Multimodal News Classification For Vietnamese

Phan Phuoc Loc Ngoc [1,2], Le Tan Loc [1,2],
[1]Science and Information Technology
[2]University of Information Technology

Abstract

Multimodal News Classification requires the integration of textual and visual information to achieve high accuracy. In this study, we propose a novel approach that employs PhoBERT to extract features from text and Vision Transformer (ViT) to extract features from images. Additionally, we develop a Vietnamese dataset comprising text-image pairs across various categories and fields.

Besides introducing and building this new dataset, we evaluate the performance of our approach on the OK-ViWikiText-ImageCaps dataset that we curated. This study focuses on modeling contextual information from the collected text and images, aiming to provide suitable recommendations for multimodal classification tasks in Vietnamese. Experimental results demonstrate that our model significantly outperforms methods relying solely on text or images, particularly benefiting news-related tasks that demand visually informative content.

This research represents a novel direction in applying PhoBERT and ViT to natural language processing tasks, specifically for Vietnamese data. The diversity and comprehensiveness of the new dataset provide broad coverage for multimodal news classification tasks and open new avenues for research in artificial intelligence and machine learning.

1 Introduction

With the rapid advancement of artificial intelligence and new natural language processing models, multimodal news classification has emerged as a fascinating and highly prioritized research topic in recent times. This task demands the development and configuration of models capable of fully exploiting the textual and visual information provided.

However, previous studies have often gone astray in framing the problem. They focus predominantly on textual features while neglecting crucial aspects in images, thereby overlooking significant information. This results in a substantial drop in the effectiveness of contemporary news classification systems.

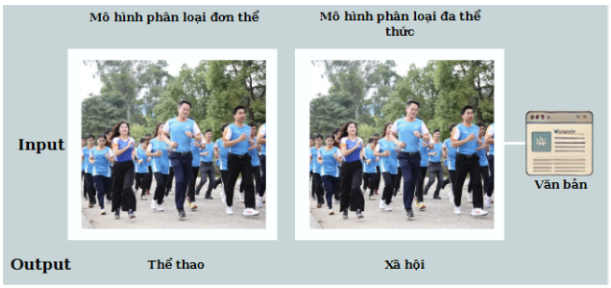

Figure 1: The difference between the text-based method and the combined method

In this study, we propose a completely novel approach by integrating a feature extraction model specialized for Vietnamese language—PhoBERT—with ViT, which extracts features from images. PhoBERT, built upon the Transformer architecture and fine-tuned for Vietnamese, effectively captures complex semantic features in text. ViT leverages self-attention mechanisms to analyze interrelationships within images. This novel combination demonstrates superior potential over previous models by enhancing the extraction of meaningful features for precise recommendations.

The OK-ViNewsText dataset, a self-curated resource, supports this research by encompassing a rich and diverse repository of Vietnamese news data. This dataset ensures coverage of complex and rare cases applicable to our study. We conduct experiments, evaluations, and comparisons using this dataset to validate the efficacy of our proposed model against text-only or image-only approaches.

Our results underline the significant advantages of integrating PhoBERT and ViT for news classification tasks, particularly in addressing visual and contextual requirements. This research paves the way for future advancements in multimodal natu-

ral language processing and artificial intelligence tailored to Vietnamese applications.

Our objectives include:

- Collecting data on journalism and electronic news from reputable Vietnamese online news platforms to create the comprehensive OK-ViNewsText dataset, which serves multiple tasks in the field of natural language processing.
- Utilizing the collected Vietnamese dataset to evaluate and measure the effectiveness of various training approaches.

2 Related work

Multimodal News Classification is one of the research fields that has garnered considerable attention from researchers worldwide due to its immense potential and significance in natural language processing and computer vision. Existing studies predominantly focus on the English language, with most language processing models being primarily designed for English. These studies typically concentrate on two main aspects: datasets and model design.

