

# Multimodal Hate Speech Detection via Cross-Domain Knowledge Transfer

Chuanpeng Yang  
Institute of Information Engineering,  
Chinese Academy of Sciences &  
School of Cyber Security, University  
of Chinese Academy of Sciences  
Beijing, China  
yangchuanpeng@iie.ac.cn

Fuqing Zhu\*  
Institute of Information Engineering,  
Chinese Academy of Sciences &  
School of Cyber Security, University  
of Chinese Academy of Sciences  
Beijing, China  
zhufuqing@iie.ac.cn

Guihua Liu  
Institute of Information Engineering,  
Chinese Academy of Sciences &  
School of Cyber Security, University  
of Chinese Academy of Sciences  
Beijing, China  
liuguihua@iie.ac.cn

Jizhong Han  
Institute of Information Engineering,  
Chinese Academy of Sciences  
Beijing, China  
hanjizhong@iie.ac.cn

Songlin Hu  
Institute of Information Engineering,  
Chinese Academy of Sciences &  
School of Cyber Security, University  
of Chinese Academy of Sciences  
Beijing, China  
husonglin@iie.ac.cn

## ABSTRACT

Nowadays, the hate speech diffusion of texts and images in social network has become the mainstream compared with the diffusion of texts-only, raising the pressing needs of multimodal hate speech detection task. Current research on this task mainly focuses on the construction of multimodal models without considering the influence of the unbalanced and widely distributed samples for various attacks in hate speech. In this situation, introducing enhanced knowledge is necessary for understanding the attack category of hate speech comprehensively. Due to the high correlation between hate speech detection and sarcasm detection tasks, this paper makes an initial attempt of common knowledge transfer based on the above two tasks, where hate speech detection and sarcasm detection are defined as primary and auxiliary tasks, respectively. A scalable cross-domain knowledge transfer (CDKT) framework is proposed, where the mainstream vision-language transformer could be employed as backbone flexibly. Three modules are included, bridging the semantic, definition and domain gaps simultaneously between primary and auxiliary tasks. Specifically, semantic adaptation module formulates the irrelevant parts between image and text in primary and auxiliary tasks, and disentangles with the text representation to align the visual and word tokens. Definition adaptation module assigns different weights to the training samples of auxiliary task by measuring the correlation between samples of the auxiliary and primary task. Domain

adaptation module minimizes the feature distribution gap of samples in two tasks. Extensive experiments show that the proposed CDKT provides a stable improvement compared with baselines and produces a competitive performance compared with some existing multimodal hate speech detection methods.

## CCS CONCEPTS

• **Computing methodologies** → **Natural language processing;**  
**Computer vision representations;**

## KEYWORDS

multimodal, hate speech, knowledge transfer, vision-language

### ACM Reference Format:

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal Hate Speech Detection via Cross-Domain Knowledge Transfer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3503161.3548255>

**Disclaimer:** *This paper uses the publicly hate/sarcasm datasets for academic research only, which contain the distasteful content that may be disturbing to some readers.*

## 1 INTRODUCTION

With the social media platforms development, it is becoming universal and popular for people uploading multimodal information (e.g., text, image, video, etc.) to express attitudes or emotions to some specific events. A large amount of images or videos with the corresponding texts appear on the platform, which promotes the diffusion of hate speech [8, 41]. The hate speech attacks people directly or indirectly based on the characteristics (including race, religion, gender, etc.) via multimodal data in most cases. Nowadays, multimodal is a trend in hate speech detection and has become urgent. The diversity and interactivity of modality information make the traditional detection based on a single modality insufficient to identify hate speech. Multimodal hate speech detection task [20]

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

MM '22, October 10–14, 2022, Lisboa, Portugal.

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9203-7/22/10...\$15.00

<https://doi.org/10.1145/3503161.3548255>

	Definition Gap		Domain Gap	
<b>Hate</b>	<p><b>H1 Label: 1</b></p>	<p><b>H2 Label: 0</b></p>	<p><b>H3 Label: 1</b></p>	<p><b>H4 Label: 1</b></p>
<b>Sarcasm</b>	<p><b>S1 Label: 1</b></p>	<p><b>S2 Label: 0</b></p>	<p><b>S3 Label: 1</b></p>	<p><b>S4 Label: 1</b></p>

Semantic Gap

**Figure 1: The examples of semantic gap, definition gap and domain gap. The semantic gap is marked in red, the definition gap is shown on the left, and the domain gap is shown on the right.**

plays a paramount and meaningful part in offensive information detection community.

Recent multimodal hate speech detection methods mainly focus on textual and visual features for fusion [20, 23] or directly fine-tuning pre-trained multimodal models [6, 20, 27, 35, 52, 58, 59]. The performance depends on the training data quality and backbone model discriminative ability. While the architecture of the backbone model is gradually complex for improving the discriminative ability, both the number and diversity of samples in the training set are limited, falling behind the model development. Besides, the samples are obviously unbalanced and widely distributed for various attacks in hate speech. As shown in [20], we observe more hate speech about *race or ethnicity* and *religion*, and the aspect of the attack on *socioeconomic class*, *gender identity*, *sexual orientation* and *immigration status* is relatively small. This obviously unbalanced data distribution makes the challenge for training a scalable and discriminative model. In this situation, it is necessary to introduce the enhanced knowledge in other domains to provide additional discriminative information for the relatively small attack samples, so that the attack category of hate speech could be understood comprehensively. The selection of the enhanced knowledge is critical.

Through extensive research and data observation, we find several common knowledge (*i.e.*, attribute or behavior) between Facebook Hateful Memes (denote as **Hate dataset**) [20] and Twitter Sarcasm Detection datasets (denote as **Sarcasm dataset**) [3]. As shown in Figure 1 (“H” denotes hate speech, while “S” denotes sarcasm), H3 indicates disrespect to women. While expressing hate, it satirizes that the kitchen should be for women. S3 calls on men to pay attention to gender equality, and the text in the image simultaneously expresses hate towards gender. Both of the above are defined as positive samples at the gender level. Similarly, H4 expresses disrespect towards Asia while implying irony, and S4 expresses sarcasm towards Japan. Both of the above are defined as positive samples at the national level. Therefore, cross-domain knowledge transfer should be feasible between hate speech and sarcasm communities.

