Towards Multi-modal Sarcasm Detection via Disentangled Multi-grained Multi-modal Distilling

Zhihong Zhu[1,†] **Xuxin Cheng**[1,†] **Guimin Hu**[2,†]
Yaowei Li[1] **Zhiqi Huang**[1] **Yuexian Zou**[1]
[1]Peking University [2]University of Copenhagen
{zhihongzhu, chengxx, ywl}@stu.pku.edu.cn rice.hu.x@gmail.com {zhiqihuang, zouyx}@pku.edu.cn

Abstract

Multi-modal sarcasm detection aims to identify whether a given sample with multi-modal information (*i.e.*, text and image) is sarcastic, which has received increasing attention due to the rapid growth of multi-modal posts on social media. Existing mainstream methods (1) process the input of each modality holistically, resulting in redundant and unrefined information; (2) entangle different modalities to perform complex cross-modal interactions, neglecting the heterogeneity and distribution gap between them. To address these issues, we propose a new framework dubbed DMMD (Disentangled Multi-grained Multi-modal Distilling) for multi-modal sarcasm detection, which conducts multi-grained knowledge distilling (intra- and inter-subspace) based on disentangled multi-modal representations. Concretely, the representations of each modality are first disentangled explicitly into modality-agnostic/specific subspaces. Then we transfer cross-modal knowledge by conducting intra-subspace knowledge distilling in a self-adaptive pattern. Based on this, we apply mutual learning to transfer the underlying inter-subspace knowledge. Extensive experiments demonstrate the effectiveness of our DMMD over state-of-the-art baselines. More encouragingly, visualization results indicate the multi-modal representations display meaningful distributional patterns.

Keywords: Multi-modal Sarcasm Detection, Disentangled Representation Learning, Multi-modal Knowledge Distilling

1. Introduction

Sarcasm constitutes a distinctive mode of affective expression that permits individuals to convey a sentiment or intention, which is typically incongruous with their authentic or overtly expressed emotional state (Dews and Winner, 1995; Gibbs, 2007). Due to the pervasiveness of sarcastic utterances on contemporary social media platforms such as *X* and *Reddit*, sarcasm detection has received considerable critical attention. Early sarcasm detection methods solely focus on the textual modality and the intra-modal incongruity (Poria et al., 2016; Zhang et al., 2016; Felbo et al., 2017; Xiong et al., 2019). With the rapid expansion of social media platforms, multi-modal messages have become ubiquitous (Lu et al., 2019; Liu et al., 2021; Sun et al., 2022). As shown in Figure 1(a), there is a text conveying positive sentiment (*gorgeous day*). Whereas, the image accompanying the post portrays a *rainstorm*, which counteracts the positive sentiment expressed by the text. As another example shown in Figure 1(b), there are words "*beautiful*" and "*trees*" in the text, which correspond to the image depiction. In this context, research on sarcasm detection has shifted from text-only modality to multi-modal information, whose core is to draw intricate sentiment connections across modalities for detecting sarcastic clues. Following previous works, this paper focuses on multi-modal sarcasm detection containing textual and visual data.

(a) what a gorgeous day \# summer \# weather

(b) the trees are so beautiful i shed a tear

Figure 1: Two examples of multi-modal sarcasm detection from Cai et al. (2019). (a) is a sarcastic example, while (b) is a non-sarcastic example.

Inspired by the successful pre-trained models like ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021) in CV and BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) in NLP, several related research efforts have been conducted to learn representations of each modality in a divide-and-conquer manner. Thereafter, a bunch of sophisticated modules (*e.g.*, graph neural networks (Liang et al., 2021, 2022a), attention mechanism (Pan et al., 2020)) have been proposed to learn correlations between elements across modalities.

Despite significant progress existing methods achieved, most of them treat the representation of each modality as a whole, ignoring the fact that not all elements in the image contribute to mining sarcastic clues (*e.g.*, the wind chimes in Figure 1(a)), and vice versa. Motivated by this, Liang et al. (2022a) explored to extract object-level fea-

[†]Equal contribution.
[*]Corresponding author.

LREC-COLING 2024, pages 16581–16591

tures from the image as visual input, which clarifies refined interactions among modalities and moves a step further. However, it heavily relies on an external object detection model, and there is still information redundancy in the input of textual modality.

Besides, the representations of multi-modalities are typically entangled in one common latent space to perform cross-modal interactions subsequently, where modality heterogeneity and distribution gap are neglected. Intuitively, different modalities contain different ways of conveying sarcastic information, *i.e.*, the language modality consists of limited texts and has more abstract and fruitful semantics than nonverbal behaviors (Zhuang et al., 2024). As such, it is not appropriate to directly divide-and-conquer processing visual and textual information.

Following these premises, we propose a new framework termed DMMD (short for Disentangled Multi-grained Multi-modal Distilling) for multi-modal sarcasm detection. Concretely, we first adopt a common encoder and two private encoders to disentangle the features of each modality into modality-agnostic/-specific subspace. To guarantee consistency for modality-agnostic representations and diversity for modality-specific representations, we introduce two subspace constraints to consolidate the feature disentangling. Based on the disentangled multi-modal representations, we then conduct multi-grained knowledge distilling to obtain refined representations for detecting sarcastic clues.

To conduct intra-subspace knowledge distilling, we construct a multi-modal distillation graph consisting of textual and visual modalities for each subspace. Beyond previous work (Gupta et al., 2016), our graph distillation could adaptively capture the direction and weight of knowledge transfer, which allows cross-modal knowledge transfer to be performed more flexibly and efficiently. To conduct inter-subspace knowledge distilling, we introduce mutual learning (Qiao et al., 2023; Cheng et al., 2023a) to effectively utilize the subspace representations concatenated by the modality-agnostic/-specific representation of each modality to jointly conduct sarcasm detection. The two subspace representations are conducted knowledge transferring for each other, which could share certain consistency regarding the detection results.

