Borrowing Human Senses: Comment-Aware Self-Training for Social Media Multimodal Classification

Chunpu Xu, Jing Li*
Department of Computing, The Hong Kong Polytechnic University, China
chun-pu.xu@connect.polyu.hk
jing-amelia.li@polyu.edu.hk

Abstract

Social media is daily creating massive multimedia content with paired image and text, presenting the pressing need to automate the vision and language understanding for various multimodal classification tasks. Compared to the commonly researched visual-lingual data, social media posts tend to exhibit more implicit image-text relations. To better glue the cross-modal semantics therein, we capture hinting features from user comments, which are retrieved via jointly leveraging visual and lingual similarity. Afterwards, the classification tasks are explored via self-training in a teacher-student framework, motivated by the usually limited labeled data scales in existing benchmarks. Substantial experiments are conducted on four multimodal social media benchmarks for image-text relation classification, sarcasm detection, sentiment classification, and hate speech detection. The results show that our method further advances the performance of previous state-of-the-art models, which do not employ comment modeling or self-training.

1 Introduction

Interpersonal communications in multimedia are gaining growing popularity on social media. More and more social media users are turning to pair images to text and vice versa to better voice opinions, exchange information, and share ideas, exhibiting rich and ever-updating resources in multimedia. While potentially benefiting people's everyday decision making, the huge volume of multimedia content might also challenge users in finding what they need. Towards a more efficient and effective way to process the online multimodal data, substantial efforts have been made to automatically understand the vision and language on social media through a broad range of multimodal classification tasks for predicting image-text relations (Vempala and Preotiuc-Pietro, 2019), sarcasm (Cai et al., 2019), metaphor (Zhang et al., 2021), point-of-interest (Villegas and Aletras, 2021), hate speech (Botelho et al., 2021), sentiment (Yu and Jiang, 2019), etc.

* Corresponding author

Text:
was going to take mollie for a walk in the park today . pennsylvania weather cooperated as per usual

Retrieved comments:
1. damn ! that 's a lot of snow to still be around for this time of year , no
2. snow in your area we 'd love to see pictures of it !
3. heavier snow now shifting to your east
…

Figure 1: A sample tweet with its image on the left. On the right, the tweet text is shown on the top, followed by the comments retrieved from similar tweets. The word "snow" (in blue) in comments helpfully hint the implicitly shared semantics between image and text.

Despite the success of visual-lingual understanding witnessed in common domains (Huang et al., 2020; Shi et al., 2020; Wang et al., 2021), existing models' performance is likely to be compromised on social media posts. The possible reason lies in the relatively more implicit and obscure image-text relations therein (Vempala and Preotiuc-Pietro, 2019), whereas the image-text pairs in the widely-used datasets outside social media (e.g., COCO dataset (Lin et al., 2014), VQA dataset (Antol et al., 2015), VCR dataset (Zellers et al., 2019)) tend to present explicit information overlap. Such issue is nevertheless ignored in many previous solutions, which follow the common practice to fuse visual and lingual features (Vempala and Preotiuc-Pietro, 2019; Zhang et al., 2021; Hessel and Lee, 2020; Botelho et al., 2021), making it hard for a multimodal model to well align cross-modal semantics attributed to their weak correlations (Fei et al., 2022).

Nonetheless, human readers seem to have no problem in digesting the cross-modal meanings on social media; in response to what they capture

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5644–5656
December 7-11, 2022 ©2022 Association for Computational Linguistics

from a post, some readers may drop user comments, where some clues to hint cross-modal understanding may be hidden, e.g., an echo of the keypoints. For instance, in Figure 1, "snow" in retrieved comments might strengthen the connection of "weather" in text with the snowing visual scene. The helpfulness of user comments has also been demonstrated in the previous NLP practice for text-only posts (Wang et al., 2019). Inspired by that, we propose "borrowing" the senses from human readers and modeling user comments to learn the hinting features therein to bridge the image-text gap. To further benefit posts without user comments, a comment retrieval algorithm is designed to gather comments from other posts in similarity, which is measured via balancing the visual and lingual semantics (henceforth **cross-modal similarity**). For the related experimental studies, a large-scale dataset is constructed to mimic the open environment (henceforth **wild dataset**). It contains over 27M multimodal tweets, each with 3 comments on average.

Then, we explore how to leverage the retrieved comments in multimodal classification and exploit a self-training framework to identify comments' hints which shape the cross-modal understanding (henceforth **comment-aware self-training**). This considers method feasibility in scenarios where large-scale labeled data is unavailable, which commonly appears in the realistic practice, because the annotation for multimodal data from social media is extremely expensive (Ma et al., 2019). Concretely, we adopt a teacher-student prototype (Meng et al., 2020; Shen et al., 2021) and tailor-make it to learn multimodal understanding with the help of user comments. A teacher model is first trained with the labeled data and pseudo-label the similar posts with comments retrieved from the wild dataset. Then, a student model is trained in guidance of both the knowledge gained by the teacher model and hinting features offered by user comments.

