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 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/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

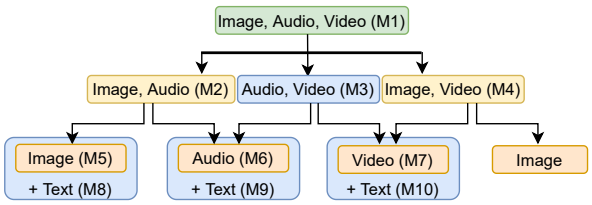

Figure 1: Different modalities and their combinations. Each of the individual modalities: *image* (I), *audio* (A) and *video* (V) are combined with *text* to create IT, AT and IV, respectively. We further group their combination pairs as *(image, audio)* (IA), *(audio, video)* (AV), and *(image, video)* (IV). We further group all the modalities as *(image, audio, video)*. The text (T) track can be added to the combined modalities, too.

for 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 re-using and re-furbishing existing multimodal datasets

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 mutlimodal datasets. The task of text summarization projects significant information in abstract way and is recently 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 in 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 datasets now cover German (DE) (Cevher et al., 2019; Alaçam et al., 2020), French (FR) (Ringeval et al., 2013), Spanish (ES) (García-Vegaa 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; Pérez-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-

Dataset	Language	Modality	Samples	Avail	Source	#Cit.
AFEW (Dhall et al., 2012)	EN	A, V	1645	OR	Movies	437
AMMER (Cevher et al., 2019)	DE	T, A, V	288	OR	Drivers	18
CH SIMS (Yu et al., 2020a)	ZH	T, I, V	2281	Yes	Movie, TV series/ shows	18
CMU-MOSEAS (Zadeh et al., 2020)	FR, ES, PT, DE	T, A, V	40000	OR	Youtube (YT)	5
CMU-MOSEI (Zadeh et al., 2018)	EN	T, A, V	23453	Yes	Youtube	242
Creep-Image (Menini et al., 2020)	EN	T, I	17912	Yes	CREENDER tool	3
CMU-MOSI (Zadeh et al., 2016)	EN	T, A, V	93	Yes	Youtube	154
EmoDB (Burkhardt et al., 2005)	DE	T, A	800	Yes	Recordings	2134
Entheos (Viegas and Alikhani, 2021)	EN	T, A, V	2351	Yes	TED talks	1
Fakeddit (Nakamura et al., 2020)	EN	T, I	1 mn	Yes	Reddit	43
HUMAINE (Douglas-Cowie et al., 2011)	EN	A, V	50	Yes	TV Recording	30
ICT-MMMO (Wøllmer et al., 2013)	EN	T, A, V	370	Yes	YT & ExpoTV	286
IEMOCAP (Busso et al., 2008)	EN	T,A, V	10000	Yes	At university	1624
Large Movie (Maas et al., 2011)	EN	T, I	25000	Yes	IMDB	
MEISD (Firdaus et al., 2020b)	EN	T, A, V	407	Yes	TV Series Friends	5
MELD (Poria et al., 2019)	EN	T, A	13000	Yes	TV Series Friends	227
Mimicry (Sun et al., 2011)	EN	A, V	54	Yes	Recorded	52
MOUD (Pérez-Rosas et al., 2013)	ES	T, A, V	400	Yes	Youtube	172
MultiOFF (Suryawanshi et al., 2020a)	EN	T, I	743	Yes	Social media	30
POM (Park et al., 2016)	EN	T, V	903	Yes	Movies	6
RECOLA (Ringeval et al., 2013)	FR	A, V	46	OR	Recorded	524
SEMAINE (McKeown et al., 2011)	EN	A, V	80	OR	Recorded	626
SEWA (Cevher et al., 2019)	EN	A, V	538	Yes	Existing DB	82
SST (Socher et al., 2013)	EN	T, I	11855	Yes	rottentomatoes.com	5837
TASS (García-Vegaa et al., 2020)	ES	T, I	3413	OR	Twitter	9
VAM (Grimm et al., 2008)	EN	A, V	499	Yes	TV Talk Show	444
Youtube D (Morency et al., 2011)	EN	T, A, V	47	Yes	Youtube	350

Table 1: Sentiment Analysis

Dataset	Language	Modality	Samples	Avail	Source	#Cit.
Flickr30K- EN- (hi-IN) (Chowdhury et al., 2018)	EN, HI-IN	T, I	155,070	OR	Flickr30K	12
Hindi Visual Genome (Parida et al., 2019)	EN, HI-IN	T, I	31525	Yes	Visual Genome	17
How2	EN, PT	T, I, V	79114	Yes	Youtube	119
HowTo100M (Huang et al., 2021a)	9 language	T, V	138 mn clips	Yes	YouTube	9
IKEA (Zhou et al., 2018)	EN, FR, DE	T, I	3600	Yes	IKEA, UNIQLO	41
MLT (Lala and Specia, 2018)	EN, FR, DE	T, I	98647	Yes	Multi30K	20
Multi30K (Elliott et al., 2016)	EN, DE	T, I	155070	Yes	Flickr	323
VATEX (Wang et al., 2019b)	EN-ZH	T, V	206,000	Yes	YouTube	117

Table 2: Machine Translation

ropean language translation (Lala and Specia, 2018; Huang et al., 2021a; Elliott et al., 2016) and Asian language translations (Wang et al., 2019b; Parida et al., 2019), there is limited contribution for other low-resourced languages.

