Multimodality for NLP-Centered Applications: Resources, Advances and Frontiers

Muskan Garg*,**, Seema Wazarkar**, Muskaan Singh§#, Ondřej Bojar§
* University of Florida, USA,
** Speech and Audio Processing Group, IDIAP Research Institute, Switzerland
** Thapar Institute of Engineering & Technology, India,
§ Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Czechia
muskangarg@ufl.edu, msingh@idiap.ch, seema.wazarkar@thapar.edu, bojar@ufal.cuni.cz

Abstract

With advancements in the methods of Natural Language Processing (NLP) by explicitly considering other modalities than just text, the resurgence of multimodal datasets has attracted significant attention. However, there remains lack of a comprehensive survey on available datasets. To this end, we take the first step and present a thorough review of publicly available datasets with different modalities for NLP tasks which they may cater. Our survey shall enable the research community to re-use, re-furnish and re-annotate the existing datasets with new modalities for multiple NLP tasks. Furthermore, we discuss new frontiers and challenges, and hope this survey will provide the community with a general picture of available multimodal datasets for various NLP applications, facilitate quick access to them and motivate future research. In this context, we release the collection of links to all multimodal datasets we discover as an easily accessible and updatable repository: [https://github.com/dr muskangarg/Multimodal-datasets](https://github.com/drmuskangarg/Multimodal-datasets).

1. Introduction

Multimodality refers to the capability of a system or method to process input of different types (or "modalities"), primarily text, image, sound or video. Embraced with multiple streams of participants’ physical responses (eyetracking, EEG, etc.) or environmental conditions (temperature, pressure etc.), multimodality plays pivotal role in enhancing intelligence of a system. These multiple modalities (Parcalabescu et al., 2021) develop as a strong research enhancement in recent years to support downstream NLP tasks. We focus on the most common modalities in current NLP tasks and speak of 10 different permutations of four modalities as summarized in Figure 1. It is interesting that v (video) modality automatically leads to multiple combination of all other modalities for analysis.

Research in this novel direction primarily aims to process textual content using visual information (e.g., images and possibly video) to support various tasks (e.g., machine translation). Its motivation derives mainly from two linguistic challenges: lexical ambiguity and out of vocabulary words which may be resolved by using multiple modalities or stand for the missing information in a way. In practice, the non-textual context provided implicitly by the additional modalities is extremely influential (“an image is worth a thousand words”, and a “sound illustration” can easily explain why, e.g., a person is having difficulties in expressing themselves). Recent studies show that visual information helps in reaching modest but encouraging improvements in quality (Elliott et al., 2016) [Caglayan et al., 2018, Libovický and Helcl, 2017]. Very recent work documents the use of visual information for interpreting implicit language (Collell et al., 2018). Our work summarizes the available multimodal datasets for seven big NLP tasks (sentiment analysis, machine translation, information retrieval, question answering, automatic summarization, human-computer interaction and semantic analysis) and other miscellaneous tasks. We hope this work will help to promote the use of available multimodal datasets and augment new annotated modalities in existing ones to push the research towards developing further interesting applications.

2. Background

The limitation of existing literature is two-fold: (i) 100+ multimodal language resources are available for many under-explored NLP tasks; (ii) developing a multimodal dataset with ground truth information is always a big investment that limits the possibilities of research. In this context, we carry out a comprehensive survey on multimodal datasets to handle these limitations. This survey will enable researchers to save efforts by reusing and re-furbishing existing multimodal datasets.

Figure 1: Different modalities and their combinations. Each of the individual modalities: image (I), audio (A) and video (V) are combined with text to create IT, AT and IV, respectively. We further group their combination pairs as (image, audio) (IA), (audio, video) (AV), and (image, video) (IV). We further group all the modalities as (image, audio, video). The text (T) track can be added to the combined modalities, too.
for diverse set of NLP applications. To the best of our knowledge, this is the first survey of its kind, and we further describe recent advancements and discuss new frontiers.

We focus on (1) finding multilingual and low-resourced datasets, and (2) reducing redundancy by grouping together datasets that evolves from one source. We describe the availability of multimodal datasets for different NLP applications and focus specifically on their modalities, language(s) and the source of collection. The major contributions of this survey are: (1) a comprehensive survey of existing multimodal datasets for different NLP applications, (2) the summary of recently developed publicly available benchmark datasets for the tuple < a, l, s > (application, language, source), (3) new frontiers and open research directions in the area of multimodal analysis.

