Multi-modal CheapFakes Detection: Cross-Encoder for Fusing Visual and Textual Features

Thao Nguyen, My Dang, Suong Hoang, Dac Nguyen

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Multi-modal CheapFakes Detection: Cross-Encoder for Fusing Visual and Textual Features

Thao T. T. Nguyen[1,2,3,†], My T. Dang[1,2,3,†], Suong N. Hoang[1,2], Dac H. NGUYEN[3]

[1]University of Science, Ho Chi Minh City 700000, Vietnam
[2]Vietnam National University, Ho Chi Minh City 700000, Vietnam
[3]AISIA Research Laboratory, Ho Chi Minh City 700000, Vietnam
Correspondence: 20280087@student.hcmus.edu.vn, 20280067@student.hcmus.edu.vn

Abstract

Detecting CheapFakes, a critical challenge in the era of misinformation, necessitates robust models capable of effectively combining multi-modal information. We present a novel approach that enhances model generalization and accuracy by curating a specialized dataset and introducing an end-to-end framework tailored for this task. Our contributions are as follows: proposing a new dataset emphasizing the specific challenges of CheapFakes detection, developing a Textual Tokens Weighted (TTW) Pooling method, which improves semantic extraction from textual data and boosts classification accuracy, optimizing the multi-head attention mechanism by applying a shared LayerNorm before feature integration, and finally, constructing a Cross-modal Encoder incorporating a co-attention mechanism to effectively fuse visual and textual representations, thereby improving contextual understanding and classification accuracy.

Leveraging Transformer-based architectures, our approach achieves promising results, with an accuracy of 83.80%, F1 score of 84.54%, and recall of 88.60% in classifying the authenticity of image-caption pairs. These findings highlight the potential of our method in advancing multi-modal analysis for misinformation detection.

1 Introduction

The proliferation of CheapFakes, where authentic images are paired with misleading captions, poses a critical challenge in the battle against misinformation. While recent efforts have made strides in fake news detection, such as feature-based machine learning models (Castillo et al., 2011; Kwon et al., 2013; Liu et al., 2015; Biyani et al., 2016) and deep learning methods (Ma et al., 2016; Rashkin et al., 2017; Chen et al., 2018), challenges remain in effectively aligning and combining multi-modal features to enhance classification accuracy.

The emergence of CheapFakes demands new methodologies that extend beyond uni-modal analysis. The COSMOS model (Aneja et al., 2021) marked a significant step forward in out-of-context (OOC) detection by matching captions to image regions and comparing semantic similarities between captions. Building on COSMOS, Tran et al., 2022; La et al., 2022 proposed models that extend the COSMOS framework to tackle both the OOC/NOOC detection (task 1) and the distinction between genuine and fake image-caption pairs (task 2). However, these models rely on rule-based and heuristic approaches and often fail to leverage the full potential of multi-modal data due to a text-side uni-modal bias.

In this work, we introduce a novel end-to-end model that leverages a cross-encoder architecture combined with a co-attention mechanism to enhance the fusion of image and text features. Our model achieves an 83.8% accuracy, marking a 25% improvement over baseline methods that use simple feature concatenation. Thus, it provides a more nuanced understanding of context.

This paper makes several pivotal contributions to the field of CheapFakes detection, highlighted as follows:

1. A specialized dataset is constructed, derived from a detailed analysis of the COSMOS dataset (Aneja et al., 2021), targeting the detection of CheapFakes. This dataset is tailored to capture the intricacies of misleading image-caption pairs, providing a robust foundation for training and evaluation.

2. We introduce a TTW Pooling method that assigns weights to individual tokens, enhancing the extraction of semantic features. Unlike conventional methods, which either focus on

[†]These authors contributed equally to this work. All authors want to thank AISIA Research Lab for supporting us during this paper.

a single token such as the [CLS] token or use mean pooling that treats all tokens equally, our approach captures both local and global contexts, resulting in richer and more nuanced sentence representations.

