

MIntRec: A New Dataset for Multimodal Intent Recognition

Hanlei Zhang

State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing, China
zhang-hl20@mails.tsinghua.edu.cn

Hua Xu*

State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing, China
xuhua@tsinghua.edu.cn

Xin Wang

Department of Computer Science and Technology, Tsinghua University; School of Information Science and Engineering, Hebei University of Science and Technology
wx_hebust@163.com

Qianrui Zhou

State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing, China
zhouqr18@mails.tsinghua.edu.cn

Shaojie Zhao

Department of Computer Science and Technology, Tsinghua University; School of Information Science and Engineering, Hebei University of Science and Technology
murrayzhao@163.com

Jiayan Teng

State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing, China
tengjy20@mails.tsinghua.edu.cn

ABSTRACT

Multimodal intent recognition is a significant task for understanding human language in real-world multimodal scenes. Most existing intent recognition methods have limitations in leveraging the multimodal information due to the restrictions of the benchmark datasets with only text information. This paper introduces a novel dataset for multimodal intent recognition (MIntRec) to address this issue. It formulates coarse-grained and fine-grained intent taxonomies based on the data collected from the TV series Superstore. The dataset consists of 2,224 high-quality samples with text, video, and audio modalities and has multimodal annotations among twenty intent categories. Furthermore, we provide annotated bounding boxes of speakers in each video segment and achieve an automatic process for speaker annotation. MIntRec is helpful for researchers to mine relationships between different modalities to enhance the capability of intent recognition. We extract features from each modality and model cross-modal interactions by adapting three powerful multimodal fusion methods to build baselines. Extensive experiments show that employing the non-verbal modalities achieves substantial improvements compared with the text-only modality, demonstrating the effectiveness of using multimodal information for intent recognition. The gap between the best-performing methods and humans indicates the challenge and importance of this task for the community. The full dataset and codes are available for use at <https://github.com/thuiar/MIntRec>.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → *Object detection*.

*Corresponding author.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike International 4.0 License.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9203-7/22/10.
<https://doi.org/10.1145/3503161.3547906>

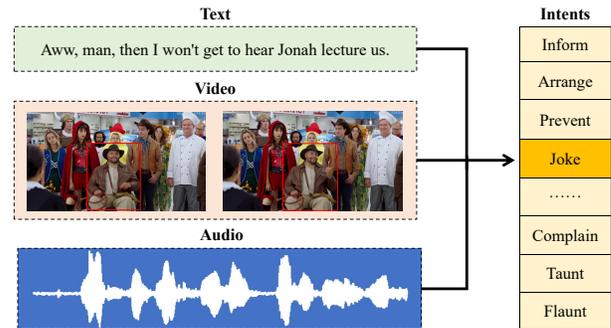


Figure 1: An example of multimodal intent recognition.

KEYWORDS

multimodal intent recognition; datasets; intent taxonomies; multimodal fusion networks; feature extraction

ACM Reference Format:

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, Jiayan Teng. 2022. MIntRec: A New Dataset for Multimodal Intent Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, Oct. 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3503161.3547906>

1 INTRODUCTION

Intent recognition is crucial in natural language understanding (NLU), which aims to leverage the text information to determine the intent categories for better conversational interactions. Though text-based intent recognition has achieved remarkable performance [11, 35, 61], it mainly focuses on goal-oriented tasks in specific domains. The intents of these tasks usually come from orders or queries with clear semantic features [8, 13], which are different from the real-world multimodal language with rich emotional, attitudinal and behavioral information. Combining natural language with non-verbal signals (e.g., expressions, body movements, and tone of speech)

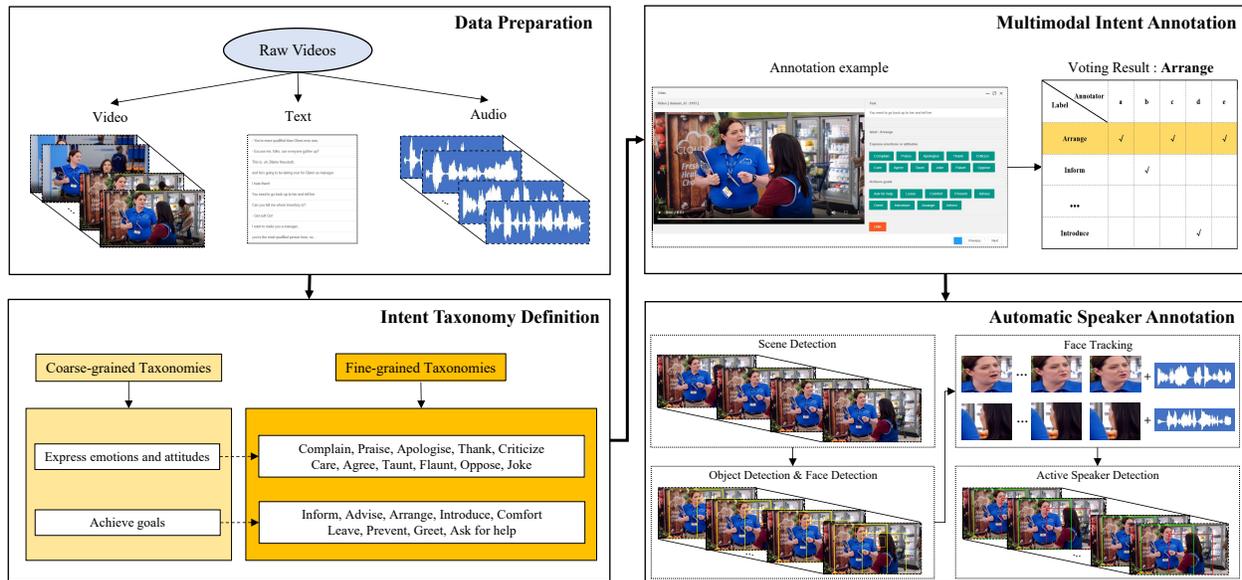


Figure 2: The process of building the MIntRec dataset.

may be beneficial in analyzing human intentions from multiple perspectives and provide more friendly services.

