machine learning & knowledge extraction

MDPI

Systematic Review

Navigating the Multimodal Landscape: A Review on Integration of Text and Image Data in Machine Learning Architectures

Maisha Binte Rashid [1], Md Shahidur Rahaman [2] and Pablo Rivas [1,*]

[1] Department of Computer Science, Baylor University, Waco, TX 76706, USA; maisha_rashid1@baylor.edu
[2] Department of Computer Science, Texas A&M University, College Station, TX 77843, USA; mdshahidur_rahaman@tamu.edu
* Correspondence: pablo_rivas@baylor.edu

Abstract: Images and text have become essential parts of the multimodal machine learning (MMML) framework in today's world because data are always available, and technological breakthroughs bring disparate forms together, and while text adds semantic richness and narrative to images, images capture visual subtleties and emotions. Together, these two media improve knowledge beyond what would be possible with just one revolutionary application. This paper investigates feature extraction and advancement from text and image data using pre-trained models in MMML. It offers a thorough analysis of fusion architectures, outlining text and image data integration and evaluating their overall advantages and effects. Furthermore, it draws attention to the shortcomings and difficulties that MMML currently faces and guides areas that need more research and development. We have gathered 341 research articles from five digital library databases to accomplish this. Following a thorough assessment procedure, we have 88 research papers that enable us to evaluate MMML in detail. Our findings demonstrate that pre-trained models, such as BERT for text and ResNet for images, are predominantly employed for feature extraction due to their robust performance in diverse applications. Fusion techniques, ranging from simple concatenation to advanced attention mechanisms, are extensively adopted to enhance the representation of multimodal data. Despite these advancements, MMML models face significant challenges, including handling noisy data, optimizing dataset size, and ensuring robustness against adversarial attacks. Our findings highlight the necessity for further research to address these challenges, particularly in developing methods to improve the robustness of MMML models.

Keywords: image; text; neural network; multimodality; machine learning; adversarial attack; fusion

Citation: Binte Rashid, M.; Rahaman, M.S.; Rivas, P. Navigating the Multimodal Landscape: A Review on Integration of Text and Image Data in Machine Learning Architectures. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1545–1563. https://doi.org/10.3390/make6030074

Academic Editor: Ahmad Taher Azar

Received: 13 March 2024
Revised: 19 June 2024
Accepted: 3 July 2024
Published: 9 July 2024

1. Introduction

The rapid advancement in digital technologies has precipitated an unprecedented increase in data across a multitude of fields, heralding a significant transformation in our comprehension of intricate systems [1,2]. This surge in data spans several modalities, encompassing visual elements in photographs, the semantic aspects of text, and auditory signals, thus offering a holistic view of the environment [3,4]. This complex environment has paved the way for the emergence of multimodal machine learning (MMML), which seeks to forge computational models that can assimilate information from varied modalities, thereby enhancing prediction accuracy and the efficacy of decision-making processes [2,5].

The rationale behind integrating multiple modalities stems from the inherent shortcomings of relying solely on single-mode data. Despite their detailed visual content, images may miss the contextual richness achievable through text [6]. Conversely, text, while semantically dense, often falls short of conveying the entirety of visual or auditory experiences [7]. The amalgamation of these modalities fosters the creation of models that are both intricate and nuanced, mirroring the perceptual abilities of humans [8,9].

The introduction of deep learning frameworks has notably advanced MMML's potential, facilitating the intricate extraction and integration of features from diverse data

streams [10,11]. Nonetheless, the task of crafting effective multimodal frameworks is fraught with challenges, including reducing overfitting, managing data disparities, and filtering out data noise [12,13]. Successful frameworks are those that deftly maintain the distinct characteristics of each modality while capitalizing on the synergies between them to enhance overall model performance [14,15].

In this era marked by the omnipresence of data and the melding of technologies, the modalities of text and imagery stand at the forefront of the MMML field. Images capture visual intricacies and convey emotional subtleties, whereas text offers semantic depth and narrative coherence [16,17]. Integrating these modalities unveils insights that surpass their parts, transforming a variety of application areas [18,19]. This study makes the following contributions:

- Exploration of how MMML leverages pre-trained models to extract features from both textual and visual data, highlighting methods that enhance data representation.
- A comprehensive review of fusion techniques, detailing approaches for integrating text and image data, along with an analysis of their benefits and impacts.
- Discussion of the limitations and challenges encountered in MMML.
- Examination of the resilience of MMML models against noisy and adversarial data to determine their adaptability and practicality in real-world scenarios.

The structure of the remainder of this paper is as follows: Section 2 outlines the research methodology employed. Subsequent sections delve into the research questions more thoroughly.

2. Methodology

This Section 2 delineates the comprehensive approach adopted to scrutinize various facets of multimodal machine learning (MMML). The process initiates with the formulation of precise research questions, proceeds with detailed search strategies, and culminates in the systematic extraction and assimilation of data, incorporating a stringent quality evaluation. This scoping review was reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) , ensuring a rigorous and transparent approach; see Figure 1 for additional details.

2.1. Research Questions

This section introduces a structured approach to navigate the intricacies of MMML. This begins with carefully crafting specific research questions that aim to guide our investigation into the nuanced aspects of MMML. These questions focus on key areas such as applying pre-trained models for feature extraction, the diversity and effectiveness of fusion methodologies, the challenges inherent to these architectures, and the resilience of MMML models in the face of noisy or adversarial data. Through a detailed examination, we formulated the following research queries:

- **RQ1:** Are well-established, pre-existing architectures utilized in multimodal machine learning models?
  - $RQ_{1.1}$ Which pre-trained models are predominantly employed for the processing and learning of image and text data?
  - $RQ_{1.2}$ Which datasets are commonly utilized for benchmarking these models?
- **RQ2:** Which fusion techniques are prevalently adopted in MMML?
- **RQ3:** What limitations or obstacles are encountered when using these architectures?
- **RQ4:** In what way can MMML models be robust against noise and adversarial data?