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 (Meetei et al., 2019) and Slovenian (Pesek 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 (Meetei 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; Kafle 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 TDIUC (Kafle 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 MSRVTT QA (Ye et al., 2017); and for TV shows, movies and gameplays 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

Dataset	Language	Modality	Samples	Avail	Source	#Cit.
ALF-200k (Zangerle et al., 2018)	EN	T, A	200000	Yes	Spotify	6
Moodo (Pesek et al., 2017)	EN, SI	A, V	6999	Yes	Film music, recorded	22
MQA (Sheng et al., 2019)	EN	T, I	12595	Yes	Egyptian Art	2
MTD (Zalkow et al., 2020)	EN	T, A	2067	Yes	CD album collection	5
MUSICLEF (Orio et al., 2011)	EN	T, A	1355	Yes	Songs	22
MusiClef (Zalkow et al., 2020)	EN	T, A	1355	Yes	mertolyrics	28
Schubert Winterreise (Weiß et al., 2021)	EN	T, I, V	24	Yes	Performances of Winterreise	12
ViTT (Huang et al., 2020a)	EN	I, V	8850	Yes	YouTube	10
WAT2019 (Meetei et al., 2019)	EN, Hi-IN	T, I	31525	Yes	Visual Genome	13
WIT (Srinivasan et al., 2021)	100+ lang.	T, I	4400	Yes	Wikipedia	23

Table 3: Information Retrieval

Dataset	Language	Modality	Samples	Avail	Source	#Cit.
GQA (Hudson and Manning, 2019)	EN	I, T	1,13,018	Yes	COCO, Flikr	344
MarioQA (Mun et al., 2017)	EN	V, T	1,87,757	Yes	Gameplays- Super Mario Bros	69
MemexQA (Jiang et al., 2017)	EN	V, T, I	13,591	Yes	Flickr	24
MIMOQA (Singh et al., 2021)	EN	T, I	200	No	Existing: Unimodal	2
MovieFIB (Maharaj et al., 2017)	EN	V, T	3,48,998	Yes	-	56
MovieQA (Tapaswi et al., 2016)	EN	V, T	14944	Yes	Diverse sources: Wikipedia,imdb	485
MQA (Sheng et al., 2016)	EN	T, I	206	No	Wiki & Online	4
MSVD QA, MSRVTT QA (Ye et al., 2017)	EN	V, T	1987	Yes	Youtube	70
PororoQA (Kim et al., 2017)	EN	V, I	16,066	Yes	cartoon videos series 'Pororo'	116
RecipeQA (Yagcioglu et al., 2018)	EN	T, I	36000	Yes	Instructable	83
Social IQ (Zadeh et al., 2019)	EN	V, T	1,250	Yes	YouTube	36
TDIUC (Kafle and Kanan, 2017)	EN	T, I	1,654,167	Yes	VQA	156
TGIF-QA (Jang et al., 2017)	EN	V, T	1,65,165	Yes	Social Media	242
TVQA (Lei et al., 2018)	EN	V, T	21,793	Yes	6 popular TV shows	229
Video Context QA (Zhu et al., 2017)	EN	V, T	1,09,895	Yes	TACoS, MPII-MD, MEDTest 14 datasets	182
YouTube2Text (Xu et al., 2017)	EN	V, T	243k	Yes	Youtube	125
VQA (Goyal et al., 2017)	EN	T, I	265,016	Yes	Amazon Mechanical Turk (AMT)	1120

Table 4: Question Answering

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).

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., 2013). 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 intialized 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; Mousselly-Sergieh 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).

Dataset	Language	Modality	Samples	Avail	Source	#Cit.
AVIATE (Atri et al., 2021)	EN	T, V	8201	No	Academic conferences	0
Dev AM (Curtis et al., 2018)	EN	A, V	172	No	Conference presentations	2
MMSS (Li et al., 2018)	EN	I, T	66,000	No	Yahoo	30
MSMO (Zhu et al., 2018)	EN	I, T	314581	No	Daily Mail	53
MM-AVS (Fu et al., 2021)	EN	T, A, V	2173	Yes	CNN & Daily Mail	0
Multimodal Microblog Summarization (Saini et al., 2021)	EN	I, T	9567	No	Twitter	0
QFVS (Sharghi et al., 2017)	EN	V, T	46	Yes	MTurk	87
Screen2Words (Wang et al., 2021)	EN	I, T	22,417	Yes	Opensource dataset: Rico-SCA	1
SumMe (Gygli et al., 2014)	EN	V, T	25	Yes	Recorded	525
TVSum (Song et al., 2015)	EN	V, I, T	50	Yes	YouTube	396

