# Scoping Review on Image-Text Multimodal Machine Learning Models

**Anonymous authors**
**Paper under double-blind review**

## 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, learning about the fusion techniques used to combine both modalities and analyze their performance 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 88 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.

## 1 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 (Vaswani et al., 2017a; Baltrušaitis et al., 2019). 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 (Talukder et al., 2020; Gao et al., 2020). 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 (Baltrušaitis et al., 2019; Siriwardhana et al., 2020).

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 (Chai & Wang, 2022). On the other hand, textual data, despite its semantic richness, may not capture the full spectrum of visual or auditory experiences (Choi & Lee, 2019). Fusing these modalities enables constructing more robust and nuanced models that approximate human-like perception (Kline et al., 2022; Bayoudh et al., 2021a).

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 (Aggarwal et al., 2022; Barua et al., 2021). However, designing effective multimodal architectures presents unique challenges, such as mitigating overfitting, addressing data imbalance, and handling noisy data (Lv et al., 2021; Kumaresan et al., 2021). Successful models strike a delicate balance between preserving the unique attributes of each modality and leveraging their inter-modal interactions to optimize performance (Zhang et al., 2020; Li et al., 2020a).

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 (Zhu et al., 2020; Singh et al., 2020). The

fusion of these modalities yields insights that are greater than the sum of their individual contributions, revolutionizing various application domains (Cai et al., 2020; Schillaci et al., 2020).

The contributions of this study are as follows:

- Examine how multimodal machine learning (MMML) uses pretrained models to extract features from text and image data, illustrating the techniques that improve data representation.
- A detailed look at of fusion architectures that clarifies the methods for fusing text and image data, as well as an evaluation of their advantages and impacts.
- Identifying limitations and challenges present in MMML.
- Investigating the robustness of MMML models when exposed to noisy and adversarial data might shed light on how adaptable and useful they are in the real world.

The remainder of the research paper is structured as follows: Section 2 describes the methodology used for this research. The later sections describe the research questions in depth.

## 2 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.

### 2.1 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, inherent limitations, and the robustness of MMML models against noisy data. After rigorous analysis, we came up with the following research questions:

- **RQ1:** Do multimodal machine learning models use well-known, previously established architectures?
  - $RQ_{1.1}$ What are the most used pre-trained architectures for extracting and training image and text data?
  - $RQ_{1.2}$ What datasets are used to compare the architectures?
- **RQ2:** What fusion strategies are usually used in MMML?
  - $RQ_{2.1}$ What are the impacts of this fusion strategy in MMML models?
- **RQ3:** What are the limitations or challenges to face using these architectures?
- **RQ4:** In what way (if any) MMML models can be robust against noise and adversarial data?
  - $RQ_{4.1}$ What type of noise or adversary can occur in MMML models?

### 2.2 Searching Methodology

To answer our research questions, we exhaustively searched through several digital libraries, looking for relevant academic publications. We constructed a comprehensive collection of pertinent literature from our thorough search across numerous academic archives. The digital library database that we used is as follows:

- Scopus
- IEEE Explorer

- Springer Link
- ACM Digital Library
- Semantic Scholar

To strategically locate 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.

Table 1: 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 | `Kewords: 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. |

### 2.3 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 that are not related to MMML or worked with modalities other than image and text are excluded from our process.

#### 2.3.1 Inclusion Criteria

- Papers that worked with both text and image data
- Papers that discussed multimodal machine learning model based on neural networks
- Papers that discussed performance of multimodal machine learning models
- Papers that are in English

#### 2.3.2 Exclusion Criteria

- Papers that have length less than 5 pages
- Papers that are not in English
- Papers that are not peer reviewed
- Articles with full text not available in the specified database
- Opinion papers
- Papers that worked with data other than image and text

After using the search queries mentioned in Table 1, we got 341 research papers. We applied inclusion and exclusion criteria to those papers and finalized 80 papers that helped us answer the research questions we wanted to address. Also, after we started working on this study and finalized our paper later this year(2023), we came up with a few papers that talked about advanced multimodal models, which we considered relevant for our paper, so we added those ten papers as well. Table 2 displays the total number of papers in each database both before and after applying the selection criteria.

Table 2: Papers from each database before and after selection criteria

| Database Name | Before | After |
|---|---|---|
| Scopus | 57 | 14 |
| IEEE Explorer | 114 | 29 |
| Springer Link | 32 | 12 |
| ACM Digital Library | 108 | 14 |
| Semantic Scholar | 30 | 9 |
| Others | - | 10 |

### 2.4 Data Extraction and Synthesis

With a methodical technique, we make sure to extract the relevant information that is crucial for answering our research questions. We meticulously scanned every article to collect information that we considered relevant to answer RQ1, RQ2, RQ3, and RQ4. We encoded information about pre-trained deep learning architectures, fusion techniques, their performance and limitations, and datasets used in those applications. To get answers to the research questions, we looked into different sections of the articles. The relevant sections for each research question are discussed in Table 3.

Table 3: Data Extraction for research questions from different sections

| Research Question | Preferred Section |
|---|---|
| RQ1, RQ2 | Methodology/Model Description/Dataset/ResultsS |
| RQ3 | Limitations/Future Work/ Research Gap |
| RQ4 | Limitations/Dataset/ Data Preprocessing |

## 3 RQ1: Do multimodal machine learning models use well-known previously established architectures?

In this research question we aim to explore the type of architectures used for MMML models. To train MMML models for text and image data, we were interested in finding out if there were any single neural network architectures available. After rigorously going through the papers we finalized, we realized MMML models use previously well-established pre-trained architectures to train image and text data.