Overall, the main contributions of this work are:

- We model multi-modal sarcasm detection based on feature disentanglement. We perform multi-grained knowledge distilling based on disentangled multi-modal representations.
- For intra-subspace knowledge distilling, we tailored modality-agnostic/-specific graph distilling in different subspaces. Within both graphs, the distillation directions and weights can be learned automatically.
- For inter-subspace knowledge distilling, we consider the intrinsic consistency between the two subspaces and adopt mutual learning to encourage distinct subspace representations to learn from each other.
- Extensive experiments demonstrate the effectiveness of our proposed model over state-of-the-art (SOTA) methods. Further analyses show the superiority of our model.

2. Related Work

2.1. Multi-Modal Sarcasm Detection

Multi-modal sarcasm detection aims to identify the sarcastic expression among different modalities (Castro et al., 2019). In particular, detecting sarcasm for both text and image modalities has increased research attention. Schifanella et al. (2016) first used both textual and visual information to tackle multi-modal sarcasm detection task. Cai et al. (2019) created a multi-modal sarcasm detection dataset based on *X* and proposed a hierarchical fusion model for the task. Thereafter, Xu et al. (2020) and Pan et al. (2020) captured both intra-modality and inter-modality incongruities based on their proposed model, respectively. Liang et al. (2021) and Liang et al. (2022a) built cross-modal graph models for drawing incongruous relations across modalities. Most recently, Liu et al. (2022) was concerned about the inconsistency of textual and visual modalities and adopted hierarchical congruity modeling in representations of multi-modalities.

However, they still tend to project multiple modalities into a common latent space and learn the hybrid representations in a holistic manner, which neglects the inherent heterogeneity and information redundancy across modalities. Besides, some research has focused on exploring the characteristics and commonalities of multi-modal representations through feature disentangling to obtain effective representations, leading to promising results in several areas (Yang et al., 2022b; Hu et al., 2024; Wang et al., 2023; Hu et al., 2022). This work fundamentally differs from them as our proposed model can transfer effective cross-modal knowledge within and between the disentangled feature subspaces.

2.2. Knowledge Distillation

The concept of knowledge distillation (KD) was first proposed by Hinton et al. (2015). KD defines a learning manner where a bigger teacher network is employed to guide the training of a smaller student network for many tasks (Li et al., 2017, 2021). KD methods primarily concentrate on transferring knowledge from teachers to students (Wang et al.,

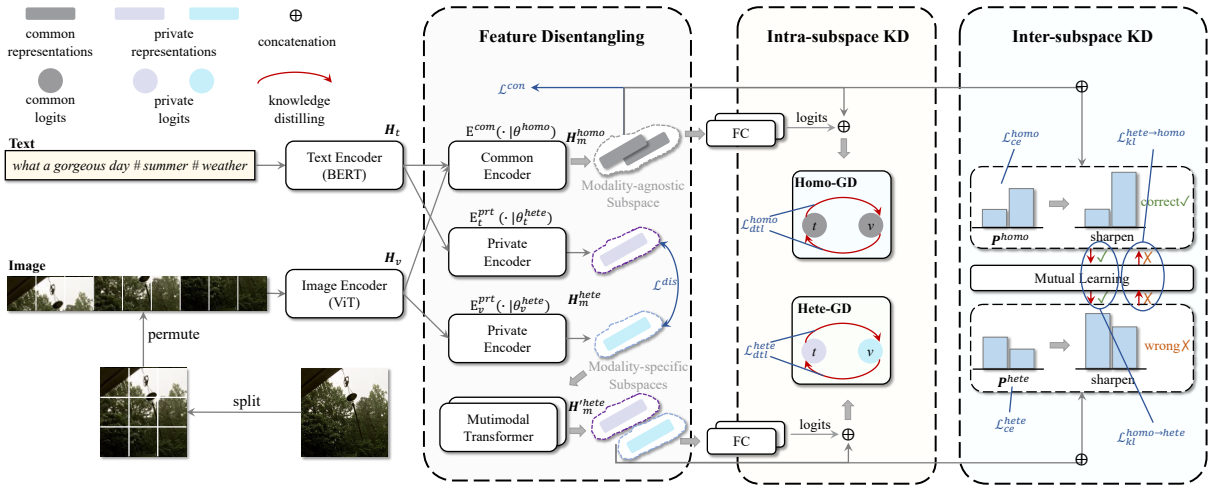


Figure 2: The overall architecture of our proposed Disentangled Multi-grained Multi-modal Distilling (DMMD) framework.

2021), while some recent studies have explored the potential of graph structures to facilitate efficient message-passing mechanisms between multiple teachers and students, thereby enabling the transfer of multiple instances of knowledge (Luo et al., 2018; Zhang and Peng, 2018; Ma et al., 2022; Zou et al., 2023; Cheng et al., 2023b). Unlike existing approaches, our objective is to utilize exclusive graph distilling components within the disentangled feature spaces in a self-adaptive manner, which is denoted as intra-subspace knowledge distilling.

Zhang et al. (2018) extended knowledge distillation and designed mutual learning, which has been proposed to leverage information from multiple models and enable effective knowledge transfer in image processing. Wen et al. (2021) resorted to mutual learning to compose the multi-modal query to retrieve the target image. In NLP, Zhao et al. (2021) employed mutual learning for speech translation to transfer knowledge between speech translation and machine translation. In this work, we adopt mutual learning to encourage consistency between the two subspace representations concatenated by modality-agnostic/-specific representations of each modality, which can be called inter-subspace knowledge distilling.

3. Problem Formulation

We first define the problem of multi-modal sarcasm detection. Suppose that we have a set of N training samples $\mathcal{D} = \{s^i\}_{i=1}^N$, where each sample $s^i = \{\boldsymbol{X}_t^i, \boldsymbol{X}_v^i, Y^i\}$ involves three elements. Thereinto, $\boldsymbol{X}_t^i$ and $\boldsymbol{X}_v^i$ denote the sentence (textual information) and image (visual information) of the i-th sample, respectively. Y^i is the ground truth label, where $Y^i = 1$ if the i-th sample is sarcastic, and $Y^i = 0$ otherwise. In a sense, we aim to devise a novel multi-modal sarcasm detection model $f(\cdot)$ which can precisely identify whether a given sentence and its attached image deliver the sarcasm,

$$f(\boldsymbol{X}_t^i, \boldsymbol{X}_v^i \,|\, \Theta) \rightarrow \hat{Y}^i, \tag{1}$$

where Θ denotes the parameters of $f(\cdot)$, $\hat{Y}^i$ is the binary classification prediction result of the model $f(\cdot)$. In the following section, we temporally omit the superscript i that indexes the training samples.

