

Understanding Social Media Cross-Modality Discourse in Linguistic Space

Chunpu Xu¹, Hanzhuo Tan¹, Jing Li^{1*}, Piji Li²

¹ Department of Computing, The Hong Kong Polytechnic University, China

² College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China

¹{chun-pu.xu, han-zhuo.tan}@connect.polyu.hk

¹jing-amelia.li@polyu.edu.hk; ²pjli@nuaa.edu.cn

Abstract

The multimedia communications with texts and images are popular on social media. However, limited studies concern how images are structured with texts to form coherent meanings in human cognition. To fill in the gap, we present a novel concept of cross-modality discourse, reflecting how human readers couple image and text understandings. Text descriptions are first derived from images (named as **subtitles**) in the multimedia contexts. Five labels – entity-level *insertion*, *projection* and *concretization* and scene-level *restatement* and *extension* — are further employed to shape the structure of subtitles and texts and present their joint meanings. As a pilot study, we also build the very first dataset containing 16K multimedia tweets with manually annotated discourse labels. The experimental results show that the multimedia encoder based on multi-head attention with captions is able to obtain the-state-of-the-art results.

1 Introduction

The growing popularity of multimedia is revolutionizing the communications on social media. The conventional text-only form has been expanded to cross modalities involving texts and images in information exchange. For multimedia messages, the language understanding acquires more than making sense of both visual and textual semantics; it also matters to figure out what glues them together to exhibit the coherent senses in human’s mind.

Nevertheless, most progress made in social media language understanding relies on texts to learn the message-level semantics (Shen et al., 2018; Nguyen et al., 2020), largely ignoring the rich meanings conveyed in images (Cai et al., 2019a; Wang et al., 2020b). Other recent multimodal studies focus on model designs to combine visual and textual signals (Park et al., 2019; Li et al., 2020;

Yu et al., 2021), ignoring the insights from how humans understand the implicit structure underlying a multimedia post.

In light of these concerns, we consider images as an integral part of social media language and propose a novel concept of *cross-modality discourse*, which defines how human readers structure the coherent meanings from image and text modalities. Our work is inspired by Vempala and Preotiuc-Pietro (2019) examining the information overlap between images and texts, whereas we take a step further to characterize how multimedia messages make sense to humans, which is beyond a simple yes-or-no prediction to whether new thing is observed. To the best of our knowledge, *we are the first to extend discourse — a pure linguistic concept — to define the linguistic roles played by images and their pragmatic relations with texts to shape the coherent meanings.*

In general, cross-modality discourse is defined by the operations adopted in human perception to couple image and text semantics. Readers may first extract the information from the images acquired to complete the cross-modality understanding, either in form of the local objects (entities) or global scenes (Rayner, 2009). Then, the extracted entities or scenes are represented in texts, named as the images’ **subtitles**, which can further contribute to structure the entity-level or scene-level discourse with the matching texts in the multimedia contexts. Concretely, for entity-level discourse, it is detailed into *insertion*, *projection*, and *concretization*, according to whether the entity is omitted, described, or mapped; similarly, scene-level *restatement* and *extension* are employed to reflect whether the story in one modality recurs or continues in the other.

To illustrate the definitions above, Figure 1 shows five multimedia Twitter posts. As can be seen from (a), readers may concentrate on the object “strawberry” and *insert* its name into the texts omitting the entity. As for (b), the “coffee” object

* Corresponding author

Tweet text	<i>Freshly picked off my allotment (strawberry) today, well chuffed</i>	<i>Happy Sunday! My best friend and I have coffee in the sunshine</i>	<i>Cartel leader whose arrest sparked killings is sentenced to prison in Dallas court</i>	<i>This dog has to hold hands on the car</i>	<i>One step closer to summer (track towards beach)</i>
Tweet image					
	(a)	(b)	(c)	(d)	(e)
Vempala et al. (2019) labels	Image adds meaning Text not represented	Image not adds meaning Text represented	Image not adds meaning Text not represented	Image not adds meaning Text represented	Image adds meaning Text not represented
Discourse labels	Entity-level: insertion (strawberry)	Entity-level: concretization (coffee)	Entity-level: projection (court)	Scene-level: restatement (dog hold hands in car)	Scene-level: extension (track towards beach)

