Multi-source Multi-modal Domain Adaptation for Visual-textual Sentiment Classification

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Learning from multiple modalities has recently been paid increasing attention in sentiment analysis tasks because of its ability to capture the complementary representation of the intrinsic multi-modal world. Recent deep learning-based multi-modal sentiment analysis methods trained on large-scale labeled data cannot guarantee good generalization to another target domain, because of the presence of domain shift. Multi-modal domain adaptation (MMDA) aims to address this issue by learning a transferable model with specific alignment across domains. However, existing MMDA methods only focus on the single-source scenario with only one labeled source domain. When labeled data is collected practically from multiple sources with different distributions, naive application of these single-source MMDA methods would fail without considering the domain shift among different sources. In this paper, we propose to study multi-source MMDA for visual-textual sentiment classification and design a novel multi-source multi-modal contrastive adversarial network, termed M2CAN, to learn domain-invariant multi-modal representations. Specifically, the designed M2CAN jointly optimizes three different alignment strategies: cross-modal contrastive alignment within each domain, cross-domain constrastive alignment for each modality, and cross-domain adversarial alignment on the fused multi-modal representations. After such alignments, different source and target domains are mapped into a shared multi-modal representation space. We conduct extensive experiments on a benchmark dataset with three domains and the results demonstrate that the proposed M2CAN significantly outperforms state-of-the-art domain adaptation approaches for visual-textual sentiment classification. Our source code will be released.

1 Introduction

Customers have become used to sharing their experiences and opinions of the products and service they purchase by posting reviews or comments on social networks [1, 2, 3, 4, 5, 6]. Sentiment analysis of the large-scale user-generated multimedia data plays a vitally important role in both customers' product selection and enterprises' product improvement. On the one hand, it can influence customers' decision-making when selecting what they want. For example, if the feedback from other customers is dominated by negative comments, it is highly probable that the current customers change their attitudes to another brand. On the other hand, it can help enterprises to analyze the drawbacks revealed by customers and correspondingly improve the quality of their products and services [1, 7]. Although text is one direct and popular modality to express customers' opinions [2], sentiment analysis solely from text may not well reflect the customers' actual feelings. For example, if we see a comment like "what a good restaurant!", we may conclude that the customer is satisfied with the dining; but if there is an also an affiliated image showing a dirty and disorderly environment, we can infer that the text is actually sarcasm and that the customer is upset about the dining environment. Therefore, sentiment analysis from multiple modalities, such as image and text, has attracted increasing research attention with the help of easy photographing on mobile devices.

Recently, deep neural networks (DNNs) have achieved the state-of-the-art performances on visual-textual sentiment classification by effectively exploring the abundant and complementary content knowledge from different modalities [8, 9, 10, 11, 12, 13]. To train a DNN well, large-scale annotations are often required; however, these are not always available, since labeling multi-modal data is time-consuming and even difficult. One may consider transferring the trained DNN on a labeled source domain to the unlabeled target domain as an alternate solution. Obviously, direct transfer cannot guarantee good generalization and often results in significant performance decay [14, 15, 16], because of the presence of domain shift [17], *i.e.* the distributions of observed multi-modal data and sentiment are different between the source and target domains. Aiming at minimizing the domain gap, domain adaptation (DA) [18, 19, 20, 21, 22] tries to learn a model on the labeled source domain that can generalize well to the target domain through specific alignment across domains, such as discrepancy-based, adversarial, and self-supervision-based methods.

Current DA methods for visual-textual sentiment classification and other multi-modal learning tasks only focus on the single-source unsupervised setting [23, 24, 25, 26, 27], by assuming that the labeled source data is collected from the same distribution. However, in practice, it is more practical that the labeled multi-modal data comes from different source distributions [19, 21]. For example, user-generated reviews can be collected from Yelp, Twitter, and Amazon. We can naively combine different sources into one source and directly apply existing single-source DA algorithms. However, because of the neglect of mis-alignment across different sources, such methods may lead to sub-optimal results [21] (see the comparison between single-best and source-combined MM-SADA in Table 1). Therefore, effective multi-source domain adaptation (MSDA) techniques [19, 21] are required to sufficiently leverage the complementary information from different sources.

Recently, some deep MSDA methods have been proposed. Based on different alignment strategies, Zhao et al. classified them into two categories [21], *i.e.* latent space transformation [28, 29, 30, 31, 32, 33, 34, 35, 36, 16, 37, 38, 39] and intermediate domain generation [40, 41, 42, 43, 44]. All these MSDA methods only consider a single modality, such as text or image. When extending them to a multi-modal setting, they usually fail since they cannot deal well with the heterogeneity gap, *i.e.* the semantic difference between data in different modalities (*e.g.* heterogeneity of the feature space of each modality and data content) [23]. Therefore, ineffectively aligning feature representations and mining cross-modal information may result in interference among different modalities, leading classification models to fail to capture accurate and stable sentiment-related patterns.

In this paper, we generalize the single-source MMDA and single-modal MSDA problems to multi-source multi-modal domain adaptation (MS-MMDA) problem, and design a novel multi-source multi-modal contrastive adversarial network, termed M2CAN, for visual-textual sentiment classification. First, we use a pair of pre-trained image and text encoders in order to project images and texts from different domains into a continuous latent feature space. Second, we perform different alignments to learn domain-invariant multi-modal representations, including (1) cross-modal contrastive alignment on the transformed lower-dimensional representations obtained by a non-linear transformation layer within each domain, (2) cross-domain constrastive alignment on the original representations for each modality, and (3) cross-domain adversarial alignment on the fused multi-modal representations obtained by multi-modal low-rank bi-linear pooling. Finally, we train a transferable task sentiment classifier based on the aligned multi-modal feature representations and corresponding source labels. Extensive experiments are conducted on a combined dataset consisting of three domains, *i.e.* Yelp [12], Twitter [45], and MVSA [46]. The results demonstrate that M2CAN significantly outperforms the state-of-the-art DA methods for visual-textual sentiment classification.

