



Scoping Review on Image-Text Multimodal Machine Learning Models

Maisha Binte Rashid 
School of Eng. & Computer Science
Dept. of Computer Science
Baylor University
Email: Maisha_Rashid1@Baylor.edu

Pablo Rivas , *Senior, IEEE*
School of Eng. & Computer Science
Dept. of Computer Science
Baylor University
Email: Pablo_Rivas@Baylor.edu

Abstract—Multimodal Machine Learning (MMML) has emerged as a promising topic with the ability to jointly utilize data from several data modalities to improve performance and address difficult real-world problems. Large-scale multimodal datasets and the availability of powerful computing resources have sped up the development of sophisticated deep learning architectures that are designed for multimodal data. In this paper, we conducted a systematic literature review focusing on the deep learning architectures used in MMML that combines image and text modalities. The objective of this paper includes looking at various models and deep learning architectures used in MMML with image and text data, learning about the fusion techniques used to combine both modalities, datasets that are used to train the models and limitations of these models. For this purpose, we have garnered 341 research articles from 5 digital library database and after an extensive review process, we have 89 research papers that allow us to thoroughly assess MMML. Our findings from these papers shed light on providing new directions for further study in this evolving and interdisciplinary domain.

Index Terms—model distillation, word embeddings, bert, natural language processing, machine learning, deep learning

I. INTRODUCTION

The advent of digital technologies has led to an exponential growth in data across various disciplines, resulting in a paradigm shift in our understanding of complex systems [1], [2]. This proliferation of data encompasses multiple modalities, including visual cues in images, textual semantics, and auditory signals, which collectively provide a more comprehensive representation of the world [3], [4]. This multifaceted landscape has given rise to the field of Multimodal Machine Learning (MMML), which aims to develop computational models capable of integrating data from diverse modalities to improve predictive accuracy and decision-making capabilities [2], [5].

The motivation for multimodal integration arises from the limitations associated with unimodal data. While images offer rich visual information, they often lack the contextual depth that can be provided by accompanying text [6]. On the other hand, textual data, despite its semantic richness, may not capture the full spectrum of visual or auditory experiences [7]. Fusing these modalities enables constructing more robust and nuanced models that approximate human-like perception [8], [9].

The advent of deep learning architectures has further accelerated the capabilities of MMML, allowing for the extraction and fusion of complex features from multiple data sources [10], [11].

However, designing effective multimodal architectures presents unique challenges, such as mitigating overfitting, addressing data imbalance, and handling noisy data [12], [13]. Successful models strike a delicate balance between preserving the unique attributes of each modality and leveraging their inter-modal interactions to optimize performance [14], [15].

In the current era of data ubiquity and technological convergence, text and image modalities have emerged as pivotal elements in the MMML landscape. Images encapsulate visual complexity and emotional nuance, while text provides semantic context and narrative structure [16], [17]. The fusion of these modalities yields insights that are greater than the sum of their individual contributions, revolutionizing various application domains [18], [19].

The objective of this systematic literature review is to provide a comprehensive analysis of state-of-the-art MMML architectures that leverage text and image data. Specifically, this study aims to:

- Investigate the utilization of pre-trained models in MMML for feature extraction from text and image data, elucidating the techniques that enhance data representation.
- Offer an in-depth examination of fusion architectures, evaluating their efficacy and impact on multimodal data integration.
- Identify the existing limitations and challenges in MMML, paving the way for future research directions.

II. METHODOLOGY

The methodology section explains the thorough technique we used to investigate different aspects of MMML. We begin by developing specific research questions and continue with exhaustive search queries followed by systematic data extraction and integration of a rigorous quality assessment.

A. Research Questions

Our approach begins with the meticulous formulation of precise research questions intended to direct our exploration of the complexities of MMML. These inquiries steer our research toward crucial issues, including using pre-trained models for feature extraction, the variety, and influence of fusion topologies and inherent limitations. After rigorous analysis we come up with the following research questions:

TABLE I: Digital Database Search Queries

Database Name	Search Query	Volume Filters
Scopus	(ABS (machine AND learning) AND TITLE (multimodal) AND ABS (image) AND ABS (text) AND (TITLE-ABS (deep AND learning) OR TITLE-ABS (neural AND network)))	None.
IEEE Explorer	((("Document Title": multimodal) AND (("Document Title": "deep") OR ("Document Title": "machine learning") OR ("Abstract": "deep") OR ("Abstract": "machine learning") OR ("Abstract": "neural network"))) AND ("Abstract": text) AND ("Abstract": image)) NOT ("Document Title": "audiovisual") NOT ("Document Title": "video"))	None.
Springer Link	Where the title contains: multimodal; Query: text AND image AND ("deep learning" OR "machine learning" OR "neural network"); Sort by relevance	Pick top 80 of most relevant.
ACM Digital Library	Abstract: (neural) AND Title: (multimodal) AND Abstract: (deep learning) AND NOT Title: (video) AND NOT Title: (audio) AND E-Publication Date: (06/27/2018 TO 06/27/2023)	None.
Semantic Scholar	Keywords: multimodal machine learning deep learning image text. Dates: (01/01/2018 To 4/31/2023) Sort by relevance.	Pick top 13 relevant documents by TL;DR visual inspection.