Figure 1: The five cross-modality discourse labels and their examples. The rows from top to bottom display their texts, images, the image-text relation labels in Vempala and Preotiuc-Pietro (2019), and our cross-modality discourse categories. The labels in Vempala and Preotiuc-Pietro (2019) concern whether new meanings are added by images to texts, whereas ours define the linguistic roles of images and their pragmatic relations with texts for coherence.

should be extracted from the image to *concretize* the word “coffee” in the text. In (c), the word “court” in text is linked with the “gavel” object. The image in (d) helps *restate* the texts scene (a dog holds hands in the car). In (e), the global scene works as an *extension* to texts and completes the story: “We are one step closer to summer following the track towards beach.”

On the contrary, the image-text relations in Vempala and Preotiuc-Pietro (2019) are limited to whether images add new meanings to texts, which is nonetheless insufficient to reflect how language is understood in multimedia contexts.

As a pilot study of cross-modality discourse, we also present the very first dataset to explore the task. It is collected from Twitter and contains 16K high-quality multimedia posts with manual annotations on their discourse labels.¹ We believe our task and the associated dataset, being the first of its kind, will be potentially beneficial to help machines gain the ability to understand social media language with multimodal elements.

To that end, we present a framework to learn the discourse structure across texts and images. Inspired by the recent advances in multimodal learning (Wang et al., 2020b; Yu et al., 2020), we employ the multi-head attention mechanism (Vaswani et al., 2017) to explore the visual-textual representations reflecting cross-modality interactions. Besides, to characterize subtitles for discourse learning, image captions generated from model trained on COCO captioning dataset (Lin et al., 2014b) are leveraged as additional features.

For empirical studies on cross-modality discourse, we conduct comprehensive experiments

¹The dataset and code are released at https://github.com/cpaax/Multimodal_Discourse.

on our dataset. The comparison results on classification show the challenges for machines to infer discourse structure and it is beyond the capability of advanced multimodal encoders to well handle our task. Nevertheless, exploring correlations of texts, captions, and visual-textual interactions helps exhibit the state-of-the-art performance in both the intra-class and overall evaluation. We further examine the effects of varying modalities and text length and find that text signals are crucial for discourse inference while joint effects of texts, images, and captions present the best results. At last, the qualitative analysis demonstrates how the multi-head attention in our model interprets discourse structure.

2 Related Work

Our paper crosses the lines of multimedia learning and discourse analysis in natural language processing. Here comes more details.

Multimedia Learning. Our paper is in the line with cross-media research that attempts to fuse textual and visual features. There are various deep learning methods proposed to leverage crossmodal features, either based on advanced neural architectures like co-attentions (Xu and Saenko, 2016; Lu et al., 2016) and multi-head attentions (Vaswani et al., 2017; Wang et al., 2020a), or pre-trained visual-lingual representations (Lu et al., 2019; Su et al., 2020; Zhang et al., 2021). Their effectiveness are demonstrated in both conventional vision-language tasks, such as image captioning (Park et al., 2019; Zhou et al., 2020; Shi et al., 2021) and visual question answering (VQA) (Yu et al., 2019; Tan and Bansal, 2019; Si et al., 2021), and social media applications, such as sarcasm detection (Cai

et al., 2019a), event tracking (Li et al., 2020; Abavisani et al., 2020b), keyphrase prediction (Zhang et al., 2019; Wang et al., 2020b).

It is seen that most progress to date made in this line focus on advancing methodology designs for general purposes (Su et al., 2020; Zhou et al., 2020) or specific applications (Wang et al., 2020b) to better capture the matched semantics across varying modalities. However, their effectiveness over social media data would be inevitably compromised resulted from the intricate image-text interactions (Vempala and Preotiuc-Pietro, 2019). We thus borrow the insights from human perception to interpret image-text relations from the linguistic viewpoints and propose the task to learn discourse structure in multimedia contexts. It is a fundamental research exhibiting the potential to help the models gather cross-modality understanding capability and might benefit various downstream applications.