In summary, the contributions of this paper are threefold: (1) We propose to study a novel and practical DA setting, *i.e.* multi-source multi-modal domain adaptation (MS-MMDA), for visual-textual sentiment classification. To the best of our knowledge, this is the first work that investigates MMDA with multiple sources. (2) We propose a novel MS-MMDA method, termed M2CAN, by contrastive and adversarial learning. Through both cross-modal alignment and cross-domain alignment, M2CAN can learn domain invariant multi-modal representations and thus minimizes the domain gap among multiple sources and the target. (3) We conduct extensive experiments on a benchmark dataset with three different domains. As compared to the best baseline, the proposed M2CAN achieves 2.7% performance gains on the average classification accuracy.

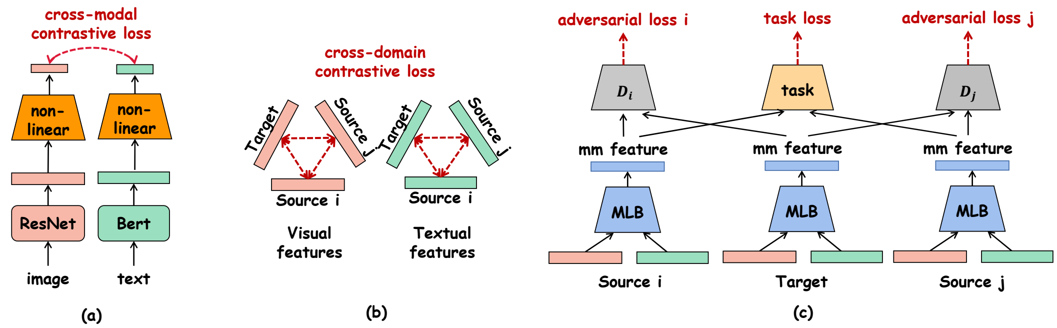


Figure 1: Illustration of the proposed M2CAN framework: (a) image-text feature encoding and cross-model contrastive alignment, (b) cross-domain contrastive alignment, and (c) cross-domain adversarial alignment and task classifier learning. All images and texts are encoded with encoders (ResNet50 [47] and BERT [48]) to a latent continuous feature space. Three different alignments are then performed to learn domain-invariant multi-modal representations, including cross-modal contrastive alignment on the transformed lower-dimensional representations obtained by non-linear transformation, cross-domain constrastive alignment on the original representations for each modality, and cross-domain adversarial alignment on the fused multi-modal representations obtained by multi-modal low-rank bi-linear pooling (MLB). A transferable task sentiment classifier is finally trained based on the aligned multi-modal (mm) feature representations and corresponding source labels.

2 Multi-source Multi-modal Domain Adaptation Network

We consider the multi-source domain adaptation setup for visual-textual sentiment classification, under the *covariate shift* assumption [18]. Assume access to K source domains $\{\mathcal{S}_i\}_{i=1}^K$ with labeled training data and a target domain $\mathcal{T}$ with unlabeled training data consisting of two modalities, *i.e.* image and text. Each domain $\mathcal{S}_i$ contains a set of examples drawn from a joint distribution $p^{(S_i)}(\mathrm{x}_{text}, \mathrm{x}_{image}, \mathrm{y})$ on the input space $\mathcal{X}_{text} \times \mathcal{X}_{image}$ and the output space $\mathcal{Y}$, and we seek to learn a sentiment classifier $f : \mathcal{X}_{text} \times \mathcal{X}_{image} \to \mathcal{Y}$ that is transferable to a target domain $\mathcal{T}$, where only unlabeled data is available. In this section, we give an overview of M2CAN, present each component of M2CAN in detail, and finally introduce the joint learning process.

2.1 Overview

The proposed M2CAN bridges the domain gap by performing both contrastive and adversarial alignments among the source and target domains. The framework is shown in Figure 1. In addition to the pre-trained encoders to encode texts and images from different domains into a semantic-preserving latent continuous feature space and the task classifier to train the final sentiment classification model based on the aligned multi-modal features, it consists of three primary alignment components:

Cross-modal contrastive alignment (CMCA): Align the encoded lower-dimensional representations between different modalities within each domain. The visual and textual representations are projected into a lower-dimensional space with a non-linear transformation layer to extract data transformation-invariant features. A contrastive loss is employed to align visual and textual representations by minimizing the spatial distance between related image and text and maximizing the distance between unrelated pairs.

Cross-domain contrastive alignment (CDCA): Align the encoded original representations between different domains for each modality. Considering that domain gap exists in each modality, the discrepancy between different domains is decreased for each modality through contrastive learning. Due to the fact that there are no negative pairs of samples, applying the same mechanism with CMCA can lead to mode collapse of non-linear layer, *i.e.* the non-linear layer might tend to project dissimilar higher dimensional features into similar lower dimensional features. Therefore, the CDCA is constructed on the original feature space.

Cross-domain adversarial alignment (CDAA): Align the fused multi-modal representations between different domains. A fused multi-modal feature space $\mathcal{X}_{mm}$ is created by using a bi-linear pooling layer, which learns a semantic-preserving and semantic-relevant projection $\mathcal{X}_{text} \times \mathcal{X}_{image} \rightarrow \mathcal{X}_{mm}$. Adversarial learning is employed to align the fused multi-modal features from different domains.