4. Approach

4.1. Feature Extraction

For a given textual sentence $\boldsymbol{X}_t = \{x_{[1,t]}, x_{[2,t]}, \cdots, x_{[L_t,t]}\}$ consisting of L_t words, we adopt the pre-trained BERT model (Devlin et al., 2019), to map each word $x_{[*,t]}$ into d-dimensional embedding[1], denoted as $\boldsymbol{H}_t \in \mathbb{R}^{L_t \times d}$. For a given image $\boldsymbol{X}_v \in \mathbb{R}^{L_h \times L_w}$, following Xu et al. (2020); Liang et al. (2021); Liu et al. (2022), we first resize it to 224×224 pixels, *i.e.*, $L = L_h = L_w = 224$. Then the image is divided into $r = p \times p$ patches, w.r.t. $\boldsymbol{X}_v \in \mathbb{R}^{r \times (L/p \times L/p)}$. Next, we feed the sequence of r image patches into a pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2021) with an MLP layer subsequently to acquire the visual representation $\boldsymbol{H}_v \in \mathbb{R}^{L_v \times d}$.

4.2. Feature Disentangling

4.2.1. Existing Solution

Since existing representations in a divide-and-conquer manner could introduce information redundancy and distribution gap (Liang et al., 2022b;

[1]Following previous works Liang et al. (2021, 2022a), we adopt the first sub-token's representation as the whole word representation.

Yang et al., 2022a) across modalities, we propose to embed initial representations of each modality into modality-agnostic and modality-specific subspaces, which has achieved success in several cross-modal tasks (Yang et al., 2022b,a; Wang et al., 2023; Xu et al., 2022), and we utilize it as the foundation for subsequent distilling. Formally, we utilize a common encoder $\mathrm{E}^{com}(\cdot \mid \theta^{homo})$ and two private encoders $\mathrm{E}_m^{prt}(\cdot \mid \theta_m^{hete})$ $(m \in \{t, v\})$ to obtain the disentangled homogeneous and heterogeneous representations,

$$\boldsymbol{H}_m^{homo} = \mathrm{E}^{com}(\boldsymbol{H}_m \mid \theta^{homo}), \tag{2}$$

$$\boldsymbol{H}_m^{hete} = \mathrm{E}_m^{prt}(\boldsymbol{H}_m \mid \theta_m^{hete}). \tag{3}$$

The common encoder $\mathrm{E}^{com}(\cdot \mid \theta^{homo})$ shares the parameters θ^{homo} across all modalities, and the private encoders $\mathrm{E}_m^{prt}(\cdot \mid \theta_m^{hete})$ learn the parameters θ_m^{hete} for each modality.

4.2.2. Subspace Constraint

Despite performing the aforementioned process, feature disentangling cannot be thoroughly guaranteed. There exists the potential for information to freely permeate between feature representations, whereby all modality information may be solely encoded in $\boldsymbol{H}_m^{hete}$, which renders homogeneous (modality-agnostic) multi-modal features meaningless. Inspired by Li et al. (2023), we introduce a consistency constraint in the modality-agnostic subspace to strengthen the commonality across modalities, which is formulated as follows,

$$\mathcal{L}^{con} = \frac{1}{|S|} \sum_{(i,j,k) \in S} \max(0, \alpha - \cos(\boldsymbol{H}_{m[i]}^{homo}, \boldsymbol{H}_{m[j]}^{homo}) + \cos(\boldsymbol{H}_{m[i]}^{homo}, \boldsymbol{H}_{m[k]}^{homo})), \tag{4}$$

where the triple tuple set[2] $S = \{(i, j, k) \mid m[i] \neq m[j], m[i] = m[k], Y^i = Y^j, Y^i \neq Y^k\}$. α is the distance margin, $m[*]$ denotes the modality of sample $*$, and $\cos(\cdot, \cdot)$ refers to the cosine similarity. The margin α enforces the distances between positive samples (*same* label; *different* modalities) to be smaller than those between negative ones (*same* modality; *different* labels).

To ensure the modality-specific representations capture different aspects of multi-modal data and reduce information redundancy across different modalities, we further introduce a disparity constraint in each modality-specific subspace,

$$\mathcal{L}^{dis} = \sum_{m \in \{t, v\}} \cos(\boldsymbol{H}_m^{homo}, \boldsymbol{H}_m^{hete}). \tag{5}$$

Thus, the formulated soft orthogonality constraint in Equation 5 can reduce information redundancy between modality-agnostic and modality-specific representations.

[2]Note that the triple tuple set is formed within each batch.

4.3. Intra-subspace Knowledge Distilling

4.3.1. Homogeneous Graph Distilling

For disentangled homogeneous representations, we first construct a homogeneous graph $\mathcal{G}^{homo}$, whose node is denoted as v_i $(i \in \{t, v\})$ w.r.t a modality. Without loss of generality, the edge $\boldsymbol{e}_{t \to v}^{homo}$ in graph $\mathcal{G}^{homo}$ denotes the difference between corresponding logits, which is represented for the distillation from v_t to v_v, and vice versa. Denote $\boldsymbol{W}^{homo}$ and $\boldsymbol{E}^{homo}$ as the edge weights matrix and distillation matrix of $\mathcal{G}^{homo}$, respectively. The weighted distillation loss can be constructed as,

$$\mathcal{L}_{dtl}^{t,homo} = \boldsymbol{w}_{v \to t}^{homo} \times \boldsymbol{e}_{v \to t}^{homo}, \tag{6}$$

$$\mathcal{L}_{dtl}^{v,homo} = \boldsymbol{w}_{t \to v}^{homo} \times \boldsymbol{e}_{t \to v}^{homo}, \tag{7}$$

where $\boldsymbol{w}_{v \to t}^{homo}$ and $\boldsymbol{w}_{t \to v}^{homo}$ refer to the distillation strength from v_v to v_t and v_t to v_v, respectively.