3. Multi-Modality in NLP

We examine the evolution of multimodal datasets for different applications. In sentiment analysis or opinion mining, the research community use data to find mental state of a user such as positive, negative or neutral. In machine translation the machine translates contents from one language to the another (understandable) language to interpret the information well. One of the major challenge of natural language understanding is to perform search operations in natural language document. It is important to retrieve appropriate information from data shared in different modalities to accomplish various real world tasks. Question answering task is development of a machine which automatically provides answers for questions asked by a user, recently, from multimodal datasets. The task of text summarization projects significant information in abstract way and is currently buffered with new modalities. Semantic analysis examines the sense of dataset to enable machine understandable activities which contributes towards better decision making. In this section we discuss major domains like sentiment analysis (3.1), machine translation (3.2), information retrieval (3.3), question answering (3.4), summarization (3.5), human-computer interaction (3.6), semantic analysis (3.7) and other miscellaneous (3.8) applications.

3.1. Sentiment Analysis

Sentiment analysis is one of the most widely studied applications of text classification. We investigate publicly available datasets and datasets available On Request (OR) to pack up available datasets in Table 1. We classify more than 25 potential sentiment analysis/opinion mining datasets based on language, modalities, and sources. We organize the datasets according to two criteria: the language they cover [3.1.1] and source where they come from [3.1.2].

3.1.1. Language-Specific Sentiment Analysis

The benchmark multimodal dataset for well-formed English language and non-English languages are (Grimm et al., 2008) and (Burkhardt et al., 2005), respectively. The language-specific multimodal datasets are available for English (EN), Indo-Asian, and European languages. Starting with the German dataset EmoDB (Burkhardt et al., 2005), the European-language multimodal languages now cover German (DE) (Cevher et al., 2019; Alaçam et al., 2020), French (FR) (Ringeval et al., 2013), Spanish (ES) (García-Vega et al., 2020), and Portuguese (PT) (Zadeh et al., 2020). Datasets use IAV: image-audio-visual, AV: audio-visual, I: image, and AT: audio-text modalities. The European datasets use recorded files or YouTube videos, and the Indo-Asian datasets like CH-SIMS (Yu et al., 2020a) use Movies, TV series or shows as the potential source of information. A recently developed European language dataset, CMU-MOSEAS dataset (Zadeh et al., 2020) (AV), has set a benchmark with 40, 000 samples of 1645 speakers with more than 68 hours of duration and is available OR.

3.1.2. Sources for Sentiment Analysis

As observed from existing literature, one of the most important sources of multimodal sentiment analysis is YouTube videos (Zadeh et al., 2018) [Morency et al., 2011] [Perez-Rosas et al., 2013]. These videos have voice (A), frames (I) and title (T) suitable for using all kinds of modalities. The IAV is the most widely adopted modality for multimodal sentiment analysis. The research community uses popular TV talk shows (Viegas and Alikhani, 2021) [Douglas-Cowie et al., 2011] [Grimm et al., 2008] and TV series (Yu et al., 2020a) [Firdaus et al., 2020b] [Poria et al., 2019] as potential sources of data. The social media data has shown effective results for human behavior analysis such as sentiment analysis (Suryawanshi et al., 2020b) [Nakamura et al., 2020] and offensive content classification (Singh et al., 2021). In addition to this, authors use recorded videos (Kossaifi et al., 2019) [McKeown et al., 2011] [Douglas-Cowie et al., 2011] and movies (Maas et al., 2011) [Park et al., 2016] and datasets from social media and IMDB use I modalities for sentiment analysis.