However, during the attempt of cross-domain knowledge transfer between hate speech and sarcasm communities, there are two specific gaps besides the traditional domain gap, which are the semantic gap and definition gap.

**Semantic Gap** indicates that the image-text pair correlation degree of the above two datasets is inconsistent. The inconsistency is embodied in the alignment degree of visual tokens (region-based image features) and word tokens. The main cause is the formation mechanism difference between hate speech and sarcasm. For Hate dataset, the major consideration is the complementary multimodal information construction. Both the semantic information of image and text compose hate speech. The designed multimodal models require comprehensively understanding and reasoning complementary information of each modality. As shown in Figure 1, the correlation degree of image-text pairs in H1, H2, H3 and H4 is relatively weak. But for Sarcasm dataset, the major consideration is the ironic relationship construction between image and text, which are more interdependent. As shown by the red mark in Figure 1, the image-text pairs in S1, S2 and S4 have stronger semantic correlation. If the inconsistency of image-text pair semantic correlation between the above two datasets is not considered, the degree of misalignment would be aggravated between visual tokens and word tokens, compromising the cross-domain knowledge transfer performance.

**Definition Gap** indicates the differences in the definition of positive and negative samples in Hate and Sarcasm datasets. For Hate dataset, the positive samples are strictly defined by Facebook community standards: only speech that attacks the protected categories listed in [20] would be labelled as hateful. But for Sarcasm dataset, tweets with a particular hashtag (*e.g.*, #genderequity, #sharkattacks) are considered a positive sample, which is a very vague and broad definition. As shown in the left part of Figure 1, H2 expresses malicious taunting of others. It is regarded as a negative sample in hate speech, but a positive sample in sarcasm detection. This conflict comes from different definitions. On the contrary, S2 expresses a comfortable living state and is considered a negative sample in the above two datasets. There are the same sentiment tendencies and contradictory sentiment tendencies between the two datasets. Therefore, only the transfer of more samples with the same sentiment tendency could have a positive effect, while the other samples may cause negative transfer. However, it is challenging to select only the knowledge of same sentiment tendency due to the complexity of definitions.

**Domain Gap** indicates the different feature distribution [50]. This gap is caused by different sampling sources. As shown in

the right part of Figure 1, the sampling sources of Hate dataset are Getty Images with hand-crafted hate speech text, which is clean and without any redundant symbols. However, the sampling sources of Sarcasm dataset are image-text tweets containing all kinds of images (e.g., posters, plain text, etc.) with tag-symbolized text (e.g., #, <user>, etc.). Because of the different feature distribution of the two datasets, direct knowledge transfer may lead to invalid transfer.

In this paper, a scalable Cross-Domain Knowledge Transfer (CDKT) framework is proposed to bridge semantic, definition and domain gaps simultaneously for multimodal hate speech detection task. Three modules (i.e., semantic adaptation module, definition adaptation module and domain adaptation module) are included, where hate speech detection and sarcasm detection are primary and auxiliary tasks, respectively. Specifically, semantic adaptation module formulates the irrelevant parts between image and text in primary and auxiliary tasks based on contrastive attention [44], and disentangles with the text representation to align the visual and word tokens. Definition adaptation module assigns different weights to the training samples of auxiliary task by measuring the correlation between samples of the auxiliary and primary task. Domain adaptation module minimizes the feature distribution gap of samples in two tasks.

The contributions of this paper are summarized as follows:

- A scalable Cross-Domain Knowledge Transfer (CDKT) framework is proposed to bridge semantic, definition and domain gaps simultaneously for multimodal hate speech detection, where the mainstream vision-language transformer ViLBER [31], UNITER [5] and ALBEF [24] could be employed as backbone.
- We find the semantic inconsistency between hate speech and sarcasm communities, and design a semantic adaptive module to disentangle the irrelevant parts of image and text representations for aligning the visual and word tokens. The degree of semantic inconsistencies could be decreased.
- Extensive experiments show that the proposed CDKT provides a stable improvement compared with baselines and produces a competitive performance compared with some existing multimodal hate speech detection methods.

## 2 RELATED WORK

In this paper, a cross-domain knowledge transfer framework is proposed for multimodal hate speech detection, where the semantic, definition and domain gaps are bridged simultaneously between hate speech detection and sarcasm detection tasks. Therefore, we briefly review the multimodal hate speech and sarcasm detection, contrastive attention and representation disentanglement, curriculum learning, and domain adaptation in the following subsections.

### 2.1 Multimodal Hate Speech and Sarcasm Detection

Previous work [21, 33, 34, 47, 56] has focused on text-based unimodal hate speech and sarcasm detection. With the social media platforms development, the diffusion of texts and images in social network has become the mainstream compared with the diffusion of texts-only. For multimodal hate speech detection, the large scale

pre-trained multimodal models [6, 20, 27, 35, 46, 52, 58] are directly fine-tuned for feature learning. Besides, some studies have also attempted to utilize data augmentation [23, 59, 60] and model integration methods [40, 52] to improve the hate speech detection performance. However, current research on this task mainly focuses on the construction of multimodal models without considering the influence of the unbalanced and widely distributed samples for various attacks in hate speech. For multimodal sarcasm detection, Cai *et al.*[3] release a sarcasm dataset compiled from image-text tweets and design a hierarchical fusion model as the baseline. Some models [36, 55, 57] are constructed for evaluation on sarcasm dataset. Multimodal hate speech and sarcasm detection are classification tasks based on offensive speech, using multimodal data to directly or indirectly attack people based on the characteristics (including race, religion, gender, etc.). Different from the above model construction methods in multimodal hate speech, we make an initial attempt to transfer common knowledge between hate speech and sarcasm. The purpose of common knowledge transfer is to provide additional discriminative information for the relatively small attack samples, so that the attack category of hate speech could be understood comprehensively.