### 3.1 $RQ_{1.1}$ 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.

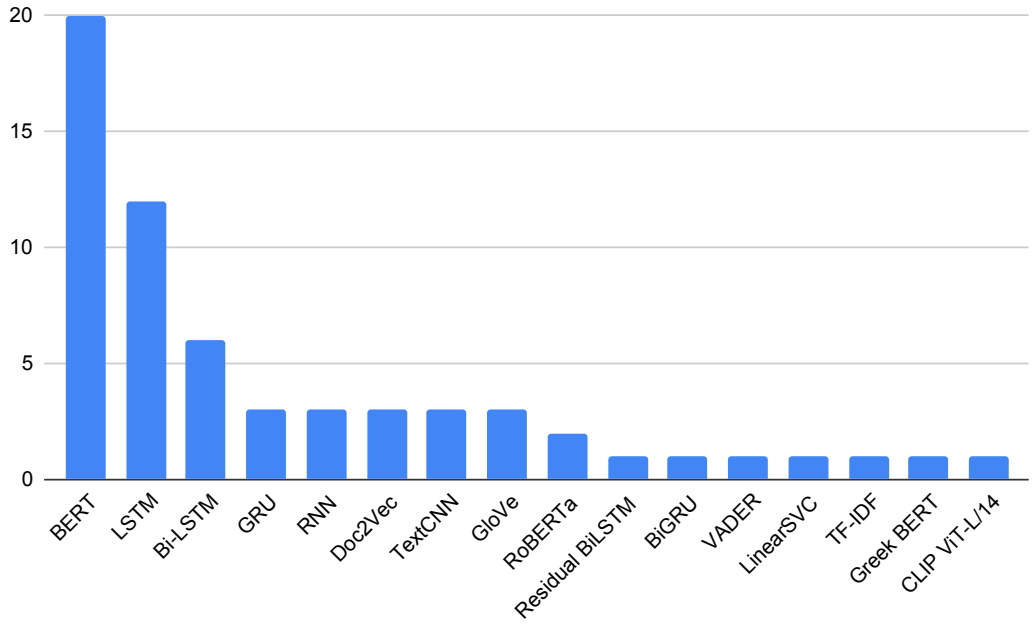


Figure 1: Most used pre-trained models for text feature extraction

#### 3.1.1 Text Feature Extractor

In Figure 1, we showed the pre-trained architectures that are mostly used to extract and train text data. From Figure 1, we see that Bilinear encoder representations from transformers (BERT) is used most to train text data. It is a pre-trained language representation model. Palani et al. (2022) 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. BERT is used in applications like detecting fake news (Palani et al., 2022), (Hangloo & Arora, 2022), rumor (Gao et al., 2023), sarcasm (Yue et al., 2023), places from social media (Lucas et al., 2022), online antisemitism (Chandra et al., 2021). It is also predicting review helpfulness (Xiao et al., 2022), tourism online reviews (Li, 2021). Though BERT is vastly used, Bhat & Chauhan (2022) and Chandra et al. (2021) used RoBERTa for detection purposes. RoBERTa is Facebook's modified version of BERT. After BERT, another architecture that is used frequently is Long-Short Term Memory (LSTM). In MMML models, LSTM is used in applications like sentiment analysis (Yadav & Vishwakarma, 2023), visual log (Chen et al., 2020), multimodal retrieval (Alsan et al., 2021), polarity detection (Ange et al., 2018). Other architectures are used for training and extracting text features but are not as popular as BERT and LSTM. These are depicted in Figure 1. In Table 4, we briefly mentioned neural network architectures that are 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 Xiao et al. (2022) converted each text into sequential embedding using BERT, with each row vector serving as a word. Gao et al. (2023) 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 (2022) also used WordPiece tokenizer to tokenize clinical data and sent it to BERT as input. To make a connection between review

Table 4: Architectures used to train text features in MMML

| Architecture Name | Article |
|---|---|
| BERT | Asgari-Chenaghlu et al. (2022), Chandra et al. (2021), Gao et al. (2023), Hangloo & Arora (2022), Li (2021), Lucas et al. (2022), Palani et al. (2022), Xiao et al. (2022), Yue et al. (2023), Zhang et al. (2023), Guo et al. (2022b), Hu et al. (2022), Ahmed et al. (2021), Agarwal (2022), Huang et al. (2022), Ban et al. (2022), Liang et al. (2022), Sahoo et al. (2023), Yu et al. (2022), Xu et al. (2023) |
| LSTM | Jácome-Galarza (2022), Kraidia et al. (2022), Kaliyar et al. (2021), Hangloo & Arora (2022), Malhotra & Jindal (2021), Yadav & Vishwakarma (2023), Ahmed et al. (2021), Ban et al. (2022), Alsan et al. (2021), Ange et al. (2018) |
| Bi-LSTM | Peña et al. (2023), Ghosal et al. (2019), Miao et al. (2021), Hossain et al. (2022), Xu et al. (2023) |
| Residual Bi-LSTM | Paul et al. (2020) |
| TF-IDF | Ha et al. (2020) |
| GRU | Rivas et al. (2022), Ban et al. (2022), Babu et al. (2022) |
| GREEK BERT | Paraskevopoulos et al. (2022) |
| RoBERTa | Chandra et al. (2021), Bhat & Chauhan (2022) |
| Text CNN | Chen et al. (2020), Wang et al. (2021), Xu & Mao (2017), Xu et al. (2023) |
| CLIP ViT-L/14 | Papadopoulos et al. (2023) |
| Bi-GRU | Karimvand et al. (2021) |
| VADER | Shirzad et al. (2020) |
| Doc2Vec | Yu et al. (2018) |
| RNN | Huang et al. (2022), Ban et al. (2022) |
| LinearSVC | Yu et al. (2018) |
| LSTM-RNN | Barveen et al. (2023) |
| GloVe | Chen & Zhang (2023), Kim et al. (2021) |
| VD-CNN | Thuseethan et al. (2020) |

comments Li (2021) proposes a new attention mechanism using BERT. Sahoo et al. (2023) 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. (2023) used BERT to extract deep semantic information from sentences as BERT uses a multi-head attention mechanism to calculate the connection between words. Lucas et al. (2022), Yu et al. (2022), Ban et al. (2022) and Liang et al. (2022) also used BERT for text embedding.

Another mostly 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 (Hochreiter & Schmidhuber, 1996). Chen et al. (2020) used LSTM to extract text features from visual logs and generate answers. Yadav & Vishwakarma (2023) used LSTM to optimize the pre-trained word embedding matrix and make high-level text features. Alsan et al. (2021) used LSTM as a text encoder to convert text into a feature vector. To take into account various emotional states, sentiments, and previous opinions for detecting polarity, Ange et al. (2018) utilized LSTM.

Bi-LSTM is an extended version of LSTM which can process long texts from forward and backward directions. To extract text information from CVs Peña et al. (2023) used Bi-LSTM which consists of 32 units and tangent activation function. Hossain et al. (2022) applied Bi-LSTM to produce contextual text representation from both forward and backward directions for input data. Ghosal et al. (2019) fed documents to Bi-LSTM and then to a Multi-Layer Perceptron (MLP-1) for text feature extractions. For emotion recognition from the

F1 dataset, Miao et al. (2021) first used GloVe for tokenizing texts and then passed the word embedding to Bi-LSTM.

Text-CNN is another architecture used for text representation. For sentiment analysis, Xu & Mao (2017) used Text-CNN with 1D convolutional network with 128 kernels each of size five and 1D MaxPooling layer of size 3. Xu et al. (2023) and Wang et al. (2021) also used Text-CNN to extract text features. For generating image description, a type of RNN is used, which is Gated Recurrent Network (GRU) in Babu et al. (2022). They passed image parameters to GRU to process and generate a sequence of words as a description of the image. For text representation and to understand the characteristics of hashtags, Ha et al. (2020) applied TF-IDF as it can capture the importance of hashtags based on their occurrences. Yu et al. (2018) used Doc2Vec for text feature extraction which extends Word2Vec. In contrast to Word2Vec, Doc2Vec turns the complete document into a fixed-length vector while also considering the document's word order. In the paper, Doc2Vec created 300-D features for each document.

