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 - March 2025

submitted in partial 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

October, 2025

Department of Computer Science and Engineering

National Institute of Technology Rourkela

Department of Computer Science and Engineering
National Institute of Technology Rourkela

October , 2025

Certificate of Examination

Roll Number: *122CS0101*
Name: *Supriya Saha*
Title of Dissertation: *Studies On Multi - Modal Fake News Detection*

We the below signed, after checking the project report mentioned above and the official record book (s) of the student, hereby state our approval of the project report submitted in partial fulfillment of the requirements of the degree of *Bachelor of Technology* in *Computer Science and Engineering* at *National Institute of Technology Rourkela*. We are satisfied with the volume, quality, correctness, and originality of the work.

Shyamapada Mukherjee
Principal Supervisor

Dr. Bibhudatta Sahoo
Head of Department

External Examiner

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Prof. Shyamapada Mukherjee
Professor

October , 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 partial 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.

October 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.

October 21, 2025 | *Supriya Saha*
NIT Rourkela | Roll Number: 122CS0101

Abstract

The detection of misinformation on social media platforms often handles text and images separately or simply combines them by simple concatenation, yielding weak cross-modal alignment and limited understanding. This weak integration leads to poor alignment between the two modalities which results in less accurate predictions and limited understanding of how those predictions are made. When the text and image in a post contradict one another then systems tend to struggle and they provide little to no explanation for their decisions.

Recent advances in multimodal learning address these challenges by improving how text and images are aligned and fused. Contrastive learning helps bring related text–image pairs closer together in the model's understanding while pushing unrelated ones apart. Cross-modal attention then allows the system to focus on whichever modality is more important for each individual post.

In order to make the process transparent, the system uses explainability techniques such as Grad-CAM, which highlights the most influential regions of an image, and token-level SHAP, which shows which words in the text contributed most to the decision.

This project proposes building an end-to-end fake news detection system for Twitter posts that combines BERT for text understanding and ResNet50 for image analysis. Both components will be pre-trained with contrastive loss to align their representations, and their outputs will be fused using cross-modal attention before classification. This approach is expected to deliver more accurate predictions and offer clear visual explanations of how each decision was made.

Keywords: *Misinformation Detection*; *Multimodal Learning*; *Contrastive Learning*; *Cross-Modal Attention*; *BERT*; *ResNet50*; *Fake News Detection*; *Explainable AI (XAI)*; *Grad-CAM*; *SHAP*; *Text–Image Fusion*.

Contents

List of Figures

List of Tables

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. Finally, **explainability techniques** such as Grad-CAM for images and token-level SHAP for text helps in finding the reasoning behind predictions, enhancing transparency and supporting more thorough error analysis. Additionally, replacing VGG19 with ResNet50 strengthens visual encoding by offering better representations and inductive biases.

This project integrates these advancements into a single, end-to-end pipeline for fake news detection on Twitter posts. The system employs BERT for textual encoding, ResNet50 for image encoding (with comparative analysis against VGG19), contrastive pre-training for cross-modal alignment, cross-modal attention for intelligent fusion, and explainability techniques to make the decision-making process interpretable.

1.2 Objectives

The primary 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. 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. But they had a hard time matching the meaning of the image with the text that accompanied it.

As datasets that include both text and images started to appear (such as Twitter and Weibo rumors, FakeNewsNet, Fakeddit, and MM-COVID), study moved toward models that could feel at both text and images. The original type of these models used late fusion, where each part (text and images) was processed separately and then joined together before making a decision. This approach performed better than previous methods and was easy to use and fast. However, late fusion treated each part equally and couldn't handle situations where the two parts conflicted or differ. It also tended to favor whichever type of data was more useful in the training data.

To fix these issues, attention-based methods and cross-modal reasoning started to be used more. Techniques like co-attention, bilinear pooling, and transformer-style cross-attention let models weigh diverse parts of the data selectively, emphasize the most important text and image elements. Vision-language transformers like VisualBERT, ViLBERT, and LXMERT showed that having a shared understanding of both text and image data outperforms just joining them directly, especially when the linkage between text and picture is important for judging truthfulness. These models similarly made it easier to question how they make decisions, leading to further explainable approach for checking misinformation.