To learn a self-adaptive weight that corresponds to the distillation strength $\boldsymbol{w}$, we propose to encode the modality logits and the representations into the graph edges. The process is formulated as follows,

$$\boldsymbol{w}_{v \to t}^{homo} = f_2((f_1(\boldsymbol{H}_v^{homo}) \oplus \boldsymbol{H}_v^{homo}) \oplus (f_1(\boldsymbol{H}_t^{homo}) \oplus \boldsymbol{H}_t^{homo})), \tag{8}$$

$$\boldsymbol{w}_{t \to v}^{homo} = f_2((f_1(\boldsymbol{H}_t^{homo}) \oplus \boldsymbol{H}_t^{homo}) \oplus (f_1(\boldsymbol{H}_v^{homo}) \oplus \boldsymbol{H}_v^{homo})), \tag{9}$$

where $\oplus$ denotes feature concatenation, $f_1(\cdot)$ is a fully connected (FC) layer for regressing logits, and $f_2(\cdot)$ is an FC layer for concatenating logits. Consequently, the graph distillation loss of two modalities can be now normalized as,

$$\mathcal{L}_{dtl}^{homo} = \left\| \boldsymbol{W}^{homo} \odot \boldsymbol{E}^{homo} \right\|_1, \tag{10}$$

where $\odot$ means element-wise product and $\|\cdot\|_1$ denotes ℓ_1-norm.

4.3.2. Heterogeneous Graph Distilling

To bridge the distribution gap between disentangled heterogeneous feature representations $\boldsymbol{H}_m^{hete}$, we utilize the multi-modal transformer (Tsai et al., 2019) to perform the modality adaptation. The core of the multi-modal transformer lies in its cross-modal attention (CA) module, which takes in features from two modalities and integrates cross-modal information. Take textual modality $\boldsymbol{H}_t^{hete}$ as the source and visual modality $\boldsymbol{H}_v^{hete}$ as the target, the cross-modal attention can be defined as $\boldsymbol{Q}_v = \boldsymbol{X}_v^{hete} \boldsymbol{M}_q$, $\boldsymbol{K}_t = \boldsymbol{X}_t^{hete} \boldsymbol{M}_k$, and $\boldsymbol{V}_t = \boldsymbol{X}_t^{hete} \boldsymbol{M}_v$ where $\boldsymbol{M}_q$, $\boldsymbol{M}_k$ and $\boldsymbol{M}_v$ are learnable parameters. The individual head of CA can be expressed as:

$$\boldsymbol{H}_{t \to v}^{\prime hete} = \mathrm{softmax}(\frac{\boldsymbol{Q}_v \boldsymbol{K}_t^\top}{\sqrt{d}}) \boldsymbol{V}_t, \tag{11}$$

where $\boldsymbol{H}_{t \to v}^{\prime hete}$ is the enhanced features from textual information to visual information, and d represents the dimension of $\boldsymbol{Q}_v$ and $\boldsymbol{K}_t$. $\boldsymbol{H}_{v \to t}^{\prime hete}$ can be derived like Equation 11, and the distillation loss

function $\mathcal{L}_{dtl}^{hete}$ can be similarly obtained like Equation 10. Next, for homogeneous/heterogeneous representations ($\boldsymbol{H}_t^{homo}$, $\boldsymbol{H}_v^{homo}$ / $\boldsymbol{H}_{t\to v}^{\prime hete}$, $\boldsymbol{H}_{v\to t}^{\prime hete}$), we concatenate both textual and visual modalities in each subspace to obtain two subspace representations, denoted as $\boldsymbol{Z}^{homo}$ and $\boldsymbol{Z}^{hete}$. For $\boldsymbol{Z}^{homo}$, we then feed it into FC layers to gain the predicted probability distributions as follows,

$$\boldsymbol{p}^{homo} = \text{softmax}(\boldsymbol{W}_1 \boldsymbol{Z}^{homo} + \boldsymbol{b}_1), \tag{12}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ are trainable parameters, $\boldsymbol{p}^{homo}$ is the predicted probability vector of homogeneous feature representations. Eventually, we calculate the cross-entropy loss to supervise homogeneous feature representations as follows,

$$\mathcal{L}_{ce}^{homo} = Y^i \log \boldsymbol{p}^{i,homo} + (1 - Y^i) \log(1 - \boldsymbol{p}^{i,homo}), \tag{13}$$

where Y^i and $\boldsymbol{p}^{i,homo}$ are the i-th elements of the ground truth Y and $\boldsymbol{p}^{homo}$, respectively.

For $\boldsymbol{Z}^{hete}$, $\boldsymbol{p}^{hete}$ and $\mathcal{L}_{ce}^{hete}$ can be derived like Equation (12) and (13), respectively.

4.4. Inter-subspace Knowledge Distilling

As both subspace representations aim to capture incongruent sarcastic information across modalities, there should be certain intrinsic consistency between the two representations. In view of this, we make the two subspace representations share knowledge with each other by adopting mutual learning (Zhang et al., 2018; Nie et al., 2018; Hong et al., 2021) following Qiao et al. (2023). Specifically, we employ the Kullback Leibler (KL) (Kullback and Leibler, 1951) Divergence between $\boldsymbol{p}^{homo}$ and $\boldsymbol{p}^{hete}$, which can measure the differences between two distributions to encourage consistency between the two subspace representations $\boldsymbol{Z}^{homo}$ and $\boldsymbol{Z}^{hete}$. To avoid incorrect knowledge being transferred, we only transfer reliable knowledge by introducing a temperature parameter controlling whether to transfer the prediction result of this sample. Formally, the objective function for inter-subspace knowledge transferring can be formulated as follows,

$$\mathcal{L}_{kl}^{homo\to hete} = \delta_{homo}\, \text{KL}(\boldsymbol{p}^{homo} || \boldsymbol{p}^{hete}), \tag{14}$$

$$\mathcal{L}_{kl}^{hete\to homo} = \delta_{hete}\, \text{KL}(\boldsymbol{p}^{hete} || \boldsymbol{p}^{homo}), \tag{15}$$