2.2 Cross-modal Contrastive Alignment

Simply extracting visual and textual features using separate encoders does not take the discrepancy between features in different modalities into account. Practically, user-generated data of visual-textual pairs might contain unrelated sentiment information. Furthermore, unaligned visual and textual features are from different feature spaces, which might affect the sentiment classification network's ability to learn appropriate patterns related to sentiment. Therefore, alignment between features from multiple modalities is necessary. For this purpose, we follow [49] and incorporate a contrastive loss into our network. By applying data augmentation in both images and texts, we construct positive and negative sample pairs of different modalities in each domain on a lower-dimensional space using a non-linear transformation layer. Since the non-linear layer is trained to be invariant to data transformation, it can remove information that may be useful for the downstream task, such as the color and orientation of objects, or tone-related words. By leveraging the non-linear transformation, more information can be formed and maintained in the original features [49]. Assuming we have a batch of visual features I, and corresponding batch of textual features T, after data augmentation, the corresponding batches of visual and textual features are I' and T', the cross-modal contrastive loss can be constructed as follows [50]:

$$\mathbb{I} = g(X_{img}), \mathbb{I}' = g(X'_{img}), \mathbb{T} = g(X_{txt}), \mathbb{T}' = g(X_{txt}), \tag{1}$$

$$\mathcal{L}_{CMCA} = -\frac{1}{n} \cdot \mathbb{1}^T \cdot log \left[\frac{e^{\mathbb{I} \circ \mathbb{T}} + e^{\mathbb{I} \circ \mathbb{T}'} + e^{\mathbb{I}' \circ \mathbb{T}} + e^{\mathbb{I}' \circ \mathbb{T}'}}{\mathbb{1}^T \cdot \left(e^{\mathbb{I} \cdot \mathbb{T}^T} + e^{\mathbb{I} \cdot \mathbb{T}'^T} + e^{\mathbb{I}' \cdot \mathbb{T}^T} + e^{\mathbb{I}' \cdot \mathbb{T}'^T} \right) \cdot \mathbb{1}} \right], \tag{2}$$

where $g : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ is a lower-dimensional projection function, d represents the dimension of original visual and textual feature, while d' represents the dimension after projection, $X_{img} \in \mathbb{R}^{n \times d}$, $X_{txt} \in \mathbb{R}^{n \times d}$, $X'_{img} \in \mathbb{R}^{n \times d}$, and $X'_{txt} \in \mathbb{R}^{n \times d}$ represent a batch of original visual and textual feature, and a batch of augmented visual and textual feature respectively, $\circ$ represents the Hadamard product, and n denotes the batch size. By minimizing the distance between visual and textual features from the same sample and maximizing the distance between visual and textual feature from different samples before and after data augmentation, our cross-modal contrastive loss is able to force the encoders to extract closer features from semantically similar samples and farther apart features from semantically different samples robustly, which achieves the purpose of CMCA.

2.3 Cross-domain Contrastive Alignment

To describe the similarity of two distributions, we introduce the Maximum Mean Discrepancy (MMD), as described below:

$$\mathcal{D}_{\mathcal{H}}(P, Q) \triangleq \sup_{f \sim \mathcal{H}} \left(\mathbb{E}_{\boldsymbol{X}^s} \left[f\left(\boldsymbol{X}^s \right) \right] - \mathbb{E}_{\boldsymbol{X}^t} \left[f\left(\boldsymbol{X}^t \right) \right] \right)_{\mathcal{H}}, \tag{3}$$

where $\boldsymbol{X}^s$ and $\boldsymbol{X}^t$ are sampled from the marginal distributions $P(X^s)$ and $Q(X^t)$ respectively, $\mathcal{H}$ is a class of function. Formally, MMD defines the difference between two distributions with their mean representations in the reproducing kernel Hilbert space (RKHS) [51]. In practice, the squared value of MMD is estimated with the empirical kernel mean representations:

$$\begin{aligned} \hat{\mathcal{D}}^{mmd} = & \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} k\left(\phi\left(\boldsymbol{x}_i^s \right), \phi\left(\boldsymbol{x}_j^s \right) \right) + \frac{1}{n_t^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} k\left(\phi\left(\boldsymbol{x}_i^t \right), \phi\left(\boldsymbol{x}_j^t \right) \right) \\ & - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} k\left(\phi\left(\boldsymbol{x}_i^s \right), \phi\left(\boldsymbol{x}_j^t \right) \right), \end{aligned} \tag{4}$$

where $\boldsymbol{x}^s \in \mathcal{S}$, $\boldsymbol{x}^t \in \mathcal{T}$, n_s and n_t denote the batch sizes, and k denotes a kernel function. We adopt the third term and ignore the first two terms in Eq. (4). Due to the existence of multiple modalities, we decompose the gap between two domains into two parts, *i.e.* visual domain gap and textual domain gap, and minimize the corresponding MMD:

$$\mathcal{L}_{CDCA} = \Sigma_{s_1,s_2}\Sigma_{\mathcal{I}^{s_1},\mathcal{I}^{s_2},\mathcal{T}^{s_1},\mathcal{T}^{s_2}} - \frac{2}{n_{s_1}n_{s_2}}\sum_{i=1}^{n_{s_1}}\sum_{j=1}^{n_{s_2}} k\left(\phi\left(\mathcal{I}_i^{s_1}\right), \phi\left(\mathcal{I}_j^{s_2}\right)\right) - \frac{2}{n_{s_1}n_{s_2}}\sum_{i=1}^{n_{s_1}}\sum_{j=1}^{n_{s_2}} k\left(\phi\left(\mathcal{T}_i^{s_1}\right), \phi\left(\mathcal{T}_j^{s_2}\right)\right),$$
$$s_1 \in Dom, s_2 \in Dom\backslash s_1, Dom = \{S_1, S_2, ..., S_K, T\}, \mathcal{I}_i^s \in \mathcal{I}^s \in X_{img}^s \cup X_{img}^{s\prime}, \mathcal{T}_i^s \in \mathcal{T}^s \in X_{txt}^s \cup X_{txt}^{s\prime}, \quad (5)$$

where X_{img}^s and $X_{img}^{s\prime}$ are all the possible visual feature batches in domain s with and without data augmentation, X_{txt}^s and $X_{txt}^{s\prime}$ are are all the possible textual feature batches in domain s with and without data augmentation, $s \in \{s_1, s_2, ..., s_K\}$. We choose the linear kernel as k. Therefore, the above cross-domain contrastive loss can be simplified as below:

$$\mathcal{L}_{CDCA} = \Sigma_{s_1,s_2}\Sigma_{\mathcal{I}^{s_1},\mathcal{I}^{s_2},\mathcal{T}^{s_1},\mathcal{T}^{s_2}} - \frac{2}{n_{s_1}n_{s_2}} \cdot \mathbb{1}^T \cdot \mathcal{I}_{s_1} \cdot \mathcal{I}_{s_2}^T \cdot \mathbb{1} - \frac{2}{n_{s_1}n_{s_2}} \cdot \mathbb{1}^T \cdot \mathcal{T}_{s_1} \cdot \mathcal{T}_{s_2}^T \cdot \mathbb{1}. \quad (6)$$

Since we cannot construct negative sample pairs as in CMCA, optimizing the above function on lower-dimensional space will cause the lower-dimensional projection function g to project all visual and textual features to all-zero features, resulting in feature encoders that are not able to learn any useful pattern. Therefore, the above optimization problem is run with on the original visual and textual features.

2.4 Cross-domain Adversarial Alignment

To better fuse the visual and textual features and produce a multi-modal feature space that includes enough sentiment-related visual-textual information, we choose a bi-linear model [52] to fuse each pair of features from different modalities into factors related with sentiment:

$$f_i = \sum_{p=1}^{N}\sum_{q=1}^{M} w_{ipq}\mathcal{I}^p\mathcal{T}^q + b_i = \mathcal{I}^T\mathbf{W}_i\mathcal{T} + b_i, \quad (7)$$

where $\mathcal{I}$ and $\mathcal{T}$ are visual and textual features, and N and M represent the dimension of feature $\mathcal{I}$ and $\mathcal{T}$, respectively. $\mathbf{W}_i \in \mathbb{R}^{N\times M}$ represents the weight matrix for output f_i, and b_i represents the bias. Assuming the dimension of output feature is L, the number of parameters of bi-linear model is $L \times (N \times M + 1)$ including bias vector b. According to a low-rank bi-linear method which is able to reduce the dimension of the weight matrix [53], the weight matrix can be decomposed into the product of two low-order matrices, which can be described as: $W_i = U_iV_i^T$, where $U_i \in \mathbb{R}^{N\times d}$ and $V_i \in \mathbb{R}^{M\times d}$. Therefore, the output feature f_i can be formalized as:

$$f_i = \mathcal{I}^T\mathbf{W}_i\mathcal{T} + b_i = \mathcal{I}^T\mathbf{U}_i\mathbf{V}_i^T\mathcal{T} + b_i = \mathbb{1}^T\left(\mathbf{U}_i^T\mathcal{I} \circ \mathbf{V}_i^T\mathcal{T}\right) + b_i, \quad (8)$$

where $\mathbb{1}$ represents a column vector consisting of component 1. Still, we need two third-order tensors, $\mathbf{U}$ and $\mathbf{V}$, for a feature vector f, whose elements are $\{f_i\}$. To reduce the order of the weight tensors by one and introduce non-linear activation function, we adopt the following bi-linear pooling function [54]:

$$\mathbf{f} = \mathbf{P}^T\left(\sigma\left(\mathbf{U}^T\mathcal{I}\right) \circ \sigma\left(\mathbf{V}^T\mathcal{T}\right)\right) + h_{img}(\mathcal{I}) + h_{txt}(\mathcal{T}) + \mathbf{b}, \quad (9)$$

where $\mathbf{f}$ represents our multi-modal feature, σ is a non-linear activate function, and h_x and h_y are shortcut mappings.

In order to bridge the domain gap across multiple source domains and the target domain in the fused multi-modal feature space, we construct cross-domain adversarial alignment. Specifically, we introduce a set of domain classifiers as discriminators, which are used to distinguish source features from target features for each source. By assuming that the above encoders and bi-linear pooling layer as a feature extractor, we can construct an adversarial loss [55] and train the feature extractor to generate indistinguishable features that aim to fool the discriminators. This gives the following cross-domain adversarial loss:

$$\begin{aligned}\mathcal{L}_{CDAA} = \Sigma_{i=1}^{K}\{&\mathbb{E}_{(x_{img},x_{txt})\sim(\mathcal{X}_{image},\mathcal{X}_{text})}\log[D_i\big(G(x_{img},x_{txt})\big)] \\ &+ \mathbb{E}_{(x_{img},x_{txt})\sim(\mathcal{X}_{image},\mathcal{X}_{text})}\log[1 - D_i(G(x_{img},x_{txt}))]\},\end{aligned} \quad (10)$$

where G denotes the feature extractor which includes the image and text encoder and the bi-linear pooling layer, we can see $G(x_{img}, x_{txt})$ as the multi-modal feature $\mathbf{f}$, and D_i denotes the discriminator belonging to source i.

