New Approach to Infer Image Content from Social Media User's Posts: Based on Fine-Tuning Multimodal AI Model

Feriel Gammoudi [0000-0002-4715-0028][1], Salma Namouri[2], Mohamed Nazih Omri [0000-0001-7803-0179][1]
[1]MARS Research Laboratory LR17ES05, University of Sousse, Sousse, Tunisia
[2]University of Sousse, Sousse, Tunisia
GammoudiFeriel@gmail.com, salma.fnamouri@gmail.com, MohamedNazih.Omri@eniso.u-sousse.tn

Abstract

Automated content analysis requires accurate inference of imagery from social media. This study describes an improved approach to image content inference that makes use of the highly adjusted LLaVA (Large Language and Vision Assistant) model. Our solution overcomes the limits of previous models in integrating textual and visual data, resulting in a more unified representation that improves user-generated image interpretation.

The fine-tuning of LLaVA solves the issues given by diverse and noisy social media data, resulting in significant increases in inference accuracy. The innovation not only improves automated content analysis and moderation but also has important implications for targeted marketing and user engagement. Our technique establishes a new standard for employing multimodal models in social media analytics, providing a comprehensive solution for analyzing complicated image-text data.

1 Introduction

In today's digital environment, social media (Gammoudi et al., 2022) platforms have evolved into critical communication hubs where users increasingly employ visual content to share experiences, express emotions, and send messages that go beyond verbal constraints (van Dijck and Poell, 2022). Images on platforms like as Instagram, TikTok, and Twitter are effective mediums for gathering rich, complicated insights regarding user interests and attitudes. The capacity to effectively understand these visual cues is no longer a luxury for corporations, researchers, and policymakers; it is a strategic necessity for decoding consumer behavior, forecasting trends, and developing individualized marketing tactics (Shao et al., 2023).

Despite the vital importance of these findings, collecting relevant information from social media images remains a considerable difficulty. While breakthroughs in natural language processing have transformed text analysis (Qiu et al., 2022), visual content poses distinct difficulties, such as different formats, situations, and cultural interpretations (Joulin et al., 2020; Ferrara et al., 2023). Traditional image analysis algorithms frequently fail to capture the subtle information buried in these pictures, leading to fragmented or inaccurate interpretations of user behavior. Furthermore, the lack of ambiguity in supporting textual context complicates determining user intent, limiting the usefulness of current techniques (Zhang and Zheng, 2022).

Integrating multimodal models is a huge step forward in marketing technology, allowing for unparalleled precision in processing and interpreting various types of data. Multimodal models, including textual and visual data, provide a more comprehensive understanding of user interactions than standard single-data-type techniques (Xu et al., 2022). This study focuses on creating and optimizing a novel multimodal solution, Pixel Speak, to enhance the monitoring and analysis of brand mentions on social media platforms.

Multimodal models' unique capacity to synthesize data from numerous sources has had transformational effects across a variety of industries, including manufacturing and insurance (Chaudhuri et al., 2021). However, its application in marketing has had a particularly significant impact, allowing for more nuanced insights into brand interactions and customer behavior (Li et al., 2023). Despite these advances, there is still a major need for more refined procedures to meet the increasing demand for improved brand mention extraction methods in industries such as call centers, public relations businesses, and marketing agencies.

To address these issues, this study proposes a novel approach to inferring (Gammoudi and Omri, 2024a,b) and interpreting the content of photographs uploaded by social media users, utilizing cutting-edge deep learning and computer vision

techniques. This project seeks to deliver a more accurate and comprehensive comprehension of visual data, altering how user-generated material is examined and used. Our technique has important implications for improving targeted marketing, increasing user engagement, and creating more personalized online experiences (Xie et al., 2023; Lee et al., 2024).

The remaining sections of this paper are organized as follows: Section 2 discusses the problem description, motivation, and driving forces behind this research. Section 3 gives an overview of the subject, followed by a discussion of related work in section 4. Section 5 describes our suggested approach and its contributions, while Section 6 provides experimental findings and an analysis of the constructed model. Finally, Section 7 wraps up the study and offers future research topics.