### 2.2 Contrastive Attention and Representation Disentanglement

The contrastive attention mechanism is the variation of self-attention mechanism [51], which the irrelevant or less relevant parts of pairwise feature vectors are extracted. It is first proposed by Song *et al.*[44] for person re-identification. Duan *et al.*[7] follow the mechanism for text summarization task. The representation disentanglement characterizes the various explanatory factors behind an observed instance in various parts of the feature vector representation [1]. Recent studies have attempted to use supervised signal to learn disentangled representations [11, 13, 17]. Ma *et al.*[32] employ representation disentanglement to retain the multiple intentions behind the edges in relation data such as social networks and user-item interaction graphs. Lee *et al.*[23] introduce representation disentanglement into multimodal hate speech detection to disentangle visual and textual representation. In this paper, we find the semantic inconsistency between hate speech and sarcasm, where the inconsistency refers to the image-text pair correlation degree. To address this problem, we combine the contrastive attention mechanism with representation disentanglement to design a semantic adaptive module. The irrelevant parts between image and text are formulated based on contrastive attention, and disentangled with the text representation to align the visual and word tokens.

### 2.3 Curriculum Learning

Curriculum learning [2] is a training strategy that mimics the human learning process, which helps to transfer knowledge from simple to difficult. Jiang *et al.*[16] propose the self-paced learning to measure the difficulty of training samples in terms of losses. Sachan *et al.*[39] apply curriculum learning for question answering task. Tay *et al.*[48] combine curriculum learning with pointer-generator networks to enrich semantic information for reading comprehension. Platanios *et al.*[38] apply curriculum learning to accelerate

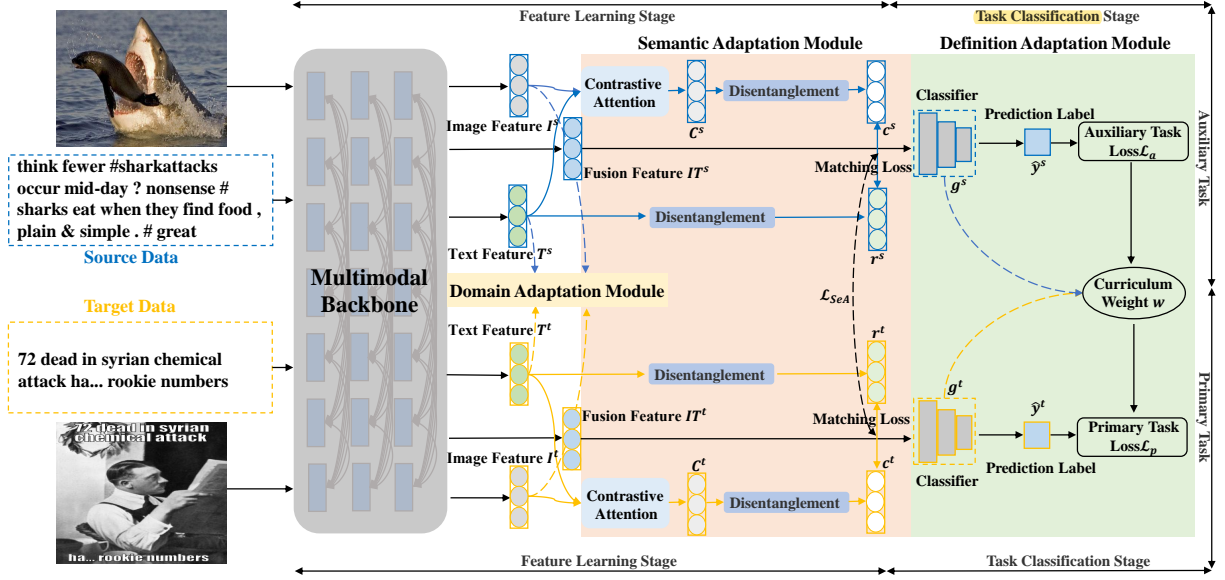


Figure 2: The architecture of the proposed CDKT framework.

model convergence for neural machine translation. In this paper, we assign different weights to the training samples of auxiliary task for bridging definition gap by judging the correlation between samples of the auxiliary and primary task inspired by [43], so that the knowledge with the same sentiment tendency could be transferred to the primary task.

## 2.4 Domain Adaptation

Domain adaptation [9, 28, 29, 37, 45, 62] has shown the superior ability for aligning the feature distribution, which mainly includes two aspects. The first method is based on statistical moment matching, *e.g.*, MMD [28, 30, 61], and the other method is based on adversarial learning, *e.g.*, domain adversarial network adaptation [10, 14, 49]. Generally, the adversarial learning based has shown superior performance to the statistic moment matching based method as described in [4, 29]. Based on the characteristics of multimodal data, we introduce conditional domain adversarial network [29] to minimize the feature distribution gap of samples in two tasks, so that the classifier trained by auxiliary task performs well on the primary task.

## 3 METHODOLOGY

The architecture of the proposed CDKT framework is shown in Figure 2, where the semantic adaptation module, definition adaptation module and domain adaptation module are included. By optimizing the above three modules, the semantic gap, definition gap and domain gap between the primary task and the auxiliary task could be bridged simultaneously. The multimodal backbone is replaceable, where ViLBERT [31], UNITER [5] and ALBEF [24] could be employed for feature learning. In this way, common knowledge is transferred from auxiliary task to primary task.