3. A shared LayerNorm is applied before integrating multi-modal features, ensuring better alignment and reducing feature dispersion, inspired by Brody et al., 2023. This step enhances the stability and effectiveness of the co-attention mechanism that follows, improving overall model performance.

4. We designed a Cross-modal Encoder with a co-attention mechanism (Lu et al., 2019) that facilitates refined interactions between image and text representations by exchanging key-value pairs in multi-headed attention. This bidirectional flow of information allows visual features to inform language representations and vice versa, effectively reducing uni-modal biases and capturing complex relationships between modalities.

2 Methodology

We propose an end-to-end model architecture comprising three main components as shown in Figure 1. We first conduct a uni-modal encoding process, introducing the TTW Pooling technique in the BERT output (Devlin et al., 2019) to transform the raw input into embeddings and extract the essential information from both inputs. Next, to fuse and align the visual and textual features, we design a Cross-modal Encoder inspired by the co-attention mechanism (Lu et al., 2019), which captures and understands the relationship between the two modalities. Finally, we utilize a classification head, specifically a Multi-Layer Perceptron (MLP) (Popescu et al., 2009) architecture with multiple dense layers. The details of our proposed model are elaborated in the following sections.

2.1 Problems statements

In detecting CheapFakes, given a pair of caption $\mathbf{S} = \{s_1, s_2, \ldots, s_n\}$ and an image $\mathbf{I}$, the objective is to identify the given caption and image are misleading information or not.

2.2 Vision-language Encoder

Our approach adopts Transformer-based architectures (Vaswani et al., 2017), harnessing both textual and visual features to detect fake captions in CheapFakes effectively.

For textual feature extraction, we use **BERT** (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), a language model that excels in generating accurate semantic representations by considering both preceding and following words in a sentence. Each input sequence $\mathbf{S}$ is tokenized using byte-level Byte Pair Encoding (BPE) (Sennrich et al., 2016), and segmented into different sentences by [CLS] and [SEP] tokens. The textual input representation is computed as follows:

$$\mathrm{T}_0 = [\mathbf{E}_{cls}; \mathbf{E}_1; \mathbf{E}_2; ...; \mathbf{E}_M; \mathbf{E}_{sep}] + \mathbf{E}_{seg} + \mathbf{E}_{pos} \quad (1)$$

where $T_0 \in \mathbb{R}^{(M+m)\times D}$, $\mathbf{E}$ is the token embedding, M is the total number of tokens, m is the number of special tokens with $m \geq 2$, and D denotes the dimension of the textual encoder. In addition, $\mathbf{E}_{\text{seg}}$, $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(M+m)\times D}$ are respectively the segment embeddings and position embeddings. The output generated by the pre-trained model in this process is the last hidden state donated as $\mathbf{T} \in \mathbb{R}^{(M+m,768)}$, which serves as a comprehensive and meaningful representation of the text content:

$$\mathbf{T} = \text{Encoder}_t(T_0) \quad (2)$$

We utilize the pre-trained **ViT-B/16-224-21k** model (Dosovitskiy et al., 2021) as our visual encoder for image feature extraction. For a 2D image input $\mathbf{I}$ with varying dimensions $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, where H and W represent the height and width of the image, and C is the number of image channels. Initially, we convert the input to an RGB image and resize it to normalized pixel dimensions. The image is then divided into smaller patches $I_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches, which are embedded and fed into the transformer model for processing. The visual input representation is computed as follows:

$$V_0 = [I_{class}; I_p^1\mathbf{E}; I_p^2\mathbf{E}; ...; I_p^N\mathbf{E}] + \mathbf{E}_{pos}, \quad (3)$$

where $V_0 \in \mathbb{R}^{(N+1)\times D}$ and $\mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D}$ is the linear projection. Similar to BERT, ViT incorporates a [class] token at the start of the patch sequence and utilizes learnable 1D positional embeddings, $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$, where D denotes the dimension of the visual encoder. The output of