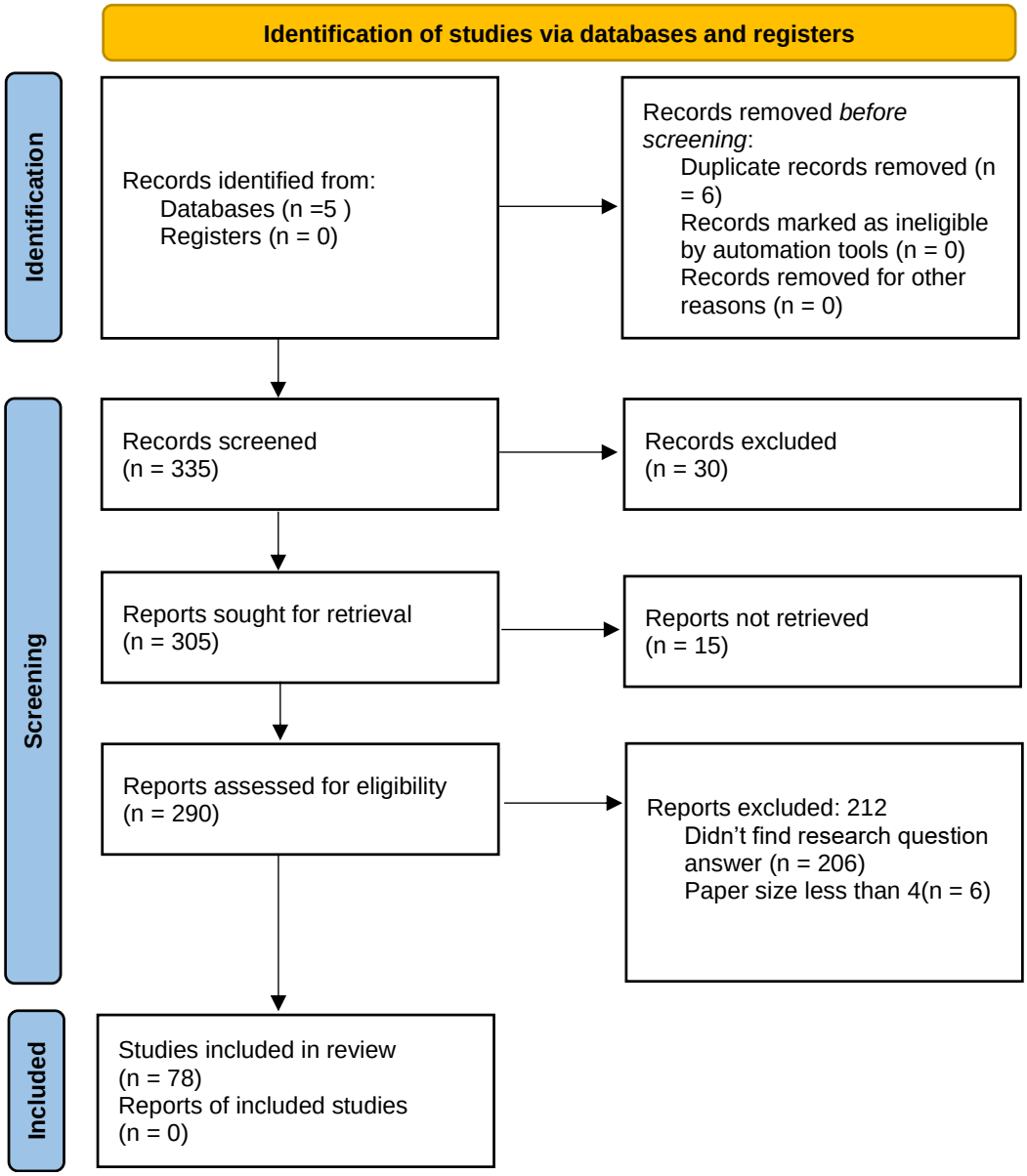

Figure 1. PRISMA 2020 flow diagram for our scoping review. See our detailed Open Science Foundation registry here: https://osf.io/vn3dt (accessed on 29 April 2024).

2.2. Searching Methodology

To address our research inquiries, we conducted a comprehensive search across multiple digital libraries to identify pertinent scholarly articles. We assembled an extensive corpus of the relevant literature through our detailed exploration of various academic databases. The digital libraries utilized for this search included the following:

- Scopus;
- IEEE Xplore;
- SpringerLink;
- ACM Digital Library;
- Semantic Scholar.

In our strategic pursuit of relevant academic materials, we employed a wide array of keywords, including **multimodality**, **deep learning**, **machine learning**, **neural network**, **image**, and **text**. This selection of keywords was meticulously crafted to encompass all topics pertinent to our study. These keywords served as the foundation for our search

queries within the aforementioned databases. The search strategies we implemented are given below:

- **Scopus**
  - Query executed: (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))).
  - Filter criteria: no filters were applied.
- **IEEE Xplore**
  - Query executed: ((((“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”))).
  - Filter criteria: no filters were applied.
- **SpringerLink**
  - Query executed: where the title contains multimodal; query: text AND image AND (“deep learning” OR “machine learning” OR “neural network”); sort by relevance.
  - Filter criteria: the top 32 most pertinent entries were selected.
- **ACM Digital Library**
  - Abstract: (neural) AND Title: (multimodal) AND Abstract: (deep learning) AND NOT Title: (video) AND NOT Title: (audio) AND E-Publication Date: (27 June 2018 TO 27 June 2023).
  - Filter criteria: sorted by relevance.
- **Semantic Scholar**
  - Query executed: keywords: multimodal machine learning deep learning image text. Dates: (1 January 2018 To 31 April 2023). Sort by relevance.
  - Filter criteria: the top 30 entries by relevance, including a ‘TL;DR’ visual summary, were chosen.

From SpringerLink, we chose 10% of the articles for initial screening from the search result because it generated nearly 300 articles that seemed irrelevant to our study. Similarly, for Semantic Scholar, we took the first 1% of articles for initial screening as we obtained thousands of papers from the search query. We did not use such filters for initial screening for the other databases because the articles generated from search queries seem relevant to our research goals. The time frame filter was added to the ACM and Semantic Scholar databases to retrieve the most recent and relevant research articles from the past five years. This was conducted because these databases were returning older publications that were less relevant to the current research objectives. We designed this selection strategy to capture a representative and a high-quality sample of the current research landscape in our field of study.

In our initial search, we obtained 341 research articles. After removing duplicates, 335 articles remained for screening. During the abstract and title screening phase, we excluded 30 articles, leaving 330 papers for full-text screening. However, due to library access limitations, we could not access the full texts of 15 papers, reducing the count to 290 papers for eligibility assessment. By applying our exclusion criteria to these 290 papers, we ultimately identified 88 relevant papers that addressed our research question and were included in our review study.

2.3. Selection Criteria

Following retrieving research papers from the databases using our search queries, we established criteria for inclusion and exclusion to refine our selection. The inclusion criteria were designed to incorporate research publications discussing multimodal machine

learning (MMML) models applied across different settings, particularly those involving image and text data. Conversely, we excluded research papers that did not pertain to MMML or dealt with modalities beyond image and text, ensuring our focus remained tightly aligned with the core objectives of our study.

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 the performance of multimodal machine learning models.
- Papers that are in English.

2.3.2. Exclusion Criteria

- Papers that have a length less than four pages.
- Papers that are not in English.
- Papers that are not peer-reviewed.
- Papers that are not published in any conference/journal.
- Articles with full text not available in the specified database.
- Opinion papers.
- Papers that worked with data other than image and text.

Following the execution of our search strategies as described above, we initially identified 341 research papers. By applying our predetermined inclusion and exclusion criteria to this collection, we could refine our selection down to 88 papers that directly contributed to addressing the research questions at the heart of our study. During our investigation and finalization of our paper, we came across several recent studies in the latter part of 2023 that delve into the latest developments in the domain of multimodal models. We have found recent innovations of MMML models in a work of Guo et al. [20] which is a survey study of these models. Finding these contributions highly pertinent, we incorporated ten more papers into our corpus. Table 1 illustrates the distribution of papers across each database, both before and after applying our selection criteria, providing a clear overview of our research process and the basis of our literature review.

Table 1. Papers from each database before and after selection criteria.

Database Name	Before Exclusion	After Exclusion
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 obtain answers to the research questions, we looked into different sections of the articles. Table 2 discusses the relevant sections for each research question.

Table 2. Data extraction for research questions from different sections.