- **RQ1-** Does multimodal machine learning models use well known previously established architectures?
 - RQ1.1- What are the most used pre-trained architectures for extracting and training image and text data?
 - RQ1.2- What datasets are used to compare the architectures?
- **RQ2-** What fusion strategies usually used in MMML?
- **RQ3-** What are the limitations or challenges to face using these architectures?

B. Searching Methodology

In an effort to answer our research questions, we exhaustively searched through several digital libraries, looking for relevant academic publications. We were able to construct a comprehensive collection of pertinent literature as a result of our thorough search across numerous academic archives. The digital library database what we used are - **Scopus, IEEE Explorer, Springer Link, ACM Digital Library** and **Semantic Scholar**. For the purpose of strategically locating relevant scholarly works, we used a broad range of keywords such as - **multimodality, deep learning, machine learning, neural network, image, text**. We created this set of keywords to cover all the topics we want to address in this study. These carefully selected keywords were then used as search queries in the mentioned databases. The search queries I used are given in Table 1.

C. Selection Criteria

We produced inclusion and exclusion criteria after getting research papers from the databases through search queries. The inclusion criteria covered research publications specifically discussing MMML models in various applications that worked with image and text data. Research papers unrelated to MMML or worked with modalities other than image and text are excluded from our process.

For the literature review, the **inclusion criteria** were carefully defined to narrow the scope of the papers considered. Specifically, the focus was on papers that worked with both text and image data, discussed multimodal machine learning models based on neural networks, and evaluated the performance of these multimodal models. Additionally, only papers written in English were included to ensure a coherent and accessible body of work. Conversely, the **exclusion criteria** were established to

eliminate papers that did not meet certain quality and relevance standards. Papers shorter than five pages were excluded to ensure depth of content. Non-English papers were also omitted, as were papers that had not undergone peer review, to maintain academic rigor. Articles with full text not available in the specified database were disregarded, as were opinion papers and papers that worked with data types other than image and text, to maintain focus on the subject matter.

Prior to applying the selection criteria, Scopus yielded 57 papers, which were reduced to 14 after selection. Similarly, IEEE Explorer had 114 papers initially, but only 34 met the criteria. Springer Link had 32 papers before and 12 after selection, while ACM Digital Library produced 108 papers before and 20 after. Lastly, Semantic Scholar contributed 30 papers initially, but only 9 remained after applying the criteria. These figures illustrate the substantial reduction in the number of papers after the rigorous application of selection criteria across the different databases.

After using the search queries mentioned in Table 1, we got total 341 research papers. We applied inclusion and exclusion criteria to those papers and finalized 89 papers that helped us answer the research questions we wanted to address.

III. RQ1 - DOES MULTIMODAL MACHINE LEARNING MODELS USE WELL KNOWN PREVIOUSLY ESTABLISHED ARCHITECTURES?

We seek to investigate the various architectures employed for MMML models in this research subject. We need to know which architectures to utilize to obtain information through multiple modalities. We discovered that MMML models extract image and text characteristics using pre-trained architectures that have been well-established in the past after carefully reading the papers we gathered.

A. RQ1.1 What are the most used pre-trained architectures for extracting and training image and text data?

This research question will help researchers find which architectures to use while developing MMML models with text and image data.

1) *Text Feature Extractor:* In Fig. 1 we showed the pre-trained architectures mostly used to extract and train text data. Fig. 1 shows that BERT (Bidirectional Encoder Representations

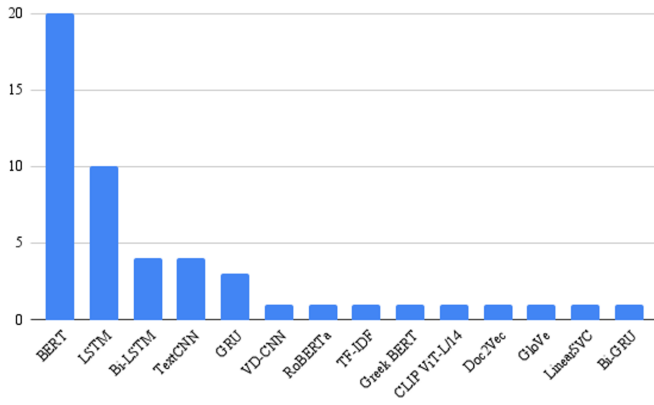


Fig. 1: Most used pre-trained models for text feature extraction

from Transformers) is used most to train text data. It is a pre-trained language representation model. Palani et al. [20] mentioned that BERT works by masking word tokens at random and expressing each mask with a vector; it can extract the underlying semantic and contextual meaning from the input words and sentences. Table 2 briefly mentions neural network architectures used in MMML models to extract text features in different articles.