Table 5: Summarization

Dataset	Language	Modality	Samples	Avail	Source	#Cit.
Chinese Whispers (Kontogiorgos et al., 2020)	ZH	A, V	34	Yes	Recorded	3
CUAVE (Patterson et al., 2002)	EN	A, V	7,000	Yes	Recorded	349
EMRE (Krishnaswamy and Pustejovsky, 2019)	EN	T, V	1500	Yes	VoxSim	9
MHAD (Ofli et al., 2013)	EN	A, V, T, I	660	Yes	Recorded	434
MHHRI (Celiktutan et al., 2017)	EN	A, V, T	746	Yes	Recorded	56
Multi-party interactions (Stefanov and Beskow, 2016)	EN	I, T	15	No	Recorded	13
Red Hen Lab (Joo et al., 2017)	RU, AR, BR, PT	A, V, I, T	350,000	No	Global TV News	20
uulmMAC (Hazer-Rau et al., 2020)	EN	A, V, T	100	Yes	Recorded	13

Table 6: Human Computer Interaction

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) (Alaçam 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; Alaçam 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) (Papasarantopoulos 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; Alaçam et al., 2020; Jaiswal et al., 2020; Gupta et al., 2019) 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 2.

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

Dataset	Language	Modality	Samples	Avail	Source	#Cit.
MDID (Kruk et al., 2019)	EN	T,I,V	1299	Yes	Instagram	39
MSDS (Al-Najjar and Hämäläinen, 2021)	ES	T, A, V		Yes	TV serials	0
MultiMET (Zhang et al., 2021)	EN	T, I	10437	Yes	Twitter, Facebook	0
MUStARD (Castro et al., 2019)	EN	T, A, V	6365	Yes	TV shows	56
Social media posts from Flickr discussing mental health (Xu et al., 2020)	EN	T, I	11828	Yes	Flickr	7
Starsem18-multimodalKB (Mousselly-Sergieh et al., 2018)	EN	T, I	100	Yes	FB15K	26
Twitter MEL (Adjali et al., 2020)	EN	T, I	14 mn	Yes	Twitter	2
YouMakeup (Wang et al., 2019a)	EN	T, V	2800	Yes	Youtube	7
Wikimedia Commons (Schamoni et al., 2018)	EN, RU, FR, DE	T, I	4 mn	Yes	Wikimedia Commons	6
WN9-IMG (Xie et al., 2017)	EN	T, I	63,225	Yes		76

Table 7: Semantic Analysis

Dataset	Lang.	Modality	Samples	Application(s)	Avail	Source	#Cit
BabelPic (Calabrese et al., 2020)	EN	T, I	10013	Events and Emotions	Yes	BabelNet- Wiki	3
Bag-of-Lies (Gupta et al., 2019)	EN	A, V	325	Deception Detection	Yes	Recorded	10
Chat-talk Corpus (Yamazaki et al., 2020)	JP	A, V	19303	Conversational Phenomena Analysis	No	Recorded	7
COGNIMUSE (Zlatintsi et al., 2017)	EN	A,V	NA	SA, Semantics, Saliency	OR	Hollywood Movies	30
EMOTyDA (Saha et al., 2020)	EN	V, T	19,365	Dialogue Act & Emotion Recognition	Yes	IEMOCAP, MELD	12
Eye4Ref (Alaçam et al., 2020)	DE	T, A, V	2024	Object detection, ASR, Multiple		Recorded	1
ILSVRC (Russakovsky et al., 2015)	EN	I, T	14,197,122	Visual Recognition	OR	Flickr & Search Engines	27558
MARC (Lin et al., 2020)	EN	T, V	150K	Cooking Recipe	Yes	Common Crawl	5
MuSE (Jaiswal et al., 2020)	EN	T, A, V	784	Personality measure	Yes	Recorded	9
MELINDA (Wu et al., 2021)	EN	I, T	5,371	Biomedical	Yes	-	2
M2H2 (Chauhan et al., 2021)	Hi-IN	T, A, V	6191	Attributes (Humour)	Yes	Hindi TV series	0
MS COCO (Lin et al., 2014)	EN	I, T	2.5 million	Object Recognition	Yes	YouTube	20735
NewsCLIPpings (Luo et al., 2021)	EN	I, T	988k	Media Generation	Yes	VisualNews	5
R2VQ (Pustejovsky et al., 2021)	EN	T, V	51331	Reciepe Comprehension	Yes	3 Reciepe Websites	0
SEMD (Firdaus et al., 2020a)	EN	T, A, V	55000	SA & Dialogue Generation		TV shows	6
SNAG (Vaidyanathan et al., 2018)	EN	T, I	100	Object Detection	Yes	Recorded	7
TrollMemes (Suryawanshi et al., 2020b)	EN	T, I	2969	Attributes (Troll)	OR	Social media	18
UR-Funny (Hasan et al., 2019)	EN	T, A, V	16514	Attribute (Humour)	Yes	TED Talks	36
UPMC Food-101 (Wang et al., 2015)	EN	I, T	100,000	Image Recipe Recognition	Yes	Google Image search	128
YFCC100M (Thomee et al., 2016)	EN	V, I, T	68,552,616	Visual and Emotion Recognition	Yes	Flickr	1008

