Multimodal for Natural Language Processing: Tasks, Methodologies and Advancements

M. Arsalan
MBZUAI
Abu Dhabi, UAE

S. AlBarri
MBZUAI
Abu Dhabi, UAE

***Abstract*—Modality is derived from the word "mode", which means the methodology of communicating or experiencing things. Artificial Intelligence (AI) mimics humans to understand the world around us by interpreting modality. The learning ability of deep neural networks uplifted AI performance in every domain. Natural Language Processing (NLP), which is an AI core application, progressed remarkably by learning semantic meaning and adapting contextual information of the text modality. Multimodal learning is the phenomenon where the essence of information is extricated from different modalities (text, image, audio, emotions, expressions) while maintaining correlation. This study presents NLP applications enhanced by multimodal data to train and make deep learning architecture more robust. The applications presented are visually rich document understanding, sentiment analysis, visual question answering, event and fake news detection, and information retrieval. For each application, recent studies, challenges, SOTA deep learning architecture and benchmark datasets are presented.**

***Index Terms*—Multimodal Data, Natural Language Processing, deep learning, representation, speech, and vision.**

I. Introduction

From information retrieval to image processing, and from text mining to recommendation systems, the research community shifted to deep learning-based approaches for complex applications [1]. Although deep learning architecture imitate the behavior of the human brain's neurons, most of the previous research is performed considering unimodal data and learns one mode or source of information to build a machine learning application. However, for an AI model to mimic the human experience accurately and improve its performance, multimodalities play vital role [2]. Multimodalities include a combination of two or more modes such as text, images, video, audio, speech, body gestures, facial expressions, physiological signals and voice tone [3].

Multimodal machine learning is a complicated task as it entails many challenges, five central challenges are listed in [4]. The five challenges are representation [5], translation [6], alignment, fusion [7], and co-learning [8]. Each challenge in multimodal applications by itself is the subject of an extensively researched field. Representation is the method by which raw data is summarized or *represented*. For instance, physiological information, audio, and voice can be represented as signals, whereas, language such as typed text can be represented symbolically. To perform cross-modality signal processing, the multimodal signals need to be projected and represented into a single vector space. Translation is the process of transforming or *translating* data from one modality to another. Translation showcases the subjectivity and open-endedness faced in representing a modality. Alignment is the process of identifying relationships between or *aligning* elements from two or more modalities. Dependencies and similarities between modalities are exploited in alignment to accomplish multimodal downstream tasks. Fusion integrates or *fuses* information from two or more unimodal data to perform prediction tasks. Fusion leverages the rich features offered in each unimodal data to build multimodals for various downstream tasks. The input of fusion is the unimodal embedding or representations and the output is implemented for a multimodal task. Finally, co-learning transmits knowledge or *learned* information between two or more different modalities or between their representations. The five challenges, as involved as they are, simplifies multimodal downstream tasks reality by assuming that all of the modalities are present, aligned, noiseless, and correctly and fully annotated during training and testing. Figure 1 comprehensively demonstrates the five challenge and the notion of transferring different unimodals into shared and common space.

For the purpose of the proposal, multimodality applications in Natural language processing (NLP) are introduced. This proposal is divided as follows, in Section II, the Literature Review introduces previous attempts to summarize and survey multimodal application. In Section III, the Plan Forward defines and expands on various key multimodal applications. Finally, the conclusion summarizes the expected future work to build the survey paper. The expected contributions of this survey are:

- Focus on main NLP tasks that can be performed efficiently using multimodality.
- Expand on several multimodal applications, note their challenges, present SOTA architectures and benchmark datasets.
- Reflect on the current literature and discuss possible improvement and future directions.