Research Question	Preferred Section
RQ1, RQ2	Methodology/Model Description/Dataset/Results
RQ3	Limitations/Future Work/Research Gap
RQ4	Limitations/Dataset/Data Pre-processing

3. RQ1: Are Well-Established, Pre-Existing Architectures Utilized in Multimodal Machine Learning Models?

In addressing this research question, our objective was to delve into the types of architectures utilized for multimodal machine learning (MMML) models, specifically focusing on training models for both text and image data. An exhaustive review of the finalized selection of papers showed that MMML models frequently employ well-established, pre-trained architectures to train image and text data. This approach underscores the reliance on proven neural network architectures pre-trained on extensive datasets, facilitating the effective learning and integration of multimodal data.

3.1. $RQ_{1.1}$ Which Pre-Trained Models Are Predominantly Employed for the Processing and Learning Image and Text Data?

This research question is designed to guide researchers in identifying which architectures are most effective for developing MMML models that process text and image data. By determining the preferred pre-trained architectures within the field, this inquiry aids in addressing the foundational structures that have demonstrated success in integrating and analyzing multimodal data.

3.1.1. Text Feature Extractor

In our exploration of pre-trained architectures for text data within MMML models, we discovered that Bidirectional Encoder Representations from Transformers (BERT) is the predominant choice. As evidenced in Table 3, BERT stands out as the most frequently utilized model for training text data. It operates by randomly masking word tokens and representing each masked token with a vector, thereby capturing the semantic and contextual essence of the input text. This capability makes BERT highly effective in a variety of applications, such as detecting fake news, identifying rumors, recognizing sarcasm, locating trending places from social media posts, combating online antisemitism, predicting the helpfulness of reviews, and analyzing tourism online reviews, as referenced in various studies [21–28].

Although BERT is heavily used, other architectures like RoBERTa, a modification of BERT by Facebook aimed at detection tasks, have also been used [26,29]. Following BERT, Long-Short Term Memory (LSTM) networks are another commonly used architecture for MMML models, particularly beneficial in applications such as sentiment analysis, creating visual logs, multimodal retrieval, and polarity detection [30–33].

While BERT and LSTM dominate the landscape for text data processing in MMML models, other architectures also contribute but to a lesser extent. Though not as popular, these models play a significant role in the diverse applications of MMML. A summary of the neural network architectures deployed for extracting text features across various studies is presented in Table 3, highlighting the versatility and range of tools available for researchers in the field.

Table 3. Architectures used to train text features in MMML.

Architecture Name	Article	Total Articles
BERT	[21–28,34–46]	21
LSTM	[22,30,32,33,38,41,47–50]	10
Bi-LSTM	[45,51–54]	5
Residual Bi-LSTM	[55]	1
TF-IDF	[56]	1
GRU	[41,57,58]	3
GREEK BERT	[59]	1
RoBERTa	[26,29]	2
Text CNN	[31,45,60,61]	4
CLIP ViT-L/14	[62]	1
Bi-GRU	[63]	1
VADER	[64]	1
Doc2Vec	[65]	1
RNN	[40,41]	2
LinearSVC	[65]	1
LSTM-RNN	[66]	1
GloVe	[67,68]	2
VD-CNN	[69]	1
Not Applicable		29

BERT has emerged as a fundamental framework in Natural Language Processing (NLP) tasks, particularly notable for its depth in text representation and interpretation within multimodal contexts. Its application extends to review helpfulness prediction, where Xiao et al. [27] employed BERT to transform texts into sequential embeddings, with each row vector denoting a word, thereby enhancing the accuracy of review helpfulness predictions. Moreover, Gao et al. [23] utilized BERT's WordPiece subword tokenization algorithm to create a word dictionary, optimizing word segmentation by selecting the most likely merges. Agarwal [39] applied the WordPiece tokenizer for processing clinical data with BERT, demonstrating its versatility across various datasets.

Li [28] introduced a novel attention mechanism through BERT to better connect review comments, thereby improving the textual analysis's relevance and interpretability. Sahoo et al. [43] highlighted BERT's ability to handle long sentences without fixed input size constraints, making it an ideal choice for extensive text feature extraction. Furthermore, Xu et al. [45] utilized BERT's multi-head attention mechanism to explore deep semantic relationships within sentences, showcasing the model's advanced analytical capabilities. The adoption of BERT for text embedding by Lucas et al. [25], Yu et al. [44], Ban et al. [41], and Liang et al. [42] further validates its effectiveness in extracting meaningful text features.

On the other hand, Long Short-Term Memory (LSTM) networks, designed to overcome the vanishing gradient problem of traditional Recurrent Neural Networks (RNNs), have also been widely used for text feature extraction. The application of LSTM ranges from extracting text features from visual logs by Chen et al. [31], optimizing pre-trained word-embedding matrices for advanced feature generation by Yadav and Vishwakarma [30], to encoding texts into feature vectors by Alsan et al. [32]. Ange et al. [33] employed LSTM to account for various emotional states, sentiments, and prior opinions in polarity detection tasks, illustrating its capacity to process complex sequential data and its importance in sentiment analysis.

These instances underscore the critical role that BERT and LSTM play in enhancing MMML models through sophisticated mechanisms for deep semantic analysis and feature extraction from text data, thereby boosting model performance across various applications. Bi-LSTM is an extended version of LSTM that can process long texts from forward and backward directions. To extract text information from CVs, Peña et al. [51] used Bi-LSTM, which consists of 32 units and a tangent activation function. Hossain et al. [54] applied Bi-LSTM to produce contextual text representation from both forward and backward directions for input data. Ghosal et al. [52] 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. [53] 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 and Mao [61] used Text-CNN with 1D convolutional network with 128 kernels each of size five and 1D MaxPooling layer of size 3. Xu et al. [45] and Wang et al. [60] also used Text-CNN to extract text features for false/fake news detection. A type of RNN is used for generating image description, which is Gated Recurrent Network (GRU) in Babu et al. [58]. They passed image parameters to GRU to process and generate a sequence of words to describe the image. For text representation and to understand the characteristics of hashtags, Ha et al. [56] applied TF-IDF as it can capture the importance of hashtags based on their occurrences. Yu et al. [65] 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. [70] introduced the VilBERT model, or Vision-and-Language BERT, 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 and Bansal [71] for tasks like image captioning and visual question answering. LXMERT employs a transformer model for the text modality, similar to BERT. It uses feed-forward neural networks and multiple layers of self-attention to process input text. As a result, LXMERT can capture the complex contextual relationships present in the text. Huang et al. [72] 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. [73] used another transformer-based model, Generative Pre-training Transformer (GPT). Multimodal Embeddings for Text and Image Representations (METER) is a multimodal model developed by Meta AI [46]. This model is used for multimodal classification tasks and image text matching. The authors used BERT, RoBERTa, and ALBERT to obtain text encoding in this model.

3.1.2. Image Feature Extractor

Just as with texts, there are specific neural network architectures designed for extracting features from and training images. Convolutional Neural Networks (CNNs) play a pivotal role in computer vision and image analysis tasks. In Table 4, we provide an overview of the neural network architectures employed in MMML models for image feature extraction, as referenced in various studies. According to Table 4, VGG-16 emerges as the most utilized architecture among others for image-related tasks. Architectures such as VGG, ResNet, AlexNet, InceptionV3, DenseNet, and SqueezeNet represent the suite of CNN models employed for deep learning tasks in imaging. VGG-16, specifically, is characterized by its 13 convolutional layers and three fully connected layers, with dropout layers following each fully connected layer to mitigate overfitting, with the exception of the last layer [65]. This configuration yields 4096-D features from each image.