where $homo \to hete$ denotes the knowledge transferring from the modality-agnostic to modality-specific subspace, and vice versa. Following Hinton et al. (2015), we sharpen the predicted distribution with a temperature parameter for knowledge transfer. In order to avoid incorrect knowledge transferring, we define a sample screening mechanism using control parameters δ_h:

$$\delta_h = \begin{cases} 1, & if\ \text{argmax}(\boldsymbol{p}^h) = Y, \\ 0, & otherwise, \end{cases} \tag{16}$$

in which argmax denotes the operation that gains the predicted labels from the predicted result $\boldsymbol{p}^h, h \in \{homo, hete\}$, and $\boldsymbol{p}^h$ is the predicted probability distribution of the modality-agnostic/-specific subspace representations which shares the knowledge (*i.e.*, $\boldsymbol{p}^{homo}$ for $\mathcal{L}_{kl}^{homo\to hete}$ and $\boldsymbol{p}^{hete}$ for $\mathcal{L}_{kl}^{hete\to homo}$).

	Training	Validating	Testing
All	19816	2410	2409
Positive	8642	959	959
Negative	11174	1451	1450

Table 1: Statistics of the experimental data.

4.5. Training Objective

Towards the optimization of the whole model, we combine all loss functions as follows,

$$\mathcal{L}^{homo} = \mathcal{L}_{ce}^{homo} + \lambda_1 \mathcal{L}_{dtl}^{homo} + \lambda_2 \mathcal{L}_{kl}^{hete\to homo}, \tag{17}$$

$$\mathcal{L}^{hete} = \mathcal{L}_{ce}^{hete} + \lambda_3 \mathcal{L}_{dtl}^{hete} + \lambda_4 \mathcal{L}_{kl}^{homo\to hete}, \tag{18}$$

$$\mathcal{L}^{all} = \mathcal{L}^{homo} + \mathcal{L}^{hete} + \lambda_5 (\mathcal{L}^{con} + \mathcal{L}^{dis}), \tag{19}$$

in which λ_* are non-negative hyper-parameters. $\mathcal{L}_{ce}^{(\cdot)}$ denote the loss function of the sarcasm detection task, $\mathcal{L}_{dtl}^{(\cdot)}$ and $\mathcal{L}_{kl}^{(\cdot)}$ denote the loss function of intra-subspace knowledge distilling and inter-subspace knowledge distilling, respectively.

Eventually, the binary classification prediction result $\hat{Y}$ is defined as follows,

$$\hat{Y} = \text{argmax}(\frac{\boldsymbol{p}^{homo} + \boldsymbol{p}^{hete}}{2}). \tag{20}$$

5. Experiments

5.1. Experimental Setup

5.1.1. Dataset

Following previous works, we evaluated our model on a publicly available multi-modal sarcasm detection benchmark dataset (Cai et al., 2019). Thereinto, tweets with some special hashtags (*e.g.* sarcasm) are positive examples and those without such hashtags are negative examples.

5.1.2. Metrics

Following Cai et al. (2019), we perform Accuracy, Precision, Recall and F1-score metrics to evaluate the performance of models.[3] Since the label distribution of the dataset is imbalanced, following Pan et al. (2020); Liang et al. (2021), we also report macro-average results.

5.1.3. Baselines

We compare our DMMD with a series of baselines, summarized as follows: 1) *Image*-Modality methods: These models use only visual information for sarcasm detection, including Image (Cai

[3]We implement the metrics by using sklearn.metrics.

Modality	Method	Accuracy (%)	F1-score			Macro-average		
			Precision(%)	Recall(%)	F1-score(%)	Precision(%)	Recall(%)	F1-score(%)
Image	Image (Cai et al., 2019)	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT (Dosovitskiy et al., 2021)	67.83	57.93	70.07	63.43	65.68	71.35	68.40
Text	TextCNN (Kim, 2014)	80.03	74.29	76.39	75.32	78.03	78.28	78.15
	Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	SIARN (Tay et al., 2018)	80.57	75.55	75.70	75.63	80.34	78.81	79.57
	SMSD (Xiong et al., 2019)	80.90	76.46	75.18	75.82	80.87	78.20	79.51
	BERT (Devlin et al., 2019)	83.85	78.72	82.27	80.22	81.31	80.87	81.09
Image+Text	HFM (Cai et al., 2019)	83.44	76.47	84.15	80.18	79.40	82.45	80.90
	D&R Net (Xu et al., 2020)	84.02	77.97	83.42	80.60	-	-	-
	Res-BERT (Pan et al., 2020)	84.80	77.80	84.15	80.85	78.87	84.46	81.57
	Att-BERT (Pan et al., 2020)	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	InCrossMGs (Liang et al., 2021)	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	CMGCN (Liang et al., 2022a)	87.55	83.63	84.69	84.16	87.02	86.97	87.00
	Liu et al. (Liu et al., 2022)	87.36	81.84	86.48	84.09	-	-	-
	MILNet (Qiao et al., 2023)	89.50	85.16	89.16	87.11	88.88	89.44	89.12
	DMMD (Ours)	**90.60**	**86.95**	**91.04**	**88.93**	**90.67**	**91.31**	**90.94**

Table 2: Comparison results between our DMMD and previous SOTA methods.

et al., 2019) and ViT (Dosovitskiy et al., 2021). 2) *Text*-Modality methods: These models use only textual information for sarcasm detection, including TextCNN (Kim, 2014); BiLSTM; SIARN (Tay et al., 2018); SMSD (Xiong et al., 2019) and BERT (Devlin et al., 2019). 3) *Multi*-Modal methods: These models take both text- and image-modality information as input for multi-modal sarcasm detection, including HFM (Cai et al., 2019); D&R Net (Xu et al., 2020); Res-BERT (Pan et al., 2020); Att-BERT (Pan et al., 2020); InCrossMGs (Liang et al., 2021); CMGCN (Liang et al., 2022a); Liu et al. (Liu et al., 2022) and MILNet (Qiao et al., 2023).

