TEAM MEMBERS:

Ahmed Hazem Hassan 221001099

Ahmed Hossam Abdelsalam 231001426

Amr Mohamed El-Hassaneen 221001520

Mohamed Ayoub Khabiry 221000728

Abstract

The pervasive dissemination of misinformation online poses a significant threat to public discourse, especially when deceptive narratives are reinforced with unrelated or manipulated images. Traditional text-based fake news detection approaches often struggle to capture the semantic dissonance between textual and visual content. In response, this work explores multimodal fake news detection by integrating visual and textual modalities through various fusion strategies. We investigate three fusion architectures: (1) an early fusion model combining CLIP and BERT embeddings, (2) an early fusion approach integrating EfficientNet for visual encoding with BERT for text representation, and (3) a late fusion architecture pairing MobileNet with LSTM, where each modality is processed independently before combining prediction scores. The extracted features are used to train a neural network classifier to distinguish between real and fake news. Our experiments demonstrate that early fusion models—particularly CLIP+BERT—achieve superior performance by effectively capturing fine-grained cross-modal correlations, while late fusion offers a more modular but slightly less expressive alternative. These findings highlight the importance of fusion strategy selection in building robust multimodal misinformation detection systems.

Keywods: FakeNews Detection, Multi-modality, Deep Learning, FakeNewsNet Dataset.

Introduction

The Problem: The Spread of Fake News

With the rapid growth of social media and online platforms, fake news has become a serious global issue. Misleading headlines and manipulated stories can spread misinformation to millions of users within minutes. This not only distorts public perception but can also harm democratic processes, public health, and social stability.

Why Fake News Detection Matters

* Impact on Society: Fake news can lead to real-world consequences like violence, panic, or distrust in governments and media.
* Difficult to Detect: Fake content often mimics real news in tone and structure, making manual detection challenging and time-consuming.
* Need for Automation: As the volume of online content increases, automated systems using AI and deep learning are crucial for detecting and filtering misinformation.

Real-World Statistics

* According to Statista, nearly 60% of users encounter fake news weekly.
* A 2020 study by MIT showed fake news spreads 6x faster than true stories on social media platforms.
* During the COVID-19 pandemic, the World Health Organization (WHO) declared an "infodemic" due to the overwhelming spread of false information.

Our Project: Multi-Model Fake News Detection

To address this problem, our team built a multi-model deep learning system that can classify news articles as real or fake. Each member focused on a different approach to evaluate the performance and effectiveness of various architectures. The models we explored include:

* LSTM (Long Short-Term Memory): Captures sequential dependencies in text.
* BERT (Bidirectional Encoder Representations from Transformers): A pre-trained language model that understands context.
* MobileNet: Used for visual features in multimodal analysis.

Clip: used for both image and text features (vit based)

Background & Related Work:

The rising surge of misinformation on the internet has presented a big challenge to the credibility of web content. Disinformation entwined facts reported as factual news—has a tendency to mix persuasive text, based stories with pictorially misleading images, thus also making it difficult to identify using the traditional, text-based approaches. Sensitive to this constraint, current research has moved in the direction of multimodal fake news detection, an approach that brings together both textual and visual modalities to reveal contradiction between saying and showing. Critical to the efficacy of multimodal models is the fusion approach employed to integrate information from heterogeneous sources.  
Two leading paradigms have been introduced in this regard. Early fusion techniques try to merge text and vision feature representations at some intermediate point in such a way that the model can learn dense cross-modal interactions. Late fusion approaches, by contrast, process each modality independently and fuse their outputs in the final prediction step. While early fusion generally offers improved performance through capturing fine-grained correlations at the cost of increased model complexity. The emergence of new Vision-Language Models (VLMs) such as CLIP has brought with it a significant leap in the alignment of images and texts. By large-scale contrastive pretraining to acquire a common embedding space, the models enable downstream tasks like identifying fake news to directly measure semantic coherence between visual and text content. Such capability is essential for detecting misinformation where unrelated or manipulated images are aligned with made-up stories.  
Previous work has established strong groundwork in this direction.  
SpotFake (Singhal et al., 2019) initially demonstrated that it is feasible to efficiently integrate BERT for text and CNN-based visual encoders such as VGG with a simple early fusion pipeline.  
MVAE (Khattar et al., 2019) introduced generative modeling for finding joint latent representations across modalities in other approaches. Whereas EANN (Wang et al., 2018) addressed the generalization issue between events by applying adversarial learning to disentangle event-specific features from truth indicators, more recent work like the Multimodal Co-Attention Network (MCAN) applied attention mechanisms to learn fine-grained alignments of image regions and text tokens, with enhanced modeling of cross-modal inconsistencies. Our paper is founded on these foundations by evaluating and comparing different fusion architectures for the fake news detection task. In particular, we consider:  
BERT + EfficientNet (early fusion),  
  