Regarding the datasets used in studies related to text news classification, such as 20NEWS (Lang, 1995) or AG News (Zhang et al., 2015), they provide a rich amount of data with a variety of text categories suitable for experimental models. However, these datasets do not contain accompanying image information with the available text pairs, an important element in modern news that requires high objectivity. Some datasets have combined images and text, such as Icecap (Hu et al., 2020a) and News Image-Text Matching (Zhao et al., 2021), which have partially helped improve performance in the process of analyzing news through both text and images. A significant limitation when using these datasets is that most of them are primarily focused on English, while Vietnamese data remains limited and lacks standardization.

In modeling, there are several studies primarily focusing on methods for fusing individual features. However, methods such as Multimodal Fusion Models (Ngiam et al., 2011) or attention mechanisms like MCAN (Yu et al., 2019) have demonstrated their effectiveness in handling both text and image information. Recently, attention-based models like those in Blip-2 (Li et al., 2023b) have significantly improved the performance of multimodal tasks. However, most of these studies have not been optimized for the Vietnamese language, or they lack a tightly integrated approach between text and images in context.

By reviewing previous studies and building on new ideas, the lack of Vietnamese datasets and optimal multimodal integration methods poses a significant challenge during implementation. This study addresses these limitations by:

- Introducing the OK-ViWikiText dataset, which includes both text and images collected from reputable news platforms in Vietnam.
- Utilizing PhoBERT to extract features from Vietnamese news text and Vision Transformer (ViT) to extract features from images.

Compared to previous methods and studies, this approach addresses more challenges and provides a more comprehensive solution. It enhances the effectiveness of analysis in multimodal news classification while paving the way for new research directions in natural language processing and computer vision for the Vietnamese language.

3 Dataset

In Vietnam, one of the major challenges in researching and developing natural language processing applications is the lack or absence of publicly available and standardized datasets for online journalism. Compared to other languages, such as English, which already have numerous electronic news datasets accessible to the research community, Vietnamese still lacks large and specialized datasets in this field. This shortfall hinders the development of machine learning models and natural language processing applications, especially in tasks such as information extraction, content analysis, or creating intelligent systems related to journalism.

The absence of publicly available datasets limits scientific research, as research teams are required to collect and build datasets from scratch, which is both time-consuming and resource-intensive. Furthermore, data from Vietnamese online news platforms are highly fragmented. Each news outlet has its own way of presenting information and formatting, from title structures and article content to the organization of sections such as comments and multimedia (images, videos). Collecting and standardizing data from these diverse sources requires a complex processing workflow to ensure

consistency across articles and avoid data mismatches. Therefore, creating a standardized Vietnamese electronic news dataset not only addresses current challenges in research but also opens up significant opportunities for the development of natural language processing technologies and artificial intelligence in Vietnam.

Additionally, constructing a dataset for journalism and electronic news requires sourcing data from reliable and reputable providers to meet the initial objectives effectively. The data collection process also faces challenges related to time constraints and the lack of historical data. Online news platforms typically only provide access to recent articles, with older content often being removed or restricted. This creates difficulties for studying long-term trends or analyzing the historical development of Vietnamese journalism.

Methodology: Constructing the dataset involves the following steps:

Phase 1: Searching and selecting online news platforms. To build a reliable dataset that aligns with the research objectives, we conducted a search and selected reputable online news platforms in Vietnam, aiming to collect approximately 40,000 images.

After thorough evaluation and selection, we decided to focus on five reputable online news websites, including Dân Trí, Đời sống và Pháp Luật, Thanh Niên, Lao Động, and VNExpress, along with their available sections.

Phase 2: Raw data collection. After identifying the online news platforms, we analyzed and reviewed the available categories on these platforms to determine the content to be collected for building the dataset. Online news platforms typically feature diverse content contributed by various journalists. Articles containing videos or advertisements were excluded during the preprocessing stage. All articles from these platforms were collected, covering data available up to 2023.

Once the objectives were defined, we examined the structure of each online news platform, including the collection of URLs for the platforms, links to specific sections, and individual articles, to determine the tools needed for data extraction. Data collection was conducted using Python, leveraging the BeautifulSoup4 library to extract information such as text, images, image captions, article categories (fields), and links to the articles. Images were stored separately, while the remaining information was organized and saved in JSON format.