3.2. Machine Translation

Multimodal Machine Translation (MMT) (Yu et al., 2020b) converts text from one language to another language using multiple modalities. Very few datasets are available for MMT, we use most of these datasets as the benchmark datasets for their respective languages. We further categorize MMT datasets into two-fold translations: (i) using IT and (ii) using IV. The Multi30k (Elliott et al., 2016) is a benchmark dataset for recent developments in image-based machine translation and the recently introduced HowTo100M (Huang et al., 2021a) has paved a concrete path for open research in video-based machine translation with nine languages. Most of the image-based datasets use Flickr images, and video-based datasets use YouTube videos. Although there is much development in English to Eu-
Table 1: Sentiment Analysis

| Dataset                      | Language | Modality | Samples | Avail | Source                  | #Ut |
|------------------------------|----------|----------|---------|-------|-------------------------|-----|
| flickr200k-EN (Chowdhary et al., 2018) | EN, HI-IN | V, I     | 155,070 | OR    | Flickr200k              | 12  |
| Hindi Visual Genome (Panda et al., 2019) | EN, HI-IN | V, I     | 31,523  | Yes   | Visual Genome           | 17  |
| HowTo100M (Huang et al., 2021a) | EN, HI-IN | V, I, V  | 79,114  | Yes   | YouTube                 | 119 |
| Multi30K (Parida et al., 2019b) | EN, HI-IN | V, I, V  | 138 mn clips | Yes | YouTube                  | 9   |
| MEI (Huang et al., 2018)       | EN, HI-IN | V, I, V  | 99,674  | Yes   | Multi30K                | 20  |
| Movi50k (Elisetti et al., 2016) | EN, DE    | V, I     | 155,070 | Yes   | Flickr                  | 323 |
| VATEX (Wang et al., 2019b)     | EN-ZH     | V, I     | 206,000 | Yes   | YouTube                 | 117 |

Table 2: Machine Translation

| Dataset                      | Language | Modality | Samples | Avail | Source                  | #Ut |
|------------------------------|----------|----------|---------|-------|-------------------------|-----|
| AFEW (Dhill et al., 2012)    | EN       | A, V     | 1045    | OR    | Movies                  | 43/ |
| AMMER (Levies et al., 2019)  | DE       | T, A, V  | 288     | OR    | Drivers                 | 18  |
| CH SIMS (Yue et al., 2020a)  | ZH       | T, I, V  | 2281    | Yes   | Movie, TV series/shows | 18  |
| CMI-MOSEA (Zadeh et al., 2020) | FR, ES, PT, DE | T, A, V | 40,000 | OR | Youtube (YT) | 5  |
| CMI-MOSEA (Zadeh et al., 2020) | EN       | T, A, V  | 23,453  | Yes   | Youtube                 | 242 |
| Creep-Image (Menon et al., 2020) | EN       | T, I     | 1,7912  | Yes   | CREEENER tool           | 3   |
| CREED (Zadeh et al., 2019)   | EN       | T, I     | 95      | Yes   | YouTube                 | 134 |
| EmoDB (Bunkhvard et al., 2005) | EN       | T, A     | 800     | Yes   | Recordings              | 2134|
| Enthook (Vegas and Atkinson, 2021) | EN       | T, A     | 2351    | Yes   | TED talks               | 1   |
| Fakedit (Nakamura et al., 2020) | EN       | T, I     | 1 mn    | Yes   | Reddit                  | 43  |
| HUMAINE (Huang et al., 2021a) | EN       | A, V     | 50      | Yes   | TV Recording            | 39  |
| R-1-MMMAO (Wolff et al., 2019) | EN       | T, A, V  | 3,170   | Yes   | YT & Exploit V          | 286 |
| XMCAP (Bustico et al., 2008) | EN       | T, A, V  | 10,000  | Yes   | At University           | 1624|
| Large Movie (Tan et al., 2014) | EN       | T, I     | 25,000  | Yes   | IMDb                    | 174 |
| MEINDA (Sriram et al., 2019) | EN       | T, A, V  | 407     | Yes   | TV Series Friends       | 5   |
| MELD (Poria et al., 2019)    | EN       | T, A     | 13,000  | Yes   | TV Series Friends       | 227 |
| Memagy (Sun et al., 2015)    | EN       | A, V     | 54      | Yes   | Recorded                | 52  |
| MOODY (Perez-Rosas et al., 2021) | EN       | T, I     | 400     | Yes   | YouTube                 | 172 |
| Multi30K (Parida et al., 2019b) | EN       | T, I     | 743     | Yes   | Social media            | 30  |
| POM (Park et al., 2016)      | EN       | T, V     | 903     | Yes   | Movies                  | 6   |
| RECOLA (Rengel et al., 2013) | FR, A     | 46       | OR      | Recorded                | 524 |
| SEMAINE (McKee et al., 2011) | EN       | T, V     | 80      | OR    | Recorded                | 82  |
| XWCAPE (Pietri et al., 2019) | EN, A, V | 538      | Yes    | Existing HB            | 82  |
| SST1 (Mou et al., 2019)      | EN       | T, I     | 11,582  | Yes   | rosettentomates.com     | 255  |
| TASS (Sacrista-Vegas et al., 2020) | EN, A     | T        | 3,413   | OR    | Twitter                 | 9   |
| VAM (Ghandeharizadeh et al., 2020) | EN, V     | A, V     | 499     | Yes   | TV Talk Show             | 444 |
| Youtube D (Morency et al., 2013) | EN       | T, A, I  | 47      | Yes   | Youtube                 | 350 |