At the similar moment, contrastive pre-training greatly enhance how well text and images match. Training methods like CLIP use a loss function called InfoNCE to bring similar text and image pairs closer together and push different ones apart in a shared space. This creates strong encoders that work well for both text and image tasks, making it easier to apply them to new problems without much training. When used for detecting misinformation, this method helps reduce reliance on misleading patterns that only work in one type of data and makes the system more robust when dealing with new or different types of information. Even small amounts of training with similar data have been shown to help improve accuracy and fairness. Alignment measures like Recall@K and mean reciprocal rank also relate to how well these models perform in real-world tasks.

Explainability has turn key for faith use. For images, methods like Grad-CAM and Grad-CAM++ show which parts of the image are important without needing additional training. For text, tools like SHAP or Integrated Gradients help distinguish which words most influence the decision. Attention maps from multimodal models likewise help by showing how different parts of the text and image relate, though they don't clarify everything on their own. Together, these tools allow for quality checks, disclose misleading patterns like over-reliance on certain words, and support human moderation by provide clear reasons for opinion.

Despite these promotion, challenges remain. Many systems still just touch text and images without properly aligning them, which can determinant problems when the data conflicts or when one is missing or noisy. Interpretability across both text and images is not widely used, and evaluating these models in terms of fairness, cross-topic performance, and handling new types of information is not as developed as measuring accuracy. Additionally, there's little comparison between different types of image models (like VGG19, ResNet50, and newer CNNs) within the same system under realistic computing 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**	**Typical Backbones/Datasets**
Early 2010s (Text-only)	TF–IDF SVM/LogReg; LSTM/GRU	Text	Unimodal supervised	Strong lexical baselines; limited visual reasoning	Twitter rumor sets (text), small veracity corpora
2015–2018 (Vision-only)	VGG16/19, ResNet50 classifiers	Image	Unimodal supervised	Visual manipulation cues; limited to image evidence	WEIBO/Twitter images, tampering datasets
2018–2021 (Late fusion)	CNN/RNN VGG/ResNet concat	Text+Image	Feature concatenation (late fusion)	First multimodal gains; simple and fast	Twitter15/16, Weibo, FakeNewsNet
2021 (Cross-modal Transformers)	VisualBERT, ViLBERT, LXMERT	Text+Image	Co-/cross-attention; joint contextualization	Learned alignment and selective fusion improve over concat	COCO/VQA pretraining; transfer to rumors
After 2022 (Advanced fusion)	MMBT, UNITER, OSCAR	Text+Image	Region-level features; transformer fusion	Finer grounding via object regions and captions	COCO/Conceptual Captions; downstream transfer

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**
Twitter Dataset			
Real News	3,324	738	4,062
Fake News	3,410	758	4,168
Subtotal	**6,734**	**1,496**	**8,230**
Weibo Dataset			
Real Events	2,313	500	2,813
Fake Events	2,351	508	2,859
Subtotal	**4,664**	**1,008**	**5,672**
Grand Total	**11,398**	**2,504**	**13,902**

Note: All posts include both text (avg. 23 BERT tokens) and images (224×224×3 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.