The BERT paradigm for text representation and interpretation has gained prominence in natural language processing. For multimodal review helpfulness prediction the authors [27] converted each text into sequential embedding using BERT, with each row vector serving as a word. Gao et al. [23] created a word dictionary with BERT utilizing the subword tokenization algorithm WordPiece, which selects the value with the highest likelihood of merging to produce word segmentation. Agarwal et al. [32] also used the WordPiece tokenizer to tokenize clinical data and send it to BERT as input. To make the connection between review comments [25] proposes a new attention mechanism using BERT. Sahoo et al. [37] implemented BERT to extract text features since it can handle long sentences as input data and has no set input size requirements. Xu et al. [39] used BERT to extract deep semantic information from sentences as BERT uses a multi-head attention mechanism to calculate the connection between words. The authors of [26], [38], [34] and [36] also used BERT for text embedding. Another most used architecture for text feature extraction is LSTM(Long short-term memory). It is one type of recurrent neural network(RNN) that deals with the vanishing gradient issue that is not solvable for RNN [65]. Chen et al. [56] used LSTM to extract text features from a visual log and generate answers. Yadav et al. [44] used LSTM to optimize the pre-trained word embedding matrix and make high-level text features. Alsan et al. [45] used LSTM as a text encoder to convert text into a feature vector. To consider various emotional states, sentiments, and previous opinions for detecting polarity, Ange et al., [46] utilized LSTM.

2) *Image Feature Extractor*: Like texts, there are neural network architectures to extract features and train images.

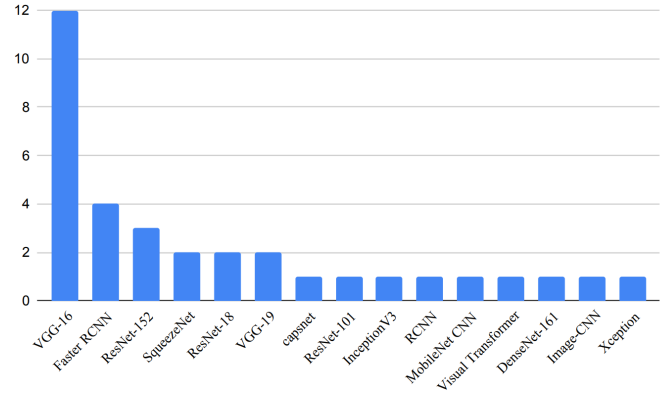


Fig. 2: Most commonly used pre-trained models for image feature extraction

Among them, Convolutional Neural Networks (CNNs), are crucial for computer vision and image analysis. In Fig. 2, we can see that VGG-16 is the most used architecture. VGG, ResNet, AlexNet, InceptionV3, DenseNet, and SqueezeNet are CNN architectures, which are deep learning models used for image-related tasks. For sentiment analysis from images, Shirzad et al. [61] used VGG-16, which is pre-trained on a Twitter dataset. They took the pre-trained model trained on the ImageNet dataset, fine-tuned it, and retrained on a Twitter dataset. Huang et al. [33] trained VGG-16 on the MNIST dataset consisting of microscopic images. Kim et al., [64] also worked with pre-trained VGG-16 but changed the last layer with a single sigmoid activation function. Babu et al. [53] combined two pre-trained models - VGG-16 and Xception for image feature extraction. Both of these models are pre-trained on the ImageNet dataset. VGG-16 consists of 16 convolutional layers, and Xception has 71 layers. Apart from CNN architectures, Faster-RCNN is another popular pre-trained architecture for image feature extraction. [29] extracted the bounding box and features of every object from each image using Faster-RCNN. Other than convolutional neural networks, transformers is also used for image feature encoding. [54] split the images into sequence patches of 16X16 pixels as a visual transformer is used for sequence processing.

B. $RQ_{1,2}$ What datasets are used to compare the architectures?

We searched the chosen articles for the datasets used in multimodal applications to answer this research question. While gathering information about datasets, we learned about some common data sources researchers use to create datasets for their research. Twitter, Flickr, IMDB, COCO(Common Objects in Context). In Fig 3, we summarized all the datasets we encountered in the articles. The authors of [52], [22], [57], [61] [55], used a Twitter dataset which consists of tweets and images. However, each employed a different kind of Twitter dataset to help them do their tasks. Fig 3 shows that the Flickr30k dataset has been used the most. Yu et al. used Flickr30k Entities, an extension of Flickr30k [38]. This dataset consists of 31,783 images with 44,518 object categories and 158k captions. For

TABLE II: Architectures used to train text features in MMML

Architecture Name	Article	Architecture Name	Article
BERT	[21], [22], [23], [24], [25], [26], [20], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39]	LSTM	[40], [41], [42], [24], [43], [44], [31], [34], [45], [46]
Bi-LSTM	[47], [48], [49], [39]	Residual Bi-LSTM	[50]
TF-IDF	[51]	GRU	[52], [34], [53]
GREEK BERT	[54]	RoBERTa	[22], [55]
Text CNN	[56], [57], [58], [39]	CLIP ViT-L/14	[59]
Bi-GRU	[60]	VADER	[61]
Doc2Vec	[62]	VD-CNN	[63]
LinearSVC	[62]	GloVe	[64]