3.3. Information Retrieval

Information retrieval is a task of identifying essential documents from dataset and ranking them in the form of a query. The task of analyzing data has recently introduced a multilingual dataset (Srinivasan et al., 2021) for IT modality using Wikipedia source. Other datasets are for the English language except Hindi (Meeitei et al., 2019) and Slovenian (Pekš et al., 2017) language. A recent music dataset of 200k samples is given as AT dataset. Author extends existing multimodal dataset (Visual Genome) for information retrieval task in Hindi language (Meeitei et al., 2019) to create Hindi Visual Genome. We use cross-domain development for other problem domains. Most of the datasets are available except that of MQA (Deng et al., 2021). Music analysis is widely explored in recent years (Zalkow et al., 2020).

3.4. Question Answering

Question Answering is a unique task of automation of help-desk by automatically answering a query. Authors choose to re-annotate the existing datasets (Agrawal et al., 2018), Singh et al., 2021, Ye et al., 2017, Zhu et al., 2017, Kaife and Kanan, 2017, Hudson and Manning, 2019) for the problem of multimodal question answering. The availability of datasets for this research area is limited to English language and there are no publicly available non-English datasets. We further investigate different modalities for this task and categorize the datasets into two different modalities: image-based question answering and video-based question answering. The image-based dataset are: VQA (Goyal et al., 2017) and TDUC (Kaffe and Kanan, 2017) and the most widely used video-based datasets are MovieFIB (Maharaj et al., 2017) and YouTube2Text (Xu et al., 2017). Domain-specific datasets for social media are GQA (Hudson and Manning, 2019), MemexQA (Jiang et al., 2017), TGIF-QA (Jang et al., 2017), SocialIQ (Zadeh et al., 2019), YouTube2Text (Xu et al., 2017), MSVD QA and MSRVT QA (Ye et al., 2017); and for TV shows, movies and gameplayes are MarioQA (Mun et al., 2017), TVQA (Lei et al., 2018).

3.5. Automatic Summarization

Automatic summarization generates a gist of the information retrieved from unstructured data of multiple modalities. In recent years, a gradual shift from text to
multimodal summarization justifies that on combining multiple modalities, they give more details about the context of data. Image-based multimodal datasets are not available in the public domain (Li et al., 2018; Zhu et al., 2018; Wang et al., 2021). A new image-based automatic summarization dataset, Screen2Words (Wang et al., 2021) is recently introduced by re-annotating the existing open-source dataset Rico-SCA and is publicly available. The video-based automatic summarization datasets are being introduced since 2014 (Gygli et al., 2014; Song et al., 2015; Sharghi et al., 2017) with few samples, but a new benchmark datasets in this domain is recently introduced with large samples for CNN and daily mail (Fu et al., 2021). The most widely used image-based dataset are SumMe (Gygli et al., 2014) and TVSum (Song et al., 2015), and the most widely used video-based datasets are MMSS (Li et al., 2018) and MSMO (Zhu et al., 2018). The source of data varies with News (Fu et al., 2021), social media (Saini et al., 2021) and academic conferences (Atri et al., 2021).