LSTM + MobileNet (late fusion),  
  
CLIP + BERT embeddings (Early Fusion).  
  
By making these comparisons, we aim to shed light on how different integration strategies influence model performance in the detection of fake news in multimodal environments.

**Methodology:**

Now, we have 4 trained different models with different parameters and accuracies. Accordingly, we will combine an image model with a text model by making both late and early fusion. MobileNetv2 and LSTM will be contributing to the late fusion model and others (*EfficientNet* and BERT) to early fusion.

**Late Fusion:**

In fact, we were inspired by Meta’s late fusion code found on their official GitHub repository. Consequently, we used their code to implement our multimodal classification system, combining predictions from independently trained models on different data modalities. We used MobileNetv2 and LSTM in late fusion. Firstly, we made our text preprocessing for model input where we applied concepts like text tokenization, making all sequences the same length (300) by padding, and encoding labels. On the other hand, we have MobileNetv2 where we used image augmentation techniques from resizing to normalizing images to increase the data sample. After that we integrated the code of late fusion of Meta within our pipeline to train the late fusion model. Sadly, it severed overfitting, but we tried our best to handle this by adding drop out layers and adding a decay weight rate in the optimizer. This approch helped in solving a bit of the problem but not completely.

**Early Fusion:**

In our early fusion strategy, we aimed to jointly model visual and textual modalities by integrating their feature representations prior to classification. Specifically, we employed *EfficientNet* (pretrained on ImageNet) to extract high-level semantic features from the image inputs, and *BERT* (pretrained base version) to generate contextualized embeddings from the associated textual data. These two feature vectors—image and text—were then concatenated to form a unified multimodal representation.

To enable joint reasoning across modalities, we introduced a **deep neural network (DNN)** architecture following the fusion step. This network consisted of multiple fully connected layers with nonlinear activations (ReLU), batch normalization, and dropout for regularization. The purpose of this network was to learn complex, nonlinear interactions between the visual and textual features in a shared representation space. The final layer of the DNN was a Sigmoid classifier tailored to our target task.

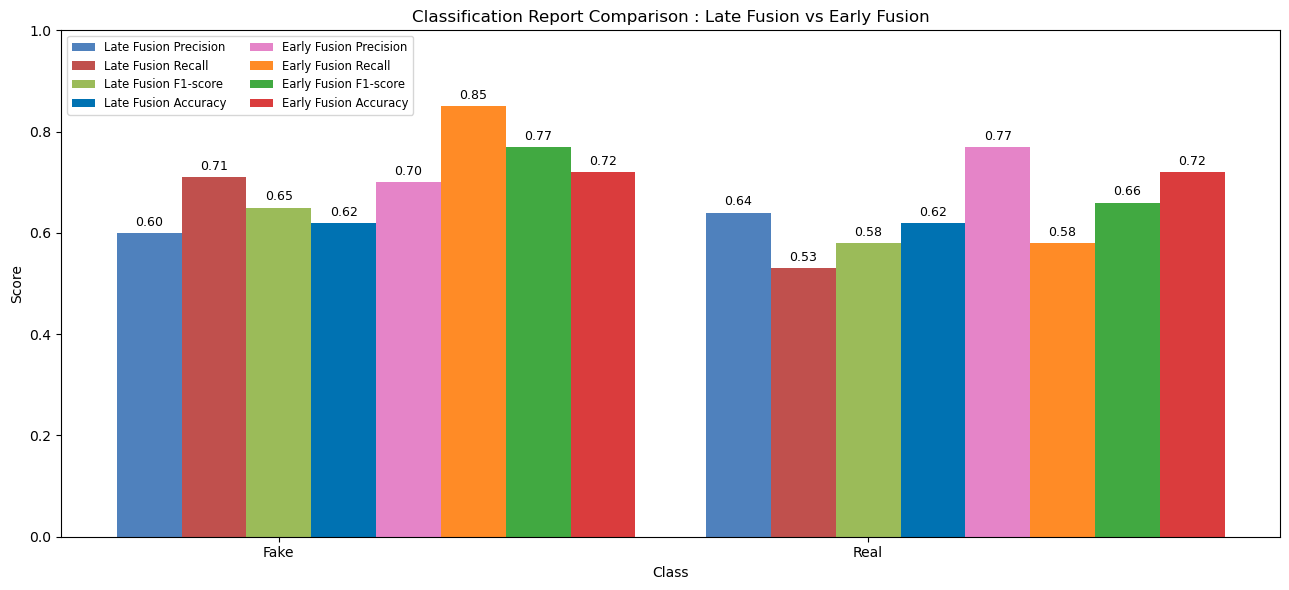
This early fusion design enabled end-to-end learning of joint representations while keeping the modality-specific encoders fixed due to computational constraints. Although this architecture allowed some interaction between modalities, its performance was limited by data scarcity and the lack of fine-tuning of pretrained encoders.