5.1.4. Settings

Following data pre-processing in Cai et al. (2019); Liang et al. (2021); Xu et al. (2020), we remove samples containing words that frequently co-occur with sarcastic utterances (*e.g.*, *sarcasm*, *sarcastic*, *irony* and *ironic*) to avoid introducing external information. We utilize pre-trained uncased BERT-base model [4] to embed each word of text-modality as a 768-dimensional embedding and employ the pre-trained ViT [5] to embed each image patch as a 768-dimensional embedding, *i.e.*, d in Section 4.1 is 768. For image pre-processing, we resize the image to 224×224 and divide it into 32×32 patches [6] (*i.e.*, $p = 7$, $r = 49$). $\lambda_{\{1-5\}}$ in Equation 17, 18 and 19 are searching in $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ via the best performance on the validation set. We use Adam optimizer (Kingma and Ba, 2015) with a learning rate of $2e$-5, weight decay of $5e$-3, batch size as 32 and dropout rate as 0.4 to train our DMMD. All experiments are conducted at an RTX 3090 GPU with 24GB memory. The experimental results of our proposed models are obtained by averaging 5 runs with random initialization.

5.2. Comparisons with State-of-the-art Methods

The comparison results are reported in Table 2. From the results, we can draw the following conclusions. **(1)** DMMD consistently outperforms all baselines across all evaluation metrics, which denotes that DMMD can significantly improve the performance over state-of-the-art methods, justifying the effectiveness of the proposed disentangled multi-grained multi-modal distilling framework. **(2)** DMMD surpasses Liu et al. (feature representations modeled hierarchically) and CMGCN (feature representations interact across modalities based on graph structure). This implies our feature disentangling approach is a step further than the conventional divide-and-conquer approach. **(3)** Methods based on text modality consistently outperform those based on image modality, which indicates that the expression of sarcastic/non-sarcastic information predominantly resides in the text modality. **(4)** Methods that leverage both image and text modalities exhibit superior performance to unimodal baselines, which demonstrates that utilizing information from both modalities is more effective for sarcasm detection. **(5)** The macro-average results outperform other commonly used metrics overall, which demonstrates that models perform better in the "*negative*" class due to the imbalanced class distribution. **(6)** To justify whether the improvement is statistically significant, we conducted t-tests between our results and the second best results and found that all the $p < 0.05$. This validates the su-

[4] https://huggingface.co/bert-base-uncased

[5] https://github.com/lukemelas/PyTorch-Pretrained-ViT

[6] We also tested other division resolutions and found that performance fluctuations were negligible across different image patch resolutions.

Section	Setting	FD	Intra-subspace KD			Inter-subspace KD		Dataset: Cai et al.		
			Homo-GD	CA	Hete-GD	ML	SS	Accuracy(%)	F1-score(%)	Marco-F1(%)
4.2	Base							85.25 (↓5.35)	83.59 (↓5.34)	84.26 (↓6.68)
	(a)	✓						86.28 (↓4.32)	84.46 (↓4.47)	86.58 (↓4.36)
4.3	(b)	✓	✓					87.40 (↓3.20)	85.61 (↓3.32)	87.95 (↓2.99)
	(c)	✓	✓		✓			87.84 (↓2.76)	86.75 (↓2.18)	88.87 (↓2.07)
	(d)	✓	✓	✓				88.25 (↓2.35)	87.12 (↓1.81)	89.30 (↓1.64)
	(e)	✓	✓	✓	✓			89.71 (↓0.89)	88.10 (↓0.83)	90.12 (↓0.82)
4.4	(f)	✓	✓	✓	✓	✓		90.26 (↓0.34)	88.58 (↓0.35)	90.62 (↓0.32)
	DMMD	✓	✓	✓	✓	✓	✓	**90.60** (-)	**88.93** (-)	**90.94** (-)

Table 3: Results of ablation study. FD: feature disentangling. Homo-GD: homogeneous graph distilling. CA: the cross-modal attention module in the multi-modal transformer. Hete-GD: heterogeneous graph distilling. ML: Mutual Learning. SS: Sample Screening. Base: conduct sarcasm detection using only the backbone models (*i.e.*, BERT and ViT).

periority of DMMD over existing methods.

5.3. Quantitative Analysis

We conduct quantitative analysis to investigate the contribution of each component of our DMMD and the results are reported in Table 3.

5.3.1. Effect of Feature Disentangling

Setting (a) in Table 3 shows that FD can successfully boost baselines with gains up to 1.03%, 0.87% and 2.32% in terms of Accuracy, F1-score and Macro-F1 respectively, demonstrating the disentangled features can reduce information redundancy and provide discriminative multi-modal features.

5.3.2. Effect of Intra-subspace Knowledge Distilling

Setting (b) shows that combining FD with Homo-GD can further improve model performance. Although the homogeneous representations of each modality are projected into the same subspace, variations in discriminative capability still exist between modalities. By Homo-GD, DMMD is able to effectively enhance weak modalities. A similar observation was made with respect to Hete-GD, as demonstrated in setting (c). However, performing Hete-GD without CA leads to degraded performance, indicating that the multimodal transformer plays a crucial role in bridging the gap between multi-modal distributions (*cf.* setting (d)). By combining CA and Hete-GD, DMMD obtains conspicuous improvements as shown in setting (e), demonstrating the importance of further exploiting heterogeneous representations of each modality for multi-modal sarcasm detection.