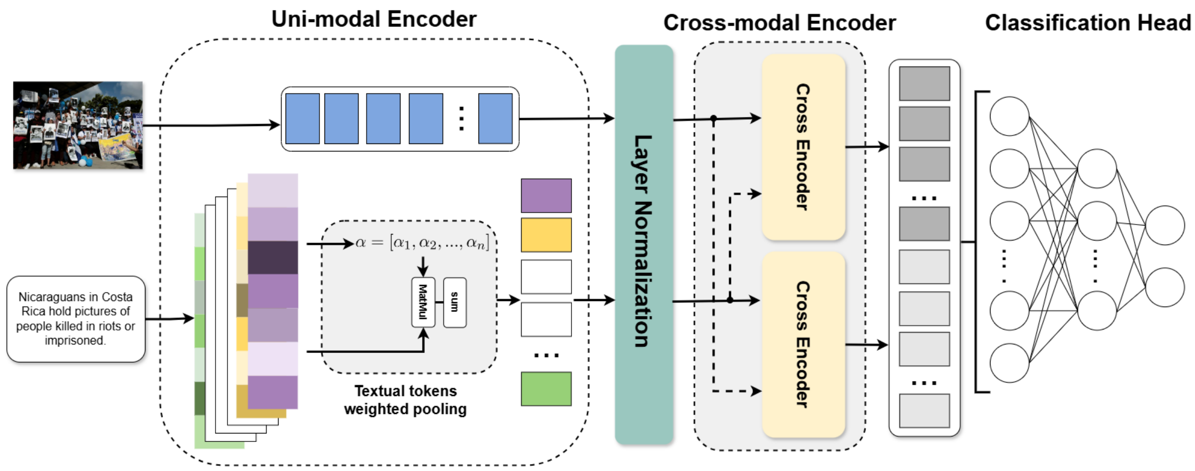

Figure 1: **Overview of our model**

the visual encoder aggregates information from all patches, producing a unique feature vector denoted as $\mathbf{V} \in \mathbb{R}^{(1,768)}$ to represent the global characteristics of the image:

$$\mathbf{V} = \text{Encoder}_v(V_0) \tag{4}$$

2.3 Textual Tokens Weighted Pooling

To synthesize a comprehensive representation of the entire sentence from the original token representations, we propose TTW Pooling as a pooling operation that captures both local and global information from the data. This approach addresses the limitations of traditional pooling techniques, which often struggle to synthesize semantic representations effectively. For instance, Pooler Output (Devlin et al., 2019) relies solely on the representation of the [CLS] token, overlooking valuable information from other tokens in the sentence. Meanwhile, Mean Pooling averages all tokens without differentiating their importance, which can result in the loss of crucial details. By employing TTW Pooling, we aim to enhance the model's ability to generate richer and more meaningful representations.

As shown in Figure 2, TTW Pooling consists of two phases: (1) performing the interpolation process to evaluate the importance of each token in a sequence and (2) aggregating the important information from the output sequence. Firstly, we transform the embeddings of each token q_i from a sequence $\mathbf{T}$ through a fully connected layer, converting the original feature space into a higher-dimensional space. After this transformation, applying the tanh activation function helps normalize the output values and smooth their distribution, mitigating the vanishing and exploding gradient problems during training. Subsequently, a linear layer is applied to compute the attention scores a_i for each token q_i. These scores a_i measure the importance of each token in the data sequence and are normalized into attention weights α_i using the softmax function, concluding the first phase:

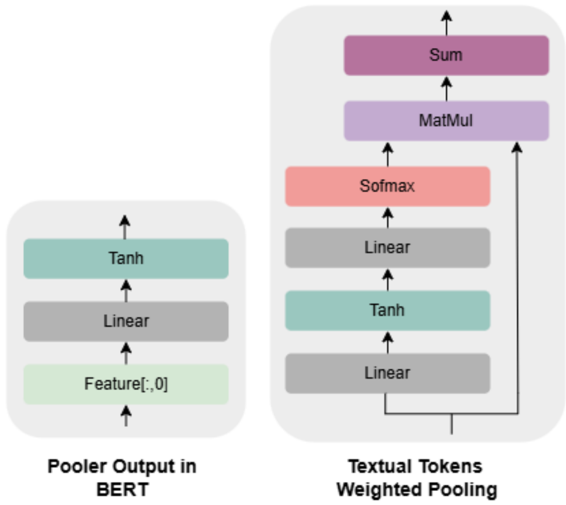

Figure 2: **Comparative Analysis of Pooler Output in BERT and Textual Tokens Weighted Pooling**.