Taking Figure 1 as an example, we might infer the speaker to be complaining about someone based on the text information. After combining the video and audio information, we find the real intention is joking, as the speaker’s expression and tone are cheerful rather than indignant. It indicates that using text alone has difficulties satisfying the requirements of identifying complex human intents in practical situations. It is essential to use complementary knowledge of different modalities to improve the performance of intent comprehension.

Multimodal language understanding has attracted much attention in recent years. A series of multimodal datasets have been proposed in many areas such as sentiment analysis [54, 57, 58], humor detection [17], sarcasm detection [9], semantic comprehension [49, 53], etc. These benchmark datasets have extensively promoted the research and application of multimodal methodologies in related fields. However, there is still a lack of multimodal datasets for intent analysis. Most of the existing intent benchmark datasets contain merely the text modality [8, 13, 26, 30] or the visual modality [23]. MDID [25] used image-caption pairs from Instagram posts to analyze multimodal intents, but the caption-based text information is different from spoken languages in the real world.

The scarcity of data has seriously restricted the development of multimodal intent recognition. Nevertheless, constructing such a multimodal intent benchmark dataset faces two main challenges. Firstly, we need to design appropriate multimodal intent categories. Current intent taxonomies are mainly based on the text information or image-caption pairs, which have limitations when applied in multimodal scenes. Secondly, it requires distinguishing the visual information of the speaker as there is usually more than one person in the same situation. However, it will take much cost to perform manual annotation.

Table 1: Statistics of the MIntRec dataset.

Total number of coarse-grained intents	2
Total number of fine-grained intents	20
Total number of videos	43
Total number of video segments	2,224
Total number of words in text utterances	15,658
Total number of unique words in text utterances	2,562
Average length of text utterances	7.04
Maximum length of text utterances	26
Average length of video segments (s)	2.38
Maximum length of video segments (s)	9.59

To solve these problems, we propose a novel dataset, MIntRec, to fill the gap in multimodal intent recognition. The process of building the dataset is shown in Figure 2. Firstly, we prepare the original multimodal data for the dataset. The TV series SuperStore is selected as the data source due to its superiority for this task. After collecting the raw videos and subtitles, we process them into text utterances with respective video and audio segments. Then, we design both coarse-grained and fine-grained intent taxonomies for the multimodal scene. The coarse-grained taxonomies contain "Express emotions or attitudes" and "Achieve goals", which are inspired by the human intention philosophy [4]. Eleven and nine fine-grained intents are respectively summarized for these two coarse-grained categories based on the video segments and high-frequency intent tags.

Next, we perform multimodal intent annotation with the prepared data and intent taxonomies. Five well-trained workers are employed for the annotation task. They label each sample among twenty intent tags with a convenient annotation platform, and the majority voting determines the multimodal labels. Finally, we obtain 2,224 high-quality samples for MIntRec. Besides, we propose

Table 2: Intent taxonomies of our MIntRec dataset with brief interpretations.

Intent Categories	Intent	Interpretations
Express emotions or attitudes	Complain	Express dissatisfaction with someone or something (e.g., saying unfair encounters with a sad expression and helpless motion).
	Praise	Express admiration for someone or something (e.g., saying with an appreciative expression).
	Apologise	Express regret for doing something wrong (e.g., saying words of apology such as sorry).
	Thank	Express gratitude in word or deed for the convenience or kindness given or offered by others (e.g., saying words of appreciation such as thank you).
	Criticize	Point out someone’s mistakes harshly (e.g., yelling out someone’s problems).
	Care	Concern about someone or be curious about something (e.g., worrying about someone’s health).
	Agree	Have the same attitude about something (e.g., saying affirmative words such as yeah and yes).
	Taunt	Use metaphors and exaggerations to accuse and ridicule (e.g., complimenting someone with a negative expression).
	Flaunt	Boast about oneself to gain admiration, envy, or praise (e.g., saying something complimentary about oneself arrogantly).
	Oppose	Have an inconsistent attitude about something (e.g., saying negative words to express disagreement).
	Joke	Say something to provoke laughter (e.g., saying something funny and exaggerated with a cheerful expression).
Achieve goals	Inform	Tell someone to make them aware of something (e.g., broadcasting something with a microphone).
	Advise	Offer suggestions for consideration (e.g., saying words that make suggestions).
	Arrange	Plan or organize something (e.g., requesting someone what they should do formally).
	Introduce	Communicate to make someone acquaintance with another or recommend something (e.g., describing a person’s identity or the properties of an object).
	Comfort	Alleviate pain with encouragement or compassion (e.g., describing something is hopeful).
	Leave	Get away from somewhere (e.g., saying where to go while turning around or getting up).
	Prevent	Make someone unable to do something (e.g., stopping someone from doing something with a hand).
	Greet	Express mutual kindness or recognition during the encounter (e.g., waving to someone and saying hello).
Ask for help	Request someone to help (e.g., asking someone to deal with the trouble).	

an automatic process for speaker annotation. The detected object bounding boxes are used to get the visual information of persons in each video frame. To identify the bounding boxes of speakers, we first detect and track faces within bounding boxes in different visual scenes and then predict the indexes of speakers with the active speaker detection algorithm. This process achieves high performance on our constructed testing set.