II. Literature Review

Researchers got inspiration from the multimodal approach used for audio-visual speech recognition [9]. Well-aligned integration of vision and hearing was a fundamental requirement of the task. Results motivated the research community to extend this idea for other tasks. The multimodal approach

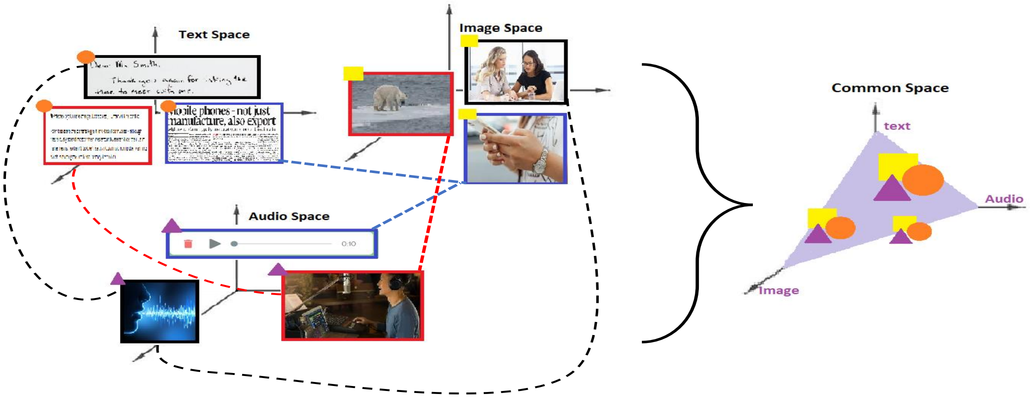

Fig. 1. Text, images and audio modalities are represented using their own features and dimensions. In multimodal applications, text, images, and audio are jointed in a common representation. The common space integrates and maps the information from each modality. Common space representation is essential and detrimental in multimodal downstream tasks.

evolved with time and has been used for different speech, vision and emotion recognition applications in the past [4]. The most recent category contains the data from multimedia resources that can be more complex than two or three modalities. The multimodal data brought up numerous applications that solve the problems more efficiently and creatively than unimodal data.

The representative application is image captioning which generate the text description of the image. Document understanding [10] extract layout, visual features, table like structural information from the documents. Sentiment Analysis [11] of reviews, tweets and posts containing Video, images, emojis and any other emotional expression. Detection includes the identification of the specific event, discovering trend, finding out emergency situation and detecting fake news on web or social media [12]. Information retrieval also rely on multimodal processing, especially in electronic health records, where the data related to patient is available in different modalities [13].

The first challenge that algorithm face while dealing multimodal data is to understand the encoding information of different modalities which is called heterogeneity gap in learning Representation [14]. The heterogeneity gap can be reduced by associating the similar semantic of different modalities in correlated manner. Fusing features from different modalities boosted the performance on cross media tasks [15].

Recent research depicts that deep learning outperformed the classical approach in understanding and processing representation due to strong learning abilities [16]. The key phenomenon is to learn representation in a hierarchical manner as general purpose learning without structural or additional information. Mentioned reasons flourished deep learning as most compatible approach for multimodal representation learning.

Numerous surveys are published for deep learning multimodal detailing the applications of computer vision and Natural language processing [17]. Various publications focus on multimodal representation, correlation, fusion, translation and co-learning [1], [4]. Some authors focused on general application related to NLP and CV and few authors addressed single application using multimodal architecture [18], [11]. Most of the previous work found is related to enhance the computer vision applications as textual features are semantically rich [19].

In this study, effort is put to discuss Natural Language Processing tasks considering multimodal data. We aim to present a comprehensive survey that consider several modalities including text, images, speech, expressions, gestures and other signals. Additionally, variations of deep neural networks that are utilized to deal multimodal data under each application are discussed. Table 1 presentsbenchmark datasets and the modalities entailed for each application.

III. Plan Forward

In this section, the taxonomy of the survey is proposed along with the subcategories of each application. All NLP applications that are enhanced by deep multimodal architectures are presented here. Details include for each multimodal application, its benchmark dataset in literature, evaluation metrics, description, and comparison for the best performing architectures. Multimodal applications are visually rich document understanding, sentiment analysis, event detection, information retrieval and processing, classification, dialogue systems, recognition, and cross-modality.

A. *Visually Rich Document Understanding*

The understanding of documents that are visually rich and contain structured information is important to analyze business and market applications [10]. Contrary to the classical information extraction methods, Visually Rich Documents (VRD) understanding considers visual layout along with the text. One of the recent research, LayoutLMv2 [20] pre-trained the Transformer-based model that learns from different modalities at the pre-training stage by integrating and aligning layout, textual and visual information. LayoutMv2 used more than 10 billion documents to achieve pre-trained model that performed SOTA on document understanding tasks. More recently, Li *et al.* [21] proposed StrucTexT, which outperformed LayoutLMv2 by pre-training the Transformer model in a self-supervised manner, consumed less number of parameters, and achieved high scores. The publicly available benchmark datasets for downstream tasks are SROIE [22] and FUNSD [23]. LayoutLMv2 achieved 97% and 84% F1 score on SROIE and FUNSD, respectively. StrucTexT accomplished 96.88% and 85.68% F1 on SROIE and FUNSD, respectively.