Lu et al. (2019) introduced the VilBERT model, or Vision-and-Language BERT, which is intended to develop task-agnostic combined representations of natural language and image content. VilBERT uses the BERT architecture for text, which consists of several layers of transformer encoders. These encoders are used for tokenization and embedding. Learning Cross-Modality Encoder Representations (LXMERT) was designed by Tan & Bansal (2019) for tasks like image captioning, Visual question answering. LXMERT employs a Transformer model for the text modality, which is similar to BERT. It uses feed-forward neural networks and multiple layers of self-attention to process input text. As a result, LXMERT is able to capture the complex contextual relationships present in the text. Huang et al. (2020) introduced a multimodal transformer called PixelBERT. The authors used BERT for text encoding by splitting the sentences into words and used WordPiece to tokenize the words. In Flamingo, Alayrac et al. (2022) used another transformer-based model, which is Generative Pre-training Transformer (GPT). Multimodal Embeddings for Text and Image Representations (METER) is a multimodal model developed by Meta AI (Dou et al., 2022). This model is used for tasks like multimodal classification tasks and image text matching. In this model, the authors used BERT, RoBERTa, and ALBERT to get text encoding. The development of the language models we covered above over time is shown in Figure 2.

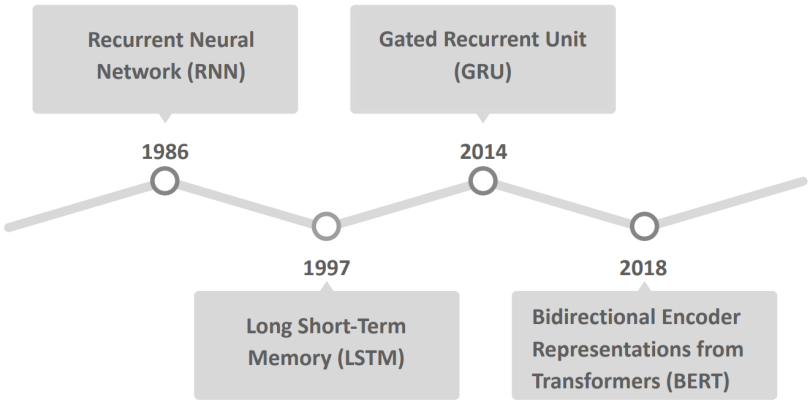


Figure 2: Evolution of machine learning models for NLP tasks

### 3.1.2 Image Feature Extractor

Like texts, there are neural network architectures to extract features and train images. Convolutional Neural Networks (CNNs) are crucial for computer vision and image analysis. In Table 5, we briefly mentioned neural network architectures used in MMML models to extract text features in different articles. In Figure 3, we can see that VGG-16 is the most used architecture among the others. VGG, ResNet, AlexNet, InceptionV3, DenseNet, and SqueezeNet are CNN architectures, which are deep learning models used for image-related tasks. VGG-16 has 13 convolutional layers with three fully connected layers. Every fully connected layer is followed by a dropout layer to prevent overfitting, except for the last layer (Yu et al., 2018). The authors

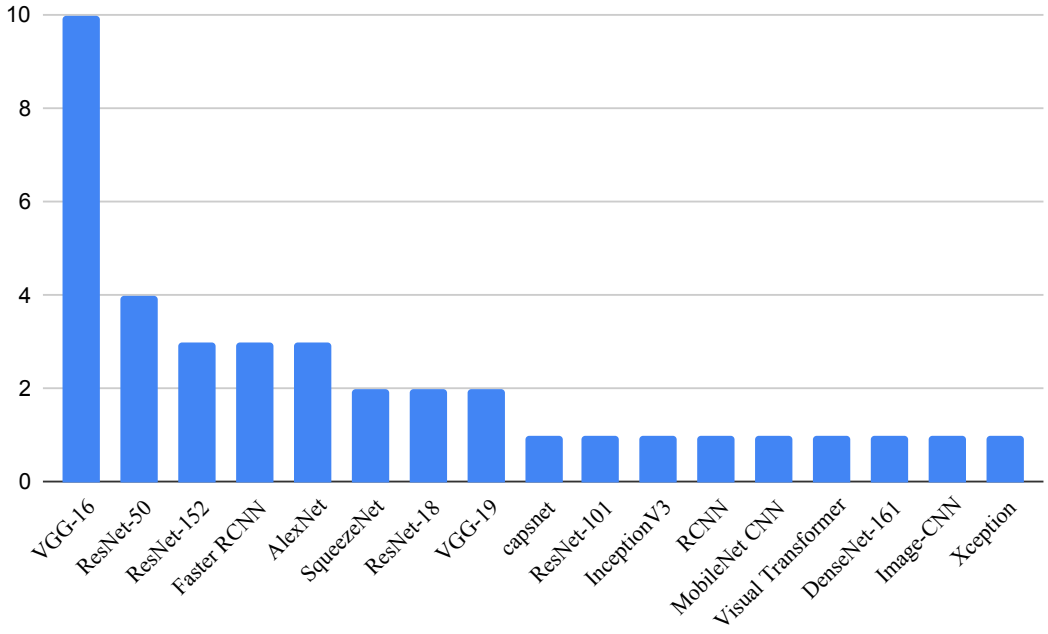


Figure 3: Most commonly used pre-trained models for image feature extraction

ended up with 4096-D features from each image in the paper. For sentiment analysis from the image, Shirzad et al. (2020) used VGG-16, which is pre-trained on the Twitter dataset. They took the pre-trained model trained on the ImageNet dataset, fine-tuned it, and retrained it on the Twitter dataset. Huang et al. (2022) trained VGG-16 on the MINT dataset, which consists of microscopic images. Kim et al. (2021) also worked with pre-trained VGG-16 but changed the last layer with a single sigmoid activation function. Babu et al. (2022) combined two pre-trained models such as 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. Another popular CNN architecture is ResNet-50. For disaster identification, Hossain et al. (2022) used pre-trained ResNet-50 with a bit modification. The authors removed the top two layers of the model. Moreover, they freeze the first 40 layers of the model to use only the weights, and for the last ten layers, they retrain the model with new weights. Rivas et al. (2022) used another version of ResNet with 152 layers and output 2048D features from each image. Another type of ResNet architecture used in multimodality is ResNet-18. Hangloo & Arora (2022) used ResNet-18 to extract visual information that can detect 1000 different categories of objects from images. Apart from the architectures of CNN, Faster-RCNN is another popular pre-trained architecture for image feature extraction. Guo et al. (2022b) extracted the bounding box and features of every object from each image using Faster-RCNN. Besides convolutional neural networks, transformers are also used for image feature encoding. Paraskevopoulos et al. (2022) split the images into sequence patches of 16X16 pixels as the visual transformer is used for sequence processing. Huang et al. (2020) used ResNet for image encoding in their multimodal transformer.