After collecting the necessary data, a thorough review and filtering process was conducted to remove irrelevant data for the research. This process included:

1. Removing articles or news items without images or with incomplete images due to outdated content or insufficient updates from journalists. Articles containing videos or those related to advertisements with unique structures were also excluded from the dataset.

2. iminating duplicate articles caused by collecting data across multiple sections. Excess information in image descriptions (e.g., "Illustration: Pixabay") was also filtered out.

3. Harmonizing the categories (fields) across different online news platforms, such as Sports, Entertainment, etc., by merging them into broader, unified categories. This resulted in a complete dataset comprising 20 major categories from all the aggregated platforms.

4 Proposed Method

In the Multimodal News Classification task, news data typically consists of two main components: descriptive text and illustrative images. Each type of data provides a different perspective of information, and effectively leveraging them can enhance classification performance. Therefore, we implemented three different methods to evaluate the impact of each data type as well as their integration:

- Using text data only (Text-Only Model).
- Using image data only (Image-Only Model).
- Combining both text and image data (Multimodal Model).

4.1 Text-Only Model

Objective: Evaluate the classification capability based solely on text.

Implementation Steps:

- The text in each data sample is processed and fed into PhoBERT—a pre-trained language model specialized for Vietnamese (Nguyen and Nguyen, 2020). PhoBERT is capable of learning and extracting semantic features from the text.

Figure 2: Data collection process

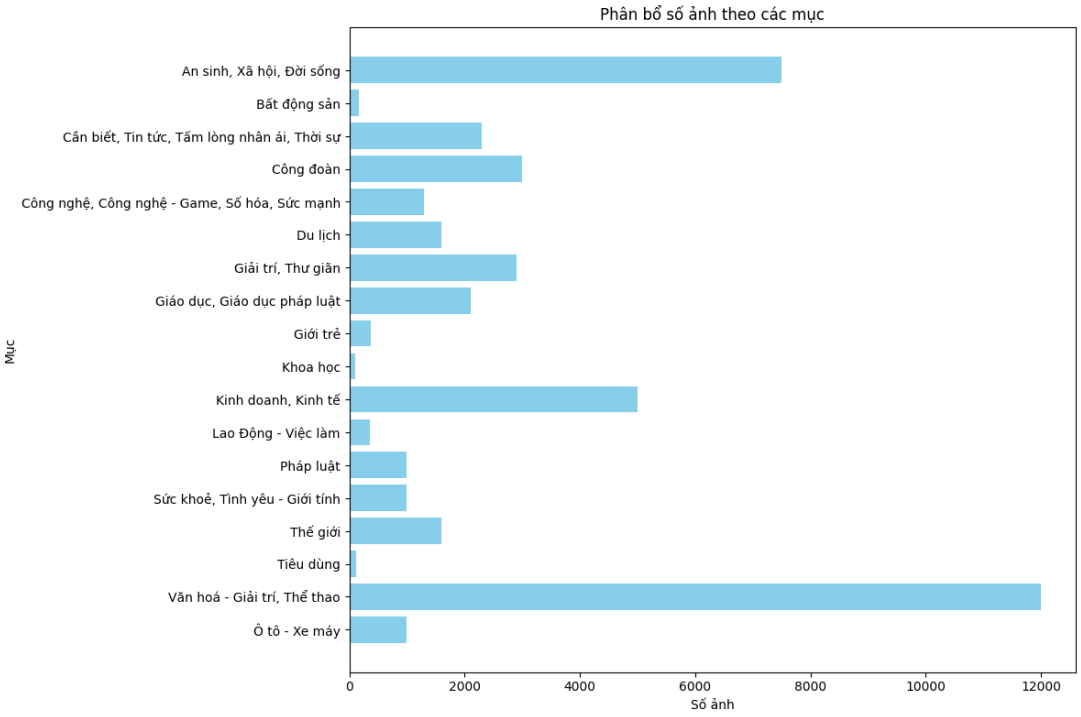

Figure 3: Summary of the collected items from reputable online newspapers in Vietnam