B. Sentiment Analysis

Sentiment analysis is the core task in NLP that extract and classify reviews, feeling, gestures and behavior toward specific entity. Sentiment Analysis understands people' perspectives for efficient decision making in multiple domains. Textual representation has been analysed widely in previous research such as [24], [25], and [26]. However, sentiment analysis task moved from text modality only to other form of modalities due to social media and internet. Chen *et al.* [27] designed deep learning architecture to extract sentiment from multimodal complex data. The author proposed two component-based methods consisting of shallow fusion and aggregation parts. The shallow fusion component extracts contextual information from the different domains using the attention mechanism and the aggregation part attains sentimental word aware fusion. The proposed architecture outperformed other methods by achieving the highest score on benchmarks multimodal datasets CMU-MOSI, CMU-MOSEI, and YouTube datasets for sentiment analysis.

C. Visual Question Answering

Visual Question Answering (VQA) attempts to answer linguistic questions by retrieving information from a visual cue [28]. VQA combines information from understanding written questions and from focusing on the related information in the high dimensional visual. Questions could vary from a simple true/false to knowledge-based and open-ended questions, whereas visuals could vary from a simple sketch or image to a whole video. Furthermore, VQA simultaneously combines several functions from NLP and CV fields such as language understanding, relation extracting, attribute and object classifying, counting, knowledge-base and common-sense reasoning [3]. VQA maps question text and visual embeddings obtained via recurrent and convolutional neural networks (RNN and CNN), respectively, to a common vector space. Mapping to the shared vector space enables VQA to tackle open-ended free-form questions. In the literature, the four general mapping approaches are joint embedding [29], attention mechanism [30], compositional model [31] and knowledge base [32]. There exist review papers targeting VQA as a research field of its own such as in [28] and [33]. In addition to surveys, benchmarks such as [5] by Carnegie Mellon University (CMU) and [34] are available. Benchmark datasets are VAQ v1.0 [33], VAQ-X [35], and VAQ-CP [36].

TABLE I
THE NLP APPLICATIONS, BENCHMARK DATASETS AND MODALITIES ARE MENTIONED.

Applications	**Benchmarks**	**Dataset Nature**
Visually Rich Document Understanding	FUNSD	Scanned Documents
Sentiment Analysis	CMU-MOSEI	Videos + Sentences
Visual Question Answering	VQA v1.0	Images + sentences
Event Detection	CrisisMMD	Tweets + Images
Fake News Detection	Weibao	Images + sentences
Information Retrieval	MIMIC	X-ray + Reports

D. Event Detection

Detection of an event, trend or situation specifically through content available on social media become more efficient and generous by considering data from different modalities [37]. A massive amount of data is generated that has a significant impact on the lives, property, and psychology of humans [38].Event detection can be mapped to other realistic scenarios, such as Emergency management, Disaster Detection, and Topic Detection [39]. The researchers are designing architectures to detect events using multimodal data that maintain comprehensive information. Even though the data from different perspectives enhance the performance, likewise, the challenges are increased for methods to deal with redundant and heterogeneous characteristics. For disaster detection, still there is no large enough benchmark dataset is available to employ deep learning architecture except CrisisMMD [40], CrisisNLP and CrisisLex. For traffic event detection, Chen *et al.* [41] , created a multimodal dataset by integrating the traffic related filtered tweets with sensor data. The author achieved 84%, 83%, 87% F1 score with CNN, RNN, and mmGAN (multimodal GAN) models, respectively.

E. Fake News Detection

Fake news utilize multimedia information to cause the panic on social media. The version of content containing text and image misinterpret the facts and manipulate human psychology which can lead to rapid false propagation of fake news [42]. The multimodal architecture can detect fake news by mismatching the pattern and considering the similarity among different modalities [43].