$$\alpha = [\alpha_1, \alpha_2, ..., \alpha_n] \text{ with } \sum_{i=1}^{n} \alpha_i = 1 \tag{5}$$

In the second phase, the attention weights α_i are multiplied by the original features q_i to obtain the context vectors v_i. These vectors v_i are then aggregated to create a composite representation for the entire sentence as $\mathbf{T}^{+} \in \mathbb{R}^{(1,768)}$. This composite representation not only integrates information from individual tokens but also encapsulates the most significant aspects of the sentence. As a result, it enhances the model's ability to capture semantic meaning, thereby improving performance in various natural language processing tasks.

$$\mathbf{T}^{+} = \sum_{i=1}^{n} \alpha_i \cdot q_i \tag{6}$$

2.4 Unimodalities Integration

2.4.1 Layer Normalization

Layer Normalization (LayerNorm) (Ba et al., 2016) is crucial in Transformer architectures, optimizing performance and ensuring stability. Recent work by Brody et al., 2023 reveals a deeper role of LayerNorm in enhancing the representational capacity of the multi-head attention mechanism. Specifically, the projection of input vectors into a $(d-1)$ dimensional space orthogonal to $[1, 1, \ldots, 1]$, and the scaling of vectors to a norm of $(\sqrt{d})$, allows the attention mechanism to evenly attend to all keys, preventing any key from becoming "un-selectable". This nuanced understanding expands beyond the conventional view of LayerNorm as a mere normalization step during forward propagation and gradient flow.

Inspired by these insights, we implement a shared LayerNorm for both text and image features before the Cross-modal Encoder. By normalizing across different domains, we align and integrate the features into a unified representation space, optimizing the attention mechanism and enhancing the model's ability to learn important relationships. This approach also reduces the number of parameters, improving performance and accelerating convergence.

2.4.2 Cross-modal Encoder

We observe that previous approaches often exhibit a bias in attention, primarily focusing on text while failing to fully exploit the potential of visual information. Therefore, we have designed a Cross-modal Encoder consisting of two main components: Image cross-encoder block and Text cross-encoder block. The core idea is to implement a co-attention mechanism (Lu et al., 2019), where these two cross-encoder blocks interact through multi-head attention. Specifically, this interaction happens when key-value pairs, possessed by multi-head attention (Vaswani et al., 2017), are exchanged between the blocks to strengthen the connection between text and image.

This structure uses distinct parameters for each modality (text and image), allowing the model to focus on the critical parts of the data and calculate attention weights for each source of information. A notable feature is its ability to share parameters between the two branches, including weights and biases. This not only enables the model to construct a shared representation space for both modalities but also allows it to automatically identify and focus on the important aspects of both text and image simultaneously, resulting in robust and informative joint representations. The superiority of the co-attention mechanism is demonstrated through comparisons with other attention mechanisms, highlighting its enhanced performance, particularly in transformer-based cross-modal encoding (Hendricks et al., 2021).