2.5 M2CAN Learning

We can train a transferable sentiment classifier over the multi-modal feature space: $f_t : \mathcal{F} \rightarrow \mathcal{Y}$, where $\mathcal{F}$ is the space of multi-modal feature $\mathbf{f}$:

$$\mathcal{L}_{task} = -\mathbb{E}_{x_{img}, x_{txt}, y) \sim (\mathcal{X}^S_{image}, \mathcal{X}^S_{text}, Y_S)} \Big[-\log P\Big(y | f_t\big(G(x_{img}, x_{txt})\big)\Big)\Big]. \tag{11}$$

The final objective function of M2CAN is a weighted combination of different losses:

$$\mathcal{L}_{M2CAN} = \mathcal{L}_{task} + \lambda_1 \cdot \mathcal{L}_{CDAA} + \lambda_2 \cdot \mathcal{L}_{CMCA} + \lambda_3 \cdot \mathcal{L}_{CDCA}, \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$ are weights for different losses. This objective function can be optimized by solving the following min-max game:

$$f_t^* = \arg\min_{f_t} \min_{G} \max_{D_1, D_2} \mathcal{L}_{\text{M2CAN}}. \tag{13}$$

3 Experiments

Here we introduce the experimental settings and compare the sentiment classification results of M2CAN and several state-of-the-art DA approaches, followed by ablation study and visualization.

3.1 Experimental Settings

Datasets. Since we are the first to study the novel MS-MMDA setting and there is no specific dataset on this task, we evaluate our approach using a combined dataset, which consists of three public datasets on visual-textual sentiment: Yelp [12], Twitter [45], and MVSA [46]. We regard the three datasets as different domains since they follow different distributions. We create multiple MS-MMDA settings by taking each domain as *target* and the rest as *sources* in each setting.

The **Yelp domain** [12] contains customer-generated reviews of food services, *e.g.* restaurants, cafeterias, and dessert shops. In total, it has more than 44,000 reviews, including 244,000 images. Each review has a piece of textual comment, at least 3 images, and a score of sentiment polarity ranging from 1 to 5. We consider those reviews with scores of 1 and 2 as carrying negative sentiment, those with scores of 3 as carrying neutral sentiment, and those with scores of 4 and 5 as carrying positive sentiment. The **Twitter domain** [45] contains 50,000 user-generated tweets with images released on Twitter. Each tweet is composed of one textual review, several images, and a three-type sentiment label: negative, neutral, and positive. The **MVSA domain** [46] is also collected from Twitter. Similar to the Twitter domain, each tweet in MVSA consists of one textual review, several images, and a three-type sentiment label: negative, neutral, and positive. Specifically, each tweet is annotated by three experts. We abandon the tweets annotated with three different labels, and keep the tweets with at least two agreements. To balance the amount of samples in different domains, we randomly choose 15,000 samples as training set and 1,500 samples as test set for all domains.

Evaluation Metrics. Following [32, 29], we employ classification accuracy to evaluate the multi-modal sentiment classification results. Larger classification accuracy indicates better performance.

Baselines. We compare M2CAN with the following baselines: (1) **Source-only**, directly training on the source domains and testing on the target domain, which includes two settings: single-best, the best test accuracy on target among all source domains; source-combined, the target accuracy of the model trained on the combined source domain. (2) **Single-source MMDA methods**, including state-of-the-art approaches MMAN [23], MM-SADA [25], and xMUDA [26] trained with both single-best and source-combined settings. (3) **Multi-source MMDA methods**, including the state-of-the-art approach MDAN [32] and the proposed M2CAN. We also report the results of an oracle setting, where the model is both trained and tested on the target domain. We can view the oracle results as an upper bound for domain adaptation.

Implementation Details. For the image encoder, we use Resnet-50 [47]. For the text encoder, we use a 12-layer "bert-base-uncased" version BERT [48]. The weights for $\mathcal{L}_{\text{task}}$, $\mathcal{L}_{\text{GAN}}$, $\mathcal{L}_{\text{CMCA}}$, and $\mathcal{L}_{\text{CDCA}}$ are 1, 0.02, 0.02 and 0.05, respectively. We use a 2-layer multi-layer perceptron (MLP) to implement the lower-dimensional projection function g, and a fully-connected layer with activation function ReLU to implement both the discriminators and the task classifier. We use Adam [56] as the

Table 1: Comparison with the state-of-the-art DA methods on the combined dataset for visual-textual sentiment classification. All results are percentages. The best and second best classification accuracies trained on the source domains are emphasized with bold and underline respectively (same in Table 2).

Standards	Models	Yelp	Twitter	MVSA	Avg
Source-only	Source-combined (text only)	57.3	62.2	55.8	58.4
	Source-combined (text & image)	56.7	59.1	57.8	57.9
	Single-best (text & image)	56.9	61.8	57.2	58.6
Single-best MMDA	MMAN [23]	57.5	62.5	58.8	59.6
	MM-SADA [25]	58.1	66.0	58.3	60.8
	xMUDA [26]	58.7	64.3	57.6	60.2
Source-combined MMDA	MMAN [23]	55.8	64.2	60.8	60.3
	MM-SADA [25]	57.5	63.2	60.3	60.3
	xMUDA [26]	56.2	63.1	**62.0**	60.4
Multi-source MMDA	MDAN [32]	58.6	64.1	61.2	61.3
	M2CAN (Ours)	**62.5**	**67.9**	61.6	**64.0**
Oracle (train on target)		65.2	68.6	68.4	67.4

optimizer with a batch size of 8. The learning rate is 0.00002 for BERT and Resnet-50, and 0.0005 for the rest. All experiments are implemented in PyTorch and conducted on a machine with a Tesla V100S-PCIE GPU with 32 GB memory. All implementation details are included in our source code.