2 PROBLEM DEFINITION, Research questions and Motivation

2.1 Problem Statement

The challenge of inferring image content from social media posts stems from the inherently sparse, noisy, and often ambiguous textual descriptions provided by users. Social media platforms are overwhelmed with vast volumes of user-generated images that are usually accompanied by minimal or unclear text, which complicates accurate interpretation. The core problem is to effectively integrate these limited textual cues with the diverse and evolving visual content, especially in the context of dynamic social media environments where the content is highly varied and context-dependent.

Traditional multimodal models often fail in these contexts due to several reasons: they struggle to balance limited textual context with the rich and complex visual features present in images; they lack robustness in dealing with highly heterogeneous and noisy data; and they are generally unable to generalize across a wide range of image types and text inputs. This results in inferior performance in real-world applications like automated content moderation targeted marketing, and user engagement strategies, where accurate content understanding is crucial.

2.2 Research questions

This section addresses key questions central to our research, aiming to provide insights and answers. These questions include:

- What are the main obstacles and limitations of effectively determining image content from social media posts when textual descriptions and image data are scarce or ambiguous?
- How can advanced multimodal models like LLaVA overcome the challenges of combining textual and visual input to increase image content inference accuracy?
- What are the unique constraints of current multimodal models in determining user interests based on image content and accompanying textual information, and how can they be overcome?
- How can machine learning and predictive modeling methods help infer image content and identify user interests in the setting of heterogeneous and noisy social media data?

2.3 Motivation

The capacity to reliably identify picture content from social media posts is critical for a variety of applications, including targeted marketing, content control, and user engagement analysis. Improved content inference allows organizations and researchers to obtain a better understanding of user preferences and behaviors, hence improving their strategies and interactions. Recent advances in multimodal models, such as LLaVA, provide promising solutions to these difficulties by using advanced algorithms for merging textual and visual input. However, there is still a significant research vacuum in modifying these models to accommodate the unique difficulties of social media information. This study seeks to close this gap by offering fresh ways that improve the accuracy of visual content inference. Addressing the limits of current models and exploiting cutting-edge technology can considerably advance the field of visual content analysis, providing considerable benefits across a wide range of disciplines such as targeted advertising, automated content analysis, and enhanced user experience.

3 Overview

3.1 LLaVA: Large Language and Vision Assistant

LLaVA (Large Language and Vision Assistant) is a big step forward in multimodal learning. It combines large-scale language models with advanced

vision models to interpret and produce insights from both text and images. This paradigm (Kim et al., 2024) seeks to bridge the gap between natural language processing (NLP) and computer vision (CV), resulting in a more integrated approach to complicated multimodal tasks. This review examines language-vision models such as LLaVa-Med, demonstrating its practical applications in biomedicine and clinical research.

3.2 Fine-Tuning

Fine-tuning is an important procedure for customizing pre-trained models to specific tasks or datasets. Fine-tuning improves the model's performance on new, related tasks by starting with a model trained on a broad, diverse dataset and then training it on a smaller, task-specific dataset. This step (Zhai et al., 2024) is necessary for adapting models to match the requirements of certain applications. This study introduces EMT (Evaluating MulTimodality), a method for assessing catastrophic forgetting in multimodal large language models (MLLM). The findings emphasize the need for improved fine-tuning strategies for MLLM.

3.3 Attention Mechanism

Attention (Niu et al., 2021) has emerged as a key term in deep learning, inspired by humans' concentration on distinguishing information. This work presents an overview of recent cutting-edge attention models and defines a unifying model that can be applied to the majority of attention structures. The attention method enables models to flexibly focus on different areas of the input data, improving their capacity to detect key elements. Attention processes are utilized in multimodal models to align and integrate input from many modalities, hence enhancing the model's performance on tasks requiring complex interactions between text and visuals.

4 State of the Art

The analysis of visual content on social media platforms has received a lot of attention in recent years since it is becoming increasingly important to understand user behavior, preferences, and trends. This section examines the most recent advances in computer vision and deep learning approaches for analyzing social media photographs, highlighting both the challenges and prospects in this field.