VilBERT uses a modified Faster R-CNN model for images, a deep neural network designed for object detection applications (Lu et al., 2019). The transformer-based architecture, similar to that used for the text, is fed with the visual attributes this network collected from the images. This enables the model to process the visual elements using self-attention in a way similar to how it processes textual data. Tan & Bansal (2019) proposed a visual language model, which is LXMERT, where the authors didn't use any CNN architecture for feature extraction. Instead, they used the object detection method and considered the features of the detected objects. The objects are represented by their bounding box positions and 2048-dimensional Region of Interest (RoI). Microsoft researchers developed Vision and Language (VinVL) and used an object detection model to get visual features. The authors extract region-based features from images using R-CNN (Zhang et al., 2021). Jia et al. (2021) introduced Large-scale Image and Noisy-text (ALIGN),

Table 5: Architectures used to train image features in MMML

| Architecture Name | Article |
|---|---|
| VGG-16 | Ghosal et al. (2019), Xiao et al. (2022), Ahmed et al. (2021), Fatichah et al. (2020), Thuseethan et al. (2020), Shirzad et al. (2020), Yu et al. (2018), Huang et al. (2022), Guo et al. (2022a), Yu et al. (2018), Kim et al. (2021), Babu et al. (2022) |
| VGG-19 | Chen & Zhang (2023), Wang et al. (2021) |
| ResNet-50 | Hossain et al. (2022), Gao et al. (2023), Peña et al. (2023) |
| ResNet-101 | Guo (2023) |
| ResNet-152 | Ban et al. (2022), Rivas et al. (2022), Chandra et al. (2021) |
| ResNet-18 | Lucas et al. (2022), Paraskevopoulos et al. (2022) |
| AlexNet | Fatichah et al. (2020), Ha et al. (2020), Hangloo & Arora (2022) |
| SqueezeNet | Li (2021), Fatichah et al. (2020) |
| DenseNet-161 | Chandra et al. (2021) |
| MobileNet | Sahoo et al. (2023) |
| InceptionV3 | Asgari-Chenaghlu et al. (2022) |
| Faster RCNN | Guo et al. (2022b), Chen et al. (2020) |
| Recurrent CNN | Paul et al. (2020) |
| Image-CNN | Xu & Mao (2017) |
| Visual Transformer | Paraskevopoulos et al. (2022) |
| Xception | Babu et al. (2022) |

where they used EfficientNet for image coding, a variation of CNN architecture. Contrastive Language Image Pre-training (CLIP) was first introduced by Radford et al. (2021) to understand various visual and text concepts. For image encoding, they used a visual transformer. Similar to this, Alayrac et al. (2022) applied a visual transformer to get image features in their model Flamingo. The visual transformer is also used in METER (Dou et al., 2022).

#### 3.1.3 Discussion of Most Popular Architectures

Based on the previous discussion, we conclude that the most commonly used architecture is BERT to extract text features. The existing language models used for natural language processing tasks were unidirectional, where predictions only considered previous tokens they've seen. It raises a problem for the tasks that need bidirectional context understanding. BERT is a pre-trained deep bidirectional model that uses a masked language model and a "next sentence prediction" task to jointly pre-train representations for text pairs (Devlin et al., 2018). BERT's model architecture is almost similar to the transformer described by Vaswani et al. (2017b), a multilayer bidirectional Transformer encoder. In the multilayer encoder, BERT uses multihead self-attention. An attention function maps a query and a set of key-value pairs and outputs the weighted sum of the values. The model can concurrently process data from various representation subspaces at multiple positions with multi-head attention [add transition]

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{1}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and the projections are the following parameter matrices:

$$\begin{aligned} W_i^Q &\in \mathbb{R}^{d_{\text{model}} \times d_k}, \\ W_i^K &\in \mathbb{R}^{d_{\text{model}} \times d_k}, \\ W_i^V &\in \mathbb{R}^{d_{\text{model}} \times d_v}, \\ W^O &\in \mathbb{R}^{h \cdot d_v \times d_{\text{model}}}. \end{aligned}$$

Here, $Q$ is the query matrix, and $K$ and $V$ are the matrices for keys and values (Vaswani et al., 2017b). The pre-training in BERT takes place by combining two tasks: Masked Language Model (MLM) and Next

Sentence Prediction (NSP). In MLM part of BERT, 15% of the input tokens are masked at random, and these masked tokens are then predicted using cross-entropy loss. A replacement technique addresses the fine-tuning challenge, which involves keeping the original tokens and using random and `[MASK]` tokens. In pre-training, a binarized next-sentence prediction task is included to improve the model's comprehension of the relationship between sentences. For example, two sentences, A and B, with a 50% chance that B is the sentence that comes after A (labeled "IsNext") and a 50% chance that B is a random sentence from the corpus (labeled "NotNext"). NSP benefits tasks like Question Answering (QA) and Natural Language Inference (NLI). In the fine-tuning part, BERT aims to tailor the model to a particular task by adapting to a smaller, task-specific dataset to train and modify the parameters of the pre-trained model. The self-attention mechanism of BERT's architecture, in particular, makes it adaptable to perform various tasks, from text classification to question answering, which makes this process efficient. In this part, BERT is fed task-specific input data and outputs accordingly.

We also discussed various techniques to extract image features, and among them, we found different variations of the Residual Network (ResNet) architectures that are primarily used. The use of ResNet architectures is preferable to others because its performance does not decrease even though the model increases the number of layers, and it is computationally efficient. This can be done when adding more layers to the network, making the added layers 'identity mapping' and the other layers are duplicate layers of the original model. This way, training accuracy will not decrease by adding more layers. He et al. (2016) first introduced Residual learning. In their paper, they defined residual block as:

$$y = F(x, \{W_i\}) + x, \tag{2}$$

where $x$ is the input layer, $y$ is the output layer, and the $F$ function is for residual mapping. He et al. (2016) first defined $H(x)$ as mapping function to fit few stacked layers. where x is the number of stacked layers. So, instead of using all stacked layers for the mapping function, the authors use another mapping function, which is $F(x) : H(x) - x$. It makes the original function as $F(x) + x$. It is possible to represent $F(x) + x$ using feedforward neural networks with what are known as "shortcut connections". By using these shortcut connections, one or more layers are skipped. We blend their outputs with the outcomes from the stacked layers, effectively maintaining the original input (identity mapping) through these shortcut connections. Interestingly, these identical shortcut links increase neither the number of parameters nor the computing complexity.

