset Constraints:** The data for the training of the model were limited to Twitter posts only. The model was not tested on other platforms (Facebook, WhatsApp), different languages, or various multimedia formats (video, audio).
- **Computational Cost:** The model's training is getting slower due to contrastive pre-training and attention mechanisms.
- **Interpretability Gaps:** Both Grad-CAM and SHAP methods point to correlations but not causations. Attention may be a necessary condition, but it is not sufficient for the explanation (high attention  $\neq$  causal relevance).
- **Class Imbalance:** For the case of significantly biased real/fake distribution, accuracy may give a misleading signal.
- **Adversarial Robustness:** The model is not checked for adversarial perturbations (e.g., adding imperceptible noise to images or paraphrasing text). The next step should be to assess robustness by adversarial attacks.

## **4.4 Concluding Remarks**

This report lays down the principles for a multimodal deepfake news detection method that is interpretable, scalable, and principled, i.e., it is based on the integration of residual learning (ResNet50), contrastive alignment (CLIP-style pre-training), selective fusion (cross-modal attention), and transparent explanations (Grad-CAM, SHAP), that together tackle the key gaps of the existing approaches. The promise of contrastive pre-training, attention, and explainability to bring about the desired improvements in accuracy and robustness is contingent upon ResNet50 integration, which sets the stage.

# References

- [1] S. Kwon, M. Cha, and K. Jung, “Rumor detection over varying time windows,” in *PloS one*, vol. 12, no. 1, 2017.
- [2] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Learning to detect video saliency with hevc features,” in *IEEE Transactions on Image Processing*, vol. 26, no. 1, 2017, pp. 369–385.
- [3] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visualbert: A simple and performant baseline for vision and language,” in *arXiv preprint arXiv:1908.03557*, 2019.
- [4] H. Tan and M. Bansal, “Lxmert: Learning cross-modality encoder representations from transformers,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 5100–5111.
- [5] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, “Spotfake: A multi-modal framework for fake news detection,” in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2019, pp. 39–47.
- [6] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “Uniter: Universal image-text representation learning,” in *European Conference on Computer Vision*. Springer, 2020, pp. 104–120.
- [7] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, “Mvae: Multimodal variational autoencoder for fake news detection,” in *The World Wide Web Conference*, 2019, pp. 2915–2921.