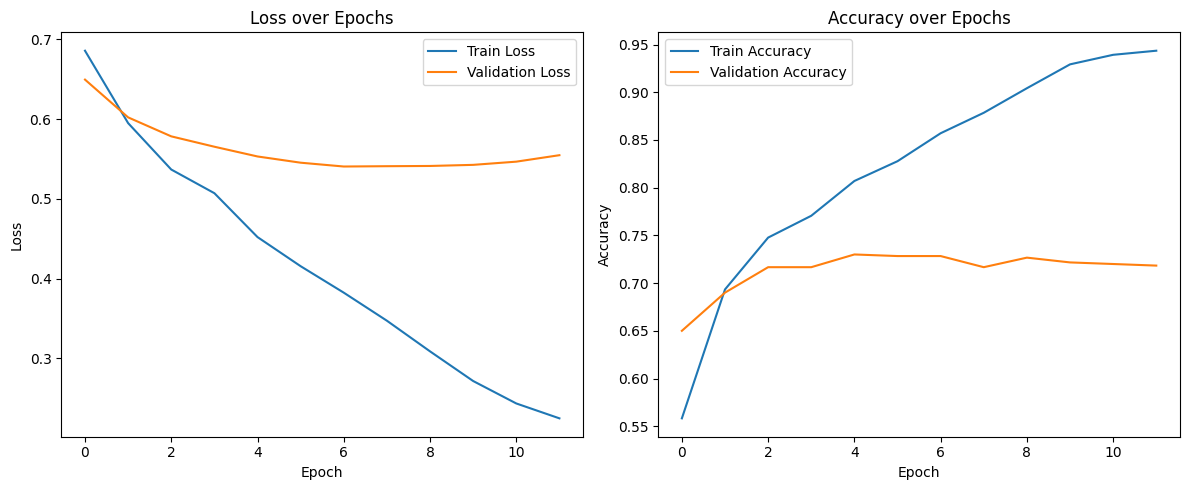
Clip + bert:

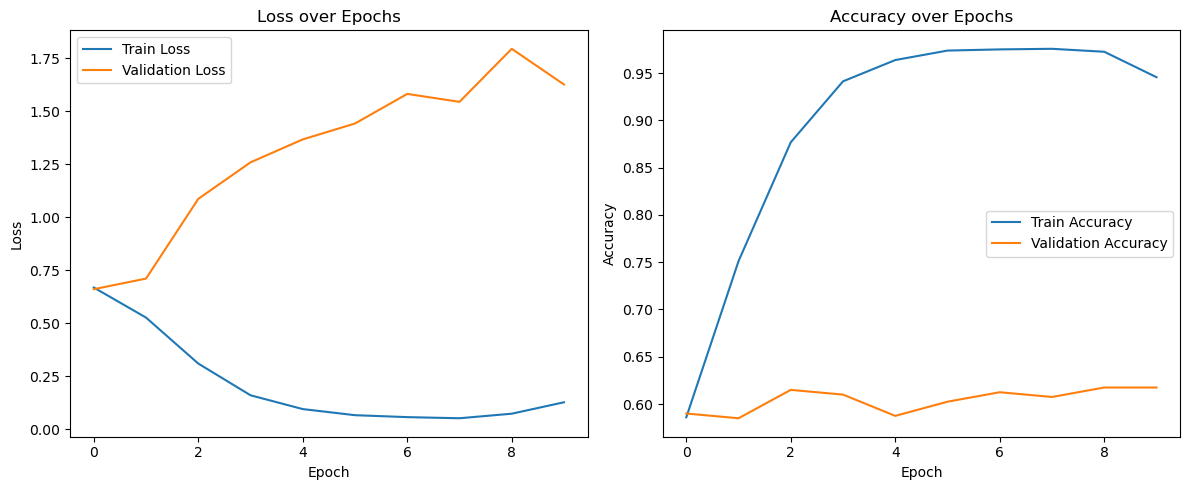
Multimodal features were extracted using two pretrained models: CLIP for images and BERT for text. Images were processed with CLIPProcessor, which automatically resized and normalized inputs before generating 512-dimensional visual embeddings. Texts were truncated to 512 characters and encoded using bert-base-uncased, with the 768-dimensional [CLS] token embedding extracted. To form a unified representation, image and text features were concatenated into a 1280-dimensional vector using an early fusion strategy.  
  
3.3 Classification:  
Fused feature vectors were saved, normalized with StandardScaler, and split into training, validation, and test sets. A feedforward neural network was trained on this data using cross-entropy loss and the Adam optimizer, with a final softmax layer for binary classification.  
  
 Evaluation:  
Model performance was assessed on the test set using standard metrics including accuracy, precision, recall, and F1-score.

Results and Discussion:

Early Fusion somehow outperformed late Fusion as shown in the classification report and loss vs accuracy graph.







Conclusion  
  
**Summary**  
  
In this research, we tackled a multimodal classification challenge by combining visual and textual data. Our preliminary method utilized early fusion, leveraging EfficientNet for extracting image features and BERT for generating textual embeddings, with a combined classifier following. We additionally investigated late fusion by merging the output probabilities from separate unimodal models (MobileNet and LSTM). Nevertheless, both approaches produced less than ideal results, primarily because of the small dataset size (around 2,000 samples), which restricted generalization and the model's resilience. To address this, we integrated a vision-language model (VLM), namely CLIP, which utilizes contrastive pretraining on extensive image-text pairs. This greatly boosted the model's ability to comprehend cross-modal semantics and improved overall precision.   
  
**Limitations**  
  
Even with the performance improvements from CLIP, our method is still limited by various factors. Most importantly, the limited sample size restricted the efficacy of both fine-tuning and training fusion-based architectures, heightening the chance of overfitting. Moreover, the computational resources we had limited our capacity to explore deeper architectures, perform extensive hyperparameter tuning, or use larger batch sizes, which might have enhanced optimization and stability.   
  
**Future Work**  
  
Future efforts should aim at increasing the dataset to facilitate more efficient training of deep multimodal models. Exploring self-supervised or few-shot learning methods might assist in alleviating data shortages. Additionally, investigating more sophisticated fusion techniques, like cross-modal transformers or gated multimodal units, may produce improved integration between modalities. Ultimately, utilizing distributed or cloud infrastructure can assist in overcoming existing computational constraints and allow for more extensive model exploration

References

KaiDMML. (2025). FakeNewsNet (Version latest) [Source code]. GitHub. <https://github.com/KaiDMML/FakeNewsNet>

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models from Natural Language Supervision*. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Singhal, T., Shah, R., & Chakraborty, T. (2019). *SpotFake: A Multi-modal Framework for Fake News Detection*. In *Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 39–47.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). *MVAE: Multimodal Variational Autoencoder for Fake News Detection*. In *Proceedings of the 2019 World Wide Web Conference*, 2915–2921.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). *EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection*. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 849–857.

Wu, Y., Wu, Y., & Huang, Y. (2021). *A Multimodal Co-Attention Network for Fake News Detection*. *IEEE Access*, 9, 22092–22101.