During the knowledge transfer, the source dataset is denoted as  $\mathcal{D}_s = \{x_i^s, y_i^s\}_{i=1}^{n^s}$ , and the target dataset is denoted as  $\mathcal{D}_t =$

$\{x_j^t, y_j^t\}_{j=1}^{n^t}$ . The samples of source dataset and target dataset are denoted as  $x_i^s$  and  $x_j^t$ , respectively. The ground truth labels of source dataset and target dataset are denoted as  $y_i^s$  and  $y_j^t$ , respectively.  $n$  is the number of samples. The objective function of the auxiliary task is defined as:

$$\mathcal{L}_a = \frac{1}{n^s} \sum_{i=1}^{n^s} L_a(y_i^s, \hat{y}_i^s), \quad (1)$$

where  $L_a$  is the cross entropy loss.  $\hat{y}_i^s = g^s(IT_i^s)$  is the predicted probability of auxiliary task.  $g^s$  represents the classifier of the auxiliary task.  $IT_i^s$  represents the fusion features in the feature learning stage. Similarly, the objective function of the primary task is defined as:

$$\mathcal{L}_p = \frac{1}{n^t} \sum_{j=1}^{n^t} L_p(y_j^t, \hat{y}_j^t). \quad (2)$$

### 3.1 Semantic Adaptation Module (SeA)

Semantic gap means that the image-text pair correlation degree of the above two datasets is inconsistent. The inconsistency is embodied in the alignment degree of visual and word tokens. To bridge the inconsistency, we combine the contrastive attention mechanism with representation disentanglement to design the semantic adaptive module. The irrelevant parts between image and text are formulated based on contrastive attention, and disentangled with the text representation to align the visual and word tokens.

First, the contrastive attention is utilized to model the irrelevant parts between images and texts. The contrastive attention mechanism is defined by Eq.(4) ~ Eq.(6). Different from the self-attention mechanism in Eq.(3), the opponent attention weights  $a_o$  is obtained through the opponent function applied on  $a_c$  followed by the softmax function as shown in Eq.(5). In this way, the most relevant part of  $Q$  and  $K$ , which receives maximum attention in the conventional

attention weights  $a_c$ , is masked in  $a_o$ . Instead, the remaining less relevant or irrelevant parts are extracted into  $a_o$  for the contrastive training. A contrastive vector is generated by Eq.(6).

$$y = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

$$a_c = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \quad (4)$$

$$a_o = \text{softmax}(1 - a_c), \quad (5)$$

$$C = a_o V, \quad (6)$$

where  $Q$  represents image features,  $K$  and  $V$  represent text features.  $d_k$  is the dimension of  $K$ .  $C$  is an inter-modal contrastive vector, representing inconsistencies between input modal variables.

Next, we apply representation disentanglement [23] to the contrastive vector  $C$  and the text representation  $T$  to calculate the inconsistent degree of the image and text. Taking target dataset as an example, the contrastive vector  $C^t \in \mathbb{R}_u$  and the text representation vector  $T^t \in \mathbb{R}_u$  are projected to the latent space  $\mathcal{S}$ , respectively:

$$c^t = W_c C^t + b_c, \quad (7)$$

$$t^t = W_t T^t + b_t, \quad (8)$$

where  $(W_c, W_t) \in \mathbb{R}_{|\mathcal{S}| \times u}$  and  $(b_c, b_t) \in \mathbb{R}_{|\mathcal{S}|}$  are learnable parameters. Since the information overlap exists between tokens of text representation in the latent space, we minimize the mutual information between latent tokens by regularization term as well as the strategy in [32]. In this situation, the likelihood of tokens appearing in latent text representation is maximized, while the absent tokens are minimized. We apply the Straight-Through Gumbel-Softmax (STGS)[15] function to sample a continuous vector  $v \in \mathbb{R}_{|\mathcal{S}|}$  from the Gumbel-Softmax distribution based on  $t^t$ :

$$v_k = \frac{\exp((\log(t_k^t) + g_k)/\tau)}{\sum_{k=1}^{|\mathcal{S}|} \exp((\log(t_k^t) + g_k)/\tau)}, \quad (9)$$

where  $g_k = -\log(-\log(u_k))$ ,  $u_k \sim \text{Uniform}(0, 1)$ .  $t_k^t$  is the  $k$ -th element in  $t^t$ .  $\tau$  is a temperature parameter. The continuous vector  $v_k$  is transformed into a one-hot vector by STGS function:

$$r_k^t = \begin{cases} 1 & \text{if } k = \arg \min_m v_m \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Finally, a matching loss function is introduced to model the degree of inconsistency between images and texts:

$$M^t = \sum_{k=1}^{|\mathcal{S}|} r_k^t \log(c_k^t) + (1 - r_k^t) \log(1 - c_k^t), \quad (11)$$

where  $c_k^t$  is the  $k$ -th element in the contrastive latent representation  $c^t$ . The degree of image-text pair inconsistency between the two datasets is defined as:

$$\mathcal{L}_{SeA} = |M^s - M^t|, \quad (12)$$

where  $M^s$  and  $M^t$  represent the degree of inconsistency between the image and text in the source and target datasets, respectively. The misalignment of visual and word tokens in the target dataset could be significantly improved by reducing semantic conflicts between auxiliary and primary tasks.

### 3.2 Definition Adaptation Module (DeA)

Definition gap means that positive and negative samples are defined differently in the two datasets. To bridge the definition gap, we design the definition adaptation module to assign different weights to the training samples of source dataset inspired by the curriculum learning [43]. In this situation, the knowledge of auxiliary task with the same sentiment tendency could be transferred to the primary task. The objective function is defined as:

$$\min_{\theta, w} E(\theta, w; \lambda) = \frac{1}{n} \sum_{i=1}^n w_i L(y_i, f(x_i; \theta)) - \epsilon \|w\|_1, \quad (13)$$

where  $\theta$  is the model parameter.  $w = [w_1, w_2, \dots, w_n]^T$ ,  $w_i \in [0, 1]$  quantifies the difficulty of the  $i$ -th samples. The hyper-parameter  $\epsilon$  constrains the learning space.