5.3.3. Effect of Inter-subspace Knowledge Distilling

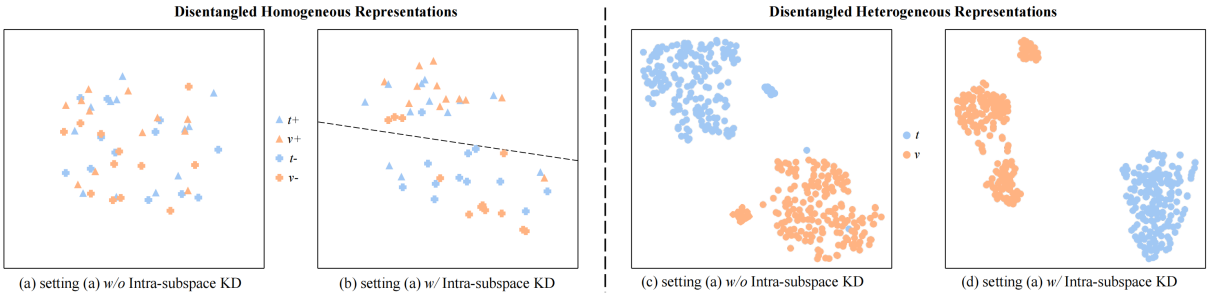
Setting (f) shows that the introduction of mutual learning can significantly boost performance. We attribute this to the fact that our DMMD enables knowledge sharing between modality-agnostic and modality-specific subspace, allowing it to leverage the underlying consistency of two subspace representations for mining sarcastic clues. Moreover, by comparing the results of setting (f) and DMMD, we conclude that choosing the right knowledge to transfer is necessary. Since Intra-subspace Knowledge Distilling and Inter-subspace Knowledge Distilling can both boost performance, combining them can lead to the most prominent improvement across all metrics (*cf.* setting DMMD), with up to 5.01%, 4.99% and 6.36% in terms of Accuracy, F1-score and Macro-F1, respectively.

Method	Dataset: Cai et al.		
	Accuracy(%)	F1-score(%)	Marco-F1(%)
MulT (*w/o* KD)	87.78 (↓2.82)	86.49 (↓2.44)	87.76 (↓3.18)
MulT (*w/* KD)	88.29 (↓2.31)	87.37 (↓1.56)	88.83 (↓2.11)
DMMD (Ours)	**90.60** (-)	**88.93** (-)	**90.94** (-)

Table 4: Architecture analysis on the MulT (multi-modal transformer) (Tsai et al., 2019) *w/o* and *w/* KD and DMMD.

5.3.4. Rationality Analysis

We compare our proposed DMMD with the `MulT` (multi-modal transformer) (Tsai et al., 2019) to further verify the rationality of the proposed framework. As shown in Table 4, where KD denotes we perform (inter-subspace) knowledge distilling on MulT to conduct adaptive knowledge transfer with the multimodal features. The core differences between MulT (*w/* KD) and DMMD are: 1) DMMD conducts feature disentangling, and 2) DMMD performs multi-grained knowledge distilling. We can observe that our DMMD gains consistent improvements than both MulT (*w/o* and *w/* KD) across all metrics, which demonstrates the rationality and feasibility of combing the feature disentangling and the multi-grained knowledge distilling.



(a) setting (a) *w/o* Intra-subspace KD (b) setting (a) *w/* Intra-subspace KD (c) setting (a) *w/o* Intra-subspace KD (d) setting (a) *w/* Intra-subspace KD

Figure 3: Visualization of the disentangled homogeneous/heterogeneous representations. $m+$ and $m-$ ($m \in \{t, v\}$) represent the modality of non-sarcastic and sarcastic samples, respectively.

5.4. Qualitative Analysis

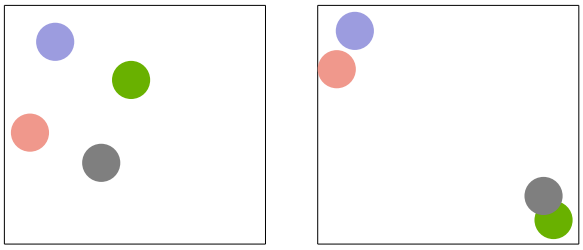
5.4.1. Visualization of Intra-subspace Representations

To thoroughly understand the proposed intra-subspace KD, we perform t-SNE (der Maaten and Hinton, 2008) to visualize homogeneous/heterogeneous representations from different subspaces. To visualize the homogeneous representations, we randomly selected 24 samples (12 samples for each label) from the testing set. To visualize the heterogeneous representations, we randomly selected 360 samples from the testing set.

From Figure 3, we can observe that: **(1)** For (a), without performing intra-subspace KD, the samples only vaguely show basic separability without explicit decision boundaries. Additionally, different modal information belonging to the same sample is entangled, making it more challenging to learn multi-modal representations. **(2)** As for (b), through performing Homo-GD in modality-agnostic subspace, our DMMD efficiently shares relevant information between modalities, leading to good discriminatory performance. **(3)** In the modality-specific subspaces, the features of different samples are expected to cluster according to their modalities because of their inter-modal heterogeneity. By comparing the results of (c) and (d), we can draw similar conclusions to those in (1) and (2) regarding the effectiveness of our approach. **(4)** The qualitative analysis further supports our motivation and verifies the effectiveness of our proposed approach in exploring intra-subspace knowledge distilling to boost the performance of sarcasm detection.

5.4.2. Visualization of Inter-subspace Representations

In Figure 4, we conduct a qualitative analysis to thoroughly understand the inter-subspace KD. We can observe that w/ inter-subspace KD, subspace representations concatenated together from different modal features in each subspace can exhibit superior binary differentiation compared to those w/o inter-subspace KD. We attribute this to the capability of our inter-subspace KD to explore properties in different subspaces and perform efficient migration of knowledge to improve performance.



(a) setting (e) *w/o* Inter-subspace KD (b) setting (e) *w/* Inter-subspace KD

Figure 4: Visualization of the subspace representations. We visualize the representations by reducing the dimension with Principal Component Analysis (PCA) (Abdi and Williams, 2010). The different colours represent different samples.

6. Conclusion

In this paper, we propose the Disentangled Multi-grained Multi-modal Distilling (DMMD) framework for multi-modal sarcasm detection, which performs intra-subspace and inter-subspace knowledge distilling based on disentangled multi-modal representations. We conducted extensive experiments on a publicly available benchmark, which demonstrated the superiority of our proposed framework.

Future work can further explore interpretable multi-modal sarcasm detection and exploit LLMs (Chen et al., 2024b,a; Xu et al., 2024) to boost performance.

7. Bibliographical References

H. Abdi and L. Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*.

S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proc. of ACL*.

Z. Chen, Z. Zhao, H. Luo, H. Yao, B. Li, and J. Zhou. 2024a. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Z. Chen, Z. Zhao, Z. Zhu, R. Zhang, X. Li, B. Raj, and H. Yao. 2024b. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *arXiv preprint arXiv:2402.11452*.