4.1 Advances in Deep Learning for Visual Content Analysis

Recent advances in deep learning have made substantial progress in the field of visual content analysis. Convolutional Neural Networks (CNNs) have emerged as the major method for feature extraction and image categorization, demonstrating effectiveness in a variety of social media scenarios. (Nadeem et al., 2019) surveyed DL applications in multimedia, focusing on end-to-end learning and solving reliability and robustness difficulties in eight problem domains such as image and video categorization. (Joo and Steinert-Threlkeld, 2018) investigate automated methods for visual content analysis in political science, utilizing deep learning and computer vision to analyze large-scale image data from social media. (Shin et al., 2020) provide a visual data analytics framework for social media, verifying innovative content elements including complexity and consistency through case studies. (Baroffio et al., 2016) provides a comprehensive review of methods for extracting, encoding, and transmitting compact visual features. These articles demonstrate the transformational potential of DL in visual content analysis across multiple areas.

4.2 Generative Models for Data Augmentation and Feature Enhancement

Generative models, like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have emerged as effective techniques for enriching datasets and improving feature extraction. (Ferrara et al., 2023) investigated the use of GANs to produce synthetic social media photos, which were then used to train models with restricted data availability, hence increasing their robustness to various visual materials. Similarly, (Xie et al., 2023) used VAEs to create latent representations of social media photos that capture both low- and high-level aspects, allowing for more accurate feature extraction and interpretation.

These generative approaches have shown helpful in mitigating the obstacles given by the highly diverse and dynamic character of social media photos, which frequently differ in quality, style, and context (Kim et al., 2023). However, integrating generative models with deep learning frameworks remains a difficult task, particularly in maximizing the balance of realism and variability in generated content (Yin et al., 2023).

4.3 Contextual Analysis and Semantic Understanding

Contextual analysis is essential for understanding photos on social media, as their meaning is frequently influenced by cultural, social, and situational aspects. (Shao et al., 2023) proposed a method for contextual image categorization that employs semantic elements to increase comprehension in a variety of social media settings.
Furthermore, developments in natural language processing (NLP) and multimodal transformers have resulted in improved semantic understanding through cross-modal interactions. (Joulin et al., 2020) introduced a transformer-based method for analyzing the interplay between text and images, which improved the extraction of useful insights from user-generated content on platforms like as Instagram and Twitter. The cross-modal approach

4.4 Multimodal Learning Approaches

To address the inherent limits of depending primarily on visual data, multimodal learning systems have gained popularity. These methods improve content interpretation by combining different data sources, such as images, text, and metadata. (Ding et al., 2023) introduced a multimodal image-text matching framework that uses contrastive learning to efficiently align visual and textual information. This approach improves the capacity to analyze social media photographs with little or unclear accompanying text, making it especially useful in situations where text data is sparse or partial.
Furthermore, (Zhang and Zheng, 2022) proposed a deep multimodal learning strategy that incorporates visual and semantic data to improve the interpretation of social media photos. Their technique uses both image pixels and contextual information from surrounding text to provide a more thorough analysis that captures the full range of user intent and sentiment. The incorporation of multimodal signals has been found to reduce ambiguity and increase the accuracy of visual content analysis on sites where users often mix photos with little or no textual explanation (Liu et al., 2023).

5 Contribution

The main contributions of this article can be summarized as follows:

- **New Multimodal Inference Approach:**We provide an innovative approach for inferring image content from social media posts that use a fine-tuned LLaVA multimodal model. This technique tackles the current limits for handling diverse and loud user-generated content.

- **Improved Model Architecture:**By integrating CLIP's visual encoding to LLaMA or Vicuna's language models via an MLP connector, we improve the model's ability to interpret complicated multimodal data.

- **Efficient Fine-Tuning Strategy:** We use LoRA (Low-Rank Adaptation) to fine-tune the LLaVA model, resulting in considerable performance benefits with few parameter updates while maintaining scalability and cost-efficiency.

- **Comprehensive Evaluation:** We validate our fine-tuned model on real-world datasets, demonstrating its efficacy through higher BLEU and ROUGE scores in applications such as content moderation and target marketing.

5.1 Proposed inferring approach

This work enhances caption prediction and brand mention extraction through multimodal models.