This is the double biconvex optimization problem, optimizing one variable at a time when the other variable is fixed. Specifically, when  $\theta$  is fixed,  $w$  is calculated by Eq.(14). The samples with loss below  $\epsilon$  are selected as ‘‘simple’’ samples ( $w_i = 1$ ). When  $w$  is fixed, gradient descent is used to update the learning parameter  $\theta$  by training ‘‘simple’’ samples only.

$$w_i = \begin{cases} 1 & \text{if } L(y_i, f(x_i; \theta)) < \epsilon \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

We apply curriculum weight  $w_i$  to indicate the sentiment tendency of the sample.  $w_i = 1$  represents the same sentiment tendency. Otherwise, it represents the contradictory sentiment tendency. Under the guidance of curriculum weight, the knowledge consistent with the sentiment tendency of the primary task is transferred from the source to the target dataset.

However, Eq.(14) does not take into account the correlation between auxiliary and primary tasks. Therefore, we improve curriculum weights according to the correlation between task definitions to further guide the learning of auxiliary tasks. Kullback-Leibler(KL) [22] divergence is introduced to calculate the predicted distribution distance of the two classifiers on the source dataset. Curriculum weight  $w_i$  is redefined as:

$$w_i = \begin{cases} 1 & \text{if } KL_{x_i^s \sim \mathcal{D}_s}(g_i^s, g_i^t) < \epsilon \\ g_i^t(y_i^s) & \text{otherwise,} \end{cases} \quad (15)$$

where  $g_i^s$  represents the predicted distribution of the source dataset on the auxiliary task classifier,  $g_i^t$  represents the predicted distribution of the source dataset on the primary task classifier.  $g_i^t(y_i^s)$  represents the probability that the source dataset is predicted to be the ground truth label on the primary task classifier. Although negative samples interfere with knowledge transfer, they also provide helpful background knowledge to some extent [53, 54]. So we only reduce the weight of negative samples. The objective function of auxiliary tasks combined with curriculum weights is redefined as:

$$\mathcal{L}_{DeA} = \frac{1}{n^s} \sum_{i=1}^{n^s} w_i(x_i^s) L_a(y_i^s, \hat{y}_i^s). \quad (16)$$

### 3.3 Domain Adaptation Module (DoA)

Domain gap means that two datasets are distributed in different feature spaces. To bridge the domain gap, we introduce Conditional Domain Adversarial Network (CDAN) [29] to minimize the feature

**Table 1: Statistics of Hate and Sarcasm datasets.**

Datasets	#Training	#Validation	#Test
Hate	Hateful(3,050)	Hateful(250)	Hateful(500)
	Non-Hateful (5,450)	Non-Hateful (250)	Non-Hateful (500)
	All(8,500)	All(500)	All(1000)
Sarcasm	Sarcasm(8,642)	Sarcasm(959)	Sarcasm(959)
	Non-Sarcasm (11,174)	Non-Sarcasm (1,451)	Non-Sarcasm (1,450)
	All(19,816)	All(2,410)	All(2,409)

distribution difference of samples in two tasks, so that the classifier trained by auxiliary task performs well on the primary task. Taking image as an example, the objective function of CDAN is defined as:

$$Cdan_I = \mathbb{E}_{x_i^s \sim \mathcal{D}_s} \log[D(\mathbf{h}(I_i^s, g_i^s))] + \mathbb{E}_{x_j^t \sim \mathcal{D}_t} \log[1 - D(\mathbf{h}(I_j^t, g_j^t))], \quad (17)$$

where  $D$  represents the domain discriminator.  $\mathbf{h}$  represents multilinear mapping, connecting  $I$  and  $g$ . By minimizing  $Cdan$ , the feature space of the two datasets is aligned to reduce the domain gap. Considering both image and text, the objective function of the domain adaptation module could be defined as:

$$\mathcal{L}_{DoA} = Cdan_I + Cdan_T. \quad (18)$$

### 3.4 Optimization

Finally, we integrate the above modules to optimize the total objective function of CDKT framework:

$$\mathcal{L}_{Loss} = \mathcal{L}_p + \alpha \mathcal{L}_{SeA} + \beta \mathcal{L}_{DeA} + \gamma \mathcal{L}_{DoA}, \quad (19)$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  are hyperparameters. By optimizing the above three modules, the semantic, definition and domain gaps between primary and auxiliary tasks could be bridged simultaneously. It is an end-to-end knowledge transfer framework for multimodal hate speech detection.

## 4 EXPERIMENTS

### 4.1 Dataset

The experiment is conducted on two publicly-available datasets: Facebook Hateful Memes dataset (denote as **Hate dataset**) and Twitter Sarcasm Detection dataset (denote as **Sarcasm dataset**). The statistics are shown in Table 1 and the details are briefly described as follows:

**Hate dataset**<sup>1</sup> is constructed as the part of Hateful Memes Challenge 2020 for Multimodal Hate Speech Detection and published in [20], which strictly follows Facebook community standard<sup>2</sup> for hate speech. The dataset contains 10K memes with binary labels (*i.e.*, hateful or non-hateful) and is divided into the training, validation, and test sets at a ratio of 85% : 5% : 10%. Only memes that attack the protected categories in [20] are considered as hateful.

<sup>1</sup><https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/>

<sup>2</sup>[https://www.facebook.com/communitystandards/hate\\_speech](https://www.facebook.com/communitystandards/hate_speech)

**Table 2: The performance comparison of CDKT on Hate dataset.**

Models	Validation		Test	
	Acc.	AUROC	Acc.	AUROC
Late Fusion	61.53	65.97	59.66	64.75
Concat BERT	58.60	65.25	59.13	65.79
MMBT-Grid	58.20	68.57	60.06	67.92
MMBT-Region	58.73	71.03	60.23	70.73
ViLBERT	62.20	71.13	62.30	70.45
Visual BERT	62.10	70.60	63.20	71.33
ViLBERT CC	61.40	70.07	61.10	70.03
Visual BERT COCO	65.06	73.97	64.73	71.41
ViLBERT	64.45	70.84	67.10	73.13
CDKT(ViLBERT)	<b>65.62</b>	<b>70.94</b>	<b>67.70</b>	<b>76.85</b>
UNITER	68.80	77.58	70.60	77.59
CDKT(UNITER)	<b>70.20</b>	<b>78.42</b>	<b>71.70</b>	<b>80.71</b>
ALBEF	68.80	<b>80.62</b>	72.10	79.91
CDKT(ALBEF)	<b>74.20</b>	79.89	<b>76.50</b>	<b>83.74</b>

**Sarcasm dataset**<sup>3</sup> consists of image-text tweets collected in [3] for Multimodal Sarcasm Detection. The dataset contains nearly 25K tweets with binary labels (*i.e.*, sarcasm or non-sarcasm) and is divided into the training, validation, and test sets at a ratio of 80% : 10% : 10%. Samples with specific hashtag (*e.g.*, #genderequity, #idiot, #funny, etc.) are considered as sarcastic.