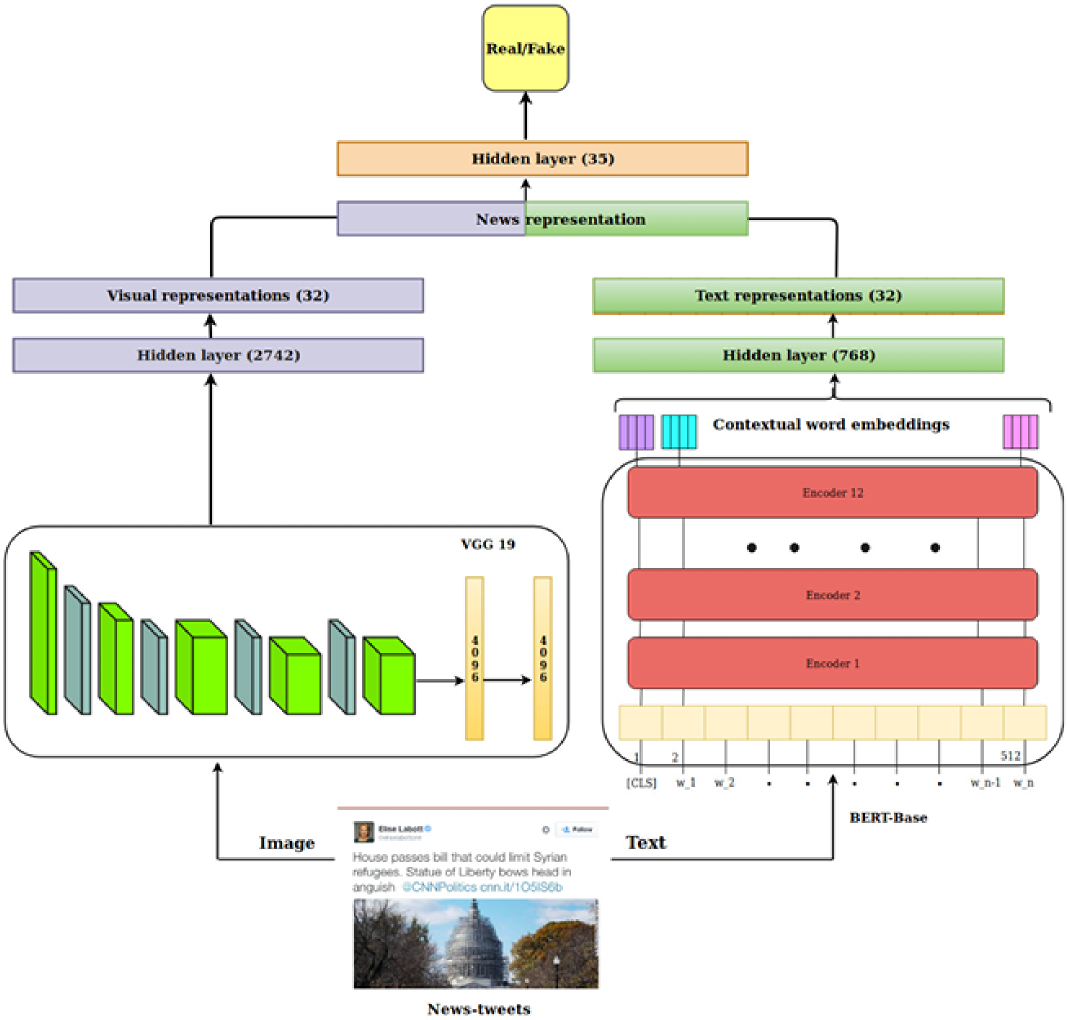

Figure 3.1: Architecture of the Proposed Multimodal Fake News Detection System.

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 7×7 convolution and a pooling layer, then four residual blocks. Each block has small 1×1 and 3×3 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 (≈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.

3.3 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:** 5×10^{-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

The Twitter fake news dataset was used with the provided train/test split. Images were filtered to retain only posts with available visuals, and text was preprocessed using the BERT tokenizer. The class distribution is approximately balanced between fake and real news posts.

3.3.3 Model Comparison

Table 3.2 presents a comparison between the baseline VGG19-BERT model and the current 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	Concat	**0.79**	**0.78**	2.8 min

The ResNet50-BERT model achieves a 2% improvement in both accuracy and F1-score over the VGG19-BERT baseline, with only a marginal increase in training time (0.3 min/epoch). This demonstrates that ResNet50's residual connections and more efficient feature extraction lead to better performance in multimodal fake news detection.

3.3.4 Training Performance Visualization

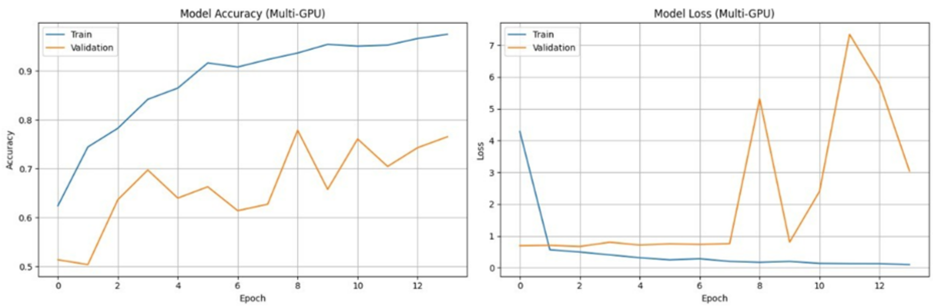

Figure 3.2: Training and validation accuracy/loss curves for VGG19-BERT baseline model. The validation loss remains relatively stable but shows fluctuations after epoch 6, indicating potential overfitting issues with the VGG19 encoder.

Figure 3.3: Training and validation accuracy/loss curves for ResNet50-BERT model. The validation loss shows better convergence with fewer fluctuations compared to VGG19, and the training accuracy steadily increases to around 90% while maintaining good generalization on validation data.