### Table 3: Information Retrieval

| Dataset          | Language | Modality | Samples  | Avail | Source       | #Cit. |
|------------------|----------|----------|----------|-------|--------------|-------|
| GQA              | EN       | T, I     | 1.13,018 | Yes   | COCO, Flickr | 344   |
| MarloQQA         | EN       | V, T     | 1.97,315 | Yes   | Gameplays: Super Mario Bros | 99   |
| MemloQQA         | EN       | V, T     | 13,391   | Yes   | Flickr       | 24    |
| MIMOQQA          | EN       | V, T     | 200      | No    | Existing, Unmooda | 2    |
| MovieFHR         | EN       | V, T     | 3.48,998 | Yes   | Youtube      | 56    |
| MovieQA          | EN       | V, T     | 14,944   | Yes   | Diverse sources: Wikipedia, imdb | 485  |
| MQA              | EN       | V, T     | 206      | No    | Wiki & Online | 4    |
| MSVD QA, MSVTV1 QA | EN      | V, T     | 198      | Yes   | Youtube      | 70    |
| PororoQA         | EN       | V, I     | 16,066   | Yes   | Cartoon videos series 'Pororo' | 116  |
| ReceiptQA        | EN       | V, T     | 36,000   | Yes   | Instructable | 83    |
| Social IQ [Zafes et al., 2019] | EN | V, I | 1,250 | Yes | Youtube | 36 |
| TDIUS (Kalle and Kanau, 2017) | EN | V, T | 1,64,167 | Yes | VQA | 156 |
| TGF-QA (Hane et al., 2017) | EN | V, T | 1,85,165 | Yes | Social Media | 242 |
| TVQA Test (Fu et al., 2018) | EN | V, T | 21,993 | Yes | 6 pippaper TV shows | 28 |
| Video Context QA (Zhu et al., 2017) | EN | V, T | 1,09,895 | Yes | TVCol, MII-MD, MGD Test 14 datasets | 182 |
| YouTube2Text (Xu et al., 2017) | EN | V, T | 243k | Yes | Youtube | 125 |
| VQA (Goyal et al., 2017) | EN | V, T | 285,016 | Yes | Amazon Mechanical Turk (AMT) | 1,160 |

### Table 4: Question Answering

3.6. Human Computer Interaction

Human-Computer Interaction (HCI) is the process of multimodal analysis for NLP tasks. It deals with the problems like topic detection and tracking (Joo et al., 2017), classifying personality traits (Celiktutan et al., 2017), affective computing (Hazer-Rau et al., 2020), speech recognition (Patterson et al., 2002) and action recognition (Ofli et al., 2019). The expression based information retrieval from 8 workers for a total of 2, 400 human intelligence tasks using VoxSim (Krishnaswamy and Pustejovsky, 2019). The largest dataset of HCI problems is a multilingual dataset (the Red Hen Lab (Joo et al., 2017)) of 350k hours which is extracted from global TV news using automated tagging tools. The most frequently used visual dataset are CAUVE (Patterson et al., 2002) and MHAD (Ofli et al., 2013). Recently introduced data collection for affective computing, uulmMAC (Hazer-Rau et al., 2020), is initialized with two homogeneous samples of 60 participants and 100 recordings. The English language HCI datasets are publicly available and can be used with all kinds of modalities from IAV to VT.

3.7. Semantic Analysis

Semantic analysis deals with a user’s intention and meaningful document representation. Such NLP task may help in solving the concept-specific problems of text mining. Two sets of languages covered for multimodal semantic analysis are English and European languages. Image-based semantic dataset are available for European languages (Schamoni et al., 2018; Al-Najjar and Hämäläinen, 2021). For English language, both image-based semantic analysis (Adjali et al., 2020; Zhang et al., 2021; Xu et al., 2020; Kruk et al., 2019) and video-based semantic dataset (Castro et al., 2019; Wang et al., 2019a) are available. The major source of information for semantic analysis are social media data (Adjali et al., 2020; Kruk et al., 2019; Xu et al., 2020; Wang et al., 2019a; Mousselle-Sveriegh et al., 2018). TV series (Castro et al., 2019; Al-Najjar and Hämäläinen, 2021) and other miscellaneous sources (Schamoni et al., 2018; Xie et al., 2017).
### 3.8. Miscellaneous

Many new NLP tasks are associated with specific domains, and multiple applications. Most of the datasets are for English (EN) language with a few exceptions of German (DE) (Alac¨am et al., 2020), Japanese (JP) (Yamazaki et al., 2020), Hindi (Hi-IN) (Chauhan et al., 2021) and some additional languages. We further investigate the benchmark datasets for object recognition (Lin et al., 2014) [Vaidyanathan et al., 2018, Ala´c¸am et al., 2020], image recipe recognition (Wang et al., 2015) and emotion recognition (Thomee et al., 2016).