To evaluate our method in practice, it is comprehensively experimented on four popular social media multimodal benchmarks for varying classification tasks. In the setup of each benchmark, our comment-aware self-training module is customized to BERT-based state-of-the-art (SOTA) architectures. Ablation studies then exhibit the individual benefits provided by comments and self-training. Then, we analyze the effects of retrieved post number and find the use of more posts would result in both the benefits and noise. Next, we probe into comment retrieval and explore the contributions of visual and lingual modality in cross-modal similarity measure, where we observe the joint effects allow the best results. At last, a case study shows how the retrieved comments mitigate cross-modal semantic gap, followed by the an error analysis to discuss the existing limitation.

In summary, our contributions are three fold.

• We demonstrate the potential to employ retrieved user comments from similar posts for a better visual-lingual understanding on social media and gather 27M cross-media tweets with comments to released to support future research in this line. [1]

• A comment-aware self-training method is proposed for cross-modal learning from both human senses underlying retrieved comments and knowledge distilled from labeled data in limited scales.

• An empirical study with substantial results is provided, where SOTA models of four popular social media benchmarks for multimodal classification perform better with the help of our comment-aware self-training module and the retrieved comments bridge social media images and text via hinting the connecting points for models to attend to.

2 Related Work

Our work is in line with multimodal learning and self-training, which are discussed below in turn.

Multimodal Learning. Previous work in this field focuses on fusing features from different modalities (e.g., vision and language) (Baltrusaitis et al., 2019) to tackle cross-modal classification tasks, such as visual question answering (VQA) (Tapaswi et al., 2016; Goyal et al., 2017; Johnson et al., 2017), visual commonsense reasoning (Zellers et al., 2019; Lu et al., 2019), and image-text retrieval (Lee et al., 2018; Li et al., 2020a,b). Most benchmark data in vision and language assumes strong image-text correlations, and many multimodal models are hence designed to explore the common semantics shared by the two modalities. However, it has been recently pointed out that many real-world scenarios, including social media, tend to present image-text pairs with weak and intricate cross-modal interactions (Vempala and Preotiuc-Pietro, 2019; Hessel et al., 2019; Fei et al., 2022).

Despite the substantial efforts made in applying multimodal learning in social media to tackle var-

[1]Our code and dataset are released at https://github.com/cpaaax/Multimodal_CAST.

ious visual-lingual tasks, e.g., sarcasm detection (Cai et al., 2019), hate speech detection (Botelho et al., 2021), metaphor detection (Zhang et al., 2021), etc., most existing methods follow the common practice to fuse visual and lingual features. It is hence challenging for them to figure out the cross-modal meanings exhibit with implicit image-text links. We thus resort to the retrieved user comments and study how models can find the hinting features therein to mitigate the cross-modal gaps.

Our work is also related to Gur et al. (2021), where the retrieved similar data show helpful in better aligning visual-lingual features for multi-modal classification. Different from them, we additional retrieve comments and learn the hints therein via self-training, which allows easy integration to most multi-modal classification archetectures.

Self-training. Our method is inspired by previous work in self-training (Scudder, 1965; Yarowsky, 1995), where labeled data is employed to train models and generate pseudo-labels for unlabeled data. It is simple yet effective to enable model robustness with limited labeled data, potentially helpful in multimodal social media tasks owing to the expensive annotation and small-scale labeled data in most benchmarks (Vempala and Preotiuc-Pietro, 2019; Botelho et al., 2021).

Here we adopt the trendy self-training paradigm in a teacher-student framework, where a teacher model is trained with labeled data and self-label unlabeled data to generate synthetic data for the student model training. It has been widely used in many tasks in CV (e.g., image classification (Xie et al., 2020; Zoph et al., 2020), object detection (Yang et al., 2021)) and NLP (e.g, question answering (Sachan and Xing, 2018; Zhang and Bansal, 2019; Rennie et al., 2020), text classification (Mukherjee and Awadallah, 2020; Meng et al., 2020; Shen et al., 2021)). However, in existing work, limited attention has been drawn to its effectiveness in social media multimodal classification and how it works with retrieved comments to gain the cross-modal understanding for noisy data.

3 Comment-Aware Self-Training

This section presents the entire comment-aware self-training workflow illustrated in Figure 2. In the following, we first describe how we gather the wild dataset for retrieval, followed by the related analysis in §3.1. Then we introduce the retrieval module to find similar posts and their comments in §3.2. Next, we describe the usage of comments in the multimodal architecture in §3.3. At last, §3.4 presents the comment-aware self-training framework based on retrieved results.

Year	Num	Text len	Com len	Com num
2014	3,178,845	12.88	8.81	3.27
2015	6,373,198	13.24	8.19	3.32
2016	6,230,437	13.69	8.98	3.13
2017	4,583,203	11.37	8.55	3.37
2018	3,370,186	10.95	9.01	3.41
2019	4,048,959	10.46	8.35	3.06
Total	27,784,828	12.31	8.62	3.25

Table 1: Statistics of the wild dataset for retrieval. **Text len** and **Com num** indicate the average length (token number) in text and average comment number per tweet. **Com len** is the average length per comment.

