A Survey on Multimodal Natural Language Processing: Foundations, Current State, and Approaches

Therese Emmanuelle Pilapil Mendoza
University of California, Berkeley
temendoz@berkeley.edu

Abstract

[1] Multimodal Natural Language Processing (NLP) is an emerging subfield that aims to process computer-generated human language through the application of modalities such as image, video, audio, and sensor data. This survey explores the subfield by synthesizing over 25 seminal and contemporary published papers, which provide an overview of historical breakthroughs in Multimodal NLP, a sense of the current state of the subfield as of May 2024, and its future approaches and applications to other interdisciplinary fields. I will discuss the captivating development of the major tasks, models, and datasets that have shaped this subfield.

1 Introduction

NLP is a field that traditionally and still broadly depends on text-based data. However, with the popular consumption of audio and visual media and ever-growing usage of digital media and technologies, multimodal NLP expands on this established approach and seeks to enrich the computer understanding of the human language with the application of visual and audio media. Having been developed throughout the 2010s with its early works coming from the 2000s, multimodal NLP has gained significant traction in just these past few years with its ever-so-growing popularization and expansion. Researchers have been putting in substantial work to explore the different innovative ways to synthesize data across many modes of data including multimodal data and text-based data, which is resulting in the fascinating unfolding of the capabilities of NLP as a whole.

2 Monumental Historical Context

The following section discusses various historical milestones in the subfield of multimodal NLP, first focusing on the early successes of image-based multimodal learning followed by video-based multimodal learning. These various tasks and datasets are highly acclaimed or commonly recognized in the subfield's current research tasks.

2.1 Early Research

In multimodal NLP, it is common to fuse multiple modalities together to help a computer understand human speech – with the most common fusion being audio-visual information – which is why multimodal deep learning is utilized in the subfield so that audio and visual data may be processed at the same time for efficiency. A pivotal moment of its application was in a 2011 study when researchers at Stanford University and University of Michigan applied deep autoencoders that examined if a shared representation could even be learned over auto and video data (Ngiam et al., 2011). This study highlighted the possibility of deep learning to be utilized to outperform the current multimodal NLP models at the time opened up the potential of other approaches in the field.

The Bilingual Evaluation Understudy (BLEU) metric, developed in 2002, is widely used in machine translation due to its modified n-gram precision approach, which had a high correlation with human judgments when first introduced. Although its relevance has been surpassed by advances in multimodal NLP, BLEU remains a critical benchmark in the field (Papineni et al., 2002).

2.2 Image Captioning

In image captioning, datasets like Microsoft Common Objects in Context (MSCOCO) and Flickr 8k play pivotal roles. MSCOCO offers a comprehensive set of over 2.5 million samples of images with complex scenes and numerous labeled instances, proving invaluable for image-based research and captioning tasks (Garg et al., 2022) (Lin et al., 2015). Similarly, the Flickr 8k dataset, utilizing CNNs and LSTM-powered RNNs, processes im-

[1] Word Count: 2092

age features alongside caption data, demonstrating capabilities that exceed those of BLEU (Cui et al., 2018).

Significant contributions include the work by Stanford University's Computer Science Department, where researchers created an image description model that aligns text and image data through multimodal deep learning, generating natural language annotations for images (Karpathy and Fei-Fei, 2015). Google's research team developed a comparable model, focusing on end-to-end training to optimize the generation of correct image descriptions (Vinyals et al., 2015).

Furthermore, a dataset extending the Flickr 30k dataset was created called the Multi 30k dataset, which is a multilingual English-German image description generator. This is a momentous feat in image captioning, as it introduces the possibility of image captioning and multimodal NLP to be done in languages other than English (Elliott et al., 2016).

With the aid of the MSCOCO and Flickr resources and the remarkable success both the Stanford and Google models have surpassing BLEU's, at the time, state-of-the-art performance, and with the development of Multi 30k, these developments illuminate the accomplishments of multimodal NLP prior to the popularization of the subfield in the 2020s and potential of deep learning as a model to use in multimodal NLP.

2.3 Video Captioning

Considerable progress has also been made in the topic of video captioning during its early development. The HowTo100M dataset was introduced in 2019 as a large-scale dataset of 136 million video clips depicting humans performing and describing over 23,000 different visual tasks. It learns using a deep neural network, and it's trained using a max-margin ranking loss that optimizes the similarity between corresponding video clips and captioning (Miech et al., 2019). VaTeX is a video captioning model, and it uses two sequence-to-sequence models: a recurrent GRU model and a transformer, which both utilize I3D action figures, and they also use a decoder conditioned on max-pooled action figures or outputs from a multi-label classifier that predicts visual entities (Caglayan et al., 2019). Both of these developments are deemed as successful in multimodal NLP and continue to be used and referred to in recent video captioning developments.