3.2 Comparison With State-of-the-art

The performance comparisons between the proposed M2CAN and the baselines for visual-textual sentiment classification, including source-only, single-source MMDA, and multi-source MMDA, are shown in Table 1. From the results, we have the following observations:

(1) Without alleviating the domain shift between the source and target domains, both source-only settings, *i.e.* single-best and source-combined, obtain poor classification accuracies, *i.e.* 58.6% and 57.8%, which are almost 10% worse than the oracle setting (67.4%). When setting Yelp and Twitter as the target domain, it is clear that adding visual modality results in performance degradation for both single-best and source-combined source-only settings as compared to using textual modality only, *e.g.* 56.9% and 56.7 % vs. 57.3% on Yelp. This indicates that the large domain gap between source and target domains results in severe interference between different modalities. These observations motivate the research on domain adaptation.

(2) When directly applying to the MS-MMDA task, both single-best and source-combined MMDA methods outperform the source-only setting. Since customers' reviews vary a lot across domains, features that are related to sentiment also differ. Therefore, these MMDA methods that can mitigate the domain gap improve the sentiment classification results.

(3) Comparing the performances of source-combined and single-best MMDA methods, we can find that naively performing single-source domain adaptation approaches on a combined dataset of different sources might produce worse result (*i.e.* 60.3% of MM-SADA) than that on the best single source (*i.e.* 60.8% of MM-SADA). This motivates our research on MS-MMDA.

(4) The proposed M2CAN performs the best (64.0%) among all adaptation settings. Compared to the best results inside the source-only, single-best MMDA, source-combined MMDA, and multi-source MMDA, M2CAN achieves 5.4%, 3.2%, 3.6%, and 2.7% performance gains, respectively. These results demonstrate that the proposed M2CAN model can achieve significantly better performance than the state-of-the-art DA methods for visual-textual sentiment classification. The superior performance of M2CAN benefits from the joint cross-modal alignment and cross-domain alignment.

3.3 Ablation Study

We conduct a series of ablation studies on the combined dataset to demonstrate the effectiveness of different components of M2CAN. The results are shown in Table 2. First, we verify the necessity of introducing extra modalities. Comparing the first two lines between with text only and with image

Table 2: Ablation study on different components of the proposed M2CAN on the combined dataset.

Models	Yelp	Twitter	MVSA	Avg
CDAA (text only)	57.2	62.8	58.1	59.4
CDAA (text & image)	58.5	64.2	<u>61.7</u>	61.5
CDAA + CMCA	61.1	64.5	61.6	62.4
CDAA + CDCA	<u>61.3</u>	<u>65.2</u>	**61.9**	<u>62.8</u>
CDAA + CMCA + CDCA (M2CAN)	**62.5**	**67.9**	61.6	**64.0**

and text for CDAA, we can see that after adding image, the average accuracy is improved by 2.1%, demonstrating the effectiveness of introducing multiple modalities. Second, we investigate whether it is necessary to construct the cross-domain adversarial alignment (CDAA). Comparing the first two lines in Table 2 and Table 1, it is clear the performance is significantly improved (*e.g.* 61.5% vs. 57.9% when both image and text are used), demonstrating the effectiveness of adversarial alignment. Third, we investigate the effectiveness of cross-modal contrastive alignment (CMCA). From the third and second lines, we can see that compared to only using CDAA, adding CMCA achieves 0.9% performance gain on the average classification accuracy, which demonstrates the necessity of the CMCA. Finally, we evaluate the influence of cross-domain contrastive alignment (CDCA). Comparing CDAA vs. CDAA+CDCA and CDAA+CMCA vs. CDAA+CMCA+CDCA, we can conclude that adding CDCA can further improve the performance, verifying that CDCA indeed contributes to the adaptation task.

3.4 Visualization

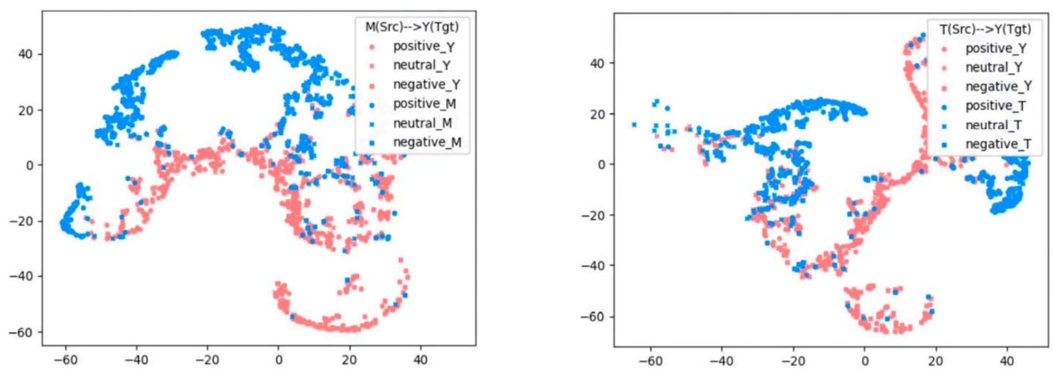
In this section, we visualize the features of source and target samples before and after adaptation using M2CAN. By using t-SNE [57] to reduce the dimensionality of samples, we plot the learned features onto a 2-dimensional plane, with the results shown in Figure 2. Figure (a) represents the feature representations before adaptation, while (b) represents the feature representations after adaptation by M2CAN. Red represents source features and blue represents target features. As we can see, before adaptation, source and target features can be obviously discriminated because of the existence of domain gap; while after adaptation, we can hardly distinguish between source and target features. Therefore, we can conclude that after adaptation the source and target features become more closely aligned, which further demonstrates the effectiveness of M2CAN. Furthermore, we also plot the loss curves in the training process. From Figure 3, we can observe that the different types of losses all decline and converge through the training process.