### 4.2 Experimental Settings

**Implementation details.** The Multimodal Backbone of CDKT framework includes ViLBERT<sup>4</sup>, UNITER<sup>5</sup> and ALBEF<sup>6</sup>. As shown in Figure 2, during the feature learning stage, parameters are shared between primary and auxiliary tasks. During the classification stage, the primary and auxiliary tasks have different classifiers. The dropout value of the classification layer MLP is set to 0.5. We use weighted Adam as the optimizer, cosine annealing and warm up strategy to control the variation of learning rate. For the above three backbone models, the initial learning rate is set to 1e-5, 6.25e-5 and 2e-5, respectively. The size of the minibatch is set to 16. We train the entire framework on NVIDIA Tesla V100 (32G memory) GPU.

**Evaluation metrics.** For Hate dataset, we follow the metrics in [20], utilizing Area Under the Receiver Operating Characteristic curve(AUROC) and accuracy (Acc.) to evaluate the effectiveness of hate speech detection. The AUROC is the primary metric. For Sarcasm dataset, we follow the evaluation method adopted by [3], using F1, precision, recall and accuracy as evaluation metrics.

### 4.3 Experimental Results

**Comparison with the baselines.** We use the Sarcasm dataset as the source domain and the Hate dataset as the target domain. The

<sup>3</sup><https://github.com/headacheboy/data-of-multimodal-sarcasm-detection>

<sup>4</sup><https://github.com/facebookresearch/mmf/tree/main/projects/vilbert>

<sup>5</sup><https://github.com/HimariO/HatefulMemesChallenge/tree/main/UNITER>

<sup>6</sup><https://github.com/salesforce/ALBEF>

**Table 3: The performance comparison of CDKT on Sarcasm dataset.**

Models	F1	Pre.	Recall	Acc.
HFM	80.18	76.57	84.15	83.44
D&R Net	80.60	77.97	83.42	84.02
Res-Bert	81.57	78.87	84.46	84.80
MIII-MMSD	82.92	80.87	85.08	86.05
ViLBERT	80.60	77.61	<b>83.84</b>	83.69
CDKT(ViLBERT)	<b>81.55</b>	<b>81.76</b>	81.33	<b>85.39</b>
UNITER	84.89	90.87	<b>79.66</b>	82.94
CDKT(UNITER)	<b>85.42</b>	<b>92.45</b>	79.38	<b>83.69</b>
ALBEF	82.30	79.06	85.82	85.31
CDKT(ALBEF)	<b>83.89</b>	<b>79.37</b>	<b>88.96</b>	<b>85.60</b>

goal is the common knowledge transfer between primary and auxiliary tasks for improving the performance of hate speech detection. ViLBERT [31], UNITER [5] and ALBEF [24] are employed as the multimodal backbone of CDKT, which are also the baselines in our experiment. It can be observed from Table 2 that the proposed CDKT outperforms the corresponding baseline under the common knowledge transfer. Specifically, when ViLBERT is the baseline, AUROC increases from 73.13% to 76.85% (+3.72%), and for Acc., from 67.10% to 67.70% (+0.60%) in test set. AUROC and Acc. have respectively increased by +3.12% and +1.10% when UNITER is the baseline. An even larger improvement can be seen when ALBEF is the baseline, AUROC has increased by +3.83% (from 79.91% to 83.74%) and Acc. has increased by +4.40% (from 72.10% to 76.50%). The above results demonstrate the effectiveness of common knowledge transfer in the proposed CDKT for multimodal hate speech detection.

**Comparison with some existing multimodal hate speech detection methods.** We compare with some existing multimodal hate speech detection methods, including Late Fusion [20] (taking the mean of the unimodal ResNet-152 [12] and BERT [18] output scores), Concat BERT [20] (concatenating ResNet-152 features with BERT and training an MLP on top), MMBT-Grid [20] (using Image-Grid features as supervised multimodal bitransformers [19]), MMBT-Region [20] (using Image-Region features as supervised multimodal bitransformers [19]), ViLBERT [31] and Visual BERT [25] (that are only unimodally pretrained and not pretrained on multimodal data), ViLBERT CC and VisualBERT COCO (that are pretrained models on CC [42] and COCO [26], respectively). Obviously, the proposed CDKT is superior to the above multimodal hate speech detection methods in Table 2. In summary, when ViLBERT, UNITER and ALBEF are the multimodal backbone of the framework, AUROC scores in test set exceed +5.44%, +9.30% and +12.33% respectively, Acc. scores also exceed +2.97%, +6.97% and +11.77% respectively.