3 Current Events

The following section examines the current state of multimodal NLP as of May 2024. With the current popular usage of popular modes of generative AI like ChatGPT, DALL-E, and Adobe's Generative AI to name a few, we are witnessing remarkable advancements being made in the subfield of multimodal NLP. We will discuss the recent advances in multimodal sentiment analysis, the usage of the GPT model, and recent interdisciplinary applications of multimodal NLP.

3.1 Multimodal Sentiment Analysis

Another key development was explored at the Multimodal Hate Speech Shared Task at CASE 2024. At this event, multiple teams were challenged with two subtasks to build a metric that would detect hate speech and identify its targets within text-embedded images (Thapa et al., 2024). The runner-up team, AAST-NLP, approached the task by incorporating Contrastive Language-Image Pre-Training (CLIP) and Bidirectional Encoder Representations from Transformers (BERT)-based models, which were reliant on feature extraction (El-Sayed and Nasr, 2024). CLTL, the winning team, utilized the Twitter-based RoBERTa and Swin Transformer V2 models in order to encode textual and visual modalities, and then additionally, it uses the Multilayer Perception (MLP) fusion mechanism for classification (Wang and Markov, 2024). The key difference that put CLTL in first place was that they were using pre-trained models, which allowed for more efficiency and effectiveness in their model. The issue with AAST-NLP's approach was that while their approach was strong, they may have not captured the full nuance of the dataset they were given as effectively as CLTL.

Events like these that explore the possibilities of multimodal NLP are happening all around the world; its expansion is worldwide! With social media being so prevalent in our everyday lives, another competition called the DravidianLangTech EACL 2024 also has competitors exploring multimodal abusive language detection but in Tamil. Binary Beasts was the runner-up team in this competition, and their approach was a multimodal fusion model, combining ConvLSTM for video features, BiLSTM for audio data, and MNB for textual content (Rahman et al., 2024). Winning team

Wit Hub utilized different individual models such as LSTM, Knearest neighbors, Linear Regression, Multinomial Naive Bayes and others to tackle specific tasks (S et al., 2024). The difference between the two approaches was where their specific focuses laid respectfully. Wit Hub utilized broad and diverse machine learning models, while Binary Beasts point of focus was more on enhancing their fusion model. The latter team most likely turned out to be the winning one due to the model's capability of capturing the complexity of social media interactions, similar to the reason CLTL won their competition. In the subfield of Multimodal NLP, complex models are what lean towards maximized accuracy because they're able to apprehend the full nuance of datasets.

Additionally, more multilingual development outside of just image-based language detection is happening in the multimodal NLP subfield. JaSPICE is a metric that evaluates Japanese captions based on scene graphs generated from dependencies and predicate-argument structures (PAS) to accurately interpret captions in Japanese and then extends the graph with synonyms (Wada et al., 2023). This metric outperforms BLEU and overall expands on what we know about multimodal captioning through its incorporation of complex semantic relationships and synonyms.

3.2 GPT

Generative pre-trained transformer (GPT) models are becoming increasingly powerful and useful, and many fall under the multimodal NLP subfield. For example, AudioGPT is a multimodal AI system that blends audio-based models with text-based models and image captioning models, facilitating tasks including speech recognition, music generation from text and image, and audio-to-text transformations (Huang et al., 2023). Additionally, OpenAI, the creators of ChatGPT, have recently developed Sora, a text-to-video generative AI model, in February 2024, which, as a pre-trained diffusion transformer, is able to generate high-quality, 1 minute-long AI-generated clips (Liu et al., 2024b).

AudioGPT and Sora are only two of the considerable amount of multimodal model developments made in just the past two years, with other notable models being the debut of GPT-4, Gemini, and so many more (Zhang et al., 2024).

3.3 Interdisciplinary Data Collection

Research done at the University of Southampton investigates the integration of audio features with text for the argumentation mining in U.S. political debates in order to assess whether incorporating audio information like pitch and intonation enhances the performance of argumentation mining models (Mestre et al., 2023). Additionally, the Aligned Multimodal Movie Treebank (AMMT) was developed at MIT, which is a dataset derived from dialogues in Hollywood movies, featuring audio, video, and text annotations aligned at the millisecond level (Yaari et al., 2022). The dataset aims to be an engaging and easy-to-use source for linguistic research and machine learning (ML) models through its incorporation of part-of-speech tags and dependency parses aligned with audiovisual data. Both of these developments are tremendous, as they expand on multimodal NLP in such unique and innovative ways.