3.5 Limitations

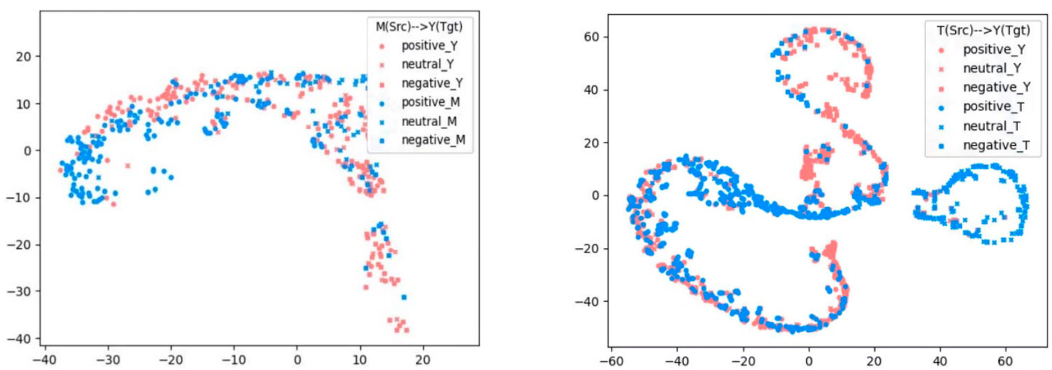
The proposed M2CAN works under the covariate shift and closed set assumptions [18, 22] with labeled source data and unlabeled target data. When other domain shifts exist, such as label shift [22] and category shift [39], we cannot guarantee satisfactory domain adaptation performances. As stated in [22], there are many different domain adaptation settings, *i.e.* multiple target domains and open-set labels. The proposed method does not explore such characteristics and thus cannot be directly applied to these settings. Incorporating existing multi-source techniques, such as adding discrimination between different sources [40], would improve the performance. Considering the constraint of hardware resources, such as GPU memory, we did not exploit such techniques. Further, because of the absence of datasets for multi-source multi-modal domain adaptation, we only verify the effectiveness of M2CAN on a combined dataset with three different domains for visual-textual sentiment classification. Extending the proposed method to other multi-modal domain adaptation with multiple sources and exploring how to perform MS-MMDA when some sources contain few labeled and sufficient unlabeled data remains our future work.

4 Conclusion

In this paper, we studied a novel and practical domain adaptation problem, *i.e.* multi-source multi-modal domain adaptation (MS-MMDA), for visual-textual sentiment classification. The designed multi-source multi-modal contrastive adversarial network (M2CAN) can learn domain-invariant multi-



(a) Before adaptation



(a) After adaptation

Figure 2: t-SNE visualization of multi-modal features before and after adaptation. Red represents source features and blue represents target features. We use Y, T, and M as abbreviations respectively for domains Yelp, Twitter, and MVSA for better visualization.

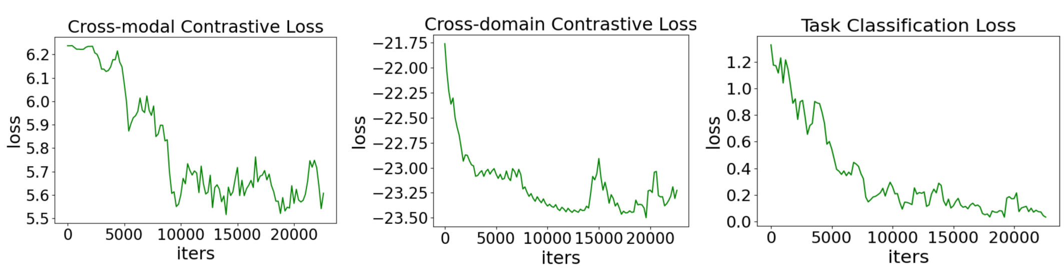


Figure 3: Visualization of different losses during the training process, including cross-modal contrastive loss, cross-domain contrastive loss and the task classification loss.

modal features by three different alignment strategies, *i.e.* cross-modal contrastive alignment within each domain, cross-domain contrastive alignment for each modality, and cross-domain adversarial alignment on the fused multi-modal representation. The cross-modal contrastive loss aligns visual and textual features, pulling semantic-related sample pairs closer and pushing semantic-unrelated sample pairs farther. The cross-domain contrastive loss together with domain adversarial loss bridge the domain gap between the source and target domains while preserving the sentiment semantics through contrastive learning and adversarial learning, respectively. Extensive experiments on a combined dataset demonstrate the superiority of the proposed M2CAN as compared to the state-of-the-art DA methods for visual-textual sentiment classification.

References

[1] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. Predicting personalized emotion perceptions of social images. In *ACM International Conference on Multimedia*, pages 1385–1394, 2016.

[2] Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *International World Wide Web Conference*, pages 1045–1052, 2017.