3.1 Wild Dataset Construction for Retrieval

To simulate retrieval from the open environment, large-scale visual-lingual tweets with comments are gathered to form a wild dataset. The detailed steps are described in the following. First, we downloaded the large-scale corpus used by Nguyen et al. (2020) to pre-train their BERTweet. The dataset contains general Twitter streams and we removed non-English tweets with fastText library (Joulin et al., 2017). Then, text-only tweets were removed and for the remaining 60 million, images and comments were gathered using Twitter stream API.[2] Next, user mentions and url links are converted to generic tags @USER and HTTPURL for privacy concern. Finally, tweets without comments were further removed, resulting in 27.8 million tweets with images, text, and comments.

The over-year statistics of wild dataset is shown in Table 1. There exhibits an interesting observation — the average text length in recent years (2017-2019) is obviously shorter than earlier. It might be because images these years allow the provision of gradually richer information, partially taking the text's roles in multimedia communications.

3.2 Post and Comment Retrieval

Given a post in image-text pair, we then discuss how to retrieve similar posts and their comments.

Retrieval of Similar Posts Post similarity is measured via balancing the effects of the image and text features. The former is learned with ResNet-152 (He et al., 2016) pre-trained on the ImageNet

[2]https://developer.twitter.com/en/docs/twitter-api

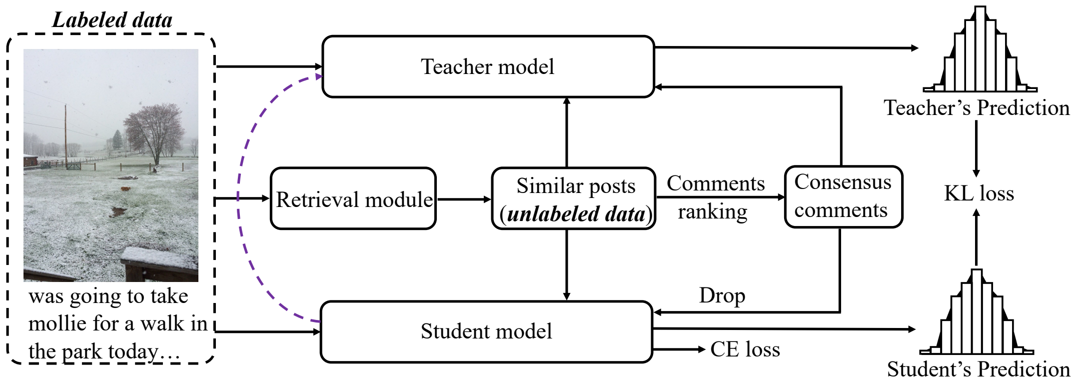

Figure 2: The workflow of comment-aware self-training. Given a post (image-text pair), we first query similar posts and their comments in a retrieval module. Then the retrieved data is employed in teacher-student training as unlabeled data, where student model is trained with CE (cross-entropy) and Kullback–Leibler (KL) divergence loss.

(Russakovsky et al., 2015) and we take the output of final pooling layer for representation. For text features, SimCSE is adopted because of its effectiveness in similarity measure (Gao et al., 2021). For any post (the query), we score its similarity to the i-th multimedia post in the wild dataset with s_i:

$$s_i = \alpha s_i^I + (1-\alpha) s_i^T \tag{1}$$

where s_i^I and s_i^T respectively indicates the image and text similarity, traded-off by the parameter α.

Here the weight α (balancing image and text effects) is empirically estimated with the averaged statistics measured over the wild datasets (for retrieval) and all experimental data (as a query set).

$$\alpha = \frac{T_{mean}}{I_{mean} + T_{mean}} \tag{2}$$

$$I_{mean} = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} p_{m,k} \tag{3}$$

$$T_{mean} = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} q_{m,k} \tag{4}$$

where $p_{m,k}$ and $q_{m,k}$ respectively refer to the image and text similarity between the m-th query and its rank-k most similar post retrieved with the corresponding modality. M is the query set size and K the cut-off number of retrieved posts to be selected.

In this way, for any query, we rank posts in wild dataset by the s_i score (leveraging image and text similarity) and the top-K most similar posts will be retrieved.

Here the Faiss library (Johnson et al., 2019) is employed for fast similarity search. Concretely, the Inverted File Index Product Quantization(IVFPQ) (Johnson et al., 2019) index is built for feature vector compression and efficient kNN queries. First, IVFPQ index is trained with the images features to perform clustering on the original feature vectors. Then, given a query image features, IVFPQ would return R-nearest[3] image indexes and related image similarity score. Similarly, R-nearest text indexes and related text similarity score could be obtained. At last, to allow an efficient search, the post related to the overlap indexes between text indexes and image indexes would be regarded as candidate similar posts, and used for searching the final similar posts by Eq.1.

Retrieval of Comments As discussed above, comments (written by human readers) potentially help in hinting the visual-lingual relations. However, while building the wild dataset, we observe only 46% multi-modal tweets contain comments. For those posts where comments are absent or inaccessible, the comments of the retrieved posts may be useful as well, because intuitively, similar posts may result in similar comments.

However, because of the noisy nature of social media data, comments may vary in their quality and effects in hinting cross-modal learning. We therefore need to shortlist high-quality comments.