X. Cheng, B. Cao, Q. Ye, Z. Zhu, H. Li, and Y. Zou. 2023a. ML-LMCL: Mutual learning and large-margin contrastive learning for improving ASR robustness in spoken language understanding. In *Proc. of ACL Findings*.

X. Cheng, Z. Zhu, W. Xu, Y. Li, H. Li, and Y. Zou. 2023b. Accelerating multiple intent detection and slot filling via targeted knowledge distillation. In *Proc. of EMNLP Findings*.

L. Van der Maaten and G. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of AACL*.

S. Dews and E. Winner. 1995. Muting the meaning a social function of irony. *Metaphor and Symbol*.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of ICLR*.

B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proc. of EMNLP*.

R. Gibbs. 2007. On the psycholinguistics of sarcasm. *Irony in language and thougt: A cognitive science reader*.

S. Gupta, J. Hoffman, and J. Malik. 2016. Cross modal distillation for supervision transfer. In *Proc. of CVPR*.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*.

G. Hinton, O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. *CoRR*.

P. Hong, T. Wu, A. Wu, X. Han, and W. Zheng. 2021. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proc. of CVPR*.

G. Hu, T. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proc. of EMNLP*.

G. Hu, Z. Zhu, D. Hershcovich, H. Seifi, and J. Xie. 2024. Unimeec: Towards unified multimodal emotion recognition and emotion cause. *arXiv preprint arXiv:2404.00403*.

Y. Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. of EMNLP*.

D. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

S. Kullback and R. Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*.

Q. Li, S. Jin, and J. Yan. 2017. Mimicking very efficient network for object detection. In *Proc. of CVPR*.

Y. Li, Y. Wang, and Z. Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *Proc. of CVPR*.

Z. Li, J. Ye, M. Song, Y. Huang, and Z. Pan. 2021. Online knowledge distillation for efficient pose estimation. In *Proc. of ICCV*.

B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*.

B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, and R. Xu. 2022a. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proc. of ACL*.

V. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou. 2022b. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Proc. of NeurIPS*.

H. Liu, W. Wang, and H. Li. 2022. Towards multimodal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proc. of EMNLP*.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*.

Y. Liu, Y. Zhang, Q. Li, B. Wang, and D. Song. 2021. What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *Proc. of EMNLP Findings*.

X. Lu, L. Zhu, Z. Cheng, L. Nie, and H. Zhang. 2019. Online multi-modal hashing with dynamic query-adaption. In *Proc. of SIGIR*.

Z. Luo, J. Hsieh, L. Jiang, J. Niebles, and L. Fei-Fei. 2018. Graph distillation for action detection with privileged modalities. In *Proc. of ECCV*.

R. Ma, G. Pang, L. Chen, and A. Hengel. 2022. Deep graph-level anomaly detection by glocal knowledge distillation. In *Proc. of WSDM*.

X. Nie, J. Feng, and S. Yan. 2018. Mutual learning to adapt for joint human parsing and pose estimation. In *Proc. of ECCV*.

H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Proc. of EMNLP Findings*.

S. Poria, E. Cambria, D. Hazarika, and P. Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proc. of COLING*.

Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, and L. Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proc. of AAAI*.

R. Schifanella, P. Juan, J. Tetreault, and L. Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*.

T. Sun, W. Wang, L. Jing, Y. Cui, X. Song, and L. Nie. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proc. of ACM MM*.

Y. Tay, A. Luu, S. Hui, and J. Su. 2018. Reasoning with sarcasm by reading in-between. In *Proc. of ACL*.

Y. Tsai, S. Bai, P. Liang, J. Kolter, L. Morency, and R. Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proc. of ACL*.

H. Wang, D. Lian, H. Tong, Q. Liu, Z. Huang, and E. Chen. 2023. Decoupled representation learning for attributed networks. *IEEE Trans. Knowl. Data Eng.*

L. Wang, J. Huang, Y. Li, K. Xu, Z. Yang, and D. Yu. 2021. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proc. of CVPR*.

H. Wen, X. Song, X. Yang, Y. Zhan, and L. Nie. 2021. Comprehensive linguistic-visual composition network for image retrieval. In *Proc. of SIGIR*.

T. Xiong, P. Zhang, H. Zhu, and Y. Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *Proc. of WWW*.

D. Xu, Z. Zhang, Z. Lin, X. Wu, Z. Zhu, T. Xu, X. Zhao, Y. Zheng, and E. Chen. 2024. Multi-perspective improvement of knowledge graph completion with large language models. *arXiv preprint arXiv:2403.01972*.

N. Xu, Z. Zeng, and W. Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proc. of ACL*.

S. Xu, Z. Zhou, and J. Shang. 2022. Asymmetric adversarial-based feature disentanglement learning for cross-database micro-expression recognition. In *Proc. of ACM MM*.

D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang. 2022a. Disentangled representation learning for multimodal emotion recognition. In *Proc. of ACM MM*.

D. Yang, H. Kuang, S. Huang, and L. Zhang. 2022b. Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In *Proc. of ACM MM*.

C. Zhang and Y. Peng. 2018. Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. In *Proc. of IJCAI*.

M. Zhang, Y. Zhang, and G. Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proc. of COLING*.

Y. Zhang, T. Xiang, T. Hospedales, and H. Lu. 2018. Deep mutual learning. In *Proc. of CVPR*.

J. Zhao, W. Luo, B. Chen, and A. Gilman. 2021. Mutual-learning improves end-to-end speech translation. In *Proc. of EMNLP*.

X. Zhuang, Z. Wang, X. Cheng, Y. Xie, L. Liang, and Y. Zou. 2024. Macsc: Towards multimodal-augmented pre-trained language models via conceptual prototypes and self-balancing calibration. In *Proc. of NAACL*.

W. Zou, X. Qi, W. Zhou, M. Sun, Z. Sun, and C. Shan. 2023. Graph flow: Cross-layer graph flow distillation for dual efficient medical image segmentation. *IEEE Trans. Medical Imaging*.

8. Language Resource References

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proc. of ACL*.