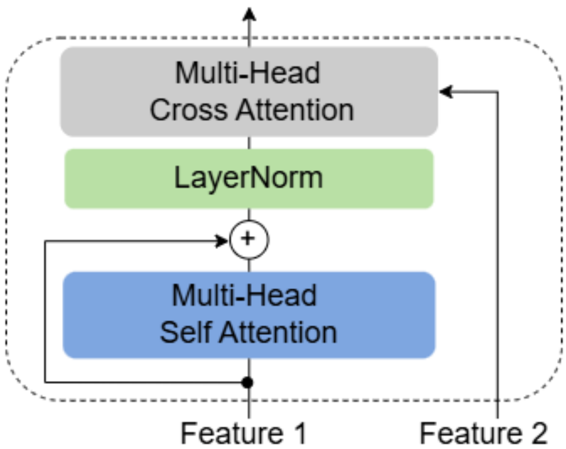

Figure 3: **Cross-encoder architecture**

In terms of functionality, the Text cross-encoder block queries textual features based on visual information, where the importance of features within the text is determined. Image cross-encoder block carries out an evaluation, utilizing language information, on the visual features. These two modules work in tandem to fully leverage the information from both text and images, enhancing the model's understanding of multi-modal data through a multi-head cross-attention mechanism. To further enhance our model's performance and optimize the operation of the multi-head cross-attention mechanism, we incorporate an additional forward pass that integrates multi-head self-attention (Vaswani et al., 2017; Luong et al., 2015). This technique allows the model to automatically identify and focus on the most critical features, thereby improving the precision of information transmission through the primary layer, which employs 24 attention heads for both layers. Moreover, we integrated a residual connection following the self-attention layer to achieve optimal convergence, as proposed by He et al., 2015. However, the result of this addition may exhibit different and inconsistent distributions. To address this, we apply layer normalization (Ba et al., 2016) to standardize the output distribution, ensuring it remains within a consistent range and uniformly distributed:

$$LayerNorm(x + att(x)), \tag{7}$$

where $att()$ is the multi-head self-attention. This approach mitigates the vanishing gradient problem and enhances gradient flow through the network, resulting in a more stable and efficient training process.

2.5 Classification Network

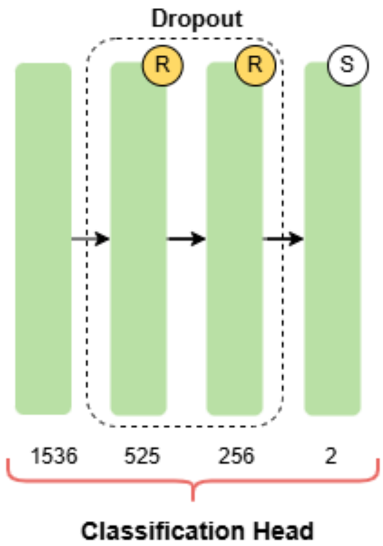

Figure 4: **Architecture of the Classification Head**. The classification head consists of linear layers, ReLU and Softmax activation functions, and dropout regularization. Note: R and S denote ReLU and Softmax activation functions, respectively

The classification head, illustrated in Figure 4, is a crucial component of a classification network, particularly for tasks such as CheapFake detection. Its primary purpose is to transform the high-dimensional features extracted by the preceding layers into actionable class probabilities. This transformation is achieved through a series of fully connected layers that process the learned representations, followed by activation functions such as ReLU (Nair and Hinton, 2010) and Softmax, which introduce non-linearity to the model. Additionally, regularization techniques like Dropout (Srivastava et al., 2014) are employed to mitigate overfitting and enhance the model's generalization capabilities. By mapping the processed features to specific class probabilities, the classification head facilitates accurate predictions, thereby playing a crucial role in the model's overall effectiveness in classifying inputs, especially in the context of detecting CheapFakes

3 Dataset

To address the novelty of this task, we have developed a specialized dataset comprising image-caption pairs where the images are authentic, but the captions are intentionally misleading. Data was gathered from different sources, ensuring consistent quality and format throughout all entries. The dataset is exclusively in English.

3.1 Data Collections

The COSMOS dataset (Aneja et al., 2021) serves as a foundational resource, consisting of 200K images and 450K textual captions obtained from various news channels and the fact-checking website Snopes. This dataset is designed to differentiate between out-of-context (OOC) and not-out-of-context (NOOC) scenarios.