Concretely, from the comment pool P gathering all retrieved posts' comments, we follow Devlin et al. (2015) to extract representative comments as consensus comments. They tend to exhibit relatively higher semantic similarity to other comments in P and can be ranked with:

$$q_i = \frac{1}{|P|} \sum_{p' \in P} Sim\left(p_i, p'\right) \tag{5}$$

where q_i is the consensus score of the i-th comment $p_i \in P$. $Sim\left(p_i, p'\right)$ indicates the p_i-p' similarity (with SimCSE). $|P|$ is P's comment size.

[3] R is set to 0.1M in the experiments.

In practice, for any post, we query and retrieve the top-N consensus comments to form a set C later used to bridge the gap between image modality and text modality (next discussed in §3.3).

3.3 Leveraging Comments in a Multimodal Classification Architecture

The previous discussions concern how to retrieve the similar posts and their (consensus) comments. In the following, we describe the implementation details of how comments can be leveraged in different multimodal classification architectures. Based on the different schemes for fusing image features and text features (Alberti et al., 2019), most multimodal classifiers could be divided into early fusion and late fusion. In the early fusion sheme, image features are embedded with text tokens on the same level (e.g., MMBT). In the late fusion sheme, image features and text features are encoded separately and interacted by the concatenation (e.g., BERT-CNN) or the attention mechanism (e.g., CoMN-BERT and MMSD-BERT).

Here we give the details of using comments to bridge the gap between images and texts without changing the original architecture of base classifiers. The comments injection methods are slightly different for early fusion and late fusion schemes, where the details come in the following.

Early Fusion Scheme The image features are projected into token space and concated with the word token embeddings as the input of multimodal bitransformer in MMBT. Then the hidden state h^f related to the [CLS] token is used as the representation of the fusion vector for the classification in MMBT. We adopt the same encoding strategy to fuse each comment and the image features, and related hidden states $\{h_1^c, ..., h_N^c\}$ are obtained. After that, the attention mechanism is used to compute the attended vector u, which is concated with the original fusion vector h^f for the classification:

$$u = \sum_{n=1}^{N} \beta_n h_n^c \tag{6}$$

$$\beta_i = \frac{\exp(z_n)}{\sum_{n=1}^{N} \exp(z_n)}; \quad z_n = \sigma(h^f, h_n^c) \tag{7}$$

where σ is a feed-forward neural network.

Late Fusion Scheme Assume the image features extracted by ResNet, text features extracted by BERT, and the original fusion vector, which is the output of the base classifier before the softmax layer, are h^v, h^t, and h^f, separately. And the comments features encoded by the BERT are denoted as $\{h_1^c, ..., h_N^c\}$. Similar to Eq.6, then attention mechanism is employed to fuse the image features and comment features and obtain the image attended vector v. Similarly, the text attended vector t which is fused by text features and comment features, could be acquired. At last, we concate the image attended vector v, text attended vector t, and original fusion vector h^f for the final classification.

3.4 Self-training with Retrieved Posts

Here we further describe how the retrieved posts (i.e., the retrieved image-text pairs) is explored in multimodal classification. Its data is commonly formulated as a labeled parallel dataset $L = \{x_i, c_i, y_i\}_{i=1}^{l}$, where x_i is an image-text pair, c_i indicates the retrieved N comments, and y_i a label specified by the task.

The labeled dataset L is usually limited in scales (Ma et al., 2019), posing the over-fitting concern. Meanwhile the retrieved posts, similar to the data in L, could form an unlabeled set ($U = \{x_i^{'}, c_i\}_{i=1}^{Kl}$) to enrich training data. Note that $x_i^{'}$, one retrieved image-text pair of x_i, shares the same consensus comments c_i with x_i. Then L and U may be integrated to allow more robust learning in a semi-supervised manner.

Here we adopt self-training based on the popular teacher-student framework (Xie et al., 2020). A teacher model (the classifier) is first trained on the labeled data L to gain task-specific knowledge and pseudo-label the unlabeled data with soft labels as "teaching samples". Then a student model, sharing the same architecture as the teacher, is trained with both the pseudo-labeled U and labeled L.

In the training of both teacher and student, their modality fusion mechanism is fed with the comment features (described in §3.4). It enables the models to explore cross-modal interactions in aware of the comments.

In practice, the student training randomly drops 50% comments while teacher employs the full comment set. It enables the student model to learn from the noised data for a better generalization instead of simply mimicking teacher's behavior.

The teacher model is trained with the cross-entropy loss for classification while KL divergence loss is additionally used for student training:

Dataset	#Train	#Val	#Test	#All
MVSA	3,611	451	451	4,511
ITR	3,575	447	449	4,471
MSD	19,816	2,410	2,409	24,635
MHP	3,998	500	502	5,000

Table 2: Statistics of the evaluation datasets.

$$\mathcal{L} = \frac{1}{|L|}\sum_{i \in L} y_i \log y_i + \frac{1}{|U|}\sum_{i \in U} \mathrm{KL}(t_i || s_i) \quad (8)$$

where $|L|$ and $|U|$ indicate L's and U's dataset size. t_i is the soft label predicted by the teacher model while s_i is the output of the student model.