### 3.2 $RQ_{1.2}$ What are the most popular dataset people used to report their performance of MMML models?

In order to answer this research question, we looked through the chosen articles to find the datasets used in multimodal applications. While gathering information about datasets we learned about some of the common data sources used by researchers to create datasets for their research. Twitter, Flicker, IMDB, COCO(Common Objects in Context).

In Figure 4, we summarized all the datasets we encountered in the articles. Rivas et al. (2022), Chandra et al. (2021), Wang et al. (2021), Shirzad et al. (2020) Bhat & Chauhan (2022) used the Twitter dataset, which consists of tweets and images. However, each employed a different Twitter dataset to help them do their tasks. Figure 4 shows that the Flickr30k dataset has been used the most. Yu et al. (2022) used Flickr30k Entities, which is an extension of Flickr30k. This dataset consists of 31,783 images with 44,518 object categories and 158k captions. Another commonly used dataset is MSCOCO. Alsan et al. (2021) used MSCOCO dataset for multimodal data retrival. MSCOCO dataset has an image and text pair and is trained on a dual encoder deep neural network. MSCOCO dataset has 80 object categories and 330k images with five descriptions per image (Babu et al., 2022).

After summarizing datasets used in the articles, we analyzed the performance of datasets in different applications, see Table 6. As a performance measure, we evaluated the F1 score in those applications. We want the F1 score because it's one of the best measures of datasets with imbalanced samples in a number of classes. From the above table, we see that for multimodal image text classification, the work by Liang et al. (2022) on the MM-IMDB dataset, gave the highest F1 score.

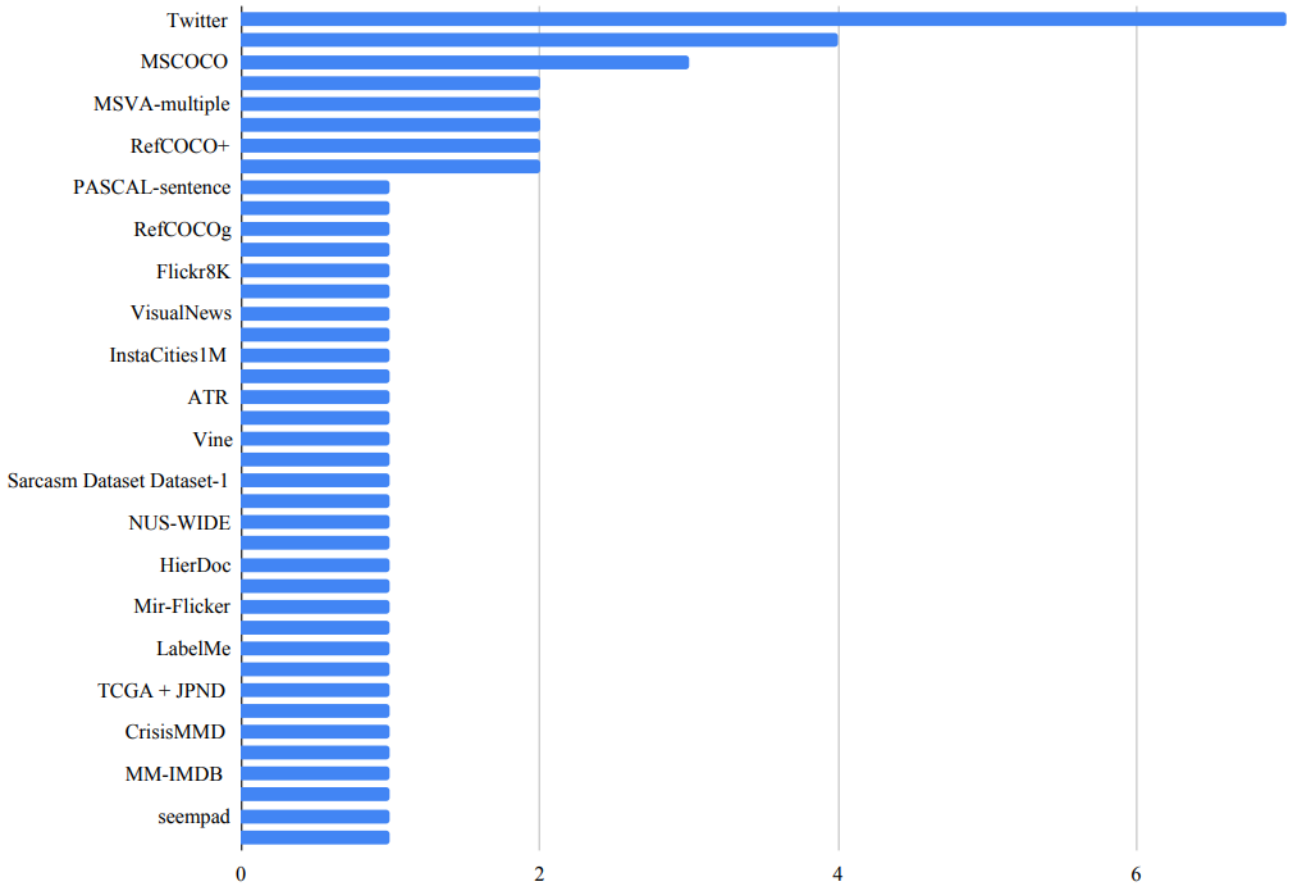


Figure 4: Mostly used dataset in MMML applications

Table 6: Performance summary of datasets in different applications

| Dataset | | | | | | |
|---|---|---|---|---|---|---|
| Weibo | MM-IMDB | FakeNewsNet | FND | Vine | Dataset-1(Sarcasm) | Reference |
| 84.1 | | | | | | Kim et al. (2021) |
| | | | | | | Hossain et al. (2022) |
| | 93.6 | | | | | Liang et al. (2022) |
| | | 92 | | | | Palani et al. (2022) |
| | | | | 78 | | Paul et al. (2020) |
| | | | 76 | | | Hangloo & Arora (2022) |
| 82.37 | | | | | | Wang et al. (2021) |
| | | | | | 86.33 | Yue et al. (2023) |

## 4 RQ2: What fusion strategies usually used in MMML?

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

- **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.
- **Ensemble Technique:** Combines predictions from text and image models.
- **Multimodal Deep Learning Architectures:** Multimodal Compact Bilinear (MCB), Multimodal DBM (Deep Boltzmann Machine), Efficient attention using Transformer, Stacked Autoencoder-Based Multimodal Data Fusion, Multimodal Fusion Architecture Search (MFAS), Bi-LSTM, RNN, GAN(Generative Adversarial Networks).