COSMOS presents a challenge for detecting misinformation because the visual content itself is not manipulated; rather, misleading or false information arises from the combination of the image and its caption. Building upon COSMOS, we have constructed a tailored dataset to assess the authenticity of image-caption pairs. This dataset is further augmented with data from Snopes.com[†], a prominent fact-checking website that combats misinformation by investigating various news stories. Our dataset includes image-caption pairs from Snopes, focusing on examples categorized as False, Miscaptioned, Mixture, and True, with each sample consisting of an image paired with its corresponding Claim statement, which serves as the caption.

To enhance the diversity and robustness of our dataset, we also generated captions using ChatGPT. After exploring methods like random selection and using the Faker package, which proved ineffective, we utilized ChatGPT by providing it with an image description and a real caption as prompts. This approach allowed us to create a wide variety of fake captions, significantly improving the overall effectiveness of the dataset.

[†]https://www.snopes.com/

3.2 Data Sources

Train Set: The training dataset was constructed through several sampling methods to ensure a diverse and representative collection of image-text pairs. We resampled from the COSMOS and Ookpik (Pham et al., 2024) datasets and collected data from Snopes.com. To enhance variability, we generated synthetic fake captions using ChatGPT, resulting in a final training set of approximately 6,348 image-text pairs.

Test Set: The test set, comprising 1,000 samples, was derived from the COSMOS test set. For our evaluation, we paired each image with Caption 1 and assigned a label of 0 (real) if the caption aligns with a NOOC (Not Out-of-Context) scenario and 1 (fake) if it corresponds to an OOC (Out-of-Context) scenario.

4 Experiments

4.1 Experimental settings

We split the data into training, validation, and test sets, with the training data divided using an **80/20** ratio for training and validation. The model was then evaluated on the test set. For preprocessing, we set the maximum sequence length for text based on the longest sequence in each batch, converted images to RGB format, and used a batch size of 32.

Our **Baseline** (Figure 5) model includes a pre-trained **BERTBASE** text encoder (110 million parameters) and a **ViT-B/16** image encoder (86.6 million parameters). We concatenated features from both encoders and used a classifier to predict labels (0 or 1).

The training was conducted using PyTorch and GPU resources, with the Adam optimizer set at a learning rate of **1e**$^{-5}$. The entire process took over 2 hours.

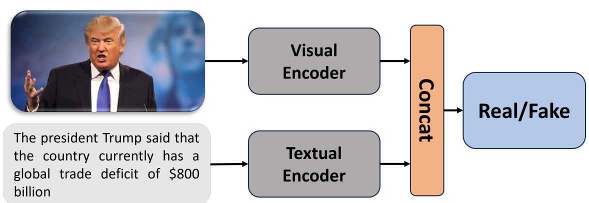

Figure 5: Baseline model

4.2 Model training

To train the model, we utilize the cross-entropy loss function (de Boer et al., 2005), defined as follows:

$$\mathcal{L}(y, \hat{y}) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_i^j \log \hat{y}_i^j, \tag{8}$$

where n represents the number of training samples, m is the number of labels (in our case, $m = 2$, y is the ground truth captions, $\hat{y}$ is the predicted captions.

4.3 Evaluation metrics

Accuracy (acc): The proportion of correctly predicted pairs (both real and fake) out of the total predictions.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Precision (pre): The ratio of correct predictions among all predictions classified as fake indicates the reliability of the model in predicting fake instances.

$$\text{precision} = \frac{TP}{TP + FP} \tag{10}$$

Recall (rec): The ratio of correct predictions among all actual fake instances reflects the model's ability to detect fake instances.

$$\text{recall} = \frac{TP}{TP + FN} \tag{11}$$

F1-Score (f1): The harmonic mean of precision and recall. It measures the model's ability to classify image-text pairs accurately while ensuring that few fake pairs are missed.

$$f1 = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{12}$$

- **TP (True Positive):** The number of image-text pairs predicted as fake and are fake.
- **TN (True Negative):** The number of image-text pairs predicted as real and are real.
- **FP (False Positive):** The number of image-text pairs predicted as fake but are real.
- **FN (False Negative):** The number of image-text pairs predicted as real but are fake.