We are also related with previous categorization tasks on social media to understand image-text relations, such as information overlap (Vempala and Preotiuc-Pietro, 2019), point-of-interest types (Villegas and Aletras, 2021), author purposes (Kruk et al., 2019), object possessions (Chinnappa et al., 2019), and so forth. Besides, interestingly, the “discourse” concept is also employed to examine the image-text relations in cooking recipes (Alikhani et al., 2019). Compared with these studies concatenating visual and textual embeddings in a “common” space, we craft text-formed subtitles to convey visual stories and explore how they shape the coherent meanings with the post texts in linguistic space. It will consequently allow deep semantic learning to capture the implicit structure holding image and text modalities, while the existing models might be incapable to gather senses of language understanding via simple feature concatenation.

Discourse Analysis. This work is related to prior studies on text-level discourse structures. The popular tasks in the styles of either RST (Rhetorical Structure Theory) (Mann and Thompson, 1988; Liu et al., 2019) or PDTB (Penn Discourse Tree Bank) (Prasad et al., 2008; Xu et al., 2018) explore the rhetorical relations of discourse units (e.g., phrases or sentences) that cohesively connect them form a sense of coherence. These studies have demonstrated their helpfulness in diverse stream of NLP applications (Choubey et al., 2020), such as sentiment analysis (Bhatia et al., 2015), text categorization (Ji and Smith, 2017), and microblog sum-

marization (Li et al., 2018). Nevertheless, limited work examines a social media image as a discourse unit of the pragmatic structure in multimedia contexts, which is a gap to be filled in this work.

3 Study Design

In this section, we first define the task to predict cross-modality discourse in §3.1. Then, we introduce how we construct the dataset in §3.2, followed by the data analysis in §3.3 and the potential applications in §3.4.

3.1 Task Definition

In our task, the input is an image-text pair from a multimedia post on social media, following the previous practice (Vempala and Preotiuc-Pietro, 2019). For each pair, the goal is to output a label from a predefined set that cover the major categories of cross-modality discourse on social media. Our intuition is that images are relatively more eye-catching and likely to be processed before the texts. For image understandings, the previous findings from psychological experiments (Rayner, 2009) point out that humans may first recognize and extract the meanings from global scenes to fill the information gap in contexts; if the gap still exists, they may go back to capture the local objects. Based on that, we first coarsely categorize the discourse label set into the level of entity (object) and scene, depending on whether object or scene is extracted to make sense of the joint meanings of images and texts.

To further elaborate the label design, the extracted information from an image (as an object or scene) is mapped to the text modality to form the **subtitle**, which allows us to formulate how human senses structure the coherent meaning with subtitles and post texts.

For entity-level discourse, three cases are examined: the entity is omitted, mentioned or linked in the texts. For the absent entity (e.g., Fig. 1(a)), the subtitle, in form of entity name, should be *inserted* into the post text to complete the meanings of a message, while the entity in Fig. 1(b) is *concretized* by the object in images. And the entity in Fig. 1(c) is implicitly projected into the relevant object. We henceforth design entity-level **insertion**, **concretization**, and **projection** to describe the above three cases, respectively.

Similarly, scene-level discourse can be separated into **restatement** and **extension** categorizes. The former refers to image serving as texts description

(e.g., Fig. 1(d)) and for the latter, posts presenting image scenes to elaborate the story left as white space in the texts (e.g., Fig. 1(e)).

3.2 Data Collection and Annotation

Our dataset is gathered from Twitter², which is drawing attentions to research digital communications (Mozafari et al., 2019; Nikolov and Radivchev, 2019; Müller et al., 2020) and exhibits prominent use of multimedia posts (Vempala and Preotiuc-Pietro, 2019; Wang et al., 2020b). We first crawled the raw data using Twitter streaming API³ and removed non-English posts and those with texts only or multiple images. Afterwards, to better model discourse from the noisy Twitter data (Vempala and Preotiuc-Pietro, 2019), we removed samples that might hinder the learning of non-trivial discourse signals. Here, four types of “bad” image-text pairs might provide tremendous noise in the learning, which are shown in Fig. 2.