Field	Description and Example
Id	The unique ID of the news item.
Images - URL - Path - Caption	A list of related images, each with the following attributes: The URL of the image. The file name of the image after download. The caption of the image.
Context	A list of text segments describing the news content. Example: • "Tại sao máy tính của bạn chạy ngày càng chậm" • "Trả lời những câu hỏi dưới đây để hiểu về các nguyên nhân khiến máy tính trở nên ì ạch và những cách khắc phục."
link	URL link to the original article. Example: https://vnexpress.net/tai-sao-may-tinh-cua-ban-chay-ngay-cang-cham-3502173.html.
section	Thể loại của bài báo. Ví dụ: Công nghệ.

Table 1: JSON file structure for Multimodal News Classification

- The embedding vector obtained from PhoBERT is passed through a simple Multi-Layer Perceptron (MLP) layer for classification.

The method utilizing PhoBERT leverages the context and linguistic structure of Vietnamese, enabling deep extraction of semantic features from news text. However, relying solely on text may overlook critical illustrative information provided by images.

4.2 Image-Only Model

Objective: Evaluate the classification capability based solely on image data.

Implementation Steps:

- The images in each data sample are processed and fed into the ViT (Vision Transformer) model—an advanced architecture for image processing. ViT works by dividing images into small patches and utilizing the self-attention mechanism to learn image features.
- The embedding vector obtained from ViT is passed through a Multi-Layer Perceptron (MLP) layer for classification.

ViT has the ability to learn global features from images, making it particularly suitable for information-rich images such as illustrations in news articles. However, in the context of article classification, ignoring the semantic information from text can still lead to errors.

4.3 Multimodal Model

Objective: Combine both textual and visual data to utilize the full multimodal information in the problem.

Implementation Steps:

- Text is fed into PhoBERT to extract embedding vectors, while images are processed through ViT to extract corresponding embeddings.
- The two embedding vectors (text and image) are combined using concatenation. This method directly merges the two vectors into a single combined vector, containing information from both modalities. The combination will be evaluated using various approaches to determine the best performance.
- This combined vector is passed through an MLP layer for classification. The MLP layer learns interactions between textual and visual features to optimize the classification process.

Combining both text and images allows for more comprehensive information extraction, minimizing the risk of information loss when relying on a single data modality. This approach is particularly suitable for the multimodal nature of news, where both text and images play crucial roles in conveying content.