Table 8: Miscellaneous

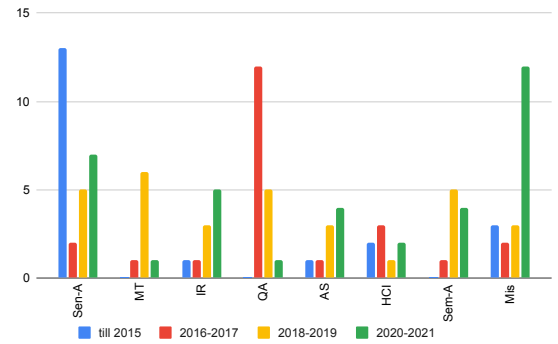

Figure 2: Year-wise distribution of research work. Sen-A: Sentiment Analysis, MT: Machine Translation, IR: Information Retrieval, QA: Question Answering, AS: Automatic Summarization, HCI: Human Computer Information, Sem-A: Semantic Analysis, Mis: Miscellaneous

(near 2016-2017) followed by insurgence of prevailing problem of machine translation (Huang et al., 2020b).

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-usage 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 (Kossaifi et al., 2019), MLT (Lala and Specia, 2018), Flickr30K- EN- (hi-IN) (Chowdhury et al., 2018), WAT2019 (Meetei et al., 2019), TDIUC (Kafle and Kanan, 2017), GQA (Hudson and Manning, 2019), Screen2Words (Wang et al., 2021), Starsem18-multimodalKB (Mousselly-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-

Dataset	Tuple	Dataset details	Dataset Annotation
CH SIMS	<SenA, ZH, IV, M/TV>	SIMS has 2,281 refined video clips collected from different movies, TV serials, and variety shows with spontaneous expressions, various head poses, occlusions, and illuminations	Fine-grained annotations of modality
CMU-MOSEI	<SenA, EN, AV, SM>	23,453 annotated sentences from more than 1000 online speakers	Annotations have been carried out by only master workers with higher than 98% approval rate to assure high quality annotations
Eye4Ref	<Mis, DE, AV, R>	86 systematically controlled sentence–image pairs and 2024 eye-movement recordings	Manually annotated data by two coders with a triplet notation as <argument, relation type, predicate>.
GQA	<QA, EN, IT, SM>	22,669,678 questions over 113,018 images from COCO, Flickr	Annotations: answer, full Answer, question
How2	<MT, EN-PT, IV, SM>	79,114 instructional videos (2,000 hours in total, with an average length of 90 seconds) with English subtitles	Descriptions of youtube videos with word-level time alignments to the ground-truth English subtitles for summarization.
M2H2	<Mis, HI-IN, AV, TV>	It contains 6,191 utterances from 13 episodes of a top-rated TV series "Shrimaan Shrimati Phir Se"	Human Annotated: 3 Ph.D. students with Fliess' Kappa score of 0.84
MDID	<SemA, EN, IT, SM>	1299 Instagram posts	labeled for three orthogonal taxonomies: the authorial intent behind the image-caption pair, the contextual relationship between the literal meanings of the image and caption, and the semiotic relationship between the signified meanings of the image and caption
MHHRI	<HCI, EN, AV, R>	audio, video, depth, EDA, temperature, 3-axis wrist acceleration from recorded data	self-acquaintance-assessed personality, self-reported engagement
Moodo	<IR, EN & SI, AV, FM/R>	200 music excerpts tagged with a genre label from Recorded emotions, free online music service Jamendo, film music dataset, collection of Slovenian folk songs, contemporary electro-acoustic music collection	6999 annotations describing their perception of emotions, colors, and music
MultiOFF	<SenA, EN, IT, SM>	445-train, 149-test, 149-Val from social media sites, such as Reddit, Facebook, Twitter and Instagram	Offensive or non-offensive labels with the help of voluntary annotators
Flickr30K (DE)	<MT, EN-DE, IT, SM>	Flickr30K dataset with German translations created by professional translators over a subset of the English descriptions	Descriptions crowdsourced independently. Human annotated by one of the authors and checked by German PhD Student
MusiClef	<IR, EN, AT, ML>	1355 total songs: 975 training+380 testing from "Rolling Stone 500 Greatest Songs of All Time" -mertolyrics	with respect to genre and mood aspects
MUStARD	<SemA, EN, AV, TV>	total of 6,365 videos from popular TV shows	audiovisual utterances annotated with sarcasm labels, f 345 videos labeled as sarcastic and 6,020 videos labeled as non-sarcastic
MultiMET	<SemA, EN, IT, SM>	Existing dataset of 64,832 image advertisements that contain both images and inside text	Crowdsourcing through CrowdFlower for sentiments and intent. Human annotated for Metaphors
R2VQ	<RC, EN, IV, RW>	18,000 AR, 25,000 EP, and 60,000 FN recipes from AllRecipes (AR), Epicurious (EP), and Food Network (FN)	Compretence based comprehension
Screen2Words	<AS, EN, IT, Ex.>	112k language summarization across ∼22k unique UI screens in opensource dataset Rico-SCA	human annotations for 22,417 Android UI screens
SocialIQ	<QA, EN, IV, SM>	1, 250 videos, 7, 500 questions, and 52, 500 answers from YouTube	Questions and answers are annotated for complexity levels: easy, intermediate and advanced, 30, 000 correct and 22, 500 incorrect answers
SumMe	<AS, EN, IV, R>	25 recorded videos	video was summarized by 15 to 18 different people
TVQA	<QA, EN, IV, TV>	152,545 QA pairs from 21,793 clips, spanning over 460 hours of video.	Faster R-CNN object detection for labels
TVSum	<AS, EN, IV, SM>	50 videos of YouTube	shotlevel importance scores annotated via crowdsourcing
TrollMemes	<Mis, EN, IT, SM>	The data was collected between November 1, 2019, until January 15, 2019, from sixteen volunteers over social media websites like WhatsApp, Facebook, Instagram, and Pinterest.	Human annotations with an inter-annotator agreement: Cohen's Kappa
uulmMAC	<HCI, EN, AV, R>	two homogeneous recorded samples of 60 participants and 100 recording	Interest, Overload, Normal, Easy, Underload, and Frustration
VATEX	<MT, EN-ZH, IV, SM>	41, 250 videos and 825, 000 captions in both English and Chinese;206, 000 English-Chinese parallel translation pairs	Human-annotated video descriptions build upon AMT
VQA	<QA, EN, IT, AMT>	265,016 images, At least 3 questions (5.4 questions on average) per image,10 ground truth answers per question, 3 plausible (but likely incorrect) answers per question	The most common answer among the 10 is the new answer