The first type refers to image portraits with some quotes to share the insights of life (henceforth **portraits**), where images and texts are not coherently related (from linguistic viewpoints) and discourse structure are unable to be defined for them. Moreover, many of them contain authors’ selfies, which might raise privacy concerns. The second type of posts, namely **background**, relies on external knowledge to capture the meanings (e.g., Fig. 2(b)), which is beyond the capability of language understanding given only the images and the matching texts. For the third, we consider **low-quality images** (e.g., low resolution and blurred ones like Fig. 2(c)), from which it is hard to capture the visual meanings. The last one refers to **OCR subtitles** (Fig. 2(d)), where the subtitles appear in the images as optical characters. It may result in a degeneration of cross-modality discourse to text-level discourse and render the learning of trivial features.

In the data annotation, we first selected 25 typical examples corresponding to each discourse label and provide them together with the annotation guidelines (with the detailed description of each label) for quality control. Then, two postgraduate students majoring in linguistics were recruited to manually label the discourse categories, given an image-text pair. “Bad” samples falling in the above four types should also be indicated in the annotation process. The inter-annotator agreement is

²twitter.com

³developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time

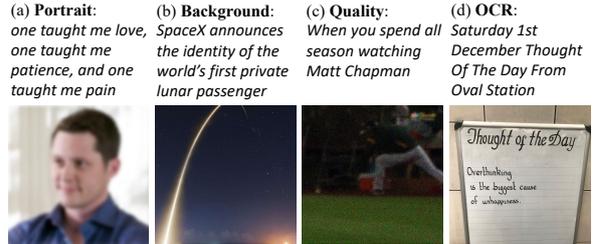


Figure 2: Examples tweets of the four “bad” types. (a) **Portrait** image with quotes in texts. (b) **Background** is externally required for understanding (rocket trajectory scenes here). (c) **Low-quality** image where objects could be barely observed. (d) **OCR Subtitle** (“Thought Of The Day”) appear in the image in optical characters.

	Total	Ins	Con	Pro	Res	Ext
Num	16,000	839	10,558	690	1,826	2,087
Len	10.69	9.11	10.85	10.98	11.24	9.92

Table 1: Statistics of the total data and that with each label: Ins: Insertion; Con: Concretization; Pro: Projection; Res: Restatement; Ext: Extension. Len: average word number in texts. Num: tweet number.

79.8% and we only kept the data with labels agreed by both annotators to ensure the feature learning quality in noisy data. At last, posts in “bad” types were taken away and the final dataset presents 16k multimedia tweets with manual labels in five discourse categories.

3.3 Data Analysis

Here we conduct a preliminary analysis of our dataset and show the statistics in Table 1. There exhibits imbalanced labels, where concretization and extension labels are relatively more popular compared to the other three. This indicates the diverse preferences of Twitter users in the way they choose to structure texts and images and the potential challenge for models to handle our task.

For the text length, it is seen that most tweets contain limited words, challenging the models to capture essential features from textual signals. Interestingly, we compare our statistics with other text-only Twitter datasets in previous work (Wang et al., 2019) and find our multimedia tweets have 30% fewer words on average. This implies that authors may tend to put less content in the text of multimedia posts, and figure the missing information in images for compensation. We also notice that insertion and extension discourse exhibit relatively shorter texts on average. It is probably because they exhibit the omitted content in texts, which presents in images.

To further characterize text length in our dataset, Fig. 3 shows the word number distribution of tweet

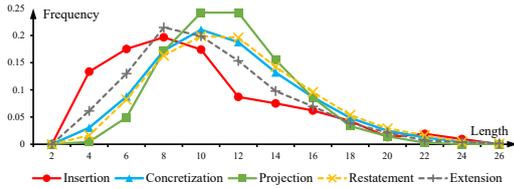


Figure 3: Text lengths (token number) distribution of posts with varying discourse labels.

texts with varying labels. All the curves demonstrate the sparse distribution over text length, owing to the freestyles of social media writings. Insertion and extension curves first peak at 8 words while the others at 10-12, then all present long tails afterwards. This again shows that texts in multimedia posts may provide limited content and those in insertion and extension contain fewer words.