4 Experimental Setup

4.1 Evaluation Datasets

Our evaluation is conducted on four Twitter classification benchmarks on multimodal sentiment classification (MVSA) (Niu et al., 2016), image-text relation (ITR) (Vempala and Preotiuc-Pietro, 2019), multimodal sarcasm detection (MSD) (Cai et al., 2019), and multimodal hate speech detection (MHP) (Botelho et al., 2021). Each data instance is an image-text pair and it is annotated with a single class label. For MVSA, MHP and MSD, we adopt the same dataset split as their original papers for fair comparisons. For ITR, we randomly split 80%, 10% and, 10% for training, validation, and test instead of their 10-fold cross-validation setup for the concern of experimental efficiency.

The statistics of evaluation datasets are shown in Table 2, where we observe the small-scale training data in MVSA, ITR, and MHP. For MVSA, though relatively larger in scales, its automatic labeling under hashtag-based distant supervision, may result in noisy labels, which further require larger data scales for robustness. These imply the annotation difficulties and potential benefits from self-training.

4.2 Implementation and Evaluation Details

All experimental models are implements with PyTorch[4] and HuggingFace Transformers[5].

Both text and comment are capped at 50 words for encoding. The batch size is set to 8, 8, 16, and 16 for ITR, MHP, MVSA, MSD. The learning rate is set to 1e-5 with a warm-up rate to 0.1. Classifiers are trained with an AdamW optimizer. The maximum of consensus comments (N) is set to 5. We run the self-training for three iterations. At each iteration, the teacher model is fine-tuned for 10 epochs on the labeled training data. The teacher model performing the best in validation is adopted to predict the pseudo-labels for the unlabeled retrieved data. The student model is then fine-tuned for 10 epochs. After that, the student model is used as teacher for the next iteration.

For evaluation metrics, we follow the benchmark practice to use precision (pre), recall (rec), and F1-score (F1) for ITR, MVSD, and MHP, and accuracy (acc) and F1 for MSVA task.

4.3 Baselines and Comparisons

To investigate our universal benefits over different classification tasks varying in SOTA methods, we integrate our comment-aware self-training module into the BERT-based SOTA architectures and examine the results following baselines and comparisons employed in the original paper for fair comparison.

MVSA Comparisons. This benchmark presents baselines of MultiSentiNet (Xu and Mao, 2017) (a deep semantic network with the visual clues guided attention), CoMN (Xu et al., 2018) (a co-memory network to learn image-text interactions), MMMU-BA (Ghosal et al., 2018) (enriching context for cross-modal fusion), and Self-MM (Yu et al., 2021) (joint training of uni-modal and multi-modal tasks to explore cross-modal consistency). CoMN-BERT is a SOTA architecture combining CoMN and the pre-trained BERT (Devlin et al., 2019), which will later be combined with our module for comparison.

ITR Comparisons. In the original paper (Vempala and Preotiuc-Pietro, 2019), LSTM-CNN performs the best via combining CNN-encoded visual features (Szegedy et al., 2015) and LSTM-encoded lingual features (Hochreiter and Schmidhuber, 1997). It is compared with the baseline ablations CNN and LSTM using uni-modal features only. To line up with the SOTA, we implement BERT-CNN to employ pre-trained BERT for text encoding instead of LSTM, which is likewise compared to a BERT classifier using lingual features only. Based on BERT-CNN, we integrate in our module to examine its effectiveness over SOTA.

MSD Comparisons. Here the MMSD baseline is introduced in the original paper (Cai et al., 2019), which employs a hierarchical fusion model to explore visual and lingual features with optical characters. We also compare with the following more

[4] https://pytorch.org/

[5] https://github.com/huggingface/transformers

Methods	Acc	F1
MultiSentiNet	69.84	69.63
CoMN	70.51	70.01
MMMU-BA	68.72	68.35
Self-MM	72.37	71.96
CoMN-BERT	71.33	70.66
CoMN-BERT (full)	**73.71**	**72.83**

Table 3: Comparison results on the MVSA dataset.

Methods	Pre	Rec	F1
LSTM	42.33	48.55	38.77
CNN	37.11	47.22	35.99
LSTM-CNN	48.21	50.78	44.58
BERT	44.65	48.78	40.39
BERT-CNN	50.31	50.60	49.72
BERT-CNN (full)	**53.69**	**54.42**	**53.38**

Table 4: Comparison results on the ITR dataset.

Methods	Pre	Rec	F1
MMSD	76.57	84.15	80.18
D&R Net	77.97	83.42	80.60
Res-BERT	78.87	84.46	81.57
Att-BERT	80.87	85.08	82.92
CMGCN	83.63	84.69	84.16
MMSD-BERT	83.57	84.52	84.04
MMSD-BERT (full)	**85.50**	**85.92**	**85.70**

Table 5: Comparison results on the MSD dataset.