TABLE III: Architectures used to train image features in MMML

Architecture Name	Article	Architecture Name	Article
VGG-16	[48], [27], [31], [66], [63], [61], [62], [33], [67], [62], [64], [53]	VGG-19	[68], [57]
ResNet-50	[69], [23], [47]	ResNet-101	[70]
ResNet-152	[34], [52], [22]	ResNet-18	[26], [54]
Xception	[53]	SqueezeNet	[25], [66]
DenseNet-161	[22]	Visual Transformer	[54]
InceptionV3	[21]	Faster RCNN	[29], [56]
Recurrent CNN	[50]	Image-CNN	[58]

modeling image-text data, Wang et al. created a dataset named MIR-Flickr using the Flickr website [57]. Another commonly used dataset is MSCOCO. Alsan et al. [45] used MSCOCO dataset for multimodal data retrieval. MSCOCO dataset has image and text pairs and is trained on a dual encoder deep neural network. MSCOCO dataset has 80 object categories 330k images with five descriptions per image [53].

IV. RQ2- WHAT FUSION STRATEGIES USUALLY USED IN MMML?

After reviewing the articles, we found different fusion techniques used in MMML models. Based on their structure and methods, we have categorized them into different categories such as the following:

- **Concatenation Technique** - concatenates textual and visual vectors.
- **Attention Technique** - calculates attention between text and image features, attention mechanism.
- **Weight-based Technique** - Early fusion, Late fusion, Intermediate fusion with different weights.
- **Multimodal Deep Learning Architectures** - Multimodal Compact Bilinear (MCB), Multimodal Deep Boltzmann Machine (DBM), Efficient attention with Transformer, Stacked Autoencoder Multimodal Data Fusion, Bi-LSTM.

A. Concatenation Technique

Concatenation implies concatenating multiple feature vectors together to get information from the features. Palani et al. [20] concatenated text and image feature vectors to get multimodal feature vectors to leverage information from both modalities. Paraskevopoulos et al. [54] used the concatenation technique to assemble text and visual encoders into a classifier model.

B. Attention Technique

To get relevant parts of each modality, the authors [48] used the attention mechanism as a fusion technique to detect appropriateness in scholarly submission. The authors mentioned that not all modalities contain equal importance. They added

an attention layer and calculated the attention score to get important modalities. Important modalities contain higher attention scores. Zhang et al. [71] used a multi-head attention mechanism for the joint representation of image and text features. The authors calculate the attention score of text and image features to integrate two modalities. They used the sigmoid function to calculate the weight of importance of images for source words. Xu et al. [39] used an attention mechanism to find a relation between each word in the sentence and the corresponding region on an image and calculated the weighted sum to ensure multimodal feature association.

C. Weight based Technique

One of the weight-based techniques is Early fusion, which is a feature-level fusion technique. It concatenates image and text feature vectors into one single vector representation, which provides heterogeneity in data [9]. To have joint representation of image and text features, [69] utilized Early fusion technique. The authors take same number of nodes from each modality's last hidden layer to give same importance to each modality. For sentiment analysis in multimodal data, to integrate two modalities, [63] applied the late fusion for sentiment analysis. Late fusion, also known as decision fusion, aggregates classifying features from each modality. Thus, each modality contributes individually to the final prediction.

D. Multimodal Deep Learning Architectures

Many deep learning architectures have been developed to accommodate multimodal feature representation, offering better information fusion and interpretation across different input modalities. One such model is Bi-LSTM, which the authors [21] used to integrate image and text features. To fuse data, Yue et al. [28] first introduced a knowledge-based network called ConceptNet. This model uses semantic similarity to calculate the similarity between image and text. In their paper, Chandra et al. [22] used the MCB pooling fusion technique. MCB performs the outer product between vector representation of image and text modality more efficiently than other techniques [72].