3.4 Potential Applications

In this subsection, we further discuss the potential downstream applications of our task and datasets, which might inspire the design of future work. A straightforward application is microblog summarization — an important task to distill the salient content from massive social media data. As many state-of-the-art summarization models only allow textual input while multimedia posts are prominent on social media, it may require the compression of these posts into text for easy processing. It is different from the traditional image captioning task (Anderson et al., 2018; Rennie et al., 2017; Huang et al., 2019), where the generated captions are translated from images. For a social media post, the text cannot trivially be seen as a “translation” of image, because of possibly ambiguous image-text interactions therein. Considering crucial roles played by discourse analysis in summarization (Xu et al., 2020), it is not hard to envision that our cross-modality discourse, describing how image and text structure coherence, would contribute to the research of multimedia summarization.

In addition, cross-modality discourse can be viewed as a fundamental task and might be helpful to other downstream tasks on social media (e.g., multimodal NER (Yu et al., 2020), multimodal crisis events classification (Abavisani et al., 2020a), multimodal sarcasm detection (Cai et al., 2019b), multimodal sentiment analysis (Truong and Lauw, 2019), and multimodal hashtag prediction (Wang et al., 2020c)). However, most previous efforts focus on the leverage of visual and lingual representations yet ignore the linguistic essence that glue the two modalities. Recently, some work pro-

pose multitask learning to consider image-text relations in multimodal learning. For example, Sun et al. (2021) investigate the relation propagation between text and image to improve the accuracy of NER in tweets. Ju et al. (2021) utilize multimodal relation types as auxiliary labels to explore multimodal aspect-sentiment analysis. The positive results from these studies imply the potential of cross-modality discourse (as a linguistic description of image-text relations) to benefit a wide range of multimodal applications. Besides, the training data of image-text relation used in (Sun et al., 2021; Ju et al., 2021) is the TRC dataset proposed by Vempala and Preotiuc-Pietro (2019). Compared to the TRC dataset, our proposed discourse dataset exhibits a tremendously larger scale (i.e., 16K VS 4.5K) and fine-grained labels for image-text relation, as shown in Fig. 1. We therefore believe our dataset would also helpfully advance the performance of various multimodal models.

4 The Discourse Learning Framework

In this section, we describe our framework that couples the signals from images and texts to predict their discourse labels. As shown in Fig. 4, the model architecture leverages representations learned from texts, images, and image captions (to reflect subtitles), which will be introduced in §4.1. Then, we will discuss how we combine multi-modality representations §4.2. At last, §4.3 presents how we predict the discourse labels and design the training processes.

4.1 Encoding Text, Image, and Captions

Texts Encoding. Here we describe how to learn text features. The text encoder is based on the bottom 6-layers of pre-trained Bertweet (Nguyen et al., 2020). It is fed with an L -length token sequence and embed its representations into a sequential hidden states $\mathbf{H}_{text} = (\mathbf{h}_1, \dots, \mathbf{h}_L)$, where each element reflects a token embedding. \mathbf{H}_{text} further goes through a max-pooling layer and produces $\bar{\mathbf{H}}_{text}$ to represent the text.

Image Encoding. To explore visual signals, images are encoded by CNN-based ResNet-101 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015). The output of the last convolutional layer in ResNet-101 is extracted as the representation of the input image. The size of the feature map is first reduced to $M \times M \times 2048$ and then reshaped into $M^2 \times 2048$. Each 1×2048 vector represents

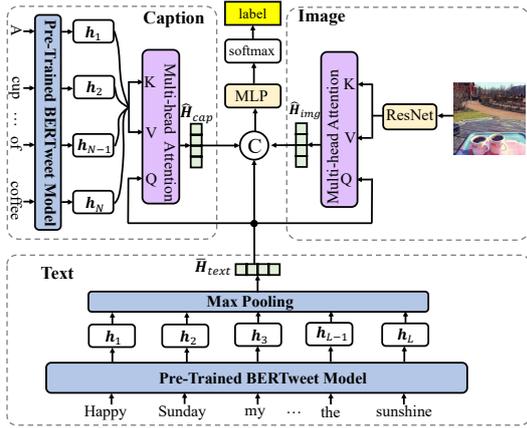


Figure 4: Our framework to learn cross-modality discourse via representations encoded from texts (bottom), captions (upper left), and images (upper right). The encoded captions and texts are compared at output layer in visual-textual contexts.