Methods	Pre	Rec	F1
Xception	56.0	54.5	54.4
LSTM	70.7	73.7	71.9
RoBERTa	75.9	76.5	75.4
MMBT	76.3	78.5	77.1
MMBT (full)	**79.15**	**79.88**	**78.76**

Table 6: Comparison results on the MHP dataset.[6]

advanced models on the benchmark: D&R Net (Xu et al., 2020) (using decomposition and relation network to learn visual-lingual interactions), Res-BERT (Pan et al., 2020) (concatenating visual features from ResNet (cite) and lingual features from BERT), Att-BERT (Pan et al., 2020) (with attention mechanism to capture image-text semantic consistency), and CMGCN (Liang et al., 2022) (building a graph to explore cross-modal interactions). MMSD-BERT is based on MMSD with a pre-trained BERT to encode texts, where we will later architect with our proposed module.

MHP Comparisons. Following the setup in Botelho et al. (2021), we consider the Xception (Chollet, 2017) baseline using visual features only. For the text-only comparison, LSTM and RoBERTa (Liu et al., 2019) are adopted. MMBT (Botelho et al., 2021) is the SOTA model learning cross-modal representations with pre-trained MultiModal BiTransformers and will be employed as the base to experiment with our module.

Integrating our Comment-aware Self-training. Based on aforementioned SOTA architectures (**base classifiers** — CoMN-BERT, BERT-CNN, MMSD-BERT, and MMBT, selected for the four benchmarks), we further employ comment-aware self-training in their training and therefore result in CoMN-BERT (full), BERT-CNN (full), MMSD-BERT (full), and MMBT (full).

To further examine the relative contributions of each sub-module in comment-aware self-training, the following ablations are considered in comparison: (1) Base+Com, integrating BERT-encoded comment features in the base classifiers. (2) Base+ST, self-training with retrieved tweets yet without comments. (3) Base+Com+ST, the full model without randomly dropping the retrieved comments in student model training.

[6]The baseline results are copied from the original paper, where the numbers are rounded to one decimal place.

5 Experimental Discussions

5.1 Main Comparison Results

Table 3∼6 shows the main comparison results on MVSA, ITR, MSD, and MHP, respectively.

The full model significantly outperforms all baselines and advances their base ablation on all test benchmarks (measured by paired t-test; $p-value < 0.05$). This indicates that our comment-aware self-training can universally benefit varying tasks and classification architectures. It enables performance gains on both the simple architecture (e.g., BERT-CNN for ITR) and other more complicated models. The possible reasons are twofold. First, retrieved comments, carrying viewpoints from human readers, may provide complementary context hinting the cross-modal semantic understanding for weakly-connected image-text pairs. Second, our self-training may enable the models to leverage both labeled data and unlabeled retrieved data, potentially mitigating the overfitting issue caused by insufficient labeled data scales.

We also observe the models with BERT encoders consistently outperform their counterparts with LSTM encoders, either in multimodal or unimodal architectures. These demonstrate the benefit of pre-training on large-scale text, where the gained generic language understanding capability may enable models to well induce cross-modal meanings.

Model	MVSA		ITR			MSD			MHP		
	Acc	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Base Classifier	71.33	70.66	50.31	50.60	49.72	83.57	84.52	84.04	76.30	78.50	77.10
Base+Com	72.34	71.57	52.08	52.67	51.64	84.76	85.19	84.98	77.31	78.29	77.67
Base+ST	73.33	71.63	51.26	51.89	50.64	84.32	85.45	84.88	77.72	78.49	77.85
Base+Com+ST	73.11	72.29	53.06	53.45	52.32	85.42	85.24	85.33	78.45	78.20	78.29
Full Model	73.71	72.83	53.69	54.42	53.38	85.50	85.92	85.70	79.15	79.88	78.76

Table 7: Ablation results on the four datasets. Our Full Model outperform all the ablations measured by all metrics.

5.2 Ablation study

The general superiority of our method has been demonstrated in §5.1 compared to previous benchmark results. Here we conduct an ablation study to further probe the relative contribution of varying components and show the results in Table 7.

The obvious performance drop of Base+Com and Base+ST, compared to the Full Model, together suggest the positive effects individually from retrieved comments and self-training. These strengthen our previous findings: the comments may enrich context with human hints to bridge visual and lingual semantics and self-training may enrich the data scales with semantically related posts and comments to allow better robustness.

For the results of Base+Com+ST, though better than other ablations, are slightly worse than the Full Model. It implies the extra benefit of modeling retrieved comments in self-training, while randomly dropping some of them may enable the student model to better catch up with the teacher, mitigating the least favorable perturbation phenomena (Xie et al., 2020) in teacher-student alignment.

5.3 Quantitative Analysis

§5.2 shows the crucial roles self-training and comment retrieval play in our method. Here we further quantify the effects of varying unlabeled data scales on self-training and those of individual modality (images or text) on comment retrieval.

Self-training w/ Varying Unlabeled Data Scales. Here we train our full model via self-training with varying number of retrieved posts (K) and show the performance gain compared to Base Classifier in Figure 3 (F1 difference of the Full Model and Base Classifier). We observe the results peak at $K = 3$ or 5, implying self-training may benefit from some similar unlabeled data while further retrieving more data may result in noise as well.