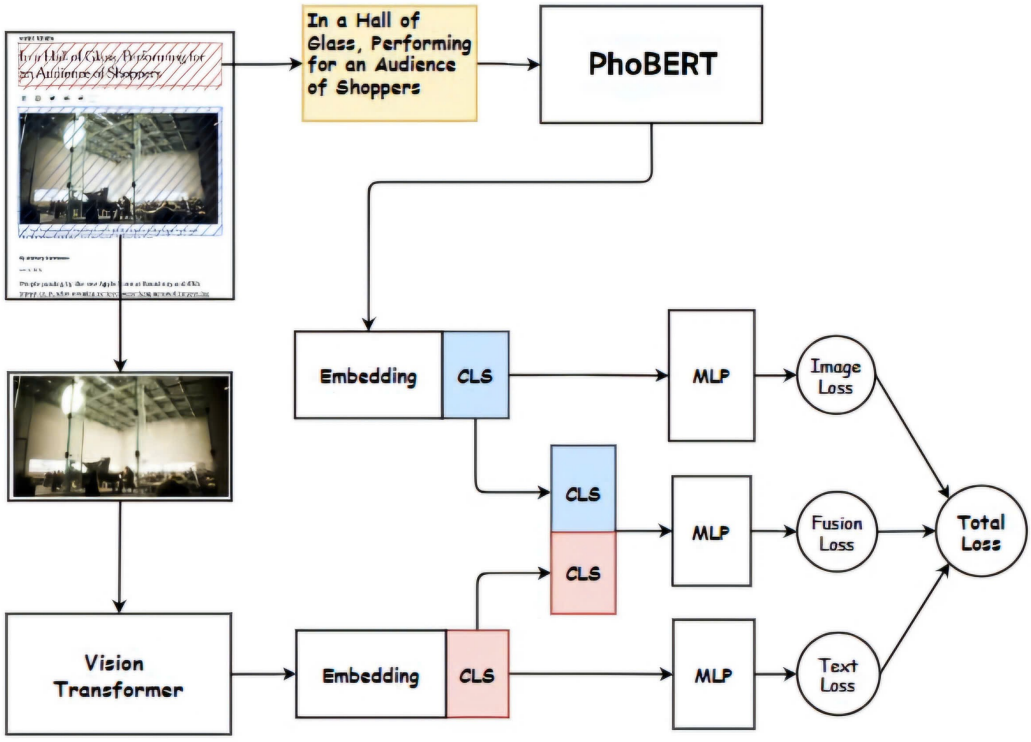

Figure 4: Architecture of the Proposed Approach

4.4 Evaluation and Comparison:

We evaluate the three methods on the same dataset using metrics such as accuracy, precision, recall, and F1-score. The analysis results will help us understand the impact of each data type on classification performance and the effectiveness of combining the two data types.

5 Experiments

5.1 Experimental Settings

To process the text, the PhoBERT model requires sentences to be segmented using the vncorenlp library. Subsequently, the article content is divided into individual sentences with the support of the underthesea library. Due to resource constraints, we only utilize the first five sentences of each article to extract text embeddings.

The OK-ViNews dataset used for evaluation is divided into training, testing, and validation sets. The distribution across these sets is proportionally maintained to ensure consistent evaluation results.

We assess three methods: first, the Text-Only model predicts outcomes using only text to evaluate classification performance. Second, the Image-Only model employs ViT. The third method, the Multimodal model, combines both text and images to evaluate the effectiveness of integrating both modalities in online news classification.

5.2 Result

The results after evaluating the three different models show a significant difference in classification performance, as presented in Table **??**. The results for the two separate tasks of image and text classification using the PhoBERT and ViT models, respectively, differ noticeably.

First, for the text-only classification model, the F1, precision, and recall scores are , , and , respectively. In contrast, the image-only classification model yields much lower scores, with F1, precision, and recall being , , and , respectively. However, the multimodal classification model, which combines both text and image data, achieves higher scores than the single-modality approaches. This demonstrates that the model effectively captures both contextual information from text and visual cues from images, leading to better performance.

Model	F1	Acc	Prec	Rec
Text-Only	0.74	0.76	0.74	0.76
Image-Only	0.55	0.58	0.54	0.58
Multimodal	**0.77**	**0.78**	**0.78**	**0.78**

Table 2: Model performance comparison.

When comparing prediction performance across categories in the dataset, the models perform well on categories such as Technology, Labor Union, Economy, Culture-Sports, and Automobiles, as these topics are easily identifiable through either text or images.

In contrast, for the Real Estate category, the text-only model may struggle to distinguish it from other categories, as the content often overlaps with economic topics. Similarly, categories such as Youth, Science, and Employment are also prone to misclassification when relying solely on text.

Category	Precision	Recall	F1-Score
Real Estate	0.00	0.00	0.00
Technology	0.81	0.75	0.78
Unions	0.94	0.96	0.95
Tourism	0.63	0.67	0.65
Education	0.00	0.00	0.00
Entertainment	0.69	0.46	0.55
Youth	0.00	0.00	0.00
Science	0.00	0.00	0.00
Economy	0.80	0.78	0.79
Law	0.78	0.77	0.78
Health - Love	0.47	0.56	0.51
World	0.68	0.82	0.75
Current Affairs	0.28	0.25	0.26
Jobs	0.00	0.00	0.00
Culture - Sports	0.89	0.85	0.87
Vehicles	0.80	0.82	0.81
Lifestyle - Society	0.62	0.74	0.68

Table 3: Performance metrics across different Vietnamese content categories with Text-Only model

For the image-only classification model, there is an improvement in accuracy for the Real Estate category, while categories such as Youth, Science, and Employment do not show significant improvement. However, in the multimodal model, which combines both text and image features, categories with previously low classification performance, such as Youth, Science, and Employment, show a significant increase. This indicates that the combination of images and text in the news articles has effectively contributed to enhancing classification performance on the Vietnamese-language dataset.