#### 3.8.1. Applications of Miscellaneous Datasets

We introduce some real-time applications such as research areas for behavioral studies are personality analysis, social well-being, and humor/trolls detection. Recently studies on humour detection (Hasan et al., 2019) [Chauhan et al., 2021] and trolls identification (Suryawanshi et al., 2020b) gives promising results with available datasets. The research community use MuSE (Jaiswal et al., 2020) dataset to solve the problem of personality measure. We further investigate behavioural analysis with deception detection (Gupta et al., 2019), emotion recognition (Thomee et al., 2016) [Saha et al., 2020, Calabrese et al., 2020], and sentiment analysis (Firdaus et al., 2020a) [Zlatintsi et al., 2017]. For analysis of digital content for cooking recipes (Pustejovsky et al., 2021) [Lin et al., 2020, Wang et al., 2015] and media generation (Luo et al., 2021) (Papaspartopoulos and Cohen, 2021), we use multimodal datasets.

#### 3.8.2. Sources for Miscellaneous Datasets

The source of information is anything ranging from video recordings (Yamazaki et al., 2020) [Alac¨am et al., 2020, Jaiswal et al., 2020] to automatic collection of social media data [Lin et al., 2014] [Russakovsky et al., 2015] [Thomee et al., 2016] or digital shows/talks. Frequently used information sources are TV series (Chauhan et al., 2021) [Firdaus et al., 2020a], movies (Zlatintsi et al., 2017), TED Talks (Hasan et al., 2019) and traditional News media. Existing studies use social media data, recipe websites and Wikipedia (Calabrese et al., 2020) to generate image-based multimodal datasets. Authors re-annotate the existing datasets (Saha et al., 2020) to enhance the existing multimodal datasets like MELD and IEMOCAP for Dialogue act and emotion recognition.

### 4. Discussion

In this section, we first study an year-wise distribution of multimodal datasets for NLP applications and briefly discuss the data availability. We enlist multimodal datasets for NLP problems as a tuple $< a, l, s >$ to discuss the cross-domain usage. We also study datasets associated with non-English languages.

#### 4.1. Year-wise Distribution

NLP Research community is playing with multimodal datasets for more than a decade now. However, there are variations in the use of such datasets. We thus investigate the evolution of multimodal datasets in Figure 4.

Many new miscellaneous NLP tasks are introduced along with multimodal datasets, and thus, recent developments for miscellaneous tasks are making progress. Before 2015, the progress in classification problem of sentiment analysis has given 13 multimodal datasets. Multimodal question answering datasets have gained attention in 2016-17 and is still being explored. We observe the equal distribution of multimodal datasets for various NLP-centered tasks in 2018-2019. We further notice that there is a subsequent shift in trends from sentiment analysis (before 2015) to question answering.
Table 7: Semantic Analysis