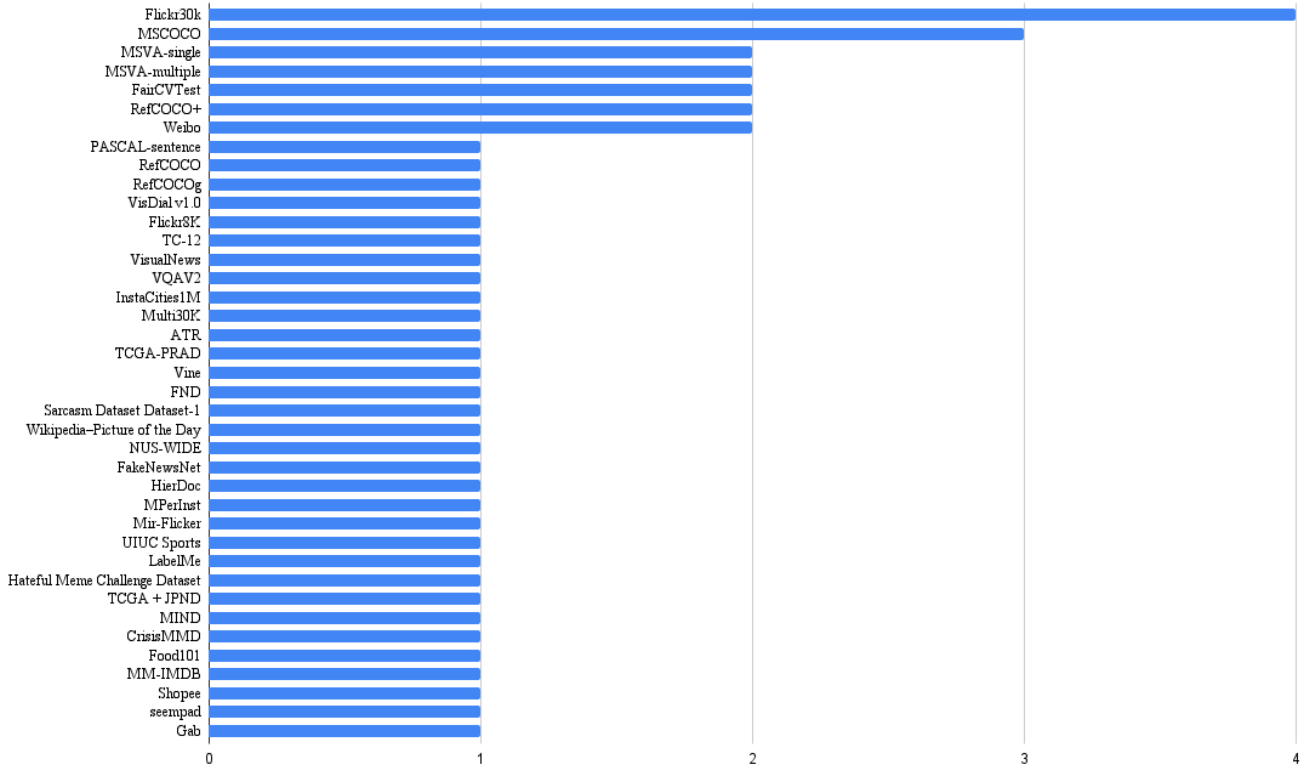


Fig. 3: Mostly used dataset in MMML applications

V. RQ3- WHAT ARE THE LIMITATIONS OR CHALLENGES TO FACE USING THESE ARCHITECTURES?

In this research question we explore the limitations or challenges occur using MMML architectures. Here we categorized the limitations and challenges that are commonly seen in MMML models.

- **Dataset Size.** One of the main challenges in MMML models is determining the ideal size for the dataset. The dataset size must be huge as MMML models work with multiple modalities' data. Data preprocessing for this huge number of data is expensive and computationally inefficient [73]. Image and text datasets vary in size and difficulty. So training them together is also challenging [74]
- **Data Annotation.** The publicly available datasets for text and images are mostly task-specific. Researchers need to make their dataset for other applications, which requires data annotation. However, large-scale data annotation is not widely available [75].
- **Noisy Data.** The noisy data in multimodality causes misclassification, as Chandra et al. [22] stated in their article. According to the authors' research, the outcome becomes inaccurate if one of the modalities is noisy.
- **Task Specific Image Feature Extractor** - For online review extraction on the multimodal features, Meng Li [25] used SqueezeNet for image feature extraction but did

not get the expected results as, according to the authors, the image feature extraction method was not appropriate for their specified task. The authors did not have their dataset trained on SqueezeNet, so image features were not fully utilized. Most pre-trained models for image feature extraction are task-specific. So, utilizing them in different tasks does not give the expected result. Jiatong Liu [35] described that for machine translation, they used ResNet-50, which is pre-trained on classification tasks. The image representation they got from using ResNet-50 needed to be more accurate.

VI. DISCUSSION AND CONCLUSION

Our scoping literature review identifies the most common methods for utilizing data from image and text modalities. We deduced from our RQ1 that the most popular pre-trained architectures for text embedding are BERT and LSTM. We observed that most researchers used various VGG and ResNet architectures for picture embedding. Furthermore, our research showed that MMML practitioners regularly use benchmark datasets like Twitter, Flickr, and the Common Objects in Context (COCO) dataset to train and assess their models. These datasets provide extensive, varied, and multimodal data sources, strengthening and broadening MMML models. As we turn our attention to the fusion methods, it becomes clear that the MMML community uses a wide range of fusion

methods, from concatenation to attention processes and neural networks. Every technique has a different set of benefits, which reflects the changing context of multimodal fusion. However, we discovered several important factors throughout our investigation of MMML's limitations and difficulties. These include computational complexity, data limitations, real-time processing difficulties, noise robustness, and the demand for bigger datasets. Researchers and practitioners must know these constraints pertaining MMML.