Table 4. Architectures used to train image features in MMML.

Architecture Name	Article	Total Articles
VGG-16	[27,38,40,52,58,64,65,65,68,69,74,75]	12
VGG-19	[60,67]	2
ResNet-50	[23,51,54]	3
ResNet-101	[76]	1
ResNet-152	[26,41,57]	3
ResNet-18	[25,59]	2
AlexNet	[22,56,74]	3
SqueezeNet	[28,74]	2
DenseNet-161	[26]	1
MobileNet	[43]	1
InceptionV3	[34]	1
Faster RCNN	[31,36]	2
Recurrent CNN	[55]	1
Image-CNN	[61]	1
Visual Transformer	[59]	1
Xception	[58]	1
Not Applicable		51

For image-based sentiment analysis, Shirzad et al. [64] utilized a pre-trained VGG-16 model, initially trained on the ImageNet dataset, then fine-tuned and retrained on a Twitter dataset. Huang et al. [40] engaged VGG-16 for training on the MINT dataset containing microscopic images. Kim et al. [68] adapted a pre-trained VGG-16 model, modifying the last layer with a sigmoid activation function. Babu et al. [58] integrated two pre-trained models, VGG-16 and Xception (both originally trained on the ImageNet dataset), for image feature extraction, where VGG-16 includes 16 convolutional layers, and Xception comprises 71 layers.

ResNet-50 is another widely adopted CNN architecture. For instance, Hossain et al. [54] employed a pre-trained ResNet-50 with modifications for disaster identification, removing the top two layers and retraining the last ten layers with new weights while freezing the first 40 layers. Rivas et al. [57] utilized a ResNet version with 152 layers, extracting 2048-D features from each image. ResNet-18 has also been used in multimodal applications; Hangloo and Arora [22] utilized ResNet-18 to extract visual information capable of identifying 1000 different object categories.

Beyond CNNs, Faster-RCNN has been employed for image feature extraction, with Guo et al. [36] using it to identify and extract features from objects within images. Additionally, transformers, typically known for sequence processing, have been adapted for image encoding. Paraskevopoulos et al. [59] divided images into 16x16 pixel patches for processing with a visual transformer, and Huang et al. [72] used ResNet within a multimodal transformer for image encoding.

VilBERT uses a modified Faster R-CNN model for images, a deep neural network designed for object detection applications [70]. 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, similar to how it processes textual data. Tan and Bansal [71] proposed a visual language model, LXMERT, where the authors did not 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 obtain visual features. The authors extract region-based features from images using R-CNN [77]. Jia et al. [78] introduced Large-Scale Image and Noisy-Text (ALIGN), where they used EfficientNet for image coding, a variation in CNN architecture. Contrastive Language Image Pre-training (CLIP) was first introduced by Radford et al. [79] to understand various visual and text concepts. For image encoding, they used a visual transformer. Similarly, Alayrac et al. [73] applied a visual transformer to obtain image features in their model Flamingo. The visual transformer is also used in METER [46].

3.1.3. Description of Language and Image Architectures

Based on the previous discussion, we found that the most commonly used architecture to extract text features is BERT. The existing language models used for natural language processing tasks were unidirectional, where predictions only considered previous tokens they have 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 [80]. BERT's model architecture is almost similar to the transformer described by Vaswani et al. [1], 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, as follows:

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

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 [1]. The pre-training in BERT takes place by combining two tasks: masked language model (MLM) and next sentence prediction (NSP). In the 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 in 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 computation-

ally efficient. This can be conducted when adding more layers to the network, making the added layers 'identity mapping' and the other layers duplicate layers of the original model. This way, training accuracy will not decrease by adding more layers. He et al. [81] 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. [81] first defined $H(x)$ as a mapping function to fit a 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. $RQ_{1.2}$ Which Datasets Are Commonly Utilized for Benchmarking These Models?

To address this research question, we meticulously reviewed the selected articles to identify the datasets employed in multimodal applications. Through this review, we uncovered several common data sources researchers frequently utilize to compile study datasets. These include social media platforms such as Twitter and Flickr, which offer rich textual and visual data sources. Additionally, we identified widely recognized datasets such as IMDB, known for its extensive collection of movie reviews and metadata, and COCO, a benchmark dataset in the field of computer vision for object detection, segmentation, and captioning tasks. This exploration highlights the diverse range of datasets that underpin research in multimodal machine learning, reflecting the broad applicability of MMML models across various domains and data types.

In our comprehensive review of the datasets encountered within the selected articles, we compiled findings into Figure 2, showcasing the diversity and frequency of dataset usage in multimodal machine learning research. Notably, the Twitter dataset, comprising tweets and images, was utilized by several researchers, including [26,29,57,60,64]. Each study selected a distinct Twitter dataset tailored to their specific research tasks.

Figure 2 highlights that the Flickr30k dataset is the most frequently used among the datasets we reviewed. An extension of this, the Flickr30k Entities dataset was employed by Yu et al. [44], encompassing 31,783 images with 44,518 object categories and 158k captions, providing a rich resource for training and testing multimodal machine learning models.

Another pivotal dataset in the field is MSCOCO, utilized by Alsan et al. [32] for multimodal data retrieval. The MSCOCO dataset, renowned for its comprehensive pairing of images and text, includes 80 object categories across 330k images, each accompanied by five descriptions, offering an extensive basis for training on dual encoder deep neural networks [58]. This assortment of datasets underscores the vast potential and applicability of MMML models across various contexts and data types, highlighting the significance of dataset selection in developing and evaluating these models.

After summarizing datasets used in the articles, we analyzed the performance of datasets in different applications; see Table 5. For performance analysis, we gathered articles that reported the F1 score as a metric for evaluating dataset performance. We want the F1 score because it is one of the best measures of datasets with imbalanced samples in a number of classes. From the above table, we see that the work by Liang et al. [42] on the MM-IMDB dataset gave the highest F1 score for multimodal image text classification.