Table 9: Summary of the most popular recent datasets. SenA: Sentiment Analysis, MT: Machine Translation, IR: Information Retrieval, QA: Question Answering, AS: Automatic Summarization, HCI: Human-Computer Interaction, SemA: Semantic Analysis, RC: Recipe Comprehension, Mis: Miscellaneous; ZH: Chinese ; M/TV: Movie/ TV series, SM: Social Media, ML: Metro-Lyrics, FM/R: Film Music/ Recorded, AMT: Amazon Mechanical Turk, Ex.: Existing dataset, RW: Recipe Websites

troduced 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 preprocess 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

Adjali, O., Besançon, R., Ferret, O., Le Borgne, H., and Grau, B. (2020). Building a multimodal entity linking dataset from tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4285–4292.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Ahmed, M. U. (2011). *A Multimodal Approach for Clinical Diagnosis and Treatment*. Ph.D. thesis, Mälardalen University.

Al-Najjar, K. and Hämäläinen, M. (2021). !qué maravilla! multimodal sarcasm detection in spanish: a dataset and a baseline. *CoRR*, abs/2105.05542.

Alaçam, Ø., Ruppert, E., Salama, A. R., Staron, T., and Menzel, W. (2020). Eye4ref: A multimodal eye movement dataset of referentially complex situations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2396–2404.

Atri, Y. K., Pramanick, S., Goyal, V., and Chakraborty, T. (2021). See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Knowledge-Based Systems*, page 107152.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Caglayan, O., Bardet, A., Bougares, F., Barrault, L., Wang, K., Masana, M., Herranz, L., and van de Weijer, J. (2018). LIUM-CVC submissions for WMT18 multimodal translation task. In Ondrej Bojar, et al., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 597–602. Association for Computational Linguistics.

Calabrese, A., Bevilacqua, M., and Navigli, R. (2020). Fatality killed the cat or: Babelpic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4686.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., and Poria, S. (2019). Towards multimodal sarcasm detection (an _obviously_ perfect paper). In Anna Korhonen, et al., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4619–4629. Association for Computational Linguistics.

Celiktutan, O., Skordos, E., and Gunes, H. (2017). Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497.

Cevher, D., Zepf, S., and Klinger, R. (2019). Towards multimodal emotion recognition in german speech events in cars using transfer learning. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

Chauhan, D. S., Singh, G. V., Majumder, N., Zadeh, A., Ekbal, A., Bhattacharyya, P., Morency, L.-p., and Poria, S. (2021). M2h2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 773–777.