This literature review has illuminated the architectural preferences and dataset selections in MMML and the adaptable fusion strategies that the community has accepted. We have given an overall overview of the state of the field today by addressing the MMML's inherent limits and difficulties. This study acts as a useful compass, directing academics and practitioners toward informed judgments and creative solutions as MMML continues to develop and broaden its applications into various disciplines. As they delve farther into the multimodal data arena, researchers and practitioners seek to deepen our understanding of the world through connected data modalities. This journey has the power to transform industries, improve decision-making, and broaden our perspective on the world. In our future work, we want to explore the behavior of MMML models under adversarial conditions. Analyzing how these models react to adversarial attacks can provide crucial insights into their security and robustness, revealing tactics to defend them from malicious manipulation.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.
- [2] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: a survey and taxonomy," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 423–443, 2019.
- [3] S. Talukder, G. Barnum, and Y. Yue, "On the benefits of early fusion in multimodal representation learning," 2020.
- [4] 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, 2020.
- [5] S. Siriwardhana, T. Kaluarachchi, M. Billingham, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *Ieee Access*, vol. 8, pp. 176 274–176 285, 2020.
- [6] W. Chai and G. Wang, "Deep vision multimodal learning: methodology, benchmark, and trend," *Applied Sciences*, vol. 12, p. 6588, 2022.
- [7] J. Choi and J. Lee, "Embracenet: a robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259–270, 2019.
- [8] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo, "Multimodal machine learning in precision health: a scoping review," *NPJ Digital Medicine*, vol. 5, 2022.
- [9] K. Bayouduh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer*, vol. 38, pp. 2939–2970, 2021.
- [10] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. Alnuaim, A. Alhadlaq, and H. Lee, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, p. 2378, 2022.
- [11] P. Barua, W. Chan, S. Dogan, M. Baygin, T. Tuncer, E. Ciaccio, M. Islam, K. Cheong, Z. Shahid, and U. Acharya, "Multilevel deep feature generation framework for automated detection of retinal abnormalities using oct images," *Entropy*, vol. 23, p. 1651, 2021.
- [12] D. Lv, H. Wang, and C. Che, "Fault diagnosis of rolling bearing based on multimodal data fusion and deep belief network," *Proc. of the Institution of Mechanical Engineers Part C Journal of Mechanical Engineering Science*, vol. 235, pp. 6577–6585, 2021.
- [13] S. Kumaresan, K. Aultrin, S. Kumar, and M. Anand, "Transfer learning with cnn for classification of weld defect," *Ieee Access*, vol. 9, pp. 95 097–95 108, 2021.
- [14] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 478–493, 2020.
- [15] J. Li, X. Yao, X. Wang, Q. Yu, and Y. Zhang, "Multiscale local features learning based on bp neural network for rolling bearing intelligent fault diagnosis," *Measurement*, vol. 153, p. 107419, 2020.
- [16] Q. Zhu, X. Xu, N. Yuan, Z. Zhang, D. Guan, S. Huang, and D. Zhang, "Latent correlation embedded discriminative multi-modal data fusion," *Signal Processing*, vol. 171, p. 107466, 2020.
- [17] J. Singh, M. Azamfar, F. Li, and J. Lee, "A systematic review of machine learning algorithms for prognostics and health management of rolling element bearings: fundamentals, concepts and applications," *Measurement Science and Technology*, vol. 32, p. 012001, 2020.
- [18] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal eeg data for depression recognition," *Information Fusion*, vol. 59, pp. 127–138, 2020.
- [19] G. Schillaci, A. Villalpando, V. Hafner, P. Hanappe, D. Colliaux, and T. Wintz, "Intrinsic motivation and episodic memories for robot exploration of high-dimensional sensory spaces," *Adaptive Behavior*, vol. 29, pp. 549–566, 2020.
- [20] B. Palani, S. Elango, and V. Viswanathan K, "Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5587–5620, 2022.
- [21] M. Asgari-Chenaghlu, M. R. Feizi-Derakhshi, L. Farzinvas, M. Balafar, and C. Motamed, "Cwi: A multimodal deep learning approach for named entity recognition from social media using character, word and image features," *Neural Computing and Applications*, pp. 1–18, 2022.
- [22] M. Chandra, D. Pailla, H. Bhatia, A. Sanchawala, M. Gupta, M. Shrivastava, and P. Kumaraguru, "'subverting the jewtocracy': Online antisemitism detection using multimodal deep learning," in *Proc. of the 13th ACM Web Science Conference 2021*, 2021, pp. 148–157.
- [23] L. Gao, Y. Gao, J. Yuan, and X. Li, "Rumor detection model based on multimodal machine learning," in *Second Intl. Conf. on Algorithms, Microchips, and Network Applications (AMNA 2023)*, vol. 12635. SPIE, 2023, pp. 359–366.
- [24] S. Hangloo and B. Arora, "Combating multimodal fake news on social media: methods, datasets, and future perspective," *Multimedia systems*, vol. 28, no. 6, pp. 2391–2422, 2022.
- [25] M. Li, "Research on extraction of useful tourism online reviews based on multimodal feature fusion," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–16, 2021.
- [26] L. Lucas, D. Tomás, and J. Garcia-Rodríguez, "Detecting and locating trending places using multimodal social network data," *Multimedia Tools and Applications*, pp. 1–20, 2022.
- [27] S. Xiao, G. Chen, C. Zhang, and X. Li, "Complementary or substitutive? a novel deep learning method to leverage text-image interactions for multimodal review helpfulness prediction," *Expert Systems with Applications*, vol. 208, p. 118138, 2022.
- [28] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, "Knowlenet: Knowledge fusion network for multimodal sarcasm detection," *Information Fusion*, vol. 100, p. 101921, 2023.
- [29] Q. Guo, K. Yao, and W. Chu, "Switch-bert: Learning to model multimodal interactions by switching attention and input," in *European Conference on Computer Vision*. Springer, 2022, pp. 330–346.
- [30] P. Hu, Z. Zhang, J. Zhang, J. Du, and J. Wu, "Multimodal tree decoder for table of contents extraction in document images," in *2022 26th Intl. Conf. on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1756–1762.
- [31] M. R. Ahmed, N. Bhadani, and I. Chakraborty, "Hateful meme prediction model using multimodal deep learning," in *2021 Intl. Conf. on Computing, Communication and Green Engineering (CCGE)*. IEEE, 2021, pp. 1–5.
- [32] S. Agarwal, "A multimodal machine learning approach to diagnosis, prognosis, and treatment prediction for neurodegenerative diseases and cancer," in *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2022, pp. 0475–0479.
- [33] P.-C. Huang, E. Shakya, M. Song, and M. Subramaniam, "Biomdse: A multimodal deep learning-based search engine framework for biofilm documents classifications," in *2022 IEEE Intl. Conf. on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 3608–3612.