### 4.1 Concatenation Technique

Concatenation technique means concatenating multiple feature vectors together to get information from the features. Palani et al. (2022) concatenated text and image feature vectors to get multimodal feature vectors to leverage information from both modalities. Paraskevopoulos et al. (2022) used the concatenation technique to concatenate text and visual encoders to assemble them into a classifier model.

### 4.2 Attention Technique

To get relevant parts of each modality Ghosal et al. (2019) used attention mechanism as fusion technique to detect appropriateness in scholarly submission. The authors mentioned that not all modalities contain equal importance. To get important modalities they added an attention layer and calculated attention score. Important modalities contain higher attention scores. Zhang et al. (2023) used a multi-head attention mechanism for joint representation of image and text features. For integrating two modalities, the authors calculate the attention score of text and image features. They used the sigmoid function to calculate the weight of importance of images for source words. Xu et al. (2023) used the attention mechanism to find the relation between each word of a sentence with the corresponding candidate region of an image and calculated the weighted sum to ensure feature association between text and image.

### 4.3 Weight based Technique

A weight based technique is Early fusion that merges data sources in the beginning of the processing. Raw data can be fused directly without any pre-processing, but usually certain features are initially extracted. These basically unimodal features are then fused by concatenating the individual data into a joint representation (Gadzicki et al., 2020). To have joint representation of image and text features, Hossain et al. (2022) 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. Early fusion is used in different multimodal tasks. For disaster identification, Hossain et al. (2022) used early fusion to combine image and text features. The authors computed the feature vectors as:

$$FF^{(i)} = V_f^{(i)} \oplus T_f^{(i)}. \tag{3}$$

Here, $V_f$ is image feature vector and $T_f$ is text feature vector. $FF^{(i)}$ is the concatenation of $i^{th}$ text and image features. Hangloo & Arora (2022) used early for fake news detection from social media posts. For sentiment analysis in multimodal data, to integrate two modalities, Thuseethan et al. (2020) applied the late fusion scheme directly on the features computed for high attention word and salient image region is the straightforward approach to construct a multimodal framework.

### 4.4 Multimodal Deep Learning Architectures

In addition to the methods outlined above, a wide range of deep learning architectures have been devised to support multimodal feature representation, providing improved information fusion and interpretation across various data modalities. One such model is Bi-LSTM which Asgari-Chenaghlu et al. (2022) used to integrate image and text features. To fuse data, Yue et al. (2023) first introduced a knowledge-based network called ConceptNet. The network calculates the pointwise mutual information of the matrix entries, which is smoothed with the contextual distribution.

### 4.5 Discussion

We have examined various methods to smoothly combine data from several data modalities as part of our research into the fusion algorithms frequently utilized in Multimodal Machine Learning (MMML). These approaches include weight-based methods (early fusion), concatenation, attention mechanisms, and other multimodal deep learning architectures. The primary objective of these techniques is to draw conclusions and representations that are insightful from the intricate interactions between text and visual data. Our investigation of fusion strategies highlights how dynamic MMML is, with a wide range of techniques meeting the complex requirements of multimodal data analysis. The particular requirements of the task at hand determine which fusion process is best, as each approach has advantages and uses of its own. Researchers and practitioners in multimodal machine learning (MMML) can completely utilize multimodal data by understanding and expertly implementing these methodologies. This will enhance our capacity to extract valuable insights and make well-informed decisions in various fields.

## 5 RQ3: What are the limitations or challenges to face using these architectures?

Multimodal Machine Learning (MMML) has reached incredible heights thanks to the search for efficient architectures and fusion methods. However, in addition to the advancements, a particular set of restrictions and difficulties have appeared, providing essential insights into the difficulties of integrating various data modalities. In this research question, we explore the limitations or challenges that 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 needs to be huge as MMML models work with data from multiple modalities. Data preprocessing for this huge number of data is both expensive and computationally inefficient (Bayoudh et al., 2021b). Image and text datasets vary in size and difficulty. So training them together is also challenging (Lu et al., 2020).

- **Data Annotation:** The publicly available datasets for text and image are mostly task-specific. Researchers need to make their own dataset for other applications, which requires data annotation. But large-scale data annotation is not widely available (Rahate et al., 2022).

- **Noisy Data:** The noisy data in multimodality causes misclassification, as Chandra et al. (2021) stated in their article. According to the authors' research, the outcome becomes inaccurate if one of the modalities has noisy data.

- **Task Specific Image Feature Extractor:** For online review extraction on the multimodal features, Li (2021) 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 own 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. Liu (2021) 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.

### 5.1 Discussion

Exploring the limitations and difficulties in MMML architectures provides insightful information about the complexities of utilizing many data modalities. When attempting to use MMML, it becomes clear that several important issues must be resolved to overcome these obstacles. To sum up, multimodal data integration is challenging, as seen by the difficulties and limitations encountered while utilizing MMML designs. Since they open the door to improved data annotation resources, task-specific model adaptations, noise reduction strategies, and more effective data preprocessing addressing these issues is crucial to the further development of MMML.

# 6 RQ4: In what way MMML models can be robust against noise and adversarial data?

Label noise and data sample noise are two types of noise that can be present in data quality: label noise refers to faults or undesirable variations in the data labels, while data sample noise is related to errors or changes in the actual data samples. Deep learning methods, particularly those based on adversarial and generative networks, have shown promise in enhancing the quality of data for machine learning tasks by effectively managing label noise and data sample noise. Label noise in datasets arises from various factors, including human mistakes, inexperience, difficult annotation jobs, low-quality data, subjective classifications, reliance on meta-data, and cost-cutting strategies on annotation processes. Label noise is a prevalent problem in real-world applications. In contrast to the ideal circumstances frequently expected in building models, label noise is common. It can result in unfavorable effects, including machine learning applications performing less well, the demand for training data increasing, and possible class imbalances. Domain knowledge can be a powerful tool to reduce label noise. For instance, ontology-based methods enhance classification tasks using hierarchical relationships between data classes. By encoding relationships between labels using a graph network, the Multi-task Graph Convolution Network (MT-GCN) model uses both well-labeled and noisy-labeled data. Auxiliary Classifier GAN (AC-GAN), Conditional GAN (cGAN), Label noise-robust GAN (rGAN), and other extensions of Generative Adversarial Networks (GANs) offer additional techniques for handling label noise (Rahate et al., 2022).

Pre-trained Vision and Language (VL) models have proven more resilient than task-specific models. By introducing noise into the embedding space of VL models, the Multimodal Adversarial Noise GeneratOr (MANGO) technique has been put forth to improve this robustness (Li et al., 2020b). The purpose of MANGO is to evaluate and enhance VL models in response to four kinds of robustness challenges: alterations in the distribution of answers over nine distinct datasets, logical reasoning, linguistic variances, and visual content manipulation. MANGO uses a neural network to produce noise, which hinders the model from readily adjusting, in contrast to techniques that provide predictable local perturbations. This method is supplemented by masking portions of photos and removing text tokens to further diversify input and influence data distribution. Using MANGO to train models has been found to enhance performance on benchmarks.