Wang *et al.* [12] proposed the fine-grained multimodal fusion networks (FMFN) to detect fake news detection in efficient way. The author presented their approach in following three steps. CNNs are used to extract visual featues and RoBERTa [44] is used to attain contextualized embedding of words in the first step. In second phase, attention mechanism is used between visual and textual features to enhance the correlation for fusing features. At final stage, binary classifier adopted to perform detection on fused features. The author evaluated the proposed model on Weibao Dataset [42] by focusing only tweets that contains text and images. The proposed model acquired 88% accuracy as compared to CARMN which scored 85%.

F. Information Retrieval

To build an Information Retrieval (IR) application, the steps are indexing, query formulation, retrieval and evaluation. In indexing and query formulation, documents and user interfaced queries are represented by their characteristic features, respectively [45]. The retrieval system then maps both representations to retrieve or extract the required useful information. The performance of the retrieval task is evaluated based on recall and precision. In multimodal information retrieval, the system searches documents with different modalities such as text, images, videos, or physiological signals and images. Employing more than one modality enriches information retrieving processes. One example of multimodal information retrieval is in electronic health records.

1) Electronic Health Records: Patients health records contain various modalities such as categorical data, text, images such MRI scans, or signals such as electrocardiogram (ECG) [45]. Information retrieval system can extract information from different modalities to report, present, and/ or predict a patient health status. For instance, Supervised Deep Patient Representation Learning Framework (SDPRL) engages different modalities information to learn patient representation [13]. SDPRL is build and tested using the benchmark dataset MIMIC-III.

IV. Conclusion

The preliminary work for the literature review is proposed. The planned future trajectory is to add additional NLP multimodal applications such as machine translation, audio-visual speech recognition tasks, information processing, and cross-modality. Moreover, further details will be presented on the NLP applications by discussing the difficulties encountered in multimodal data representation, translation, alignment, fusion, and co-learning tasks. The SOTA architectures for each application are to be presented and detailed along with the performance, available benchmarks, and benchmark datasets. Results summary for different architectures will be tabulated to enhance the information reach and condensation in literature reviews

References

[1] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[2] P. Wignell, K. Chai, S. Tan, K. O'Halloran, and R. Lange, "Natural language understanding and multimodal discourse analysis for interpreting extremist communications and the re-use of these materials online," *Terrorism and political violence*, vol. 33, no. 1, pp. 71–95, 2021.

[3] J. Summaira, X. Li, A. M. Shoib, S. Li, and J. Abdul, "Recent advances and trends in multimodal deep learning: A review," *CoRR*, vol. abs/2105.11087, 2021.

[4] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[5] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, R. Salakhutdinov, and L. Morency, "Multibench: Multiscale benchmarks for multimodal representation learning," *CoRR*, vol. abs/2107.07502, 2021.

[6] U. Sulubacak, O. Caglayan, S. Grönroos, A. Rouhe, D. Elliott, L. Specia, and J. Tiedemann, "Multimodal machine translation through visuals and speech," *CoRR*, vol. abs/1911.12798, 2019.

[7] J. Gao, P. Li, Z. Chen, and J. Zhang, "A Survey on Deep Learning for Multimodal Data Fusion," *Neural Computation*, vol. 32, pp. 829–864, 05 2020.

[8] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *CoRR*, vol. abs/2107.13782, 2021.

[9] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[10] J. Im, M. Kim, H. Lee, H. Cho, and S. Chung, "Self-supervised multimodal opinion summarization," *arXiv preprint arXiv:2105.13135*, 2021.

[11] G. Chandrasekaran, T. N. Nguyen, and J. Hemanth D, "Multimodal sentimental analysis for social media applications: A comprehensive review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1415, 2021.

[12] J. Wang, H. Mao, and H. Li, "Fmfn: Fine-grained multimodal fusion networks for fake news detection," *Applied Sciences*, vol. 12, no. 3, p. 1093, 2022.

[13] X. Zhang, B. Qian, Y. Li, Y. Liu, X. Chen, C. Guan, and C. Li, "Learning robust patient representations from multi-modal electronic health records: a supervised deep learning approach," in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 585–593, SIAM, 2021.

[14] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 251–260, 2010.

[15] A. Habibian, T. Mensink, and C. G. Snoek, "Video2vec embeddings recognize events when examples are scarce," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 10, pp. 2089–2103, 2016.