Model	**acc**	**f1**	**pre**	**re**	**params (M)**	**size (MB)**
Baseline	0.588	0.572	0.572	0.618	196	787.16
Our model	0.838	0.845	0.808	0.886	211	845.07

Table 1: Comparison with Baseline model

Model	**acc**	**pre**	**re**	**f1**	**params (M)**	**size (MB)**
num_head = 24						
Our model + pooler output	0.795	0.808	0.772	0.791	208	835.61
Our model + mean pooling	0.806	0.789	0.836	0.812	208	835.61
Our model + TTW pooling	**0.838**	0.808	0.886	**0.845**	211	845.07
num_head = 12						
Our model + mean pooling	0.797	0.870	0.698	0.775	208	835.61
Our model + TTW pooling	0.829	0.786	0.904	0.841	211	845.07

Table 2: Evaluation of different pooling methods

4.4 Evaluation of Proposed Techniques

The results in Table 1 provide compelling evidence of the effectiveness of our proposed enhancements - Textual Tokens Weighted Pooling shared LayerNorm, and the co-attention mechanism within the Cross-modal Encoder. These components collectively drive significant improvements over the baseline, which simply concatenates image and text features. Our approach achieves an accuracy of 83.8%, marking a substantial leap of 25% compared to conventional methods that rely on rudimentary feature fusion. This underscores the potency of our model in synthesizing multi-modal information with greater precision and depth.

TTW Pooling stands out by directing attention to the most salient features in both local and global contexts. Unlike traditional pooling methods, TTW selectively amplifies important tokens, refining the semantic representations and contributing to more effective sequence modeling.

The shared LayerNorm further fortifies the alignment between text and image features by normalizing them into a unified feature space. This alignment is crucial for optimizing the attention mechanism within the Cross-modal Encoder, facilitating the seamless integration of multi-modal data. Consequently, this leads to accelerated convergence during training and yields enhanced model robustness.

Lastly, the co-attention mechanism within the Cross-modal Encoder represents a vital advancement in bridging the gap between textual and visual information. By enabling direct cross-modal attention, the model constructs a cohesive representation that captures the critical aspects of both modalities, driving improved accuracy and classification performance.

These results highlight the impact of our proposed methods in advancing the state-of-the-art in multi-modal data processing, demonstrating their effectiveness in achieving superior performance over conventional approaches.

4.4.1 Impact of Textual Tokens Weighted Pooling

Our experiments, as shown in Table 2, demonstrate that TTW Pooling consistently outperforms Mean Pooling and Pooler Output by effectively emphasizing key input features. By generating a weight matrix that accentuates the importance of critical tokens, TTW Pooling delivers a richer and more precise representation of input sequences. This leads to a tangible improvement in model performance, underscoring the significance of pooling strategies in capturing and amplifying essential information within the data.

Moreover, fine-tuning the number of attention heads (num_head) emerges as a critical factor in optimizing model performance. A judicious selection of this parameter not only enhances model efficiency but also mitigates overfitting and bolsters generalization.

Model	acc	pre	re	f1
shared layer norm	0.838	0.808	0.886	0.845
non-shared layer norm	0.787	0.770	0.818	0.793

Table 3: Effect of LayerNorm on feature alignment and model performance

Feature Extraction Model	acc	pre	re	f1	params (M)	size (MB)
ResNet50 BERT	0.792	0.813	0.759	0.759	150	600.560
ResNet101 BERT	0.773	0.805	0.720	0.760	169	676.528
ResNet152 BERT	0.772	0.803	0.720	0.759	184	739.103
EfficientNet-b0 BERT	0.776	0.830	0.678	0.746	129	519.486
EfficientNet-b4 BERT	0.783	0.853	0.686	0.756	143	575.698
EfficientNet-b7 BERT	0.802	0.817	0.775	0.792	190	763.723
ViT BERT	**0.838**	0.808	**0.886**	**0.845**	211	845.07
ViT ROBERTa	0.789	0.794	0.780	0.787	226	905.73

Table 4: Comparison of Feature Extraction Models

4.4.2 Impact of LayerNorm

Table 3 highlights that the application of a shared LayerNorm significantly enhances model performance. By normalizing the different modalities together, the shared LayerNorm fosters a stronger alignment of features, thereby improving the model's ability to effectively capture and leverage the relationships between text and image data. On the other hand, the non-shared LayerNorm may impede this integration, as it treats each modality independently, potentially leading to less optimal performance.