After extracting features for each modality, we build baselines with three strong multimodal fusion methods. The experimental results show that leveraging the nonverbal information achieves 1% ~ 2% stable improvements on both binary and multi-class classification. However, the results of the best methods are still far from human performance, indicating the challenge of the multimodal intent recognition task.

Our contributions are summarized as follows:

(1) In this work, we build a novel multimodal intent recognition dataset, MIntRec, containing 2,224 high-quality samples with multimodal intent annotations. To the best of our knowledge, it is the first benchmark dataset for intent recognition in real-world multimodal scenes.

(2) New intent taxonomies are designed for this task. Concretely, we provide two coarse-grained and twenty fine-grained intent categories for the study of multimodal intent analysis.

(3) An automatic speaker annotation process is proposed to produce high-quality annotated bounding boxes for speakers under

the evaluation of over 12K human-annotated keyframes. It saves much time and laboratory and may benefit similar annotation tasks.

(4) Extensive experiments conducted on our dataset show utilizing multimodal information is superior to text-based intent recognition. The best-performing methods still have much room for improvement compared with humans.

2 MINTREC DATASET

In this section, we will introduce the process of building the MIntRec dataset, including data preparation, intent taxonomy definition, multimodal intent annotation, and automatic speaker annotation. The detailed statistics of MIntRec are shown in Table 1.

2.1 Data Preparation

Multimodal intent recognition requires plenty of nonverbal signals in real-world conversational scenes. For this purpose, we select the TV series Superstore as the source of our dataset, which has two main advantages: On the one hand, it consists of a wealth of characters (including seven prominent and twenty recurring roles) with different identities in the superstore, which is helpful to produce rich body language, expressions, and tones as multimodal information. On the other hand, it contains a mass of stories in various scenes (e.g., shopping mall, warehouse, office), which favor collecting diverse intent categories.

The raw videos and subtitles of Superstore are accessible on YouTube and OpenSubtitles¹. To obtain video segments, we first extract each utterance’s starting and ending timestamps of a speaker and then split the raw videos according to these timestamps. The corresponding audio segments are extracted from raw videos with the moviepy toolkit².

2.2 Intent Taxonomy Definition

Existing intent taxonomies are restricted in specific tasks [8, 13, 26] or from social media posts [23, 25], which are uncommon in real-world scenes. Therefore, we design new taxonomies for multimodal intent recognition, including two coarse-grained and twenty fine-grained intent categories.

In artificial intelligence research, intentions are regarded as plans or goals of an agent, accompanied by the corresponding feedback actions [4, 51]. However, Schröder [41] pointed out that the brain’s emotional evaluations of situations are also critical components of human intentions. We combine these two aspects and crawl through the raw videos to generalize two representative coarse-grained intent taxonomies for multimodal intent recognition, including "Express emotions or attitudes" and "Achieve goals".

The coarse-grained intent taxonomy is insufficient to distinguish the complex and diverse types of human intentions in the real world. Thus, we further classify it into fine-grained categories. Firstly, we analyze different video segments as many as possible and collect several rough intent tags as candidates for each coarse-grained category. Then, we discuss and divide similar tags into the same group (e.g., introducing something or someone, worrying about someone or being interested in something). Next, we collect high-frequency intent tags and organize them into twenty fine-grained categories, including eleven classes for "Express emotions or attitudes" and nine for "Achieve goals". Some intents may be cued by a single modality such as text (e.g., thank, apologise, greet, agree, praise), video (e.g., leave, prevent), or audio (e.g., complain, criticize). Other intents may be inferred by combining different modalities (e.g., comfort, care, joke, taunt, flaunt). Brief interpretations of each intent category are summarized by observing practical examples and referring to related materials [42], as shown in Table 2.

2.3 Multimodal Intent Annotation

After preparing data and defining intent taxonomies, we employ five students with an English foundation for annotation. Employees are offered interpretations and typical examples of each intent category as guidelines. Only well-trained employees are allowed for annotation. As intents usually exist in specific scenes of events [41], there are irrelevant utterances among consecutive video segments, so we add a UNK tag to the label set to distinguish them. To improve the labeling efficiency, we build a database to manage all the multimodal data and a convenient platform for annotation. Users only need to click the button of the intent tag to complete annotation for a piece of data.

Each of the five workers is required to complete the annotation task of the same set of data independently. They need to choose the most confident intent tag for each sample by combining the

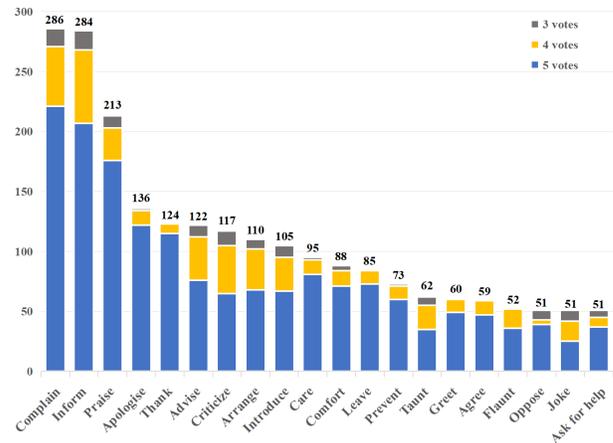


Figure 3: Voting statistics of 2,224 samples in MIntRec.

video, audio, and text information. The intent labels are determined by the majority voting (three out of five). The samples with votes larger or equal to three (not UNK) are saved. Finally, we acquired 2,224 high-quality samples to make up the MIntRec dataset.