5.1.1 LLaVA 1.5 7B Model

LLaVA-1.5 is an auto-regressive language model based on the transformer architecture, which has been fine-tuned from LLaMA/Vicuna with GPT-generated multimodal instruction-following data. The model incorporates simple yet effective modifications from its predecessor, LLaVA, enabling it to achieve state-of-the-art performance on 11 benchmarks such as Science QA.

The architecture is made up of three major components. First, the Visual Encoder is in charge of extracting features from images, using models like the ViT-336 CLIP to capture fine visual details. Second, the Language Model provides coherent text replies, relying on advanced models such as LLaMA or Vicuna to produce contextually relevant and articulate results. Finally, the MLP Connector acts as a critical link between the visual and language components, aligning feature representations from the Visual Encoder with textual replies created by the Language Model. This seamless integration promotes shared knowledge and engagement between visual and textual modalities.

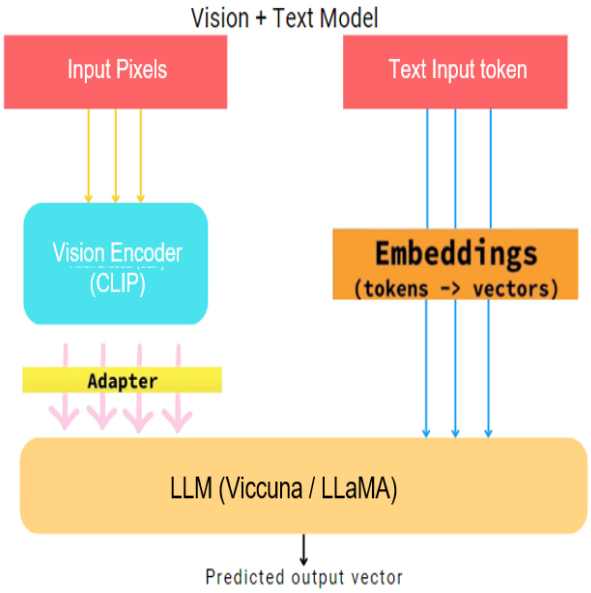

Figure 1: LLaVA Model Architecture

CLIP: Utilizes a self-attention mechanism to process images into feature vectors. This approach allows CLIP to effectively capture and represent complex visual information, transforming it into a format that can be seamlessly integrated with other components of the system.

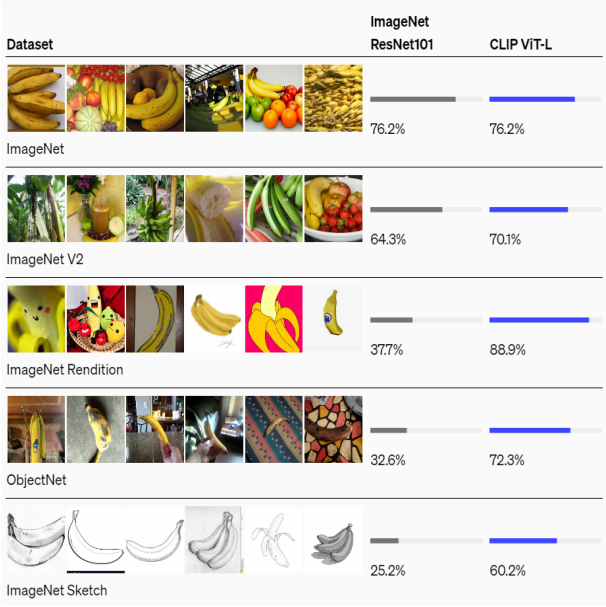

Figure 2: ImageNet vs. CLIP (**?**)

5.1.2 Advantages of LLaVA

Cost-Efficiency: Minimal training with pre-trained models.
Performance: Matches GPT-4's multimodal capabilities
Open Source: Flexible for visual and linguistic tasks

5.2 Fine-Tuning

Fine-tuning involves adapting pre-trained models for new tasks by leveraging transfer learning principles.