Chen, J. and Yang, D. (2020). Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In Bonnie Webber, et al., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4106–4118. Association for Computational Linguistics.

Chen, S., Jin, Q., and Fu, J. (2019). From words to sentences: A progressive learning approach for zero-resource machine translation with visual pivots. *International Joint Conference on Artificial Intelligence*.

Chowdhury, K. D., Hasanuzzaman, M., and Liu, Q. (2018). Multimodal neural machine translation for low-resource language pairs using synthetic data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42.

Collell, G., Van Gool, L., and Moens, M.-F. (2018). Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-second AAAI conference on artificial intelligence*.

Curtis, K., Campbell, N., and Jones, G. (2018). Development of an annotated multimodal dataset for the investigation of classification and summarisation of presentations using high-level paralinguistic features. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Deng, Y., Guo, D., Guo, X., Zhang, N., Liu, H., and Sun, F. (2021). MQA: answering the question via robotic manipulation. In Dylan A. Shell, et al., editors, *Robotics: Science and Systems XVII, Virtual Event, July 12-16, 2021*.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(03):34–41.

Douglas-Cowie, E., Cox, C., Martin, J.-C., Devillers, L., Cowie, R., Sneddon, I., McRorie, M., Pelachaud, C., Peters, C., Lowry, O., et al. (2011). The humaine database. In *Emotion-Oriented Systems*, pages 243–284. Springer.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics (ACL).

Firdaus, M., Chauhan, H., Ekbal, A., and Bhattacharyya, P. (2020a). Emosen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing*.

Firdaus, M., Chauhan, H., Ekbal, A., and Bhattacharyya, P. (2020b). Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453.

Fu, X., Wang, J., and Yang, Z. (2021). Mm-avs: A full-scale dataset for multimodal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926.

García-Vegaa, M., Díaz-Galianoa, M. C., García-Cumbrerasa, M. Á., del Arcoa, F. M. P., Montejo-Ráeza, A., Jiménez-Zafraa, S. M., Cámarab, E. M., Aguilarc, C. A., Antonio, M., Cabezudod, S., et al. (2020). Overview of tass 2020: Introducing emotion detection. In *Proceedings of Iberian Languages Evaluation Forum (IberLEF 2020)*.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, pages 865–868. IEEE.

Gupta, V., Agarwal, M., Arora, M., Chakraborty, T., Singh, R., and Vatsa, M. (2019). Bag-of-lies: A multimodal dataset for deception detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. (2014). Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer.

Hasan, M. K., Rahman, W., Zadeh, A. B., Zhong, J., Tanveer, M. I., Morency, L.-P., and Hoque, M. E. (2019). Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056.

Hazer-Rau, D., Meudt, S., Daucher, A., Spohrs, J., Hoffmann, H., Schwenker, F., and Traue, H. C. (2020). The uulmmac database—a multimodal affective corpus for affective computing in human-computer interaction. *Sensors*, 20(8):2308.

Huang, G., Pang, B., Zhu, Z., Rivera, C., and Soricut, R. (2020a). Multimodal pretraining for dense video captioning. In Kam-Fai Wong, et al., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 470–490. Association for Computational Linguistics.

Huang, P.-Y., Hu, J., Chang, X., and Hauptmann, A. (2020b). Unsupervised multimodal neural machine translation with pseudo visual pivoting. *The 58th Annual Meeting of the Association for Computational Linguistics*.

Huang, P., Patrick, M., Hu, J., Neubig, G., Metze, F., and Hauptmann, A. (2021a). Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In Kristina Toutanova, et al., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2443–2459. Association for Computational Linguistics.

Huang, Y., Feng, X., Feng, X., and Qin, B. (2021b). The factual inconsistency problem in abstractive text summarization: A survey. *CoRR*, abs/2104.14839.

Hudson, D. A. and Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Jaiswal, M., Bara, C.-P., Luo, Y., Burzo, M., Mihalcea, R., and Provost, E. M. (2020). Muse: a multimodal dataset of stressed emotion. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1499–1510.

Jang, Y., Song, Y., Yu, Y., Kim, Y., and Kim, G. (2017). Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Jiang, L., Liang, J., Cao, L., Kalantidis, Y., Farfade, S., and Hauptmann, A. G. (2017). Memexqa: Visual memex question answering. *CoRR*, abs/1708.01336.

Joo, J., Steen, F. F., and Turner, M. (2017). Red hen lab: Dataset and tools for multimodal human communication research. *KI-Künstliche Intelligenz*, 31(4):357–361.

Kafle, K. and Kanan, C. (2017). An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973.

Kim, K., Heo, M., Choi, S., and Zhang, B. (2017). Deepstory: Video story QA by deep embedded memory networks. In Carles Sierra, editor, *Pro-*

ceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pages 2016–2022. ijcai.org.