4 Challenges, Limitations, and Other Approaches

Multimodal NLP is a quick-developing subfield. These recent developments have set new standards for the future of it; subsequently, they have set new standards for the future of the subfield and have challenged scholars to think of other approaches to further expand the subfield. we must reflect on the direction we are going with it. In this section, we will be going through various ways we are challenging and expanding on what we know about multimodal NLP.

4.1 Multimodal Representation Learning

Researchers recognize the rapid development of LLMs, which is why a specific group of researchers is porposing a MultiModal Chart Benchmark (MMC-Benchmark) for evaluating LLMs, as well as the Multimodal Chart Assistance (MMCA), which is an LLM that shows superior performance on chart-related benchmarks by using instruction tuning (Liu et al., 2024a). This would be helpful for the future, as we can use these charts as a tool to provide a structure to data that can help models learn multimodal elements more effectively.

4.2 Prompt Engineering

PromptCharm is another metric that "supports text-to-image creation through multimodal prompting and image refinement for novice users" (Wang et al., 2024). It uses the Structural Similarity Index (SSIM) to evaluate its performance, and studies have shown that users recognize how well it understands their request and outputs a quality generated image (Wang et al., 2024).

4.3 Concerns

In addition to the recent developments made in current-day multimodal NLP, researchers are trying to explore even more approaches of multimodal generative AI similar to ChatGPT, and in this paper, they raise challenges of the current state of the subfield. They raise the need for extensive training data, computational resources, potential biases and metrics that can help eliminate them (GATO for example), and ethical concerns – particularly in "deep-fake" content (Gozalo-Brizuela and Garrido-Merchan, 2023). They also question what the overall purpose of prompt AI and multimodal NLP is.

Researchers at Carnegie Mellon University discuss and attempt to find solutions and suggestions for the challenges in multimodal learning, highlighting how integrating multiple communicative modalities like linguistic, acoustic, and visual inputs can pose unique computational and theoretical challenges due to their heterogeneity and interconnectedness (Liang et al., 2023).

5 Conclusion

Multimodal learning and its significant integration into our society today has posed a shift in the overall field of NLP with its amazing capabilities to interpret complex interactions between video and image-based, audio-based, and text-based data. As the field advances and more and more robust systems, metrics, models, and tasks emerge into development, the subfield will remain a critical, yet engaging topic for researchers to delve into. In all, this paper has discussed the monumental points of history in multimodal NLP, the subfield's current and fast-emerging developments, and different frameworks and concerns, accentuating the new, arising, and exciting evolution.

References

Ozan Caglayan, Zixiu Wu, Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. Imperial college london submission to vatex video captioning task.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning.

Ahmed El-Sayed and Omar Nasr. 2024. AAST-NLP at multimodal hate speech event detection 2024 : A multimodal approach for classification of text-embedded images based on CLIP and BERT-based models. pages 139–144.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions.

Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondřej Bojar. 2022. Multimodality for NLP-centered applications: Resources, advances and frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6837–6847, Marseille, France. European Language Resources Association.

Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchan. 2023. Chatgpt is not all you need. a state of the art review of large generative ai models.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2023. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024a. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. 2024b. Sora: A review on background, technology, limitations, and opportunities of large vision models.

Rafael Mestre, Stuart E. Middleton, Matt Ryan, Masood Gheasi, Timothy Norman, and Jiatong Zhu. 2023. Augmenting pre-trained language models with audio feature embedding for argumentation mining in political debates. pages 274–288.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. *ICML'11: Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 689–696.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal abusive language detection in Tamil based on integrated approach of machine learning and deep learning techniques. pages 212–217.

Anierudh S, Abhishek R, Ashwin Sundar, Amrit Krishnan, and Bharathi B. 2024. Wit hub@DravidianLangTech-2024:multimodal social media data analysis in Dravidian languages using machine learning models. pages 229–233.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during Russia-Ukraine crisis - shared task at CASE 2024. pages 221–228.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator.

Yuiga Wada, Kanta Kaneda, and Komei Sugiura. 2023. JaSPICE: Automatic evaluation metric using predicate-argument structures for image captioning models. pages 424–435.

Yeshan Wang and Ilia Markov. 2024. CLTL@multimodal hate speech event detection 2024: The winning approach to detecting multimodal hate speech and its targets. pages 73–78.

Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. Promptcharm: Text-to-image generation through multi-modal prompting and refinement.

Adam Yaari, Jan DeWitt, Henry Hu, Bennett Stankovits, Sue Felshin, Yevgeni Berzak, Helena Aparicio, Boris Katz, Ignacio Cases, and Andrei Barbu. 2022. The aligned multimodal movie treebank: An audio, video, dependency-parse treebank. pages 9531–9539.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models.