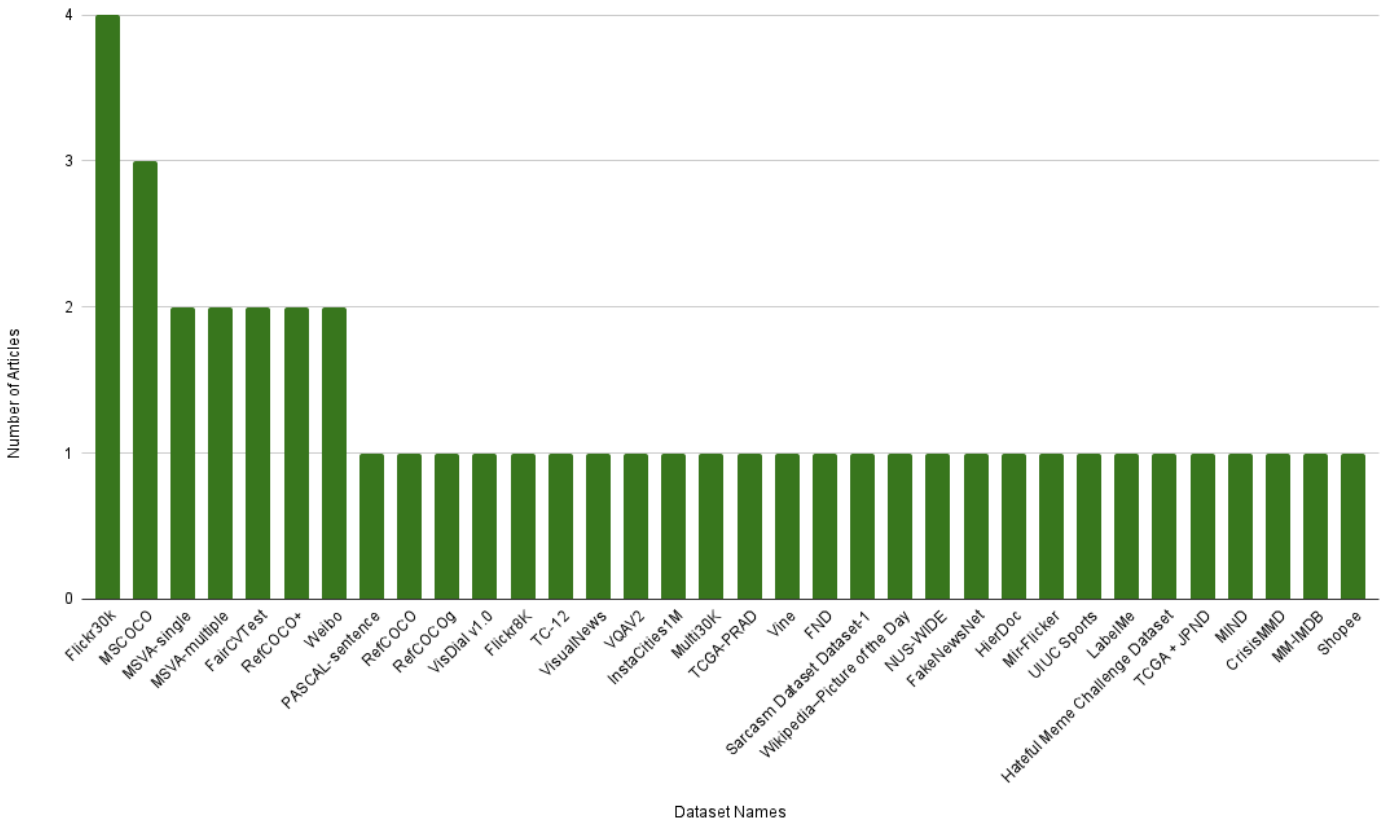

Figure 2. Mostly used dataset in MMML applications.

Table 5. Performance metrics across different datasets.

Dataset	F1 Score (%)	Reference
Weibo	84.1	[68]
Weibo	82.37	[60]
MM-IMDB	93.6	[42]
FakeNewsNet	92	[21]
FND	76	[22]
Vine	78	[55]
Dataset-1 (Sarcasm)	86.33	[24]

4. RQ2: Which Fusion Techniques Are Prevalently Adopted in MMML?

Reviewing the literature, we identified various fusion techniques employed in multimodal machine learning (MMML) models. These techniques, pivotal for integrating textual and visual data, are classified based on their structural and methodological approaches into several categories:

- **Concatenation Technique:** This method involves the straightforward combination of textual and visual vectors to create a unified representation, facilitating the simultaneous processing of both data types. For instance, Palani et al. [21] concatenated text and image feature vectors to generate multimodal feature vectors, thereby harnessing the strengths of both text and visual information. The authors performed the concatenation by averaging the vector values in each vector position. Similarly, Paraskevopoulos et al. [59] applied the concatenation technique to merge text and visual encoders, assembling them into a classifier model to enhance the model's interpretative power.

- **Attention Technique:** This approach utilizes the attention mechanism to focus on specific parts of the text and image features, enhancing the model's ability to discern relevant information from both modalities for improved decision-making. Ghosal et al. [52] utilized an attention mechanism as a fusion technique for detecting appropriateness in scholarly submissions, acknowledging that not all modalities are equally important. By introducing an attention layer and computing attention scores, the model could prioritize modalities with higher relevance, as demonstrated by Zhang et al. [35] who employed a multi-head attention mechanism for the joint representation of image and text features, calculating attention scores to weight the importance of images for source words. Xu et al. [45] further explored this technique by using the attention mechanism to discern relationships between words in a sentence and corresponding image regions, thereby ensuring a meaningful association between text and image features.
- **Weight-based Technique:** This category includes Early Fusion, Late Fusion, and Intermediate Fusion techniques, each applying different weightage strategies to the integration process, allowing for a nuanced amalgamation of modalities at various stages of the model's architecture. Hossain et al. [54] utilized Early Fusion for disaster identification by merging image and text features, ensuring equal representation from each modality by taking the same number of nodes from the last hidden layer of each modality. This technique was also applied by Hangloo and Arora [22] for detecting fake news in social media posts. Late Fusion, on the other hand, is applied after feature computation, as seen in the work of Thuseethan et al. [69] for sentiment analysis, where it directly integrates features computed for attention-heavy words and salient image regions, showcasing the versatility of weight-based fusion in constructing multimodal frameworks.
- **Deep Learning Architectures:** In the field of multimodal deep learning architectures, the development and application of diverse deep learning models have significantly advanced the area of multimodal feature representation. These architectures facilitate enhanced fusion and interpretation of information across different data modalities. A notable example is using Bi-LSTM by Asgari-Chenaghlu et al. [34] for integrating image and text features, showcasing the model's ability to handle sequential data effectively. Additionally, Yue et al. [24] introduced a knowledge-based network, ConceptNet, to fuse data. This network employs the calculation of pointwise mutual information for matrix entries, further refined by smoothing with the contextual distribution, illustrating an innovative approach to integrating multimodal data.

A summary of these techniques is given in Table 6.

Table 6. Fusion technique categories used in articles.

Category	References
Concatenation Technique	[21,59]
Attention Technique	[35,45,52]
Weight-based Technique	[22,54,69]
Deep Learning Architectures	[24,34]

5. RQ3: What Limitations or Obstacles Are Encountered When Using These Architectures?

The exploration of MMML has unveiled significant advancements in efficient architectures and fusion methods. However, challenges and limitations have emerged alongside these developments, highlighting the complexities of integrating various data modalities. After thoroughly reviewing the research papers, we observed that most should have extensively discussed the limitations or obstacles encountered when working with multimodal machine learning models. However, through a rigorous examination of the articles, we identified and categorized the limitations that other researchers have encountered. In this section, we investigated the limitations or challenges encountered when utilizing MMML architectures, categorizing common issues observed in MMML models:

- **Dataset Size:** One of the primary challenges in MMML models is determining the optimal size for datasets, as these models require large datasets due to data integration from multiple modalities. Data preprocessing for such vast amounts of data is costly and computationally intensive [9]. Furthermore, the disparity in size and complexity between image and text datasets complicate their simultaneous training [82].
- **Data Annotation:** Most publicly available datasets for text and images are tailored for specific tasks, necessitating the creation of custom datasets for new applications. This process involves data annotation, which, on a large scale, is often not readily accessible [83].
- **Noisy Data:** The presence of noisy data within multimodal contexts can lead to misclassification [26]. The accuracy of outcomes diminishes if one of the modalities contains noisy data, underscoring the importance of data quality in MMML models.
- **Task-Specific Image Feature Extractor:** The effectiveness of MMML models can be limited by task-specific image feature extractors. Challenges in extracting relevant features due to the inappropriateness of the method for specific tasks highlight the need for task-aligned model selection [28,84].