| Dataset                          | Lang. | Modality | Samples | Applications(s) | Avail | Source                  | #Ct. |
|----------------------------------|-------|----------|---------|-----------------|-------|-------------------------|------|
| BabyIFC (Kalabrese et al., 2020) | EN    | T, T     | 10013   | Events and Emotions | Yes   | BabyIBNet-Wiki          | 3    |
| Bag-of-Laes (Kapta et al., 2019) | EN    | A, V     | 325     | Deception Detection | Yes   | Recorded                | 10   |
| Chat-talk Corpus (Yamamoto et al., 2020) | EN | T, A, V | 119303  | Conversational Phenomena Analysis | No     | Recorded                | 7    |
| COGNIMUSE (Kamitsa et al., 2017) | EN    | A, V, NA | 19,365  | SA, Semantics, Salience | OR     | Hollywood Movies        | 30   |
| EMOTyDA (Saha et al., 2020)      | EN    | T, V     | 39,356  | Dialogue Act & Emotion Recognition | Yes   | IEMOCAP-MELD            | 12   |
| EyeRef (Allaˇ{c}ni´c et al., 2020) | DE    | A, V     | 2024    | Object detection, ASR, Multiple | Recorded | 1                  |      |
| ILSVRC (Razmakovsky et al., 2015) | EN    | T, T     | 14,197,122 | Visual Recognition | OR     | Flickr & Search Engines | 27,585 |
| MARC (Las et al., 2020)          | EN    | T, V     | 150K    | Cooking Recipe | Yes   | Common Crawl            | 5    |
| Mine: (Jaswal et al., 2020)      | EN    | T, A, V  | 784     | Personality measure | Yes   | Recorded                | 9    |
| MELINEX (Wu et al., 2021)        | EN    | I, T     | 5,371   | Biomedical       | Yes   | 2                      |      |
| DJT3 (Takahashi et al., 2019)    | HI-IN | T, A, V  | 6391    | Attributes (Humour) | Yes   | Hindi TV series        | 3    |
| MS COCO (Fan et al., 2014)       | EN    | I, T     | 2.5 million | Object Recognition | Yes   | YouTube                 | 20,735 |
| NewoCLIPings (Luo et al., 2021)  | EN    | I, T     | 988k    | Media Generation | Yes   | VisualNews              | 5    |
| RCVQ (Pustejovsky et al., 2021)  | EN    | T, A, V  | 51331   | Recipe Comprehension | Yes   | 3 Recipe Websites | 0    |
| SEMD (Fardians et al., 2018)     | EN    | T, A, V  | 55000   | SA & Dialogue Generation | Yes   | Recorded                | 4    |
| SNAX (Vaidyanathan et al., 2018) | EN    | T, I     | 100     | Object Detection | Yes   | Recorded                | 7    |
| TrollMemes (Suryawanshi et al., 2020) | EN | T, I     | 2967    | Attributes (Troll) | OR     | Social media            | 20   |
| UK Famili (Hendrick et al., 2019) | EN    | T, A, V  | 10,014  | Attribute (Humour) | Yes   | TED Talks               | 36   |
| UPARC Fod (Wang et al., 2015)    | EN    | I, T     | 100,000 | Image Recognition | Yes   | Google Image search     | 125  |
| YPCC100M (Honee et al., 2016)    | EN    | V, T     | 68,552,616 | Visual and Emotion Recognition | Yes   | Flickr                  | 1008 |

Table 8: Miscellaneous

| Dataset                          | Lang. | Modality | Samples | Applications(s) | Avail | Source                  | #Ct. |
|----------------------------------|-------|----------|---------|-----------------|-------|-------------------------|------|
|                                |       |          |         |                 |       |                         |      |

4.2. Cross-Domain Usage

Multimodal datasets are either created, re-annotated, or re-used for different NLP tasks. We further emphasize this cross-domain usage by mapping the existing and newly introduced datasets as the re-annotation, and re-use helps reduce time, cost, and efforts. Some of the most widely used datasets: Multi30K (Elliott et al., 2016), Flickr30K, Visual Genome, VQA (Goyal et al., 2017), COCO, Rico-SCA, FB15K, IEMOCAP & MELD, are used to re-annotate and generate new datasets: SEWA (Kossafi et al., 2019), MTL (Laia and Specia, 2018), Flickr30K EN (hi-IN) (Chowdhury et al., 2018), WAT2019 (Meeeti et al., 2019), TDIUC (Kafle and Kanam, 2017), GQA (Hudson and Manning, 2019), Screen2Words (Wang et al., 2021), Starsem18-multimodalKB (Moussely-Sergieh et al., 2018), EMOTyDA (Saha et al., 2020), respectively. We found that this cross-domain usage helps to enhance the scope of the multimodal datasets.

4.3. Benchmark Datasets and Their Availability

It is difficult to obtain datasets due to ethical constraints. We choose to determine the benchmark multimodal datasets which are given for different applications, multiple languages, and discrete set of sources. In this context, we give datasets for tuple < a, l, m, s > and enlist some new permutations for which multimodal datasets are still unavailable. To handle this, we have enlisted this information in Table 9.

As per our investigation, there are minimal studies for multimodal machine translation in low-resourced languages (Chen et al., 2019) as there is no available dataset. There are minimal studies with non-English language for multimodal question answering and automatic summarization. We do not enlist the miscellaneous datasets as various multimodal datasets are in-
produced for a unique set of tuple \(<a, l, m, s>\) for various NLP tasks.

