Studies On Multi - Modal Fake News Detection

Research Project

Supriya Saha
(122CS0101)

Supervision of
Prof. Shyamapada Mukherjee

Department of Computer Science and Engineering
NIT Rourkela-769008, India

October 21, 2025

Outline

Introduction

- Misinformation spreads quickly on social media using both text and images which creates a multimodal detection challenge.
- Traditional models treat text and images separately or fuse them weakly, leading to poor semantic alignment and limited interpretability.
- SpotFake uses BERT (text) + VGG19 (image) with simple fusion but this limits its ability to handle complex, real-world misinformation.
- Recent advances improve this by:
  - Contrastive Learning (CLIP-style): Aligns text–image embeddings for stronger cross-modal understanding.
  - Cross-Modal Attention: Focuses on the most informative tokens or regions.
  - Explainability (Grad-CAM, SHAP): Displays decision reasoning for transparency.

Literature Review

Table 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	[? ?]
2015–2018 (Vision-only)	VGG16/19, ResNet50 classifiers	Image	Unimodal supervised	Visual manipulation cues; limited to image evidence	[? ?]
2018–2021 (Late fusion)	CNN/RNN VGG/ResNet concat	Text+Image	Feature concatenation (late fusion)	First multimodal gains; simple and fast	[? ?]
2021 (Cross-modal Transformers)	VisualBERT, ViLBERT, LXMERT	Text+Image	Co-/cross-attention; joint contextualization	Learned alignment and selective fusion improve over concat	[? ? ?]
After 2022 (Advanced fusion)	MMBT, UNITER, OSCAR	Text+Image	Region-level features; transformer fusion	Finer grounding via object regions and captions	[? ? ?]

Objectives

- Develop a BERT + ResNet50 multimodal model (compare with VGG19).
- Apply CLIP-style contrastive pre-training for text–image alignment.
- Introduce cross-modal attention for selective, context-aware fusion.
- Integrate explainability tools — Grad-CAM, SHAP and attention heatmaps for interpretable predictions.

Dataset Overview

Table **??** summarizes the datasets used for multimodal fake news detection in this study.

Table 2: 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).

Methodology

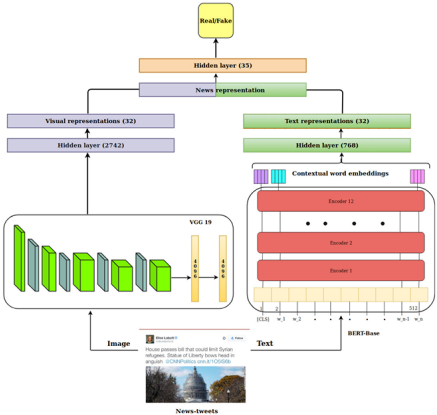

Figure 1: Proposed multimodal fake news detection architecture using VGG-19 for visual feature extraction and BERT-Base for text embeddings, with fusion layers for final classification.

Methodology (Cont.)

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.

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

Performance Comparison

Table **??** presents a comparison between the baseline VGG19-BERT model and the current ResNet50-BERT model on the Twitter fake news dataset.

Table 3: Performance Comparison: VGG19-BERT vs ResNet50-BERT on Twitter Dataset.

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

Training Curves - VGG19-BERT Baseline

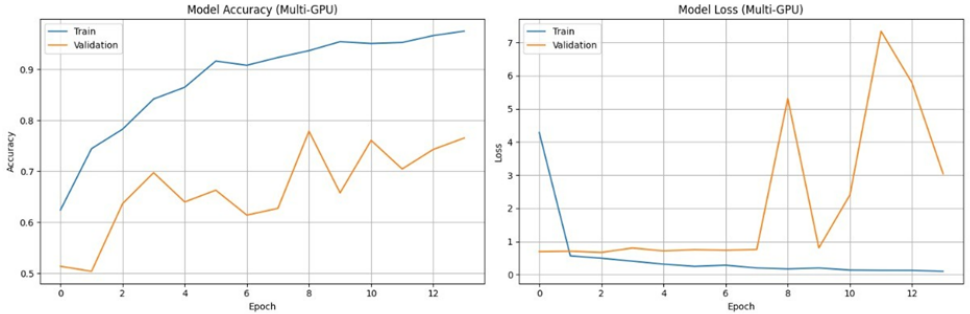

Figure 2: Training and validation accuracy/loss curves for VGG19-BERT baseline model. The validation loss remains relatively stable but shows fluctuations after epoch 6, indicating potential overfitting issues with the VGG19 encoder.

Training Curves - ResNet50-BERT

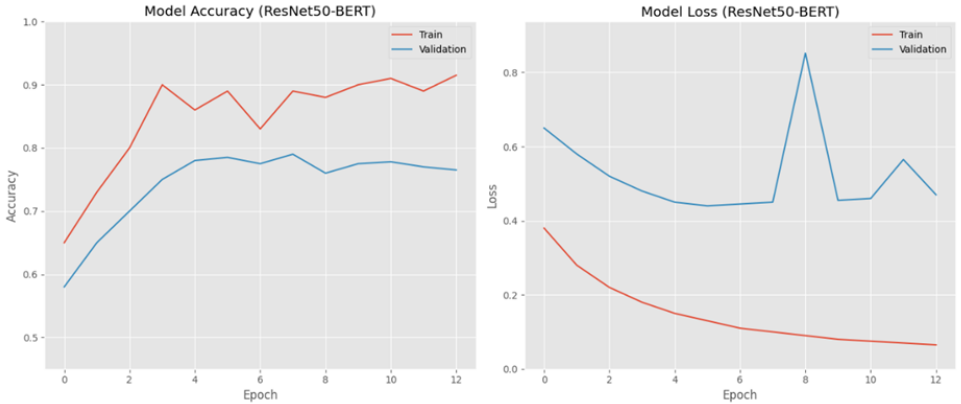

Figure 3: Training and validation accuracy/loss curves for ResNet50-BERT model. The validation loss shows better convergence with fewer fluctuations compared to VGG19, and the training accuracy steadily increases to around 90% while maintaining good generalization on validation data.

Conclusion

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.

Future Work

Contrastive Learning for Alignment

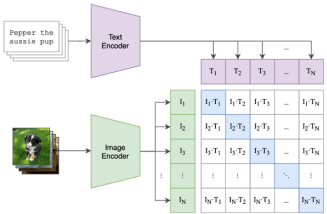

Figure 4: Contrastive learning organizes embeddings by moving similar items closer and pushing dissimilar items apart.

Cross-Modal Attention Mechanism

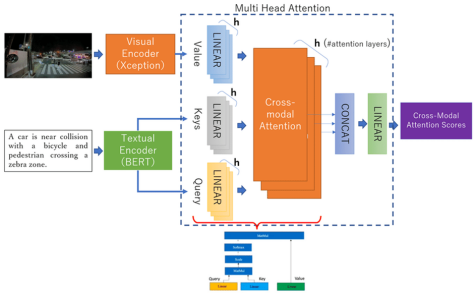

Figure 5: Cross-modal attention dynamically weights modality contributions based on content relevance.

Explainability: Grad-CAM highlights key image regions, SHAP reveals important textual features, and Attention Heatmaps visualize how the model links text with corresponding image regions.

References I

Thank You