## 6.1 Discussion

From our search queries and after snowballing, we have found very few papers that discussed noise and adversarial attacks in the multimodal machine learning model. In MMML, the study of robustness and adversarial attacks is still in its infancy, with little research on these complex problems. The potential for adversarial weaknesses may be particularly substantial but understudied, given the inherent intricacy of MMML models, which integrate and correlate information from a variety of input kinds, including text, images, and audio. Research on the adversarial resilience of MMML systems needs to be more critical, as seen by the scarcity of work in this area. This gap offers a chance to do new research to create novel protection mechanisms while delving further into the subtleties of hostile threats in multimodal situations. Expanding research efforts to strengthen MMML models against adversarial attacks is essential as they become more complex to ensure their dependability and credibility in practical applications. Developments in this area may result in multimodal systems that are more resilient and can endure a broader range of hostile strategies.

# 7 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.

#### Acknowledgments

Removed for blind review.

## References

Shobhit 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)*, pp. 0475–0479. IEEE, 2022.

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*, 22:2378, 2022. doi: 10.3390/s22062378.

Md Rekib Ahmed, Neeraj Bhadani, and Ishita Chakraborty. Hateful meme prediction model using multimodal deep learning. In *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*, pp. 1–5. IEEE, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Hüseyin Fuat Alsan, Ekrem Yıldız, Ege Burak Safdil, Furkan Arslan, and Taner Arsan. Multimodal retrieval with contrastive pretraining. In *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pp. 1–5. IEEE, 2021.

Tato Ange, Nkambou Roger, Dufresne Aude, and Frasson Claude. Semi-supervised multimodal deep learning model for polarity detection in arguments. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2018.

Meysam Asgari-Chenaghlu, M Reza Feizi-Derakhshi, Leili Farzinvash, MA Balafar, and Cina 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.

Gali Tanishk Venkat Mahesh Babu, Selvani Deepthi Kavila, and Rajesh Bandaru. Multimodal framework using cnn architectures and gru for generating image description. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 2116–2121. IEEE, 2022.

T. Baltrušaitis, C. Ahuja, and L. Morency. Multimodal machine learning: a survey and taxonomy. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 41:423–443, 2019. doi: 10.1109/tpami.2018.2798607.

Minchao Ban, Liansong Zong, Jie Zhou, and Zhiming Xiao. Multimodal aspect-level sentiment analysis based on deep neural networks. In *2022 8th International Symposium on System Security, Safety, and Reliability (ISSSR)*, pp. 184–188. IEEE, 2022.

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*, 23:1651, 2021. doi: 10.3390/e23121651.

A Barveen, S Geetha, and MK Mohamed Faizal. Meme expressive classification in multimodal state with feature extraction in deep learning. In *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pp. 1–10. IEEE, 2023.

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*, 38:2939–2970, 2021a. doi: 10.1007/s00371-021-02166-7.

Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pp. 1–32, 2021b.

Aruna Bhat and Aditya Chauhan. A deep learning based approach for multimodal sarcasm detection. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 2523–2528. IEEE, 2022.

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*, 59:127–138, 2020. doi: 10.1016/j.inffus.2020.01.008.

W. Chai and G. Wang. Deep vision multimodal learning: methodology, benchmark, and trend. *Applied Sciences*, 12:6588, 2022. doi: 10.3390/app12136588.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, pp. 148–157, 2021.

Donghua Chen and Runtong Zhang. Building multimodal knowledge bases with multimodal computational sequences and generative adversarial networks. *IEEE Transactions on Multimedia*, 2023.

Xiaofan Chen, Songyang Lao, and Ting Duan. Multimodal fusion of visual dialog: A survey. In *Proceedings of the 2020 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence*, pp. 302–308, 2020.

J. Choi and J. Lee. Embracenet: a robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019. doi: 10.1016/j.inffus.2019.02.010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18166–18176, 2022.

Chastine Fatichah, Petrus Damianus Sammy Wiyadi, Dini Adni Navastara, Nanik Suciati, and Abdul Munif. Incident detection based on multimodal data from social media using deep learning methods. In *2020 International conference on ICT for smart society (ICISS)*, pp. 1–6. IEEE, 2020.

Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pp. 1–6, 2020. doi: 10.23919/FUSION45008.2020.9190246.

J. Gao, P. Li, Z. Chen, and J. Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32:829–864, 2020. doi: 10.1162/neco_a_01273.

Lulu Gao, Yali Gao, Jie Yuan, and Xiaoyong Li. Rumor detection model based on multimodal machine learning. In *Second International Conference on Algorithms, Microchips, and Network Applications (AMNA 2023)*, volume 12635, pp. 359–366. SPIE, 2023.

Tirthankar Ghosal, Ashish Raj, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. A deep multimodal investigation to determine the appropriateness of scholarly submissions. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 227–236. IEEE, 2019.

Liye Guo. Art teaching interaction based on multimodal information fusion under the background of deep learning. *Soft Computing*, pp. 1–9, 2023.

Nan Guo, Zhangpeng Fu, and Qihui Zhao. Multimodal news recommendation based on deep reinforcement learning. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 279–284. IEEE, 2022a.

Qingpei Guo, Kaisheng Yao, and Wei Chu. Switch-bert: Learning to model multimodal interactions by switching attention and input. In *European Conference on Computer Vision*, pp. 330–346. Springer, 2022b.

Yui Ha, Kunwoo Park, Su Jung Kim, Jungseock Joo, and Meeyoung Cha. Automatically detecting image–text mismatch on instagram with deep learning. *Journal of Advertising*, 50(1):52–62, 2020.

Sakshini Hangloo and Bhavna Arora. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia systems*, 28(6):2391–2422, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jurgen Schmidhuber. Bridging long time lags by weight guessing and "long short-term memory". *Spatiotemporal models in biological and artificial systems*, 37(65-72):11, 1996.

Eftekhar Hossain, Mohammed Moshiul Hoque, Enamul Hoque, and Md Saiful Islam. A deep attentive multimodal learning approach for disaster identification from social media posts. *IEEE Access*, 10:46538–46551, 2022.

Pengfei Hu, Zhenrong Zhang, Jianshu Zhang, Jun Du, and Jiajia Wu. Multimodal tree decoder for table of contents extraction in document images. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 1756–1762. IEEE, 2022.