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 metadata, 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.

To address this research question, we examined the articles' methodology and discussion sections, seeking information about adversarial attacks, noisy data, and adversarial robustness. We aimed to identify any discussions or analyses related to these topics that could impact the performance and reliability of multimodal machine learning models. From our review, we first found that 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 [83].

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 [85]. 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.

7. Discussion

Based on a thorough literature review, we have concluded that BERT, LSTM, and their variations are the most popular language models among researchers in multimodal machine learning. Architectures such as ResNet and VGG, which are variations in CNN, are commonly used for image-processing tasks. Through our investigation into fusion techniques commonly employed in MMML, we have looked into various methods designed to merge data from multiple modalities. Our exploration covers a spectrum from weight-based methods like Early Fusion, concatenation, and attention mechanisms to cutting-edge multimodal deep learning architectures. These techniques aim to generate insightful conclusions and representations from the complex interplay between text and visual data. By showcasing numerous methods that address the complex requirements of multimodal data analysis, this review emphasizes the dynamic character of MMML. The choice of fusion technique is dictated by the specific needs of the task, with each method offering distinct benefits and applications.

Investigating the limitations and challenges within MMML architectures offers valuable insights into the complexities of employing multiple data modalities. It becomes evident that addressing these critical issues is paramount for overcoming the obstacles inherent in MMML designs. Enhancing data annotation resources, adapting models to specific tasks, devising strategies for noise reduction, and improving data preprocessing techniques are crucial steps for the further development of MMML.

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 primary phase, with little research on these complex problems. The potential for adversarial robustness 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 attack of MMML systems needs to be more critical, as seen by the scarcity of work in this area. This gap offers a chance to conduct new research to create novel protection mechanisms while looking further into the subtleties of adversarial 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.

8. Conclusions

Our scoping literature review has systematically identified prevalent methods for leveraging data from image and text modalities. From our investigation into RQ1, we found that BERT and LSTM stand out as the leading pre-trained architectures for text embedding. In contrast, various VGG and ResNet architectures are predominantly utilized for image embedding. Our study further reveals that MMML practitioners frequently employ benchmark datasets such as Twitter, Flickr, and COCO to train and evaluate their models. These datasets offer rich, diverse, and multimodal data sources, enhancing and expanding MMML models' capabilities.

As we delve into fusion methods, it is evident that the MMML community employs a broad spectrum of techniques, ranging from concatenation to attention mechanisms and advanced neural networks. Each method brings distinct advantages, reflecting the dynamic nature of multimodal fusion. However, our exploration of MMML's limitations and challenges uncovered several critical issues, including computational complexity, data constraints, real-time processing challenges, noise resilience, and larger datasets. Awareness of these limitations is crucial for researchers and practitioners engaged in MMML.

This literature review sheds light on the architectural preferences, dataset selections, and flexible fusion strategies embraced by the MMML community. By addressing the inherent limitations and challenges of MMML, this study serves as a valuable guide, steering

scholars and practitioners toward informed decisions and innovative solutions as MMML continues to evolve and expand its reach into various domains. As the exploration into multimodal data deepens, there is a profound opportunity to enhance our understanding of the world through integrated data modalities. This endeavor holds the potential to revolutionize industries, improve decision-making processes, and enrich our perspective on the world. In our future work, we aim to investigate the behavior of MMML models under adversarial conditions. Analyzing how these models respond to adversarial attacks will offer vital insights into their security and robustness, uncovering strategies to shield them from malicious interference.

Author Contributions: Conceptualization, M.B.R. and P.R.; methodology, M.B.R and P.R.; software, none; validation, none; formal analysis, M.B.R. and P.R.; investigation, M.B.R. and M.S.R.; resources, M.B.R. and P.R.; data curation, M.B.R. and M.S.R.; writing—original draft preparation, M.B.R.; writing—review and editing, M.B.R. and P.R.; visualization, none; supervision, P.R.; project administration, P.R.; funding acquisition, P.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was executed while P.R. and M.B.R. were funded by the National Science Foundation under grants NSF CISE—CNS Award 2136961 and 2210091.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762. https://doi.org/10.48550/arxiv.1706.03762.
2. Baltrušaitis, T.; Ahuja, C.; Morency, L. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
3. Talukder, S.; Barnum, G.; Yue, Y. On the benefits of early fusion in multimodal representation learning. *arXiv* **2020**, arXiv:2011.07191. https://doi.org/10.48550/arxiv.2011.07191.
4. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Comput.* **2020**, *32*, 829–864. [CrossRef] [PubMed]
5. Siriwardhana, S.; Kaluarachchi, T.; Billinghurst, M.; Nanayakkara, S. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access* **2020**, *8*, 176274–176285. [CrossRef]
6. Chai, W.; Wang, G. Deep vision multimodal learning: Methodology, benchmark, and trend. *Appl. Sci.* **2022**, *12*, 6588. [CrossRef]
7. Choi, J.; Lee, J. Embracenet: A robust deep learning architecture for multimodal classification. *Inf. Fusion* **2019**, *51*, 259–270. [CrossRef]
8. Kline, A.; Wang, H.; Li, Y.; Dennis, S.; Hutch, M.; Xu, Z.; Wang, F.; Cheng, F.; Luo, Y. Multimodal machine learning in precision health: A scoping review. *npj Digit. Med.* **2022**, *5*, 171. [CrossRef]
9. Bayoudh, K.; Knani, R.; Hamdaoui, F.; Mtibaa, A. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.* **2021**, *38*, 2939–2970. [CrossRef]
10. Aggarwal, A.; Srivastava, A.; Agarwal, A.; Chahal, N.; Singh, D.; Alnuaim, A.; Alhadlaq, A.; Lee, H. Two-way feature extraction for speech emotion recognition using deep learning. *Sensors* **2022**, *22*, 2378. [CrossRef]
11. Barua, P.; Chan, W.; Dogan, S.; Baygin, M.; Tuncer, T.; Ciaccio, E.; Islam, M.; Cheong, K.; Shahid, Z.; Acharya, U. Multilevel deep feature generation framework for automated detection of retinal abnormalities using oct images. *Entropy* **2021**, *23*, 1651. [CrossRef] [PubMed]
12. Lv, D.; Wang, H.; Che, C. Fault diagnosis of rolling bearing based on multimodal data fusion and deep belief network. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2021**, *235*, 6577–6585. [CrossRef]
13. Kumaresan, S.; Aultrin, K.; Kumar, S.; Anand, M. Transfer learning with cnn for classification of weld defect. *IEEE Access* **2021**, *9*, 95097–95108. [CrossRef]
14. Zhang, C.; Yang, Z.; He, X.; Deng, L. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 478–493. [CrossRef]
15. Li, J.; Yao, X.; Wang, X.; Yu, Q.; Zhang, Y. Multiscale local features learning based on bp neural network for rolling bearing intelligent fault diagnosis. *Measurement* **2020**, *153*, 107419. [CrossRef]
16. Zhu, Q.; Xu, X.; Yuan, N.; Zhang, Z.; Guan, D.; Huang, S.; Zhang, D. Latent correlation embedded discriminative multi-modal data fusion. *Signal Process.* **2020**, *171*, 107466. [CrossRef]
17. Singh, J.; Azamfar, M.; Li, F.; Lee, J. A systematic review of machine learning algorithms for prognostics and health management of rolling element bearings: Fundamentals, concepts and applications. *Meas. Sci. Technol.* **2020**, *32*, 012001. [CrossRef]