Chapter 4
Conclusion and Future Work

4.1 Summary

This report presents the design and partial implementation of an enhanced multimodal fake news detection system that integrates advanced vision–language techniques to address fundamental limitations in existing pipelines. The motivation stems from the inadequacies of late fusion strategies, weak cross-modal alignment, and opacity in decision-making. Preliminary architecture choices have been validated through ablation planning: comparing ResNet50 against the VGG19 baseline will quantify gains in accuracy.

4.2 Future Work

4.2.1 Contrastive Pre-training for Cross-Modal Alignment

Addition of Contrastive learning to pre-train text and image encoders jointly before supervised fine-tuning. Current concatenation-based fusion operates on embeddings learned independently, which occupy incompatible latent regions and fail to model semantic correspondence. Contrastive learning addresses this by optimizing an InfoNCE loss that maximizes cosine similarity for matched text-image pairs (from the same post) while minimizing similarity for randomly sampled negative pairs within each mini-batch.

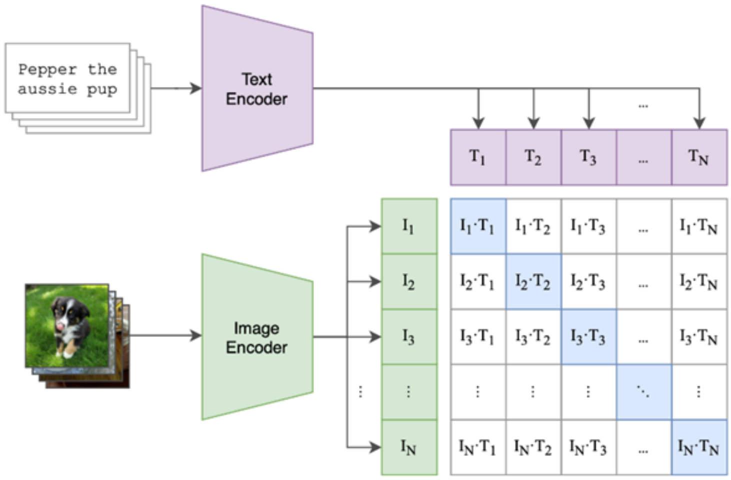

Figure 4.1: Contrastive Learning for Cross-Modal Alignment. The diagram illustrates how contrastive learning organizes the embedding space by moving similar items (anchor and positives) closer together while pushing dissimilar items (negatives) further apart, creating distinct clusters for different categories.

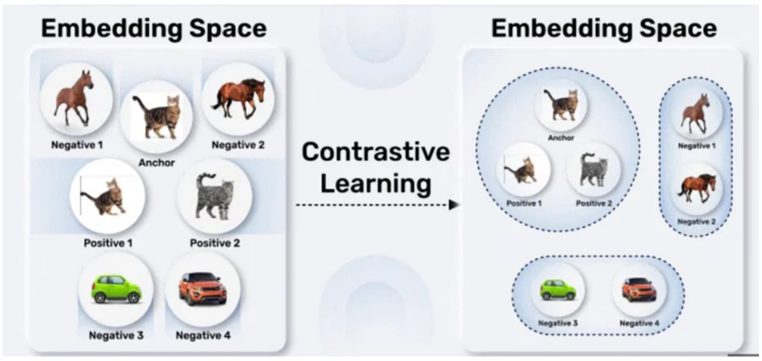

Figure 4.2: Contrastive Learning Usage Example. This figure demonstrates the practical application of contrastive learning where text and image pairs are encoded and aligned in a shared embedding space, enabling better cross-modal understanding.

4.2.2 Cross-Modal Attention for Selective Fusion

While contrastive learning aligns modalities, concatenation still treats them uniformly at the instance level. Cross-modal attention resolves this by enabling the model to dynamically weight modality contributions based on content. For instance, if text is vague ("Breaking news!") but the image shows clear evidence (manipulated photo), attention up-weights visual features; conversely, if the image is uninformative (generic landscape), attention focuses on text.