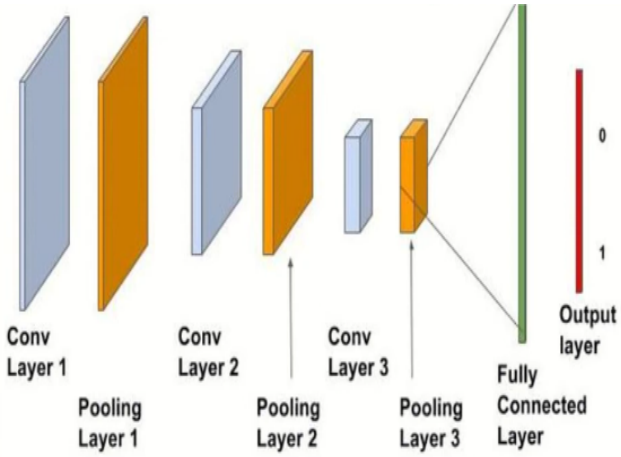

Figure 3: CNN Layer Structure

5.2.1 Transfer Learning

Transfer learning adapts pre-trained models to new tasks by freezing some layers and retraining others.

5.2.2 Fine-Tuning Benefits

- **Cost-Efficiency:** More affordable than training from scratch.
- **Effectiveness:** Improves both general and task-specific learning.
- **Accessibility:** Enables advanced models in resource-limited settings.

5.2.3 Fine-Tuning Steps

- **Data Preparation:** Clean and format the data.
- **Model Selection:** Choose a pre-trained model.
- **Parameter Configuration:** Set hyperparameters like learning rate and epochs.
- **Validation:** Evaluate with relevant metrics.
- **Iteration:** Refine based on evaluation results.
- **Deployment:** Deploy the fine-tuned model.

6 Experimental Study

In this section, we present how we fine-tuned the model and its implementation. We also show as well some screenshots of the fine-tuning code and explain the main terms in the fine-tuning script.

6.1 Simulation setup

We fine-tuned our model with LoRA (Low-Rank Adaptation), which efficiently updates a small number of new parameters while leaving the old model parameters unchanged. This strategy is less costly than fine-tuning the entire model.

6.1.1 Datasets

The LLaVA model was fine-tuned with two datasets:

- **SROIE 2019 Text Recognition:** This dataset features 973 scanned English receipts, processed into 33,626 training images and 18,704 test images to enhance text recognition capabilities.
- **OK-VQA:** Derived from the COCO dataset, this dataset includes 5,000 samples of images, questions, answers, and question IDs, aimed at improving visual question answering.

6.1.2 Training Process

The fine-tuning process followed these steps:

- **Data Preparation:** Captions were formatted into JSON to simulate a conversation between GPT and the user.
- **Repository Setup:** The LLaVA model repository was cloned, and necessary dependencies were installed.
- **Monitoring:** We utilized Weights and Biases to track training metrics, such as GPU efficiency and loss rates, to monitor performance and avoid issues like overfitting.
- **Parameter Configuration:** We configured LoRA to fine-tune specific layers, updating only 0.4% of parameters. The learning rate was set to 2e-4, and training was performed over 5 epochs to achieve optimal results.
- **Acceleration:** Deepspeed was employed to accelerate training by utilizing multiple GPU cores in parallel.
- **Model Finalization:** Post-training, we merged the updated weights into the base model, resulting in a new fine-tuned version, as depicted in Figure 4.

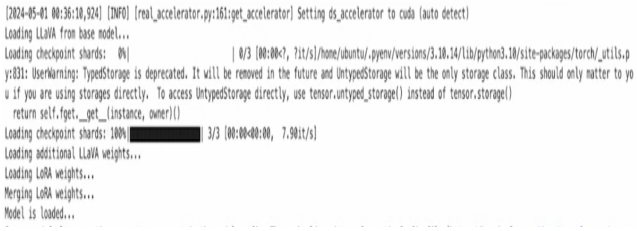

Figure 4: Merge New Weights to the Open Source Model Result Screenshot

6.2 Model Evaluation

This section discusses how we evaluated our model, providing evaluation curve graphics and a comparison of our fine-tuned model to the open-source baseline.

6.2.1 Evaluation Metrics

We assess our model using BLEU and ROUGE scores, which are standard metrics in natural language processing (NLP).