18. Cai, H.; Qu, Z.; Li, Z.; Zhang, Y.; Hu, X.; Hu, B. Feature-level fusion approaches based on multimodal eeg data for depression recognition. *Inf. Fusion* **2020**, *59*, 127–138. [CrossRef]
19. Schillaci, G.; Villalpando, A.; Hafner, V.; Hanappe, P.; Colliaux, D.; Wintz, T. Intrinsic motivation and episodic memories for robot exploration of high-dimensional sensory spaces. *Adapt. Behav.* **2020**, *29*, 549–566. [CrossRef]
20. Guo, R.; Wei, J.; Sun, L.; Yu, B.; Chang, G.; Liu, D.; Zhang, S.; Yao, Z.; Xu, M.; Bu, L. A Survey on Image-text Multimodal Models. *arXiv* **2023**, arXiv:2309.15857.
21. Palani, B.; Elango, S.; Viswanathan K, V. CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT. *Multimed. Tools Appl.* **2022**, *81*, 5587–5620. [CrossRef] [PubMed]
22. Hangloo, S.; Arora, B. Combating multimodal fake news on social media: Methods, datasets, and future perspective. *Multimed. Syst.* **2022**, *28*, 2391–2422. [CrossRef] [PubMed]
23. Gao, L.; Gao, Y.; Yuan, J.; Li, X. Rumor detection model based on multimodal machine learning. In Proceedings of the Second International Conference on Algorithms, Microchips, and Network Applications (AMNA 2023), Zhengzhou, China, 13–15 January 2023; SPIE: Bellingham, WA, USA, 2023; Volume 12635, pp. 359–366.
24. Yue, T.; Mao, R.; Wang, H.; Hu, Z.; Cambria, E. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Inf. Fusion* **2023**, *100*, 101921. [CrossRef]
25. Lucas, L.; Tomás, D.; Garcia-Rodriguez, J. Detecting and locating trending places using multimodal social network data. *Multimed. Tools Appl.* **2023**, *82*, 38097–38116. [CrossRef]
26. Chandra, M.; Pailla, D.; Bhatia, H.; Sanchawala, A.; Gupta, M.; Shrivastava, M.; Kumaraguru, P. "Subverting the Jewtocracy": Online antisemitism detection using multimodal deep learning. In Proceedings of the 13th ACM Web Science Conference 2021, Virtual Event, 21–25 June 2021; pp. 148–157.
27. Xiao, S.; Chen, G.; Zhang, C.; Li, X. Complementary or substitutive? A novel deep learning method to leverage text-image interactions for multimodal review helpfulness prediction. *Expert Syst. Appl.* **2022**, *208*, 118138. [CrossRef]
28. Li, M. Research on extraction of useful tourism online reviews based on multimodal feature fusion. *Trans. Asian Low-Resour. Lang. Inf. Process.* **2021**, *20*, 1–16. [CrossRef]
29. Bhat, A.; Chauhan, A. A Deep Learning based approach for MultiModal Sarcasm Detection. In Proceedings of the 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 16–17 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 2523–2528.
30. Yadav, A.; Vishwakarma, D.K. A deep multi-level attentive network for multimodal sentiment analysis. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–19. [CrossRef]
31. Chen, X.; Lao, S.; Duan, T. Multimodal fusion of visual dialog: A survey. In Proceedings of the 2020 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence, Shanghai, China, 17–19 October 2020; pp. 302–308.
32. Alsan, H.F.; Yıldız, E.; Safdil, E.B.; Arslan, F.; Arsan, T. Multimodal retrieval with contrastive pretraining. In Proceedings of the 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Kocaeli, Turkey, 25–27 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
33. Ange, T.; Roger, N.; Aude, D.; Claude, F. Semi-supervised multimodal deep learning model for polarity detection in arguments. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–8.
34. Asgari-Chenaghlu, M.; Feizi-Derakhshi, M.R.; Farzinvash, L.; Balafar, M.; Motamed, C. CWI: A multimodal deep learning approach for named entity recognition from social media using character, word and image features. *Neural Comput. Appl.* **2022**, *34*, 1905–1922. [CrossRef]
35. Zhang, Z.; Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Li, Z.; Zhao, H. Universal Multimodal Representation for Language Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9169–9185. [CrossRef]
36. Guo, Q.; Yao, K.; Chu, W. Switch-BERT: Learning to Model Multimodal Interactions by Switching Attention and Input. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 330–346.
37. Hu, P.; Zhang, Z.; Zhang, J.; Du, J.; Wu, J. Multimodal Tree Decoder for Table of Contents Extraction in Document Images. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1756–1762.
38. Ahmed, M.R.; Bhadani, N.; Chakraborty, I. Hateful Meme Prediction Model Using Multimodal Deep Learning. In Proceedings of the 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 23–25 September 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
39. Agarwal, S. A Multimodal Machine Learning Approach to Diagnosis, Prognosis, and Treatment Prediction for Neurodegenerative Diseases and Cancer. In Proceedings of the 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 26–29 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 0475–0479.
40. Huang, P.C.; Shakya, E.; Song, M.; Subramaniam, M. BioMDSE: A Multimodal Deep Learning-Based Search Engine Framework for Biofilm Documents Classifications. In Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 6–8 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 3608–3612.