Kontogiorgos, D., Sibirtseva, E., and Gustafson, J. (2020). Chinese Whispers: A Multimodal Dataset for Embodied Language Grounding. In *Language Resources and Evaluation Conference LREC 2020*.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B. W., et al. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*.

Krishnaswamy, N. and Pustejovsky, J. (2019). Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pages 44–51.

Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., and Divakaran, A. (2019). Integrating text and image: Determining multimodal document intent in instagram posts. In Kentaro Inui, et al., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4621–4631. Association for Computational Linguistics.

Lala, C. and Specia, L. (2018). Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lei, J., Yu, L., Bansal, M., and Berg, T. L. (2018). TVQA: localized, compositional video question answering. In Ellen Riloff, et al., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1369–1379. Association for Computational Linguistics.

Li, H., Zhu, J., Liu, T., Zhang, J., Zong, C., et al. (2018). Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158.

Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., and He, B. (2021). A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*.

Libovický, J. and Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning. In Regina Barzilay et al., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 196–202. Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lin, A. S., Rao, S., Celikyilmaz, A., Nouri, E., Brockett, C., Dey, D., and Dolan, B. (2020). A recipe for creating multimodal aligned datasets for sequential tasks. In Dan Jurafsky, et al., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4871–4884. Association for Computational Linguistics.

Lücking, A. and Pfeiffer, T. (2012). Framing multimodal technical communication. In *Handbook of Technical Communication*, pages 591–644. De Gruyter Mouton.

Luo, G., Darrell, T., and Rohrbach, A. (2021). Newsclippings: Automatic generation of out-of-context multimodal media. In Marie-Francine Moens, et al., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6801–6817. Association for Computational Linguistics.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Maharaj, T., Ballas, N., Rohrbach, A., Courville, A., and Pal, C. (2017). A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

Meetei, L. S., Singh, T. D., and Bandyopadhyay, S. (2019). Wat2019: English-hindi translation on hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188.

Menini, S., Aprosio, A. P., and Tonelli, S. (2020). A multimodal dataset of images and text to study abusive language. In *CLiC-it*.

Morency, L.-P., Mihalcea, R., and Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176.

Mousselly-Sergieh, H., Botschen, T., Gurevych, I., and Roth, S. (2018). A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 225–234.

Mun, J., Hongsuck Seo, P., Jung, I., and Han, B. (2017). Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875.

Nakamura, K., Levy, S., and Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In Nicoletta Calzolari, et al., editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6149–6157. European Language Resources Association.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2013). Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60. IEEE.

Orio, N., Rizo, D., Miotto, R., Schedl, M., Montecchio, N., and Lartillot, O. (2011). Musiclef: a benchmark activity in multimodal music information retrieval. In *ISMIR*, pages 603–608. Citeseer.

Papasarantopoulos, N. and Cohen, S. B. (2021). Narration generation for cartoon videos. *arXiv preprint arXiv:2101.06803*.

Parcalabescu, L., Trost, N., and Frank, A. (2021). What is multimodality? *CoRR*, abs/2103.06304.

Parida, S., Bojar, O., and Dash, S. R. (2019). Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.

Park, S., Shim, H. S., Chatterjee, M., Sagae, K., and Morency, L.-P. (2016). Multimodal analysis and prediction of persuasiveness in online social multimedia. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(3):1–25.

Patterson, E. K., Gurbuz, S., Tufekci, Z., and Gowdy, J. N. (2002). Cuave: A new audio-visual database for multimodal human-computer interface research. In *2002 IEEE International conference on acoustics, speech, and signal processing*, volume 2, pages II–2017. IEEE.

Patwardhan, A. S. and Knapp, G. M. (2017). Multimodal affect analysis for product feedback assessment. *CoRR*, abs/1705.02694.

Pérez-Rosas, V., Mihalcea, R., and Morency, L.-P. (2013). Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982.

Pesek, M., Strle, G., Kavčič, A., and Marolt, M. (2017). The moodo dataset: Integrating user context with emotional and color perception of music for affective music information retrieval. *Journal of New Music Research*, 46(3):246–260.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, et al., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

Pustejovsky, J., Holderness, E., Tu, J., Glenn, P., Rim, K., Lynch, K., and Brutti, R. (2021). Designing multimodal datasets for NLP challenges. *CoRR*, abs/2105.05999.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Saha, T., Patra, A., Saha, S., and Bhattacharyya, P. (2020). Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Saini, N., Saha, S., Bhattacharyya, P., Mrinal, S., and Mishra, S. K. (2021). On multimodal microblog summarization. *IEEE Transactions on Computational Social Systems*.

Schamoni, S., Hitschler, J., and Riezler, S. (2018). A dataset and reranking method for multimodal mt of user-generated image captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 140–153.

Sharghi, A., Laurel, J. S., and Gong, B. (2017). Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4788–4797.