### 4.4 Challenges

There are several challenges faced by multimodality:

- **Joint or Coordinated Representation** Combining two modalities for exploiting the redundancy of multiple modalities. The heterogeneous nature of multimodal data makes it challenging to procure complete information in their vector representation.

- **Translation** or mapping the data from one modality to another is subjective and often open-ended. For instance, there are several ways to describe an image but not one way for perfect Translation.
• Alignment or identifying relation between subelements of different modalities. For instance, we want to map the meeting minutes to the video recording. To tackle this challenge, we need to measure similarity between different modalities and deal with possible long-range dependency and contact switching.

• Fusion joining information from two or more modalities to perform prediction (Lücking and Pfeiffer, 2012). For instance, for audio-visual speech recognition, the visual description of the lip motion is fused with the speech signal to predict the spoken words. The information coming from different modalities may have varying predictive power and noise topology, possibly missing data.

• Co-learning or transfer learning between modalities, their representation, and their predictive models. This is exemplified by algorithms of co-training, conceptual grounding, and zero-shot learning.

4.5. New Frontiers

This section will discuss some new frontiers that meet the actual NLP application needs and fit in with real-world scenarios. Besides verbal information, non-verbal information either supplements existing information or provides further information, which enriches the textual representation.

Synchronous multimodal dialogues refer textual, audio, and video recordings for the same event. Alignment of audio and video may enhance text representation and may provide new insights, such as an emotion, behavior, facial expressions, or person’s presence. However, facial features and voiceprints are of supreme privacy for individuals, making them hard and sensitive to be acquired. Future works can consider multimodal data processing for various applications under the federal learning framework (Li et al., 2021).

Asynchronous multi-modal dialogues refer different modalities that happen at different times. For instance, with the development of communication technology, multi-modal messages, such as voice messages, and pictures are frequently used in chat dialogues via applications like Messenger, WhatsApp, and WeChat. These messages provide richer information, serving as the part of a dialogue flow. Future works should consider textual information of voice messages via ASR systems, new entities provided by pictures, and emotions associated with text, image frames and audio to produce meaningful summaries and to retrieve information.

Customer service aims to address questions raised or feedback provided by agents. Therefore, it naturally has strong motivations, assisting this process with multimodal effect recognition and capturing consumer facial expressions, body postures, and gestures after product usage (Patwardhan and Knapp, 2017). In future, multiple modalities could be added such as, eye-tracking, nodding of head.

Medical AI assistance aims at quickly finishing electronic health records and medically aiding for faithful rather than creative decision making. AI methods combine text and images (say MRI images) to generate a complete customer assistance (Ahmed, 2011). Even though current multimodal systems have made a significant progress, they suffer from the problem of fabricating some factual information from the text which are called hallucinations (Huang et al., 2021b). (Chen and Yang, 2020) point out that the wrong reference is one of the main errors made by the dialogue summarization model, which means the generated summaries contain information which is not faithful to the original dialogue (e.g., associate one’s actions or locations with a wrong speaker). This error primarily hinders the application of dialogue summarization systems. We argue that this problem is mainly caused by the multiple participants and diverse references in the dialogue.

In the future, we can enhance it with the coreference resolution model with features and simplicity using contextual and discourse information. It can also be utilized to map the fake news detection applications. Multi-modality can ease the domain adaption across various domains and languages in different application such as conversational agents, social media, machine translation, medical imaging.

5. Conclusion

We provided an extensive survey of multi-modal datasets in the hope that it will reduce the efforts put in by researchers to obtain, manually clean, and pre-process datasets for their use in multimodal analysis. We found that some datasets contain annotations of different types, making them rather versatile for various NLP tasks. We map the tuple < a, l, s > (application, language, source > across all the multimodal datasets. We have released the entire collection of all multimodal datasets for NLP applications publicly to the community for re-usability and continuous updates. We formulate inferences, challenges, and new frontiers in this context. We also enumerate the detailed annotations of the benchmark multimodal datasets. As future work, we plan to conduct surveys related to some more tasks like image captioning, speech synthesis, explainable AI and others.

6. Acknowledgment

This work has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR), and 19-26934X (NEUREM3) of the Czech Science Foundation.

7. References