The detailed voting statistics are shown in Figure 3. It can be observed that intent categories with clear text or video cues (e.g., apologise, thank, leave, prevent, greet, agree) are easier to reach a higher agreement with votes larger than three. We also notice that the dataset is imbalanced. The reason is that it satisfies the distribution of different intents in real-world scenarios. Some intents occur more frequently (e.g., complain, inform, and praise), while others do not (e.g., joke and ask for help). Nevertheless, each intent category still contains at least fifty samples.

2.4 Automatic Speaker Annotation

As suggested in [54, 57, 58], we first extract frames from each video segment to represent the video information. Then, we aim to annotate the visual contents related to the speakers, which are the objects of multimodal intent recognition.

We perform object detection on the video frames to obtain rich visual information containing facial and body features. Specifically, we use a Faster R-CNN [37] with the backbone ResNet-50 [21] pre-trained on the MS COCO dataset [29] (containing 250,000 person instances with seventeen keypoints) to predict the bounding boxes of persons in each frame.

However, there are still two challenges for speaker annotation. For one thing, there may be no or little visual information about the speaker in the extracted frame (e.g., most of the body is covered, or the speaker only appears on the back of the body). For another thing, there are usually multiple persons with detected bounding boxes in each frame, making it hard to distinguish the speaker. To solve these problems, we propose an automatic process to perform speaker annotation with the aid of the audio-visual active speaker detection algorithm [12, 45]. It contains the following four steps:

Firstly, we use the scene detection toolkit³ to distinguish different visual scenes in a video segment, as there may be a change in

¹<https://www.opensubtitles.org/>

²<https://pypi.org/project/moviepy/>

³<https://pypi.org/project/scenedetect/>

contents between adjacent frames. Secondly, a pre-trained Faster R-CNN is used to detect the bounding boxes of persons for each frame in a visual scene. Since facial motion (e.g., lips movement) is critical for detecting speakers, we further use the S³FD [64] algorithm to detect faces in the bounding boxes and establish the one-to-one mapping between faces and object bindings. Note that this step also filters the keyframes with clear facial features. Thirdly, a simple and effective method is used to perform face tracking. Concretely, we compute intersection-over-union (IoU) for two faces in adjacent frames and consider two faces are from the same person if IoU is above 0.5. Given that accidental objects may block faces, we tolerate up to ten consecutive frames of missing faces. Finally, we use a pre-trained TalkNet model with tracked faces and corresponding audio information to predict speakers and determine respective bounding boxes with the mapping obtained in the second step. With the aid of this process, we automatically generate more than 120K keyframes with speaker annotations of bounding boxes free from any manual intervention.

To evaluate the quality of keyframes and bounding box information, we construct a testing set with more than 12K human-annotated keyframes. Specifically, we first uniformly extract one shot every ten frames and manually select keyframes with clear visual information. Then, a pre-trained Faster R-CNN is used to predict the object bindings of persons in each keyframe, and annotators label the speakers by choosing the indexes of corresponding bounding boxes.

Compared with the human-annotated keyframes, the missing rate of generated keyframes is only 2.3%. Among the hit keyframes, the proportion of high-quality predicted bounding boxes (IoU > 0.9) is 90.9%. The evaluation results demonstrate the reliability of the automatic speaker annotation process. Besides, it is much more efficient without labor-intensive and time-consuming manual annotations.

3 METHODOLOGY

After preparing the corresponding text, video, and audio data of speakers, we extract features of each modality and use them for multimodal fusion.

3.1 Feature Extraction

3.1.1 Text. Due to the excellent performance of the pre-trained BERT language model in the Natural Language Processing (NLP) community [24], we use it to extract text features. For each text utterance, we obtain the token embeddings $z^T \in \mathbb{R}^{L_T \times H_T}$ from the last hidden layer of BERT, where L_T is the sequence length of text utterances, and H_T is the feature dimension 768.

3.1.2 Vision. The object detection method is used for extracting vision features. For each video segment, we first leverage a pre-trained Faster R-CNN with the backbone ResNet-50 to extract the feature representations x of all keyframes. Then, we map x into the regions with the annotated bounding boxes B to obtain the vision feature embeddings $z^V \in \mathbb{R}^{L_V \times H_V}$:

$$z^V = \text{AvgPool}(\text{RoIAlign}(x, B)), \quad (1)$$

where RoIAlign [20] is used to extract the fixed size feature maps (e.g., 7×7). AvgPool is used to reduce both weight and height sizes

Table 3: Dataset splits in MIntRec. The training, validation and testing sets are split into 3:1:1.

Item	Total	Express emotions or attitudes	Achieve goals
Train	1,334	749	585
Valid	445	249	196
Test	445	248	197

to the unit size. L_V is the sequence length of video segments, and H_V is the feature dimension 256.