[16] Y. LeCun, Y. Bengio, G. Hinton, *et al.*, "Deep learning. nature, 521 (7553), 436-444," *Google Scholar Google Scholar Cross Ref Cross Ref*, 2015.

[17] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[18] L. Stappen, A. Baird, L. Schumann, and S. Bjorn, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *IEEE Transactions on Affective Computing*, 2021.

[19] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer*, pp. 1–32, 2021.

[20] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, *et al.*, "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding," *arXiv preprint arXiv:2012.14740*, 2020.

[21] Y. Li, Y. Qian, Y. Yu, X. Qin, C. Zhang, Y. Liu, K. Yao, J. Han, J. Liu, and E. Ding, "Structext: Structured text understanding with multi-modal transformers," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1912–1920, 2021.

[22] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. Jawahar, "Icdar2019 competition on scanned receipt ocr and information extraction," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520, IEEE, 2019.

[23] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, pp. 1–6, IEEE, 2019.

[24] E. H.-J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang, and M. Song, "Topic-based content and sentiment analysis of ebola virus on twitter and in the news," *Journal of Information Science*, vol. 42, no. 6, pp. 763–781, 2016.

[25] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," *arXiv preprint arXiv:1707.01780*, 2017.

[26] B. Wood, O. Williams, V. Nagarajan, and G. Sacks, "Market strategies used by processed food manufacturers to increase and consolidate their power: a systematic review and document analysis," *Globalization and health*, vol. 17, no. 1, pp. 1–23, 2021.

[27] M. Chen and X. Li, "Swafn: Sentimental words aware fusion network for multimodal sentiment analysis," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1067–1077, 2020.

[28] Q. Wu, D. Teney, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *CoRR*, vol. abs/1607.05910, 2016.

[29] C. Chappuis, S. Lobry, B. Kellenberger, B. L. Saux, and D. Tuia, "How to find a good image-text embedding for remote sensing visual question answering?," *CoRR*, vol. abs/2109.11848, 2021.

[30] T. Rahman, S. Chou, L. Sigal, and G. Carenini, "An improved attention for visual question answering," *CoRR*, vol. abs/2011.02164, 2020.

[31] S. Subramanian, S. Singh, and M. Gardner, "Analyzing compositionality in visual question answering," in *ViGIL@NeurIPS*, 2019.

[32] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," *CoRR*, vol. abs/1906.00067, 2019.

[33] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," *CoRR*, vol. abs/1505.00468, 2015.

[34] X. Shi, J. Mueller, N. Erickson, M. Li, and A. J. Smola, "Benchmarking multimodal automl for tabular data with text fields," *CoRR*, vol. abs/2111.02705, 2021.

[35] D. H. Park, L. A. Hendricks, Z. Akata, B. Schiele, T. Darrell, and M. Rohrbach, "Attentive explanations: Justifying decisions and pointing to the evidence," *CoRR*, vol. abs/1612.04757, 2016.

[36] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," *CoRR*, vol. abs/1712.00377, 2017.

[37] L. Hu, B. Zhang, L. Hou, and J. Li, "Adaptive online event detection in news streams," *Knowledge-Based Systems*, vol. 138, pp. 105–112, 2017.

[38] N. Algiriyage, R. Prasanna, K. Stock, E. E. Doyle, and D. Johnston, "Multi-source multimodal data and deep learning for disaster response: A systematic review," *SN Computer Science*, vol. 3, no. 1, pp. 1–29, 2022.

[39] K. Xiao, Z. Qian, and B. Qin, "A survey of data representation for multi-modality event detection and evolution," *Applied Sciences*, vol. 12, no. 4, p. 2204, 2022.

[40] F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," in *Twelfth international AAAI conference on web and social media*, 2018.

[41] Q. Chen, W. Wang, K. Huang, S. De, and F. Coenen, "Multi-modal generative adversarial networks for traffic event detection in smart cities," *Expert Systems with Applications*, vol. 177, p. 114939, 2021.

[42] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 795–816, 2017.

[43] X. Zhou, J. Wu, and R. Zafarani, "Safe: similarity-aware multi-modal fake news detection (2020)," *Preprint. arXiv*, vol. 200304981, 2020.

[44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[45] W. Weng and P. Szolovits, "Representation learning for electronic health records," *CoRR*, vol. abs/1909.09248, 2019.