#### 4.4 Hate to Sarcasm

To further verify the scalability of the common knowledge transfer in the proposed CDKT framework, we conduct the opposite experiment: we use the Hate dataset as the source domain and the Sarcasm

**Table 4: Ablation study tested on Hate dataset.**

Models	Acc.	AUROC
ALBEF	72.10	79.91
CDKT(ALBEF)	76.50	83.74
CDKT(ALBEF) w/o DoA	72.10	81.10
CDKT(ALBEF) w/o DeA	73.30	82.53
CDKT(ALBEF) w/o SeA	72.70	80.59

**Table 5: Ablation study tested on Sarcasm dataset.**

Models	F1	Pre.	Recall	Acc.
ALBEF	82.30	79.06	85.82	85.31
CDKT(ALBEF)	83.89	79.37	88.96	85.60
CDKT(ALBEF) w/o DoA	82.25	78.09	86.88	84.94
CDKT(ALBEF) w/o DeA	82.85	79.36	86.67	85.35
CDKT(ALBEF) w/o SeA	82.34	79.01	85.96	84.35

dataset as the target domain, *i.e.*, evaluating the effectiveness of whether the common knowledge in hate speech detection could be transferred to sarcasm detection task for improving the performance. As shown in table 3, CDKT provides a stable improvement over all metrics (especially in F1, Acc.) compared with baselines. Specifically, F1 increases by +0.53% ~ +1.59% and Acc. increases by +0.29% ~ +1.70% in test set. The common knowledge transfer has superior scalability on sarcasm detection task.

We compare with some existing multimodal sarcasm detection methods, including HFM [3] (taking text, image, and attribute feature as modalities), D&R Net [57] (using adjective-noun pairs and semantic association between image and text), Res-bert [36] (concatenating the output of image features from ResNet and text features from BERT), MIII-MMSD [36] (using self-attention and co-attention mechanisms to capture inter and intra-modality incongruity, respectively). It can be observed from Table 3 that the proposed CDKT is also superior to the above methods in some indicators. Extensive experiments demonstrate the effectiveness of CDKT again.

#### 4.5 Ablation Study

In this paper, the most discriminative ALBEF model is utilized as the multimodal backbone of CDKT to conduct the ablation study for evaluating the effectiveness of each module. Table 4 and 5 show the results on Hate and Sarcasm datasets, respectively.

**CDKT w/o DoA.** The domain adaptation module is used to reduce the feature distribution difference between Hate and Sarcasm datasets. After removing it, the knowledge transferred is treated as noise on the primary task classifier due to the domain gap. Specifically, in hate speech detection, Acc. is equal to the baseline, while in sarcasm detection, F1, Pre. and Acc. are all lower than the baseline. The results show that direct knowledge transfer may lead to invalid transfer without considering domain gap. Domain adaptation module is the cornerstone of cross-domain knowledge transfer.

**CDKT w/o DeA.** The definition adaptation module is used to select samples with same sentiment tendencies. After removing it, samples with contradictory sentiment tendencies are also treated






<p>Nationality</p>	 <p><b>american</b> kids all across <b>american</b> storefronts!!! freeway off n on ramps!!! street corners!!! begging for \$\$\$ for the're alcohol!!! n drugs!!!</p>	 <p>no little <b>asian</b> don't eat them raw</p>	 <p>to see better, <b>asians</b> sometime switch to fullscreen veiw</p>	 <p>had to load these by hand today felt like a <b>syrian</b> guy looking for his wife</p>
<p>Disability or Disease</p>	 <p>this is the worst <b>cancer</b> i've ever seen</p>	 <p>kid with <b>cancer</b>: so when i get out of the hospital i'm going to join you? the avengers visiting him</p>	 <p>"nice watch must have <b>cost an arm and a leg</b>"</p>	 <p>dark humour is like a child with <b>cancer</b> it doesn't get old</p>
<p>Sexual Orientation / Gender Identity</p>	 <p>when did you decide to <b>gender swap</b>? it started when i swapped my mustang</p>	 <p>me convincing <b>homophobes</b> that i'm just like them and not a <b>sexually deviant pervert</b></p>	 <p>nowadays chicks feel like showing skin is a sign of confidence but in reality its a form of insecurity because they have nothing else to offer other than <b>sexuality</b></p>	 <p>back in my day there were only two genders <b>male and not male</b></p>
<p>Immigration Status</p>	 <p>so you say 12 russians can influence an election but 25 million <b>illegal aliens</b> in the usa cant ?</p>	 <p>stop <b>illegal immigrants</b> they're taking the land we stole</p>	 <p>so obama imports 70,000 <b>somali immigrants</b> and parks them in minnesota, where almost all cluster in a single area...</p>	 <p><b>illegal immigration</b> the cowboy way we'll know them if they try to come back in!</p>

Figure 3: Case study for CDKT. The categories of hate speech are listed on the left, while the samples correctly classified by CDKT are shown on the right.

in the same way. The results show that negative samples with conflicting sentiment tendency could weaken the effect of knowledge transfer without considering the definition gap. Definition adaptation module is indispensable for cross-domain knowledge transfer.

**CDKT w/o SeA.** The semantic adaptation module is used to align visual and word tokens. After removing it, the model performance decreases the most. Specifically, in hate speech detection, AUROC decreases by -3.15%, and in sarcasm detection, performance drops by -1.54% on average. The results show that the introduction of another dataset could aggravate the degree of misalignment between visual and word tokens in the target dataset. The insufficient fusion between image and text modalities leads to the discriminative ability degradation of the framework. In this case, semantic adaptive module is particularly significant for cross-domain knowledge transfer. Ablation study results demonstrate that the combination of the above three modules could provide optimal performance.

#### 4.6 Case Study

The ability of CDKT is to transfer common knowledge from sarcasm detection to hate speech detection, addressing the unbalanced and widely distributed samples in hate speech. To have an intuitive understanding of the proposed CDKT, we show the case in Figure 3. The left part of the figure represents the hate categories that are relatively less distributed in hate speech, with keywords highlighted in red. We can observe that more samples with smaller proportions of hate categories are classified correctly, including *Nationality*, *Disability or Disease*, *Sexual orientation*, *Gender identity* and *Immigration status*. Taking *Nationality* as an example, in the first sample, the image information is closely related to the text information, indicating hate towards America. With the help of CDKT, the model

could align and fuse the tokens of the two modalities more accurately, making the model easier to infer the hate information. In the second and third samples, neither text nor image could identify the hate of Asia. However, with the promotion of common knowledge, the facial expressions and actions in the images are complementary to the information in the text, which enables the model to mine the potential hate. The last sample expresses hate towards the Syrian. With the help of knowledge transfer, models might combine Syria with war as the prior knowledge to deduce hate of nationality. The above cases illustrate that the limitation of model discriminative ability caused by data imbalance has been improved, and the common knowledge related to hate speech has been transferred from sarcasm dataset to hate speech detection.