3.1.3 Audio. The speech toolkit librosa [32] is first used to acquire audio time series at 16,000 Hz. Then, the pre-trained model wav2vec 2.0 [3] is used to extract audio features, which learns powerful representations for speech recognition with self-supervised learning. We obtain the acoustic feature embeddings $z^A \in \mathbb{R}^{L_A \times H_A}$ from the last hidden layer of wav2vec 2.0, where L_A is the sequence length of audio segments, and H_A is the feature dimension 768.

3.2 Benchmark Multimodal Fusion Methods

After feature extraction, we benchmark three powerful multimodal fusion methods for the MIntRec dataset. These methods aim to learn the interactions between different modalities with the extracted features and obtain friendly representations for multimodal fusion.

3.2.1 MulT. The Multimodal Transformer (MulT) [46] is an end-to-end method to deal with non-aligned multimodal sequences. It extends the vanilla Transformer [47] to the cross-modal Transformer with the pairwise inter-modal attention mechanism, which helps to capture the adaptation knowledge between different modalities in the latent space.

3.2.2 MISA. Hazarika et al. [19] proposed the framework MISA to learn multimodal representations with modality-invariant and modality-specific properties. On the one hand, a shared subspace is utilized to learn common features of all modalities. On the other hand, distinct subspaces are designed to capture the unique attributes of each modality. For this purpose, the training objectives contain four aspects: similarity loss, difference loss, reconstruction loss, and task-specific loss.

3.2.3 MAG-BERT. Rahman et al. [36] integrated two nonverbal modalities into BERT with an additional multimodal adaptation gate (MAG) module. MAG can produce a position shift in the semantic space adaptive to acoustic and visual information. It can be flexibly placed between layers of BERT to receive inputs from nonverbal modalities.

In this work, the features of each modality z^T, z^V , and z^A can be directly used as the inputs of MulT and MISA. As MAG-BERT needs aligned multimodal data, we pass the features of video and audio (z^V and z^A) through the Connectionist Temporal Classification (CTC) [16] module to align with the text feature z^T in the word-level as suggested in [46]. For each method, we use the multimodal annotations as targets and perform the classification task under the supervision of the softmax loss.

Table 4: Multimodal intent recognition results on the MIntRec dataset. "Twenty-class" and "Binary" denote the multi-class and binary classification over fine-grained and coarse-grained intent taxonomies. Δ denotes the most improvement over the text-classifier baseline in the current evaluation metric of each method.

Methods	Modalities	Twenty-class				Binary			
		ACC	F1	P	R	ACC	F1	P	R
Classifier	Text	70.88	67.40	68.07	67.44	88.09	87.96	87.95	88.09
MAG-BERT	Text + Audio	72.16	68.28	68.88	68.88	88.83	88.71	88.67	88.85
	Text + Video	72.09	67.92	69.09	68.73	88.45	88.28	88.36	88.27
	Text + Audio + Video	72.65	68.64	69.08	69.28	89.24	89.10	89.10	89.13
	Δ	1.77 \uparrow	1.24 \uparrow	1.02 \uparrow	1.84 \uparrow	1.15 \uparrow	1.14 \uparrow	1.15 \uparrow	1.04 \uparrow
MuT	Text + Audio	71.80	67.95	69.18	67.96	88.74	88.61	88.59	88.68
	Text + Video	71.98	68.76	69.68	68.79	88.79	88.66	88.63	88.77
	Text + Audio + Video	72.52	69.25	70.25	69.24	89.19	89.07	89.02	89.18
	Δ	1.64 \uparrow	1.85 \uparrow	2.18 \uparrow	1.80 \uparrow	1.10 \uparrow	1.11 \uparrow	1.07 \uparrow	1.09 \uparrow
MISA	Text + Audio	71.60	68.37	69.57	68.30	88.45	88.31	88.32	88.35
	Text + Video	71.53	68.34	69.68	68.19	88.74	88.60	88.63	88.65
	Text + Audio + Video	72.29	69.32	70.85	69.24	89.21	89.06	89.12	89.06
	Δ	1.41 \uparrow	1.92 \uparrow	2.78 \uparrow	1.80 \uparrow	1.12 \uparrow	1.10 \uparrow	1.17 \uparrow	0.97 \uparrow
Human	-	85.51	85.07	86.37	85.74	94.72	94.67	94.74	94.82
	Δ	14.63 \uparrow	17.67 \uparrow	18.30 \uparrow	18.30 \uparrow	6.63 \uparrow	6.71 \uparrow	6.79 \uparrow	6.73 \uparrow

4 EXPERIMENTS

This section introduces the experimental setup, baselines, and experimental results.

4.1 Experimental Setup

4.1.1 Dataset Splits. We shuffle the video segments in random and split training, validation, and testing sets by multimodal annotations in 3:1:1. The detailed statistics are shown in Table 3.

4.1.2 Evaluation Metrics. Four metrics are used to evaluate the model performance: accuracy (ACC), F1-score (F1), precision (P), and recall (R). We report the macro score over all classes for the last three metrics. The higher values indicate better performance of all metrics.

4.1.3 Implementation Details. For the text and audio modalities, we employ the pre-trained BERT model (bert-base-uncased, with 12 Transformer layers) and pre-trained wav2vec 2.0 model implemented in PyTorch [50]. For the video modality, we use a pre-trained Faster R-CNN with ResNet-50 backbone implemented in MMDetection Toolbox [10].

As sequence lengths of the segments in each modality need to be fixed, we use zero-padding for shorter sequences. L_T , L_V , and L_A are 30, 230, and 480, respectively. For all methods, the training batch size is 16, and the number of training epochs is 100. We adjust the hyper-parameters with macro F1-score. For a fair comparison, we report the average performance over ten runs of experiments with random seeds 0-9.