- [34] M. Ban, L. Zong, J. Zhou, and Z. Xiao, "Multimodal aspect-level sentiment analysis based on deep neural networks," in *2022 8th International Symposium on System Security, Safety, and Reliability (ISSSR)*. IEEE, 2022, pp. 184–188.
- [35] J. Liu, "Multimodal machine translation," *IEEE Access*, pp. 1–1, 2021.
- [36] T. Liang, G. Lin, M. Wan, T. Li, G. Ma, and F. Lv, "Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 492–15 501.
- [37] C. C. Sahoo, D. S. Tomar, and J. Bharti, "Transformer based multimodal similarity search method for e-commerce platforms," in *2023 IEEE Guwahati Subsection Conference (GCON)*. IEEE, 2023, pp. 01–06.
- [38] Z. Yu, M. Lu, and R. Li, "Multimodal co-attention mechanism for one-stage visual grounding," in *2022 IEEE 8th Intl. Conf. on Cloud Computing and Intelligent Systems (CCIS)*. IEEE, 2022, pp. 288–292.
- [39] J. Xu, H. Zhao, W. Liu, and X. Ding, "Research on false information detection based on multimodal event memory network," in *2023 3rd Intl. Conf. on Consumer Electronics and Computer Engineering (ICCECE)*. IEEE, 2023, pp. 566–570.
- [40] L.-R. Jácome-Galarza, "Multimodal deep learning for crop yield prediction," in *Doctoral Symposium on Information and Communication Technologies*. Springer, 2022, pp. 106–117.
- [41] I. Kraïdia, A. Ghenai, and N. Zeghib, "Hst-detector: A multimodal deep learning system for twitter spam detection," in *Intl. Conf. on Computing, Intelligence and Data Analytics*. Springer, 2022, pp. 91–103.
- [42] R. K. Kaliyar, A. Mohnot, R. Raghul, V. Prathyushaa, A. Goswami, N. Singh, and P. Dash, "Multideepfake: Improving fake news detection with a deep convolutional neural network using a multimodal dataset," in *Advanced Computing: 10th Intl. Conf., IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10*. Springer, 2021, pp. 267–279.
- [43] A. Malhotra and R. Jindal, "Multimodal deep learning architecture for identifying victims of online death games," in *Data Analytics and Management: Proc. of ICDAM*. Springer, 2021, pp. 827–841.
- [44] A. Yadav and D. K. Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 1, pp. 1–19, 2023.
- [45] H. F. Alsan, E. Yildiz, E. B. Safdil, F. Arslan, and T. Arsan, "Multimodal retrieval with contrastive pretraining," in *2021 Intl. Conf. on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 2021, pp. 1–5.
- [46] T. Ange, N. Roger, D. Aude, and F. Claude, "Semi-supervised multimodal deep learning model for polarity detection in arguments," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.
- [47] A. Peña, I. Serna, A. Morales, J. Fierrez, A. Ortega, A. Herrarte, M. Alcantara, and J. Ortega-Garcia, "Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment," *SN Computer Science*, vol. 4, no. 5, p. 434, 2023.
- [48] T. Ghosal, A. Raj, A. Ekbal, S. Saha, and P. Bhattacharyya, "A deep multimodal investigation to determine the appropriateness of scholarly submissions," in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2019, pp. 227–236.
- [49] H. Miao, Y. Zhang, D. Wang, and S. Feng, "Multimodal emotion recognition with factorized bilinear pooling and adversarial learning," in *Proc. of the 5th Intl. Conf. on Computer Science and Application Engineering*, 2021, pp. 1–6.
- [50] S. Paul, S. Saha, and M. Hasanuzzaman, "Identification of cyberbullying: A deep learning based multimodal approach," *Multimedia Tools and Applications*, pp. 1–20, 2020.
- [51] Y. Ha, K. Park, S. J. Kim, J. Joo, and M. Cha, "Automatically detecting image–text mismatch on instagram with deep learning," *Journal of Advertising*, vol. 50, no. 1, pp. 52–62, 2020.
- [52] R. Rivas, S. Paul, V. Hristidis, E. E. Papalexakis, and A. K. Roy-Chowdhury, "Task-agnostic representation learning of multimodal twitter data for downstream applications," *Journal of Big Data*, vol. 9, no. 1, p. 18, 2022.
- [53] G. T. V. M. Babu, S. D. Kavila, and R. Bandaru, "Multimodal framework using cnn architectures and gru for generating image description," in *2022 2nd Intl. Conf. on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, 2022, pp. 2116–2121.