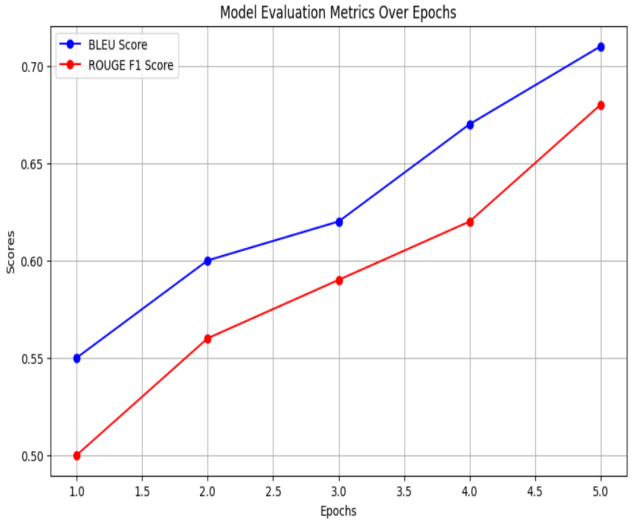

Figure 5: Evaluation Curve using BLEU and ROUGE scores

ROUGE Score

The ROUGE score evaluates the quality of machine-generated text by focusing on recall. Specifically, we use the ROUGE F1 score, calculated as follows:

$$\text{ROUGE-F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

where:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

Here, TP stands for True Positives, FP for False Positives, and FN for False Negatives.

BLEU Score

The BLEU score measures precision by comparing n-grams in the generated text with those in reference texts. It is computed as:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (4)$$

where:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \quad (5)$$

$$p_n = \frac{\text{Number of matched } n\text{-grams}}{\text{Total number of } n\text{-grams in the candidate}} \quad (6)$$

$$w_n = \frac{1}{N} \quad (7)$$

Here:

- BP = Brevity Penalty
- c = Length of the candidate translation
- r = Length of the reference corpus
- p_n = Modified precision for n-grams of size n
- w_n = Weight for each n-gram precision (usually $\frac{1}{N}$)

6.2.2 Comparison between initial model and fine-tuned model

For the following comparison, we tested the first model using a set of prompts (visual and textual instructions). Then we ran the same prompts against our fine-tuned model.

Table 1: Table of Comparison Between LLaVA and our Fine Tuned Model

	LLaVA	Fine tuned LLaVA
Object	0.6	0.65
Hand Writing	0.77	0.8

Following these processes, we compare the results for object and handwriting detection. We utilized the BLEU score to compare the two generated captions.

The results in Table 1 show that fine-tuning has resulted in considerable increases in model performance. For object detection, the BLEU score went from 0.6 to 0.65, indicating a substantial gain in accuracy. In contrast, handwriting detection improved significantly, with the BLEU score increasing from 0.77 to 0.8. This shows that fine-tuning has significantly improved the model's capabilities, particularly in identifying and understanding handwriting. Overall, the fine-tuned model performs better, with notable improvements in handwriting detection.

7 conclusion

In this article, we presented a novel way to infer image content from social media publications using multimodal models that incorporate computer vision and natural language processing approaches. Our methodology significantly improves understanding of user-generated visual content, outperforming existing single-modality algorithms in object detection and contextual analysis.
By combining textual and visual data, our technology delivers a more complete knowledge of user intent and interests, which is crucial for marketing, user engagement, and trend prediction applications. The fine-tuned model, particularly in handwriting detection, demonstrates the value of multimodal techniques for deriving greater insights from complicated data sources.
Future work could involve expanding the dataset to include a larger range of social media platforms and investigating new modalities such as audio or video to improve the model's capabilities. This study paves the door for more accurate and efficient content analysis in many social media situations, providing useful tools for businesses, scholars, and policymakers.

Acknowledgments

No organization with a direct or indirect financial interest in the topic covered in the manuscript is associated with the writers.

References

Luca Baroffio, Andrea E. C. Redondi, Matteo Tagliasacchi, et al. 2016. A survey on compact features for visual content analysis. *APSIPA Transactions on Signal and Information Processing*, 5:e13.