4.2 Baselines

We build a series of baselines for the MIntRec dataset. As the text modality is predominant in the intent recognition task, we train a classifier with the text-only modality as the primary baseline. As

suggested in [24], we use the first special token [CLS] from the last hidden layer as the sentence representation and fine-tune the pre-trained BERT model with the downstream classification task for better performance.

As introduced in section 3.2, three multimodal fusion methods, MAG-BERT, MuT, and MISA, are used to benchmark our dataset. Besides, we also modify them to use two modalities (Text + Audio and Text + Video) as inputs for ablation studies.

We have a different set of two annotators to evaluate the human performance on this task. They are provided with the training and validation sets with multimodal annotations for learning and assessment as in baselines. After that, they need to label the unseen testing set, and their average results are reported.

4.3 Results

We conduct experiments on the MIntRec dataset with several baselines and show the results in Table 4. For each multimodal fusion method, the best results are highlighted in bold. The improvements over the text-classifier are shown with Δ .

The multimodal fusion methods achieve substantial improvements on all metrics of twenty-class and binary classification compared with the text-only modality. All the multimodal fusion methods for twenty-class classification stably improve over 1% scores on all metrics. All the baselines achieve much higher performance on binary classification. We suggest the reason is that recognizing coarse-grained intent categories is much easier than distinguishing fine-grained intent categories. Nevertheless, all the multimodal fusion methods still yield over 1% improvements on almost all metrics. The results demonstrate that effectively leveraging the multimodal information helps enhance the intent recognition capability.

However, even the best-performing methods are still far away from humans. Compared with the text modality, the human performance improves by 14% ~ 19% on twenty-class classification and

Table 5: Results of each fine-grained intent category in "Express emotions and attitudes".

Methods	Complain	Praise	Apologise	Thank	Criticize	Care	Agree	Taunt	Flaunt	Oppose	Joke
Text-classifier	64.36	85.69	97.93	97.22	47.06	87.42	94.26	15.53	46.12	32.32	27.42
MAG-BERT	67.65	86.03	97.76	96.52	49.02	85.59	91.60	15.78	47.09	33.97	37.54
MuT	65.48	84.72	97.93	96.83	49.72	88.12	92.23	26.12	48.91	34.68	33.95
MISA	63.91	86.63	97.78	98.03	53.44	87.14	92.05	22.15	46.44	36.15	38.74
Δ	3.29↑	0.94↑	0.00	0.81↑	6.38↑	0.70↑	2.03↓	10.59↑	2.79↑	3.83↑	11.32↑
Human	80.08	93.44	96.15	96.90	72.21	96.09	87.21	65.55	78.10	69.04	72.22
Δ	15.72↑	7.75↑	1.78↓	0.32↓	25.15↑	8.67↑	7.05↓	50.02↑	31.98↑	36.72↑	44.80↑

Table 6: Results of each fine-grained intent category in "Achieve goals".

Methods	Inform	Advise	Arrange	Introduce	Comfort	Leave	Prevent	Greet	Ask for help
Text-classifier	67.74	67.68	64.67	68.64	77.05	73.37	82.47	84.90	66.20
MAG-BERT	71.00	69.30	63.82	67.42	76.43	75.77	85.07	91.06	64.44
MuT	70.85	69.43	65.44	71.19	76.44	75.58	81.68	86.65	69.12
MISA	70.18	69.56	67.32	67.22	78.78	77.23	83.30	82.71	67.57
Δ	3.26↑	1.88↑	2.65↑	2.55↑	1.73↑	3.86↑	2.60↑	6.16↑	2.92↑
Human	79.69	87.14	81.40	84.09	95.95	97.06	86.43	94.15	88.54
Δ	11.95↑	19.46↑	16.73↑	15.45↑	18.90↑	23.69↑	3.96↑	9.25↑	22.34↑

6% ~ 7% on binary classification. The improvements are much more significant than in multimodal fusion methods, indicating this task is very challenging for multimodal research.

5 DISCUSSION

This section analyzes the effect of nonverbal modalities and shows the performance of fine-grained intent categories with quantitative results.

5.1 Effect of Nonverbal Modalities

We conduct ablation studies for each multimodal fusion method to investigate the influence of the video and audio modalities. Specifically, we compare the tri-modality with bi-modality and show results in Table 4.

5.1.1 Bi-modality. After combining text with audio modality, the intent recognition performance achieves overall gains on both twenty-class and binary classification. Specifically, MAG-BERT, MuT, and MISA increase accuracy scores of 1.28%, 0.92%, and 0.72% on twenty-class and 0.74%, 0.65%, and 0.36% on binary classification, respectively. Combining text with video modality also leads to better performance in all settings. Similarly, MAG-BERT, MuT, and MISA increase accuracy scores of 1.21%, 1.10%, and 0.65% on twenty-class and 0.36%, 0.70%, and 0.65% on binary classification.

Due to the consistent improvements in leveraging video or audio modality, we suppose the two nonverbal modalities are critical for multimodal intent recognition. The valuable information such as tone of voice and body movements may be helpful to recognize human intents from new dimensions.

5.1.2 Tri-modality. Though multimodal fusion methods with bi-modality have achieved better performance than the text modality,

we find utilizing the tri-modality brings more gains. MAG-BERT achieves a slight advantage on the precision score but performs worse on the other metrics. The positive results demonstrate that both video and audio modalities benefit this task. The benchmark multimodal fusion methods can fully use the information from different modalities by modeling cross-modal interactions.