Pei-Chi Huang, Ejan Shakya, Myoungkyu Song, and Mahadevan Subramaniam. Biomdse: A multimodal deep learning-based search engine framework for biofilm documents classifications. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3608–3612. IEEE, 2022.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

Luis-Roberto Jácome-Galarza. Multimodal deep learning for crop yield prediction. In *Doctoral Symposium on Information and Communication Technologies*, pp. 106–117. Springer, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Rohit Kumar Kaliyar, Arjun Mohnot, R Raghhul, VK Prathyushaa, Anurag Goswami, Navya Singh, and Palavi Dash. Multideepfake: Improving fake news detection with a deep convolutional neural network using a multimodal dataset. In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10*, pp. 267–279. Springer, 2021.

Aria Naseri Karimvand, Reza Salehi Chegeni, Mohammad Ehsan Basiri, and Shahla Nemati. Sentiment analysis of persian instagram post: a multimodal deep learning approach. In *2021 7th International Conference on Web Research (ICWR)*, pp. 137–141. IEEE, 2021.

Edward Kim, Connor Onweller, and Kathleen F McCoy. Information graphic summarization using a collection of multimodal deep neural networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10188–10195. IEEE, 2021.

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*, 5, 2022. doi: 10.1038/s41746-022-00712-8.

Insaf Kraidia, Afifa Ghenai, and Nadia Zeghib. Hst-detector: A multimodal deep learning system for twitter spam detection. In *International Conference on Computing, Intelligence and Data Analytics*, pp. 91–103. Springer, 2022.

S. Kumaresan, K. Aultrin, S. Kumar, and M. Anand. Transfer learning with cnn for classification of weld defect. *Ieee Access*, 9:95097–95108, 2021. doi: 10.1109/access.2021.3093487.

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*, 153:107419, 2020a. doi: 10.1016/j.measurement.2019.107419.

Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020b.

Meng Li. Research on extraction of useful tourism online reviews based on multimodal feature fusion. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–16, 2021.

Tao Liang, Guosheng Lin, Mingyang Wan, Tianrui Li, Guojun Ma, and Fengmao Lv. Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15492–15501, 2022.

Jiatong Liu. Multimodal machine translation. *IEEE Access*, pp. 1–1, 2021. doi: 10.1109/ACCESS.2021.3115135.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10437–10446, 2020.

Luis Lucas, David Tomás, and Jose Garcia-Rodriguez. Detecting and locating trending places using multimodal social network data. *Multimedia Tools and Applications*, pp. 1–20, 2022.

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*, 235:6577–6585, 2021. doi: 10.1177/09544062211008464.

Anshu Malhotra and Rajni Jindal. Multimodal deep learning architecture for identifying victims of online death games. In *Data Analytics and Management: Proceedings of ICDAM*, pp. 827–841. Springer, 2021.

Haotian Miao, Yifei Zhang, Daling Wang, and Shi Feng. Multimodal emotion recognition with factorized bilinear pooling and adversarial learning. In *Proceedings of the 5th International Conference on Computer Science and Application Engineering*, pp. 1–6, 2021.

Balasubramanian Palani, Sivasankar Elango, and Vignesh Viswanathan K. Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert. *Multimedia Tools and Applications*, 81(4):5587–5620, 2022.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, pp. 36–44, 2023.

Georgios Paraskevopoulos, Petros Pistofidis, Georgios Banoutsos, Efthymios Georgiou, and Vassilis Katsouros. Multimodal classification of safety-report observations. *Applied Sciences*, 12(12):5781, 2022.

Sayanta Paul, Sriparna Saha, and Mohammed Hasanuzzaman. Identification of cyberbullying: A deep learning based multimodal approach. *Multimedia Tools and applications*, pp. 1–20, 2020.

Alejandro Peña, Ignacio Serna, Aythami Morales, Julian Fierrez, Alfonso Ortega, Ainhoa Herrarte, Manuel Alcantara, and Javier Ortega-Garcia. Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment. *SN Computer Science*, 4(5):434, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

Ryan Rivas, Sudipta Paul, Vagelis Hristidis, Evangelos E Papalexakis, and Amit K Roy-Chowdhury. Task-agnostic representation learning of multimodal twitter data for downstream applications. *Journal of Big Data*, 9(1):18, 2022.

Chandan Charchit Sahoo, Deepak Singh Tomar, and Jyoti Bharti. Transformer based multimodal similarity search method for e-commerce platforms. In *2023 IEEE Guwahati Subsection Conference (GCON)*, pp. 01–06. IEEE, 2023.

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*, 29:549–566, 2020. doi: 10.1177/1059712320922916.

Amirhossein Shirzad, Hadi Zare, and Mehdi Teimouri. Deep learning approach for text, image, and gif multimodal sentiment analysis. In *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 419–424. IEEE, 2020.

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*, 32:012001, 2020. doi: 10.1088/1361-6501/ab8df9.

S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara. Multimodal emotion recognition with transformer-based self supervised feature fusion. *Ieee Access*, 8:176274–176285, 2020. doi: 10.1109/access.2020.3026823.

S. Talukder, G. Barnum, and Y. Yue. On the benefits of early fusion in multimodal representation learning. 2020. doi: 10.48550/arxiv.2011.07191.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Selvarajah Thuseethan, Sivasubramaniam Janarthan, Sutharshan Rajasegarar, Priya Kumari, and John 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)*, pp. 267–274. IEEE, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. 2017a. doi: 10.48550/arxiv.1706.03762.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017b.

Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 3708–3716, 2021.

Shuaiyong Xiao, Gang Chen, Chenghong Zhang, and Xiangge Li. Complementary or substitutive? a novel deep learning method to leverage text-image interactions for multimodal review helpfulness prediction. *Expert Systems with Applications*, 208:118138, 2022.

Jinzhong Xu, Hailong Zhao, Weiguang Liu, and Xinyang Ding. Research on false information detection based on multimodal event memory network. In *2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pp. 566–570. IEEE, 2023.

Nan Xu and Wenji Mao. A residual merged neutral network for multimodal sentiment analysis. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pp. 6–10. IEEE, 2017.

Ashima Yadav and Dinesh Kumar Vishwakarma. A deep multi-level attentive network for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19 (1):1–19, 2023.

Yi Yu, Suhua Tang, Kiyoharu Aizawa, and Akiko Aizawa. Category-based deep cca for fine-grained venue discovery from multimodal data. *IEEE transactions on neural networks and learning systems*, 30(4): 1250–1258, 2018.

Zhihan Yu, Mingcong Lu, and Ruifan Li. Multimodal co-attention mechanism for one-stage visual grounding. In *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pp. 288–292. IEEE, 2022.

Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. Knowlenet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion*, 100:101921, 2023.

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*, 14:478–493, 2020. doi: 10.1109/jstsp.2020.2987728.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5579–5588, 2021.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Universal multimodal representation for language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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*, 171:107466, 2020. doi: 10.1016/j.sigpro.2020.107466.