Sheng, S., Van Gool, L., and Moens, M. F. (2016). A dataset for multimodal question answering in the cultural heritage domain. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17.

Sheng, S., Laenen, K., and Moens, M.-F. (2019). Can image captioning help passage retrieval in multimodal question answering? In *European Conference on Information Retrieval*, pages 94–101. Springer.

Singh, H., Nasery, A., Mehta, D., Agarwal, A., Lamba, J., and Srinivasan, B. V. (2021). Mimoqa: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Song, Y., Vallmitjana, J., Stent, A., and Jaimes, A. (2015). Tvsum: Summa-

rizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187.

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., and Najork, M. (2021). WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. In Fernando Diaz, et al., editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2443–2449. ACM.

Stefanov, K. and Beskow, J. (2016). A multi-party multi-modal dataset for focus of visual attention in human-human and human-robot interaction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4440–4444.

Sun, X., Lichtenauer, J., Valstar, M., Nijholt, A., and Pantic, M. (2011). A multimodal database for mimicry analysis. In *International Conference on Affective Computing and Intelligent Interaction*, pages 367–376. Springer.

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., and Buitelaar, P. (2020a). Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41.

Suryawanshi, S., Chakravarthi, B. R., Verma, P., Arcan, M., McCrae, J. P., and Buitelaar, P. (2020b). A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Vaidyanathan, P., Prud'hommeaux, E., Pelz, J. B., and Alm, C. O. (2018). Snag: Spoken narratives and gaze dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–137.

Viegas, C. and Alikhani, M. (2021). Entheos: A multimodal dataset for studying enthusiasm. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2047–2060.

Wang, X., Kumar, D., Thome, N., Cord, M., and Precioso, F. (2015). Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Wang, W., Wang, Y., Chen, S., and Jin, Q. (2019a). Youmakeup: A large-scale domain-specific multimodal dataset for fine-grained semantic comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5133–5143.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., and Wang, W. Y. (2019b). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Wang, B., Li, G., Zhou, X., Chen, Z., Grossman, T., and Li, Y. (2021). Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510.

Weiß, C., Zalkow, F., Arifi-Müller, V., Müller, M., Koops, H. V., Volk, A., and Grohganz, H. G. (2021). Schubert winterreise dataset: A multimodal scenario for music analysis. *Journal on Computing and Cultural Heritage (JOCCH)*, 14(2):1–18.

Wøllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L.-P. (2013). Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Wu, T., Singh, S., Paul, S., Burns, G. A., and Peng, N. (2021). MELINDA: A multimodal dataset for biomedical experiment method classification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14076–14084. AAAI Press.

Xie, R., Liu, Z., Luan, H., and Sun, M. (2017). Image-embodied knowledge representation learning. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3140–3146. ijcai.org.

Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. (2017). Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Xu, Z., Pérez-Rosas, V., and Mihalcea, R. (2020). Inferring social media users' mental health status from multimodal information. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6292–6299.

Yagcioglu, S., Erdem, A., Erdem, E., and Ikizler-Cinbis, N. (2018). Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368.

Yamazaki, Y., Chiba, Y., Nose, T., and Ito, A. (2020). Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 443–448.

Ye, Y., Zhao, Z., Li, Y., Chen, L., Xiao, J., and Zhuang, Y. (2017). Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 829–832.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., and Yang, K. (2020a). Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727.

Yu, Z., Yu, Z., Guo, J., Huang, Y., and Wen, Y. (2020b). Efficient low-resource neural machine translation with reread and feedback mechanism. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3):1–13.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L. (2016). MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Zadeh, A., Chan, M., Liang, P. P., Tong, E., and Morency, L.-P. (2019). Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Zadeh, A., Cao, Y. S., Hessner, S., Liang, P. P., Poria, S., and Morency, L.-P. (2020). Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 1801. NIH Public Access.

Zalkow, F., Balke, S., Arifi-Müller, V., and Müller, M. (2020). Mtd: A multimodal dataset of musical themes for mir research. *Transactions of the International Society for Music Information Retrieval*, 3(1).

Zangerle, E., Tschuggnall, M., Wurzinger, S., and Specht, G. (2018). Alf-200k: Towards extensive multimodal analyses of music tracks and playlists. In *European Conference on Information Retrieval*, pages 584–590. Springer.

Zhang, D., Zhang, M., Zhang, H., Yang, L., and Lin, H. (2021). Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225.

Zhou, M., Cheng, R., Lee, Y. J., and Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation. In *EMNLP*.

Zhu, L., Xu, Z., Yang, Y., and Hauptmann, A. G. (2017). Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421.

Zhu, J., Li, H., Liu, T., Zhou, Y., Zhang, J., and Zong, C. (2018). Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

Zlatintsi, A., Koutras, P., Evangelopoulos, G., Malandrakis, N., Efthymiou, N., Pastra, K., Potamianos, A., and Maragos, P. (2017). Cognimuse: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017(1):1–24.