5.2 Performance of Fine-grained Intent Classes

To investigate the effect of the multimodal information in each fine-grained intent category, we report the average macro F1-score of each class over ten runs of experiments for all baselines and show results in Table 5 and 6. The best results of multimodal fusion methods are highlighted in bold. Δ indicates the most improvement of multimodal fusion methods and humans over the text-classifier.

Firstly, we observe the results of each class in the coarse-grained intent category "Express emotions and attitudes" in Table 5. It is shown that multimodal fusion methods perform better than text-classifier in most classes. Notably, we find there are some classes with over 3% significant improvements (e.g., complain, criticize, taunt, oppose, joke). The success of these intents is intuitive, as they contain vivid nonverbal signals of expressions and tones, requiring the aid of visual and audio information. However, we also notice that the multimodal information is less helpful with few improvements or even degradation in some classes (e.g., apologise, thank, praise, agree). The reason is that these classes usually contain clear textual cues such as sorry, thank, yeah, etc. In this case, the pre-trained language model is good enough for intent recognition.

Secondly, we observe the results of each class in the coarse-grained intent category "Achieve goals" in Table 6. The performance of multimodal fusion methods consistently achieves over 1% ~ 6% improvements in all classes. It is reasonable because these

classes are highly associated with broad body movements, such as gestures, posture, arm behaviors, etc. By comparison, capturing this information with the text modality is rather challenging.

Finally, we observe the human performance in Table 5 and 6. As expected, humans have gained an absolute advantage over the text modality in most intent categories. However, the human performance is lower than text-classifier in three classes (apologise, thank, agree). It suggests that even humans may make mistakes in the classes biased to the text modality. In contrast, humans are good at reasoning through different modalities and show significant superiority with over 10% improvements in many intents, such as taunt, flaunt, oppose, joke, etc. The huge gap indicates the necessity of exploring an effective way to leverage nonverbal information.

6 RELATED WORKS

6.1 Multimodal Language Datasets

Multimodal language understanding is a booming area with a series of emerging benchmark datasets. For example, many datasets have been proposed in multimodal sentiment analysis [6, 54, 57, 58] and emotion recognition [34]. Some multimodal datasets also detect unique properties of human languages, such as sense of humor [17, 18], metaphor [60], sarcasm [7, 9]. Moreover, multimodal datasets are designed for a series of other tasks in NLP, such as dialogue act classification [39, 40], named entity recognition [43], comprehension and reasoning [49, 53], comments generation [48], fake news detection [33], etc. Nevertheless, there is a lack of multimodal datasets for intent analysis in real-world dialogue scenes.

6.2 Benchmark Datasets for Intent Analysis

Intent analysis is a popular research field in NLU, and many important tasks have been proposed, such as joint intent detection and slot filling [35, 59], open intent detection [27, 62] and discovery [28, 63]. The booming of this area benefits from several benchmark intent datasets proposed in recent years, such as ATIS [22], Snips [13], CLINC150 [26], HWU64 [30], and BANKING77 [8]. These datasets collected the corpus by interacting with the intelligent assistant or customers in specific domains and used the crowdsourcing task among service requests to determine intent labels. StackOverflow [52] and StackExchange [5] gathered data from technical question and answering platforms. Their intent labels are defined as the tags assigned to the questions. SWBD [15] corpus contained 42 dialogue acts (DAs) for task-independent conversations. Still, many DAs are ambiguous concepts (e.g., statement-opinion and statement-non-opinion), which are difficult to be applied in real applications. Intentionomy [23] analyzed the visual intents among social media posts and collected an image dataset. However, all these datasets merely contain information from a single modality.

MDID [25] integrated image and text information for intent recognition. However, the multimodal information from Instagram posts is limited, and the taxonomies are inappropriate in real-world scenes. In contrast, MIntRec contains rich multimodal information in dialogue scenes with text, video, and audio modalities.

6.3 Multimodal Fusion Methods

Based on the multimodal language datasets, multimodal fusion methods are proposed to capture the interactions between language

and nonverbal modalities. Traditional methods, such as MCB [14] and TFN [55] obtained representations by learning intra-modality and inter-modality relations. However, the high-dimensional representations suffer from high computational complexity. LMF [31] designed low-rank multimodal tensors to solve this problem. MFN [56] first learned view-specific interactions for every single modality and then used the attention mechanism to summarize cross-perspective interactions through the multi-perspective gated memory.

Recent methods adopt Transformer-based methods for multimodal representation learning. For example, MulT [46] managed to learn interactions between different modalities with directional cross-modal attention. MISA [19] performed multimodal fusion with multi-headed self-attention to capture the relations between modality-invariant and modality-specific representations. MAG-BERT [36] introduced the multimodal adaptation gate for pre-trained Transformers to receive information from different modalities. In this work, we adapt the above three algorithms to multimodal intent recognition as benchmark methods.

6.4 Audio-visual Active Speaker Detection

Active speaker detection (ASD) aims to detect the speaker(s) in a visual scene. In this work, we focus on ASD with audio and visual information. Some studies [1, 12] treated this problem as a binary classification task and used a multi-layer perceptron (MLP) for ASD with concatenated audio and visual features. Besides, temporal structures [38, 44] such as recurrent neural networks (RNNs) were adopted to obtain better performance with time-series information. MAAS [2] introduced graph convolutional networks (GCNs) to model interactions between audio and video modalities. TalkNet [45] introduced an audio-visual cross-attention mechanism for effectively modeling cross-modal interactions and a self-attention mechanism for capturing long-term speech dependencies. In this work, we use TalkNet for automatic speaker annotation.