Sujit Chaudhuri, Anil Shankar, and Sandeep Kumar. 2021. Multimodal integration for enhanced predictive analytics in industry. *Journal of Machine Learning Research*, 22(1):45–65.

Zheng Ding, Xue Wang, and Han Zhou. 2023. Multimodal image-text matching with contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):256–268.

Emilio Ferrara, Onur Varol, Carlos Davis, Filippo Menczer, and Alessandro Flammini. 2023. The rise of social bots: A decade of impact on social media and future directions. *Communications of the ACM*, 66(2):123–137.

Feriel Gammoudi and Mohamed Nazih Omri. 2024a. Deep learning and machine learning-based approaches to inferring social media network users' interests from a missing data issue. In *The 17th International Conference on Knowledge Science, Engineering and Management (KSEM 2024)*, volume 5.

Feriel Gammoudi and Mohamed Nazih Omri. 2024b. Generative ai and deep learning based method detecting purchasers from missing data social media. In *The 17th International Conference on Development in eSystem Engineering (DeSE)*.

Feriel Gammoudi, Mondher Sendi, and Mohamed Nazih Omri. 2022. Survey on social media influence environment and influencers identification. *Social Network Analysis and Mining*, 12(145).

Jaeho Joo and Zeynep C. Steinert-Threlkeld. 2018. Image as data: Automated visual content analysis for political science. *arXiv preprint arXiv:1810.01544*.

Armand Joulin, Edouard Grave, Tomas Mikolov, Piotr Bojanowski, and Tomas Mikolov. 2020. Bag of tricks for efficient text classification. *Journal of Machine Learning Research*, 21(74):1–27.

Donghyun Kim, Seungwoo Park, and Jihoon Lee. 2023. Advanced techniques for image content analysis on social media platforms. *Journal of Computational Social Science*, 6(1):112–130.

Kiduk Kim, Kyungjin Cho, Ryoungwoo Jang, et al. 2024. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. *Korean Journal of Radiology*, 25(3):224.

Jungwoo Lee, Hyunseok Kang, and Seungwoo Han. 2024. Visual content analysis for enhanced social media user profiling. *IEEE Transactions on Multimedia*, 26:980–993.

Jia Li, Qi Zhang, and Hong Liu. 2023. Transformative impact of multimodal models in marketing: A review. *Marketing Science*, 42(3):563–578.

Xiaohui Liu, Ying Chen, and Rui Zhao. 2023. Multimodal integration with large language and vision models: A review. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–158.

M. S. Nadeem, V. N. Franqueira, X. Zhai, et al. 2019. A survey of deep learning solutions for multimedia visual content analysis. *IEEE Access*, 7:84003–84019.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

Jianxiong Qiu, Hui Li, Yifan Wang, and Yanyan Zhao. 2022. Emerging trends in social media text analysis using nlp techniques. *Annual Review of Information Science and Technology*, 56:147–172.

Chao Shao, Gianluca L. Ciampaglia, Onur Varol, Kevin C. Yang, Alessandro Flammini, and Filippo Menczer. 2023. The spread of low-credibility content by social bots. *Nature Communications*, 14(1):179.

Dongsoo Shin, Shuo He, G. M. Lee, et al. 2020. Enhancing social media analysis with visual data analytics: A deep learning approach. *2020*.

José van Dijck and Thomas Poell. 2022. Social media and the transformation of public space. *Journal of Digital Culture*, 29(4):371–389.

Zhiwei Xie, Shuo Yan, and Zhenyu Zhuang. 2023. Contextual image classification for social media with semantic features. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(2):54.

Ling Xu, Yuxin Zhang, and Wei Chen. 2022. State-of-the-art multimodal models and their applications. *Annual Review of Computer Science*, 10:87–108.

Zhen Yin, Xiaolong Duan, and Ming Gao. 2023. Context matters: Enhancing image analysis with semantic and affective cues on social media. *IEEE Transactions on Affective Computing*, 14(2):327–336.

Yuexiang Zhai, Shengbang Tong, Xiao Li, et al. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR.

Yuxin Zhang and Qi Zheng. 2022. Interpreting social media images: A deep multimodal learning approach. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):3918–3930.