[3] Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. A feature-oriented sentiment rating for mobile app reviews. In *International World Wide Web Conference*, pages 1909–1918, 2018.

[4] Lin Gong and Hongning Wang. When sentiment analysis meets social network: A holistic user behavior modeling in opinionated data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1455–1464, 2018.

[5] Niru Maheswaranathan, Alex H Williams, Matthew D Golub, Surya Ganguli, and David Sussillo. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Advances in Neural Information Processing Systems*, pages 15670–15679, 2019.

[6] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *AAAI Conference on Artificial Intelligence*, pages 303–311, 2020.

[7] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. Emoji-powered representation learning for cross-lingual sentiment classification. In *International World Wide Web Conference*, pages 251–262, 2019.

[8] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In *ACM International Conference on Multimedia*, pages 1008–1017, 2016.

[9] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[10] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[11] Jie Xu, Feiran Huang, Xiaoming Zhang, Senzhang Wang, Chaozhuo Li, Zhoujun Li, and Yueying He. Visual-textual sentiment classification with bi-directional multi-level attention networks. *Knowledge-Based Systems*, 178:61–73, 2019.

[12] Quoc-Tuan Truong and Hady W Lauw. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence*, pages 305–312, 2019.

[13] Feiran Huang, Kaimin Wei, Jian Weng, and Zhoujun Li. Attention-based modality-gated networks for image-text sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(3):1–19, 2020.

[14] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1994–2003, 2018.

[16] Luyu Yang, Yogesh Balaji, Ser-Nam Lim, and Abhinav Shrivastava. Curriculum manager for source selection in multi-source domain adaptation. In *European Conference on Computer Vision*, pages 608–624, 2020.

[17] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[18] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.

[19] Shiliang Sun, Honglei Shi, and Yuanbin Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24:84–92, 2015.

[20] Wouter Marco Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[21] Sicheng Zhao, Bo Li, Colorado Reed, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020.

[22] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, and Kurt Keutzer. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[23] Xinhong Ma, Tianzhu Zhang, and Changsheng Xu. Deep multi-modality adversarial networks for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 21(9):2419–2431, 2019.

[24] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *ACM International Conference on Multimedia*, pages 429–437, 2018.

[25] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020.

[26] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12605–12614, 2020.

[27] Huaiwen Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. Multimodal disentangled domain adaption for social media event rumor detection. *IEEE Transactions on Multimedia*, 2020.

[28] Massimiliano Mancini, Lorenzo Porzi, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Boosting domain adaptation by discovering latent domains. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3771–3780, 2018.

[29] Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *Conference on Empirical Methods on Natural Language Processing*, pages 4694–4703, 2018.

[30] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pages 8246–8256, 2018.

[31] Yitong Li, Michael Murias, Samantha Major, Geraldine Dawson, and David E Carlson. Extracting relationships by multi-domain matching. In *Advances in Neural Information Processing Systems*, pages 6799–6810, 2018.

[32] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, pages 8568–8579, 2018.

[33] Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *AAAI Conference on Artificial Intelligence*, pages 5989–5996, 2019.

[34] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

[35] Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Multi-source domain adaptation for text classification via distancenet-bandits. In *AAAI Conference on Artificial Intelligence*, pages 7830–7838, 2020.

[36] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. In *AAAI Conference on Artificial Intelligence*, pages 12975–12983, 2020.

[37] Junfeng Wen, Russell Greiner, and Dale Schuurmans. Domain aggregation networks for multi-source domain adaptation. In *International Conference on Machine Learning*, pages 10214–10224, 2020.

[38] Naveen Venkat, Jogendra Nath Kundu, Durgesh Kumar Singh, Ambareesh Revanur, and R Venkatesh Babu. Your classifier can secretly suffice multi-source domain adaptation. In *Advances in Neural Information Processing Systems*, 2020.

[39] Ziliang Chen, Pengxu Wei, Jingyu Zhuang, Guanbin Li, and Liang Lin. Deep cocktail networks. *International Journal of Computer Vision*, pages 1–24, 2021.

[40] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 7285–7298, 2019.

[41] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua. Multi-source domain adaptation for visual sentiment classification. In *AAAI Conference on Artificial Intelligence*, pages 2661–2668, 2020.

[42] Sicheng Zhao, Yang Xiao, Jiang Guo, Xiangyu Yue, Jufeng Yang, Ravi Krishna, Pengfei Xu, and Kurt Keutzer. Curriculum cyclegan for textual sentiment domain adaptation with multiple sources. In *The Web Conference*, 2021.

[43] Jianzhong He, Xu Jia, Shuaijun Chen, and Jianzhuang Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[44] Sicheng Zhao, Bo Li, Pengfei Xu, Xiangyu Yue, Guiguang Ding, and Kurt Keutzer. Madan: multi-source adversarial domain aggregation network for domain adaptation. *International Journal of Computer Vision*, pages 1–26, 2021.

[45] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. In *AAAI Conference on Artificial Intelligence*, pages 5674–5681, 2018.

[46] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*, pages 15–27, 2016.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[48] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[49] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[50] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[51] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019.

[52] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[53] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. Bilinear classifiers for visual recognition. In *Advances in Neural Information Processing Systems*, pages 1482–1490, 2009.

[54] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations*, 2017.

[55] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[56] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[57] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 3.5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplemental material.
   (b) Did you specify all the training details (*e.g.*, data splits, hyperparameters, how they were chosen)? [Yes] See Section 3.1–Implementation Details.
   (c) Did you report error bars (*e.g.*, with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (*e.g.*, type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 3.1–Implementation Details.
4. If you are using existing assets (*e.g.*, code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]