7 CONCLUSIONS

This paper first presents a new dataset for multimodal intent recognition. It has 2,224 high-quality annotated samples with corresponding multimodal information. New taxonomies with coarse-grained and fine-grained intent categories are specifically designed for real-world multimodal scenes. We also propose an automatic process to obtain the information of object bounding boxes towards speakers, which vastly reduces the annotation burden. We make great efforts to ensure the quality of our dataset and build baselines with three multimodal fusion methods. Comprehensive experiments verify the superiority of multimodal information for intent recognition. The gap between the best-performing multimodal fusion methods and humans indicates there is still a long way to go for multimodal intent recognition.

ACKNOWLEDGMENTS

This paper is funded by National Natural Science Foundation of China (Grant No. 62173195) and Beijing Academy of Artificial Intelligence (BAAI). We would like to thank Guohui Guan, Wenrui Li, and Xiaofei Chen for their efforts during dataset construction.

REFERENCES

- [1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. 2020. Self-supervised learning of audio-visual objects from video. In *Proceedings of the European Conference on Computer Vision*. Springer, 208–224.
- [2] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. 2021. MAAS: Multi-Modal Assignment for Active Speaker Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 265–274.
- [3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 33th Advances in Neural Information Processing Systems*, Vol. 33. 12449–12460.
- [4] Michael E Bratman. 1988. Intention,–Plans,–and–Practical–Reason. *Mind* 97, 388 (1988).
- [5] Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*. 174–185.
- [6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.
- [7] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2506–2515.
- [8] Inigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 2nd Annual Meeting of the Association for Computational Linguistics Workshop*. 38–45.
- [9] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4619–4629.
- [10] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155* (2019).
- [11] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909* (2019).
- [12] Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *Proceedings of the Asian Conference on Computer Vision*. 251–263.
- [13] Alice Coucke, Alaa Saade, Adrian Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190* (2018).
- [14] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 457–468.
- [15] John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. 517–520.
- [16] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*. 369–376.
- [17] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor Knowledge Enriched Transformer for Understanding Multimodal Humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 12972–12980.
- [18] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 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*. 2046–2056.
- [19] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1122–1131.
- [20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- [22] Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania*.
- [23] Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. 2021. Intentionomy: a Dataset and Study towards Human Intent Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12986–12996.
- [24] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.
- [25] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 4622–4632.
- [26] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 1311–1316.
- [27] Ting-En Lin and Hua Xu. 2019. Deep Unknown Intent Detection with Margin Loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5491–5496.
- [28] Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering New Intents via Constrained Deep Adaptive Clustering with Cluster Refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8360–8367.
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. 740–755.
- [30] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566* (2019).
- [31] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*. 2247–2256.
- [32] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*. 18–25.
- [33] Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 6149–6157.
- [34] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 527–536.
- [35] Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2078–2087.
- [36] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2359–2369.
- [37] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 91–99.
- [38] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. 2020. Ava active speaker: An audio-visual dataset for active speaker detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 4492–4496.
- [39] Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cognitive Computation* 13, 2 (2021), 277–289.
- [40] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4361–4372.
- [41] Tobias Schröder, Terrence C Stewart, and Paul Thagard. 2014. Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive science* 38, 5 (2014), 851–880.
- [42] Esc Simpson, Ja & Weiner. 1989. Oxford english dictionary. *Dictionary, Oxford English* (1989).

- [43] Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A Large-Scale Chinese Multimodal NER Dataset with Speech Clues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. 2807–2818.
- [44] Fei Tao and Carlos Busso. 2019. End-to-end audiovisual speech activity detection with bimodal recurrent neural models. *Speech Communication* 113 (2019), 25–35.
- [45] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3927–3935.
- [46] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6558–6569.
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 30th Advances in neural information processing systems*, Vol. 30.
- [48] Weiyang Wang, Jieting Chen, and Qin Jin. 2020. VideoIC: A Video Interactive Comments Dataset and Multimodal Multitask Learning for Comments Generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2599–2607.
- [49] Weiyang Wang, Yongcheng Wang, Shizhe Chen, and Qin Jin. 2019. 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*. 5133–5143.
- [50] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.
- [51] Michael Wooldridge and Nicholas R Jennings. 1995. Intelligent agents: Theory and practice. *The knowledge engineering review* 10, 2 (1995), 115–152.
- [52] Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short Text Clustering via Convolutional Neural Networks. In *Proceedings of the Workshop on Vector Space Modeling for Natural Language Processing*. 62–69.
- [53] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cimbis. 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1358–1368.
- [54] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. 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*. 3718–3727.
- [55] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114.
- [56] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [57] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).
- [58] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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*. 2236–2246.
- [59] Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and S Yu Philip. 2019. Joint Slot Filling and Intent Detection via Capsule Neural Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5259–5267.
- [60] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A Multimodal Dataset for Metaphor Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. 3214–3225.
- [61] Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021. TEXTOIR: An Integrated and Visualized Platform for Text Open Intent Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. 167–174.
- [62] Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021. Deep Open Intent Classification with Adaptive Decision Boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14374–14382.
- [63] Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021. Discovering New Intents with Deep Aligned Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14365–14373.
- [64] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. 2017. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*. 192–201.