Modality Effects on Comment Retrieval. Recall that in comment retrieval, we balance the visual and lingual similarity to retrieve similar posts (and

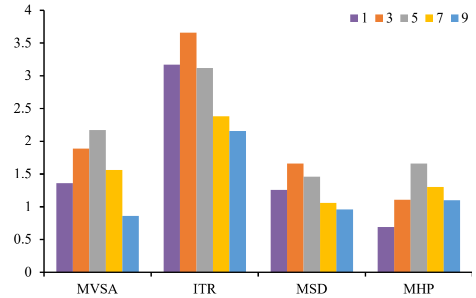

Figure 3: Performance gain observed from self-training given varying number of unlabeled retrieved posts (and their associated comments). X-axis: within each dataset, the bars from left to right indicate self-training with varying number of posts (K); y-axis: the difference in F1 between our Full Model and Base Classifier.

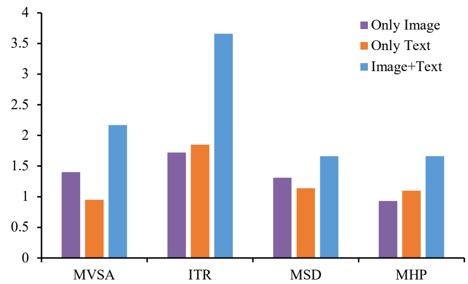

Figure 4: F1 gain compared to the Base Classifier (y-axis) over varying datasets. For each dataset, the bars from left to right indicate the retrieval with image only, text only, and both image and text (Image+Text).

obtain their comments). To further study how features in varying modalities affect comment retrieval results, we examine two ablations relying on the similarity in image (Only Image) and text (Only Text) in comparison to the full model trading-off image and text similarities (Image+Text).

The results (F1 gain compared to Base Classifier) are shown in Figure 4. Varying tasks might prefer the similarity measure with image or text semantics, whereas the full model leveraging posts' visual and lingual features achieves the best results.

5.4 Qualitative Analysis

The discussions above are from a quantitative view. To provide more insight, a case study will be presented here, followed by analyses for error cases.

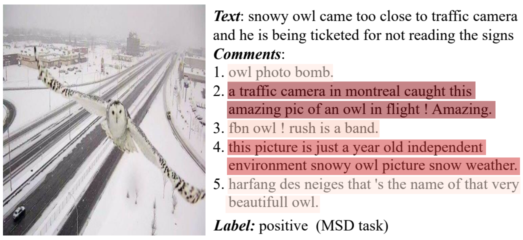

Figure 5: Visualization of attention heatmaps over the retrieved comments for the MSD benchmark. Deeper colors indicate higher attention weights.

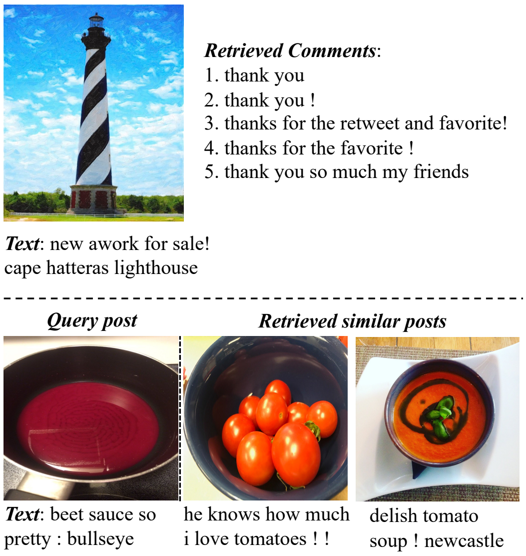

Figure 6: Examples of major error types from comment and post retrieval. The top indicates the general comments and the bottom semantically unrelated posts.

Case Study. To analyze how comments hint at cross-modal understanding, attention maps over comments are shown in Figure 5, where the case is sampled from the MSD benchmark. As can be seen, models tend to capture the salient comments mentioning the key visual objects, e.g., "owl" and "traffic camera" in the case, helpfully connecting visual semantics to lingual. It is probably because human readers are likely to echo crucial points in their comments in response to what they viewed from a post, which inspires models to explicitize the weakly-connected visual-lingual semantics.

Error Analysis. The potential benefit has been potentially observed in varying cross-modal learning scenarios; however, many errors are also related to the retrieved comments and posts used in self-training. Figure 6 summarizes the two major error types.

First, general comments, e.g., "thank you", are retrieved, useless in learning specific meanings in social media posts. This calls for a future direction for comment selection in a more effective manner.

Second, semantically unrelated posts might be retrieved due to the misunderstanding to the query and hence result in irrelevant comments. For example, the posts concerning "tomato" are wrongly retrieved because of its similar color to "beet sause". Future work may consider the advance in similarity measurement of cross-modal posts and the detection of high-quality unlabeled data (e.g., selecting the pseudo-labeled data by confidence) for self-training.

6 Conclusion

We have presented the potential of employing comments to better form visual-lingual understanding on social media, where 27M tweets with comments are contributed for the related study. A novel framework is proposed to retrieve comments from similar posts and explore comments' hinting capabilities via self-training. Experimental results on four social media benchmarks show the universal benefit of leveraging retrieved comments and conduct comment-aware self-training on various multi-modal classification tasks and architectures.

Acknowledgements

This paper is substantially supported by NSFC Young Scientists Fund (No.62006203), a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU/25200821), PolyU internal funds (1-BE2W, 4-ZZKM, and 1-ZVRH), and CCF-Baidu Open Fund (No. 2021PP15002000).