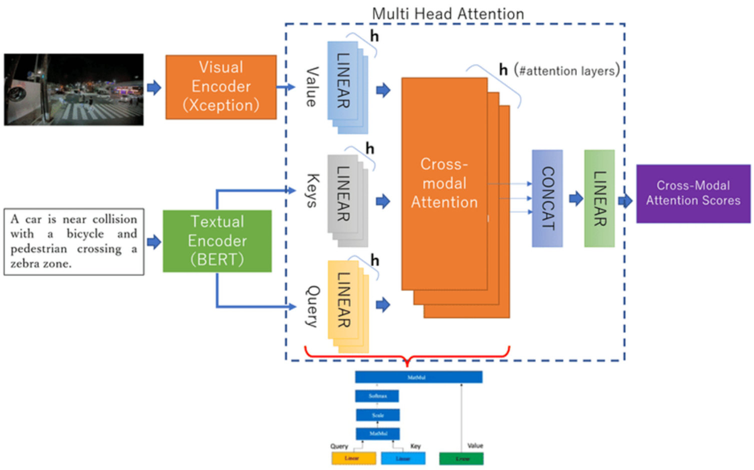

Figure 4.3: Cross-Modal Attention Mechanism. The architecture shows how visual and textual encoders feed into linear projection layers, followed by multi-head cross-modal attention that selectively fuses information from both modalities through Query-Key-Value transformations, producing cross-modal attention scores.

4.2.3 Explainability Framework: Grad-CAM and SHAP

To render the model less of a black box and more interpretable, three explainability measures Grad-CAM, SHAP, and Attention Heatmap were introduced.

Grad-CAM (Gradient-weighted Class Activation Mapping) supports the identification and visualization of the key input image areas, which in turn had the most significant impact on the model's decision. With this technique, the image becomes overlaid with a heatmap marking the regions that predominantly drive the prediction, thus providing an indication of whether the model is focusing on the relevant visual cues such as faces, objects, or manipulated areas.

SHAP (Shapley Additive Explanations) is a technology that highlights the words that have the highest contribution to the output. It comes to such a conclusion by determining the prediction each word has, thus giving the importance scores that point out the most influential textual features.

Attention heatmaps are a different modality used to establish the interaction between different modalities. Such means illustrate how a model associates' language and images by pointing out the visual parts that correspond to certain words in the text. This, in turn, assists in comprehending the cross-modal reasoning process and verifying if the model is efficiently correlating information across both modalities.

4.3 Limitations and Challenges

Even if the proposed changes cover most of the issues, a few limitations remain:

- **Dataset Constraints:** For the training of the model, only Twitter posts were used as data. The model's capability to work with other platforms (Facebook, WhatsApp), different languages, or various multimedia formats (video, audio) has not been verified.

- **Computational Cost:** The training of the model is getting slower due to contrastive pre-training and attention mechanisms.

- **Interpretability Gaps:** Both Grad-CAM and SHAP methods indicate correlations but not causations. Attention is a necessary condition, but it is not sufficient for the explanation (high attention $\neq$ causal relevance).

- **Class Imbalance:** In the case of a significantly biased real/fake distribution, the accuracy may provide a deceptive signal.

- **Adversarial Robustness:** The model has not been tested for adversarial perturbations (e.g., adding imperceptible noise to images or paraphrasing text). Determining robustness through adversarial attacks (FGSM, PGD) should be the following step.

4.4 Concluding Remarks

This report establishes the foundation for a principled, interpretable, and scalable multimodal deepfake news detection. The integration of residual learning (ResNet50), contrastive alignment (CLIP-style pre-training), selective fusion (cross-modal attention), and transparent explanations (Grad-CAM, SHAP) addresses critical gaps in existing approaches. While ResNet50 integration lays the groundwork, contrastive pre-training, attention, and explainability, will deliver the promised improvements in accuracy and robustness.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘Bert: Pre-training of deep bidirectional transformers for language understanding,’’ in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2019, pp. 4171–4186.

[2] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image recognition,’’ in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[3] K. Simonyan and A. Zisserman, ‘‘Very deep convolutional networks for large-scale image recognition,’’ *arXiv preprint arXiv:1409.1556*, 2014.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, ‘‘Learning transferable visual models from natural language supervision,’’ in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[5] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, ‘‘Mvae: Multimodal variational autoencoder for fake news detection,’’ *The World Wide Web Conference*, pp. 2915–2921, 2019.

[6] X. Zhou, J. Wu, and R. Zafarani, ‘‘Safe: Similarity-aware multi-modal fake news detection,’’ in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020, pp. 354–367.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘‘Grad-cam: Visual explanations from deep networks via gradient-based

localization,'' *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[8] S. M. Lundberg and S.-I. Lee, ''A unified approach to interpreting model predictions,'' *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, ''Multimodal fusion with recurrent neural networks for rumor detection on microblogs,'' in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 795–816.

[10] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Yao, C. Lyu, and J. Li, ''Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues,'' *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1212–1220, 2021.