41. Ban, M.; Zong, L.; Zhou, J.; Xiao, Z. Multimodal Aspect-Level Sentiment Analysis based on Deep Neural Networks. In Proceedings of the 2022 8th International Symposium on System Security, Safety, and Reliability (ISSSR), Chongqing, China, 27–28 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 184–188.
42. Liang, T.; Lin, G.; Wan, M.; Li, T.; Ma, G.; Lv, F. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 15492–15501.
43. Sahoo, C.C.; Tomar, D.S.; Bharti, J. Transformer based multimodal similarity search method for E-Commerce platforms. In Proceedings of the 2023 IEEE Guwahati Subsection Conference (GCON), Guwahati, India, 23–25 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
44. Yu, Z.; Lu, M.; Li, R. Multimodal Co-Attention Mechanism for One-stage Visual Grounding. In Proceedings of the 2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS), Chengdu, China, 26–28 November 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 288–292.
45. Xu, J.; Zhao, H.; Liu, W.; Ding, X. Research on False Information Detection Based on Multimodal Event Memory Network. In Proceedings of the 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 6–8 January 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 566–570.
46. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; 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, New Orleans, LA, USA, 18–24 June 2022; pp. 18166–18176.
47. Jácome-Galarza, L.R. Multimodal Deep Learning for Crop Yield Prediction. In Proceedings of the Doctoral Symposium on Information and Communication Technologies, Manta, Ecuador, 12–14 October 2022; Springer: Cham, Switzerland, 2022; pp. 106–117.
48. Kraidia, I.; Ghenai, A.; Zeghib, N. HST-Detector: A Multimodal Deep Learning System for Twitter Spam Detection. In Proceedings of the International Conference on Computing, Intelligence and Data Analytics, Kocaeli, Turkey, 16–17 September 2022; Springer: Cham, Switzerland, 2022; pp. 91–103.
49. Kaliyar, R.K.; Mohnot, A.; Raghhul, R.; Prathyushaa, V.; Goswami, A.; Singh, N.; Dash, P. MultiDeepFake: Improving Fake News Detection with a Deep Convolutional Neural Network Using a Multimodal Dataset. In Proceedings of the Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, 5–6 December 2020; Springer: Singapore, 2021; pp. 267–279.
50. Malhotra, A.; Jindal, R. Multimodal deep learning architecture for identifying victims of online death games. In *Data Analytics and Management, Proceedings of ICDAM, Jaipur, India, 26 June 2021*; Springer: Singapore, 2021; pp. 827–841.
51. Peña, A.; Serna, I.; Morales, A.; Fierrez, J.; Ortega, A.; Herrarte, A.; Alcantara, M.; Ortega-Garcia, J. Human-centric multimodal machine learning: Recent advances and testbed on AI-based recruitment. *SN Comput. Sci.* **2023**, *4*, 434. [CrossRef]
52. Ghosal, T.; Raj, A.; Ekbal, A.; Saha, S.; Bhattacharyya, P. A deep multimodal investigation to determine the appropriateness of scholarly submissions. In Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2–6 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 227–236.
53. Miao, H.; Zhang, Y.; Wang, D.; Feng, S. Multimodal Emotion Recognition with Factorized Bilinear Pooling and Adversarial Learning. In Proceedings of the 5th International Conference on Computer Science and Application Engineering, Sanya, China, 19–21 October 2021; pp. 1–6.
54. Hossain, E.; Hoque, M.M.; Hoque, E.; Islam, M.S. A Deep Attentive Multimodal Learning Approach for Disaster Identification From Social Media Posts. *IEEE Access* **2022**, *10*, 46538–46551. [CrossRef]
55. Paul, S.; Saha, S.; Hasanuzzaman, M. Identification of cyberbullying: A deep learning based multimodal approach. *Multimed. Tools Appl.* **2022**, *81*, 26989–27008. [CrossRef]
56. Ha, Y.; Park, K.; Kim, S.J.; Joo, J.; Cha, M. Automatically detecting image–text mismatch on Instagram with deep learning. *J. Advert.* **2020**, *50*, 52–62. [CrossRef]
57. Rivas, R.; Paul, S.; Hristidis, V.; Papalexakis, E.E.; Roy-Chowdhury, A.K. Task-agnostic representation learning of multimodal twitter data for downstream applications. *J. Big Data* **2022**, *9*, 18. [CrossRef]
58. Babu, G.T.V.M.; Kavila, S.D.; Bandaru, R. Multimodal Framework Using CNN Architectures and GRU for Generating Image Description. In Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 28–29 April 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 2116–2121.
59. Paraskevopoulos, G.; Pistofidis, P.; Banoutsos, G.; Georgiou, E.; Katsouros, V. Multimodal Classification of Safety-Report Observations. *Appl. Sci.* **2022**, *12*, 5781. [CrossRef]
60. Wang, Y.; Ma, F.; Wang, H.; Jha, K.; Gao, J. Multimodal emergent fake news detection via meta neural process networks. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021; pp. 3708–3716.
61. Xu, N.; Mao, W. A residual merged neutral network for multimodal sentiment analysis. In Proceedings of the 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 10–12 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 6–10.
62. Papadopoulos, S.I.; Koutlis, C.; Papadopoulos, S.; Petrantonakis, P. Synthetic Misinformers: Generating and Combating Multimodal Misinformation. In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation, Thessaloniki, Greece, 12–15 June 2023; pp. 36–44.

63. Karimvand, A.N.; Chegeni, R.S.; Basiri, M.E.; Nemati, S. Sentiment analysis of persian instagram post: A multimodal deep learning approach. In Proceedings of the 2021 7th International Conference on Web Research (ICWR), Tehran, Iran, 19–20 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 137–141.
64. Shirzad, A.; Zare, H.; Teimouri, M. Deep Learning approach for text, image, and GIF multimodal sentiment analysis. In Proceedings of the 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 29–30 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 419–424.
65. Yu, Y.; Tang, S.; Aizawa, K.; Aizawa, A. Category-based deep CCA for fine-grained venue discovery from multimodal data. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 1250–1258. [CrossRef]
66. Barveen, A.; Geetha, S.; Faizal, M.M. Meme Expressive Classification in Multimodal State with Feature Extraction in Deep Learning. In Proceedings of the 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 5–7 April 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–10.
67. Chen, D.; Zhang, R. Building Multimodal Knowledge Bases with Multimodal Computational Sequences and Generative Adversarial Networks. *IEEE Trans. Multimed.* **2023**, *26*, 2027–2040. [CrossRef]
68. Kim, E.; Onweller, C.; McCoy, K.F. Information graphic summarization using a collection of multimodal deep neural networks. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 10188–10195.
69. Thuseethan, S.; Janarthan, S.; Rajasegarar, S.; Kumari, P.; Yearwood, J. Multimodal deep learning framework for sentiment analysis from text-image web Data. In Proceedings of the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Melbourne, Australia, 14–17 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 267–274.
70. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [CrossRef]
71. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
72. Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; Fu, J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv* **2020**, arXiv:2004.00849.
73. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736.
74. Fatichah, C.; Wiyadi, P.D.S.; Navastara, D.A.; Suciati, N.; Munif, A. Incident detection based on multimodal data from social media using deep learning methods. In Proceedings of the 2020 International conference on ICT for smart society (ICISS), Bandung, Indonesia, 19–20 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
75. Guo, N.; Fu, Z.; Zhao, Q. Multimodal News Recommendation Based on Deep Reinforcement Learning. In Proceedings of the 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 15–17 April 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 279–284.
76. Guo, L. Art teaching interaction based on multimodal information fusion under the background of deep learning. *Soft Comput.* **2023**, 1–9. [CrossRef]
77. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5579–5588.
78. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 4904–4916.
79. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
80. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
81. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
82. Lu, J.; Goswami, V.; Rohrbach, M.; Parikh, D.; Lee, S. 12-in-1: Multi-task vision and language representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10437–10446.
83. Rahate, A.; Walambe, R.; Ramanna, S.; Kotecha, K. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Inf. Fusion* **2022**, *81*, 203–239. [CrossRef]
84. Liu, J. Multimodal Machine Translation. *IEEE Access* 2021, *early access*. [CrossRef]
85. Li, L.; Gan, Z.; Liu, J. A closer look at the robustness of vision-and-language pre-trained models. *arXiv* **2020**, arXiv:2012.08673.