- [54] G. Paraskevopoulos, P. Pistofidis, G. Banoutsos, E. Georgiou, and V. Katsouros, "Multimodal classification of safety-report observations," *Applied Sciences*, vol. 12, no. 12, p. 5781, 2022.
- [55] A. Bhat and A. Chauhan, "A deep learning based approach for multimodal sarcasm detection," in *2022 4th Intl. Conf. on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE, 2022, pp. 2523–2528.
- [56] X. Chen, S. Lao, and T. Duan, "Multimodal fusion of visual dialog: A survey," in *Proc. of the 2020 2nd Intl. Conf. on Robotics, Intelligent Control and Artificial Intelligence*, 2020, pp. 302–308.
- [57] Y. Wang, F. Ma, H. Wang, K. Jha, and J. Gao, "Multimodal emergent fake news detection via meta neural process networks," in *Proc. of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 3708–3716.
- [58] N. Xu and W. Mao, "A residual merged neural network for multimodal sentiment analysis," in *2017 IEEE 2nd Intl. Conf. on Big Data Analysis (ICBDA)*. IEEE, 2017, pp. 6–10.
- [59] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. Petrantonakis, "Synthetic misinformers: Generating and combating multimodal misinformation," in *Proc. of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, 2023, pp. 36–44.
- [60] A. N. Karimvand, R. S. Chegeni, M. E. Basiri, and S. Nemati, "Sentiment analysis of persian instagram post: a multimodal deep learning approach," in *2021 7th Intl. Conf. on Web Research (ICWR)*. IEEE, 2021, pp. 137–141.
- [61] A. Shirzad, H. Zare, and M. Teimouri, "Deep learning approach for text, image, and gif multimodal sentiment analysis," in *2020 10th Intl. Conf. on Computer and Knowledge Engineering (ICCKE)*. IEEE, 2020, pp. 419–424.
- [62] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, "Category-based deep cca for fine-grained venue discovery from multimodal data," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 4, pp. 1250–1258, 2018.
- [63] S. Thuseethan, S. Janarthan, S. Rajasegarar, P. Kumari, and J. Yearwood, "Multimodal deep learning framework for sentiment analysis from text-image web data," in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2020, pp. 267–274.
- [64] E. Kim, C. Onweller, and K. F. McCoy, "Information graphic summarization using a collection of multimodal deep neural networks," in *2020 25th Intl. Conf. on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 10 188–10 195.
- [65] S. Hochreiter and J. Schmidhuber, "Bridging long time lags by weight guessing and "long short-term memory"," *Spatiotemporal models in biological and artificial systems*, vol. 37, no. 65-72, p. 11, 1996.
- [66] C. Fatchah, P. D. S. Wiyadi, D. A. Navastara, N. Suciati, and A. Munif, "Incident detection based on multimodal data from social media using deep learning methods," in *2020 Intl. Conf. on ICT for smart society (ICISS)*. IEEE, 2020, pp. 1–6.
- [67] N. Guo, Z. Fu, and Q. Zhao, "Multimodal news recommendation based on deep reinforcement learning," in *2022 7th Intl. Conf. on Intelligent Computing and Signal Processing (ICSP)*. IEEE, 2022, pp. 279–284.
- [68] D. Chen and R. Zhang, "Building multimodal knowledge bases with multimodal computational sequences and generative adversarial networks," *IEEE Transactions on Multimedia*, 2023.
- [69] E. Hossain, M. M. Hoque, E. Hoque, and M. S. Islam, "A deep attentive multimodal learning approach for disaster identification from social media posts," *IEEE Access*, vol. 10, pp. 46 538–46 551, 2022.
- [70] L. Guo, "Art teaching interaction based on multimodal information fusion under the background of deep learning," *Soft Computing*, pp. 1–9, 2023.
- [71] Z. Zhang, K. Chen, R. Wang, M. Utiyama, E. Sumita, Z. Li, and H. Zhao, "Universal multimodal representation for language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [72] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.
- [73] K. Bayouh, 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.
- [74] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 437–10 446.
- [75] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Information Fusion*, vol. 81, pp. 203–239, 2022.