4.4.3 Impact of feature extraction models

In the evaluation of feature extraction models (Table 4), the combination of ViT-B/16 and BERT-BASE demonstrated the best performance, achieving an accuracy of 83.80% and an F1 score of 84.54%. The Vision Transformer (ViT) excels in processing entire images through self-attention mechanisms and enables a more comprehensive understanding of spatial relationships in images, surpassing CNN-based models like ResNet and EfficientNet, which are limited by localized convolutional operations.

Additionally, while RoBERTa is a larger and more powerful model compared to BERT, its combination with ViT did not yield superior performance. This suggests that moderate-sized models like ViT and BERT may offer better performance in many scenarios due to their optimized balance of complexity, generalization capabilities, and reduced risk of overfitting.

5 Conclusion

In this paper, we address the challenge of CheapFakes detection by introducing an advanced end-to-end model that effectively integrates image and text features through a Cross-modal Encoder with a co-attention mechanism. This allows for refined interactions between visual and textual data. To further enhance the extraction of fine-grained and comprehensive information from text, we introduce TTW Pooling within BERT's output. We also clarified the role of LayerNorm in the Transformer's attention mechanism. By applying LayerNorm before multi-modal feature fusion, we standardize the uni-modal features into a coherent space, enhancing the model's ability to discern critical relationships between modalities. Ultimately, we have constructed a new dataset that encompasses a broader range of fake caption cases. This dataset expansion improves the model's performance and provides a richer resource for future research in this domain.

While the test set results remain limited, we believe our contributions offer valuable insights and advancements in the field of CheapFakes detection. In the future, we intend to further refine our approach, investigate cutting-edge techniques and large language models (LLMs), and expand our evaluation framework to enhance the effectiveness and robustness of CheapFakes detection methods. A significant aspect of our future work involves expanding the dataset to include additional languages, such as Vietnamese, to ensure the model's applicability across diverse linguistic contexts.

References

Shivangi Aneja, Christoph Bregler, and Matthias Nießner. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *ArXiv*, abs/2101.06278.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. 2016. " 8 amazing secrets for getting more clicks": detecting clickbaits in news streams using article informality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Shaked Brody, Uri Alon, and Eran Yahav. 2023. On the expressivity role of LayerNorm in transformers' attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14211–14221, Toronto, Canada. Association for Computational Linguistics.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22*, pages 40–52. Springer.

P. T. de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134:19–67.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Lisa Anne Hendricks, John F. J. Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE.

Tuan-Vinh La, Minh-Son Dao, Quang-Tien Tran, Thanh-Phuc Tran, Anh-Duy Tran, and Duc-Tien Dang-Nguyen. 2022. A combination of visual-semantic reasoning and text entailment-based boosting algorithm for cheapfake detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7140–7144.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1867–1870.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*.

Kha-Luan Pham, Minh-Khoi Nguyen-Nhat, Anh-Huy Dinh, Quang-Tri Le, Manh-Thien Nguyen, Anh-Duy Tran, Minh-Triet Tran, and Duc-Tien Dang-Nguyen. 2024. Ookpik- a collection of out-of-context image-caption pairs. In *MultiMedia Modeling*, pages 132–144, Cham. Springer Nature Switzerland.

Marius-Constantin Popescu, Valentina Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. 2009. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Quang-Tien Tran, Thanh-Phuc Tran, Minh-Son Dao, Tuan-Vinh La, Anh-Duy Tran, and Duc Tien Dang Nguyen. 2022. A textual-visual-entailment-based unsupervised algorithm for cheapfake detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7145–7149.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.