### 5 CONCLUSION

In this paper, a scalable Cross-Domain Knowledge Transfer (CDKT) framework is proposed to transfer common knowledge from sarcasm detection to hate speech detection. The unbalanced distribution phenomenon of attack categories is alleviated in hate speech detection. Three modules (*i.e.*, semantic adaptation module, definition adaptation module and domain adaptation module) are included to bridge semantic, definition and domain gaps simultaneously. Moreover, the mainstream vision-language transformer could be flexibly employed as backbone, verifying the scalability of CDKT. With the promotion of common knowledge, the discriminative ability is improved significantly, and attack categories of hate speech could be understood comprehensively. Experimental results show that CDKT provides a stable improvement compared with baselines and produces a competitive performance compared with some methods. The ablation and case studies further demonstrate the effectiveness of each module in CDKT.



## REFERENCES

- [1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013), 1798–1828.
- [2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning*. 41–48.
- [3] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the Association for Computational Linguistics*. 2506–2515.
- [4] Zhangjie Cao, Kaichao You, Ziyang Zhang, Jianmin Wang, and Mingsheng Long. 2022. From Big to Small: Adaptive Learning to Partial-Set Domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*. 104–120.
- [6] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891* (2020).
- [7] Xiangyu Duan, Hongfei Yu, Mingming Yin, Min Zhang, Weihua Luo, and Yue Zhang. 2019. Contrastive Attention Mechanism for Abstractive Sentence Summarization. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3044–3053.
- [8] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* (2018), 1–30.
- [9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. 1180–1189.
- [10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* (2016), 2096–2030.
- [11] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is "love" evading hate speech detection. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*. 2–12.
- [12] 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.
- [13] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*. 44–51.
- [14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*. 1989–1998.
- [15] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
- [16] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. 2015. Self-paced curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2694–2700.
- [17] Theofanis Karaletos, Serge Belongie, and Gunnar Rätsch. 2015. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011* (2015).
- [18] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
- [19] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950* (2019).
- [20] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the International Conference on Neural Information Processing Systems*. 2611–2624.
- [21] Myung Jong Kim, Younggwon Kim, JaeDeok Lim, and Hoirin Kim. 2010. Automatic detection of malicious sound using segmental two-dimensional mel-frequency cepstral coefficients and histograms of oriented gradients. In *Proceedings of ACM International Conference on Multimedia*. 887–890.
- [22] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* (1951), 79–86.
- [23] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling Hate in Online Memes. In *Proceedings of ACM International Conference on Multimedia*. 5138–5147.
- [24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Proceedings of the International Conference on Neural Information Processing Systems*. 1978–1992.
- [25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).
- [26] 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 *European Conference on Computer Vision*. 740–755.
- [27] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871* (2020).
- [28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*. 97–105.
- [29] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In *Proceedings of the International Conference on Neural Information Processing Systems*. 1647–1657.
- [30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*. 2208–2217.
- [31] Jiaseen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the International Conference on Neural Information Processing Systems*. 13–23.
- [32] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled self-supervision in sequential recommenders. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 483–491.
- [33] Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427* (2017).
- [34] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 14867–14875.
- [35] Niklas Muennighoff. 2020. Vilio: state-of-the-art Visio-Linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788* (2020).
- [36] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP*. 1383–1392.
- [37] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-Adversarial Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [38] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. 2019. Competence-based Curriculum Learning for Neural Machine Translation. In *Proceedings of NAACL-HLT*. 1162–1172.
- [39] Mrinmaya Sachan and Eric Xing. 2016. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the Association for Computational Linguistics*. 453–463.
- [40] Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *arXiv preprint arXiv:2012.13235* (2020).
- [41] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the International Workshop on Natural Language Processing for Social Media*. 1–10.
- [42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Association for Computational Linguistics*. 2556–2565.
- [43] Yang Shu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2019. Transferable curriculum for weakly-supervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4951–4958.
- [44] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1179–1188.
- [45] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*. 443–450.
- [46] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying*. 32–41.
- [47] Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with Sarcasm by Reading In-Between. In *Proceedings of the Association for Computational Linguistics*. 1010–1020.
- [48] Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and Effective Curriculum Pointer-Generator Networks for Reading Comprehension over Long Narratives. In *Proceedings of the Association for Computational Linguistics*. 4922–4931.
- [49] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7167–7176.
- [50] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* (2014).
- [51] 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 International Conference on Neural Information Processing Systems*. 6000–6010.
- [52] Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975* (2020).
- [53] Jinpeng Wang, Jieming Zhu, and Xiuqiang He. 2021. Cross-Batch Negative Sampling for Training Two-Tower Recommenders. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1632–1636.
- [54] Peifeng Wang, Shuangyin Li, and Rong Pan. 2018. Incorporating gan for negative sampling in knowledge representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [55] Xinyu Wang, Xiaowen Sun, Tan Yang, and Hongbo Wang. 2020. Building a bridge: a method for image-text sarcasm detection without pretraining on image-text data. In *Proceedings of the International Workshop on Natural Language Processing Beyond Text*. 19–29.
- [56] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*. 88–93.
- [57] Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the Association for Computational Linguistics*. 3777–3786.
- [58] Weibo Zhang, Guihua Liu, Zhuohua Li, and Fuqing Zhu. 2020. Hateful memes detection via complementary visual and linguistic networks. *arXiv preprint arXiv:2012.04977* (2020).
- [59] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–6.
- [60] Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290* (2020).
- [61] Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. 2019. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 5989–5996.
- [62] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Jingwu Chen, Zhiping Shi, Wenjuan Wu, and Qing He. 2019. Multi-representation adaptation network for cross-domain image classification. *Neural Networks* (2019), 214–221.