Limitations

Here we point out more limitations in addition to what we have discussed in §5.4.

First, the post and comment retrieval should be built upon a large-scale corpus (wild dataset). Substantial efforts might be needed to gather a likewise corpus, if applying our work to other a different social media platform. However, for our follow-up work exploring Twitter as well, the pre-trained retrieval module, based on the Faiss library, can work in an efficient manner. For example, we test the retrieval module to search similar posts for 5,000 cross-media posts by using the Faiss library on one single 2080Ti GPU, and it would cost 231.66s for image modality and 77.68s for text modality.[7]

[7]The reason of different retrieval time is due to the different dimension of extracted features (i.e., the dimension is 2048 for image feature while 768 for text)

Second, the time and size of the retrieval corpus would result in another limitation. As shown in Table 1, we build the dataset with multimodal tweets posted from 2014 to 2019. While a timely update might be needed if the task requires fresher data, e.g., the research of COVID-19 because the event becomes trendy in 2020. Nevertheless, dynamic dataset update might also explosively scale up the data quantity, and how to enable feasible real-time dataset management calls for another research question, which is beyond the scope of this paper and is valuable to be explored in future studies.

Ethical Considerations

The benchmark datasets we experiment for classification are publicly available with previous work, where the ethical concerns have been addressed by the authors of these original papers.

Our paper contributes a large-scale Twitter corpus (wild dataset) for post and comment retrieval. We collected the data therein following the terms of use of standard data acquisition process regularized by Twitter API. The data is downloaded only for the purpose of academic research. Following the Twitter policy for datasets open-access, only the tweet IDs will be released. Data requestors will be asked to sign a declaration form before accessing the data, making sure that the dataset will only be reused for research, under Twitter's policy compliance, and not for collecting anything possibly raising ethical concerns, such as the sensitive and personal information.

For our experiments, we have pre-processed to anonymize the data for privacy concern, e.g., removing authors' names and changing @mention and URL links to generic tags.

References

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2131–2140. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Austin Botelho, Scott A. Hale, and Bertie Vidgen. 2021. Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1896–1907. Association for Computational Linguistics.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2506–2515. Association for Computational Linguistics.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. *CoRR*, abs/1505.04467.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):1–13.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Deepanway Ghosal, Md. Shad Akhtar, Dushyant Singh Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods* *in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3454–3466. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Shir Gur, Natalia Neverova, Christopher Stauffer, Ser-Nam Lim, Douwe Kiela, and Austin Reiter. 2021. Cross-modal retrieval augmentation for multi-modal classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 111–123. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 861–877. Association for Computational Linguistics.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2034–2045. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7166–7176. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 427–431. Association for Computational Linguistics.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 212–228. Springer.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1767–1777. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Yunshan Ma, Xun Yang, Lizi Liao, Yixin Cao, and Tat-Seng Chua. 2019. Who, where, and what to wear?: Extracting fashion knowledge from social media. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 257–265. ACM.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9006–9017. Association for Computational Linguistics.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II*, volume 9517 of *Lecture Notes in Computer Science*, pages 15–27. Springer.

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 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1383–1392. Association for Computational Linguistics.

Steven J. Rennie, Etienne Marcheret, Neil Mallinar, David Nahamoo, and Vaibhava Goel. 2020. Unsupervised adaptation of question answering systems via generative self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1148–1157. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252.

Mrinmaya Sachan and Eric P. Xing. 2018. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 629–640. Association for Computational Linguistics.

H. J. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Inf. Theory*, 11(3):363–371.

Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. Taxoclass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4239–4249. Association for Computational Linguistics.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7454–7464, Online. Association for Computational Linguistics.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4631–4640. IEEE Computer Society.

Alakananda Vempala and Daniel Preotiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2830–2840. Association for Computational Linguistics.

Danae Sánchez Villegas and Nikolaos Aletras. 2021. Point-of-interest type prediction using text and images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7785–7797. Association for Computational Linguistics.

Junjie Wang, Yatai Ji, Jiaqi Sun, Yujiu Yang, and Tetsuya Sakai. 2021. MIRTT: learning multimodal interaction representations from trilinear transformers for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2280–2292. Association for Computational Linguistics.

Yue Wang, Jing Li, Irwin King, Michael R. Lyu, and Shuming Shi. 2019. Microblog hashtag generation via encoding conversation contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1624–1633. Association for Computational Linguistics.

Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. Computer Vision Foundation / IEEE.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 2399–2402. ACM.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 929–932. ACM.

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 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3777–3786. Association for Computational Linguistics.

Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. 2021. Interactive self-training with mean teachers for semi-supervised object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5941–5950. Computer Vision Foundation / IEEE.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics, 26-30 June 1995, MIT, Cambridge, Massachusetts, USA, Proceedings*, pages 189–196. Morgan Kaufmann Publishers / ACL.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5408–5414. ijcai.org.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10790–10797. AAAI Press.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. Multimet: A multi-modal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3214–3225. Association for Computational Linguistics.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2495–2509. Association for Computational Linguistics.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. 2020. Rethinking pre-training and self-training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.