Category	Precision	Recall	F1-Score
Real Estate	0.33	0.08	0.13
Technology	0.56	0.45	0.50
Unions	0.70	0.84	0.76
Tourism	0.50	0.43	0.46
Education	0.00	0.00	0.00
Entertainment	0.37	0.28	0.32
Youth	0.00	0.00	0.00
Science	0.00	0.00	0.00
Economy	0.53	0.60	0.57
Law	0.40	0.28	0.33
Health - Love	0.36	0.27	0.31
World	0.48	0.43	0.46
Current Affairs	0.21	0.10	0.14
Jobs	0.00	0.00	0.00
Culture - Sports	0.68	0.84	0.75
Vehicles	0.69	0.80	0.74
Lifestyle - Society	0.57	0.26	0.35

Table 4: Performance metrics across different Vietnamese content categories with Image-Only model

Category	Precision	Recall	F1-Score
Real Estate	0.23	0.25	0.24
Technology	0.74	0.80	0.77
Unions	0.94	0.97	0.95
Tourism	0.61	0.75	0.67
Education	0.22	0.15	0.18
Entertainment	0.68	0.77	0.72
Youth	0.23	0.40	0.29
Science	0.50	0.79	0.61
Economy	0.84	0.73	0.78
Law	0.77	0.73	0.75
Health - Love	0.66	0.75	0.70
World	0.70	0.81	0.75
Current Affairs	0.77	0.44	0.50
Jobs	0.20	0.18	0.19
Culture - Sports	0.89	0.93	0.91
Vehicles	0.71	0.88	0.79
Lifestyle - Society	0.63	0.68	0.65

Table 5: Performance metrics across different Vietnamese content categories with Multimodal model

6 Conclusion

Through this project, we have successfully constructed the OK-ViNewsText dataset, which includes news article content, images, and related captions. This dataset is designed to be a suitable resource for tasks in the field of Natural Language Processing (NLP) for the Vietnamese language. Using the collected dataset, we conducted evaluations on the task of news category classification from various perspectives, demonstrating that combining both text and images significantly improves the accuracy and effectiveness of electronic news classification with the proposed method.

However, there are still some limitations. The collected dataset suffers from an imbalance across categories, which affects the classification performance, as the method has not yet shown clear re-

sults for categorizing topics with fewer instances. Additionally, due to resource and equipment constraints, the evaluation could only be performed on a small portion of the collected data.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Hop Van Bui, Hoang Huy Ha, Phuc Van Phan, and Oanh Ngoc Tran. 2024. Vistral v.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Khang T. Doan, Bao G. Huynh, Dung T. Hoang, Thuc D. Pham, Nhat H. Pham, Quan T. M. Nguyen, Bang Q. Vo, and Suong N. Hoang. 2024. Vintern-1b: An efficient multimodal large language model for vietnamese.

Anwen Hu, Shizhe Chen, and Qin Jin. 2020a. Icecap: Information concentrated entity-aware image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4217–4225.

Anwen Hu, Shizhe Chen, and Qin Jin. 2020b. Icecap: Information concentrated entity-aware image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 4217–4225, New York, NY, USA. Association for Computing Machinery.

Ken Lang. 1995. 20news: A dataset for text classification and clustering.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Deep learning for multimodal data fusion. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. 2024. Visually-aware context modeling for news image captioning.

Anthropic Team. 2024. Introducing contextual retrieval.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning.

Chi Tran and Huong Le Thanh. 2024. Lavy: Vietnamese multimodal large language model.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.

Zhao Yumeng, Yun Jing, Gao Shuo, and Liu Limin. 2021. News image-text matching with news knowledge graph. *IEEE Access*, 9:108017–108027.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *arXiv preprint arXiv:1509.01626*.

Yumeng Zhao, Yun Jing, Gao Shuo, and Liu Limin. 2021. News image-text matching with news knowledge graph. *IEEE Access*, 9:108017–108027.