the visual features in a corresponding image area and is projected to the same dimension of text feature \mathbf{h} by linear layer. The post-level visual feature is denoted as $\mathbf{H}_{img} = (\mathbf{v}_1, \dots, \mathbf{v}_{M_2})$, where \mathbf{v}_i refers to an 1×2048 vector that represents the feature of an area in the image.

Image Caption Encoding In order to capture more semantic information from images, we further exploit image captions (henceforth captions) as an additional modality. Our intuition is that captions may inject essential visual semantics underlying images into a descriptive language in texts (Xu et al., 2015). They are potentially helpful to reflect the rich interactions between image objects and discover subtitle-style clues as essential discourse indicators. We first employ the model presented by Anderson et al. (2018) to predict the captions of each image. The captioning model is pre-trained on the COCO captioning dataset (Lin et al., 2014b), which mostly consists of natural pictures outside social media domain. Then, we encode the token sequence of captions following the same process of text encoding (discussed above) and yield caption representation: $\mathbf{H}_{cap} = (\mathbf{h}_1, \dots, \mathbf{h}_N)$. Here N indicates the number of tokens in the caption, \mathbf{h}_i refers to the i -th hidden state of the Bertweet encoder.

4.2 Integrating Multimodal Representations

As pointed out in previous work (Wang et al., 2020b), modalities on social media data exhibit much more intricate interactions compared with the widely-studied vision-language datasets (Lin et al., 2014a; Young et al., 2014). To allow the framework to attend various types of cross-modality interactions, we employ multi-head attentions (Vaswani

et al., 2017) to comprehensively explore the interactions between the encoded image features (\mathbf{H}_{img}) and max-pooled text representations ($\bar{\mathbf{H}}_{text}$).

Concretely, we set text features as the query \mathbf{Q} , image features as the key and value \mathbf{K}, \mathbf{V} , and compute the multi-head attention $MA(\cdot)$ as follows:

$$MA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [hd_1; \dots; hd_n] \mathbf{W}^O \quad (1)$$

where n is the number of heads, $[\cdot]$ indicates the concatenation operations, and the attention of the j -th head is:

$$hd_j = A(\mathbf{Q}\mathbf{W}_j^Q, \mathbf{K}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V) \quad (2)$$

$$A(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \theta\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (3)$$

Here d_k is the normalization factor, $\theta(\cdot)$ means softmax function. $\mathbf{W}^O, \mathbf{W}_j^Q, \mathbf{W}_j^K, \mathbf{W}_j^V$ are learnable variables. The attended images (in aware of texts) are denoted as $\hat{\mathbf{H}}_{img}$, which further serves as the context to help explore the discourse clues from captions and texts.

For discourse modeling, the encoded texts ($\bar{\mathbf{H}}_{text}$) are compared with captions (carrying subtitle-style features) to infer how the subtitles can be structured with texts. To that end, we first employ a multi-head attention mechanism to encode text-aware attended caption $\hat{\mathbf{H}}_{cap}$, which captures salient contents from captions to indicate discourse categories. Furthermore, $\hat{\mathbf{H}}_{cap}$ are concatenated with $\bar{\mathbf{H}}_{text}$ to model their structure; also concatenated are the attended images $\hat{\mathbf{H}}_{img}$ as the image-text interaction contexts for cross-modality discourse learning.

4.3 Discourse Prediction and Model Training

The discourse labels are predicted with a multi-layer perceptron (MLP) fed with $\mathbf{H} = [\hat{\mathbf{H}}_{cap}; \bar{\mathbf{H}}_{text}; \hat{\mathbf{H}}_{img}]$, the integrated feature vectors, which is further activated with a softmax function to predict the likelihood over the four discourse labels. For training, recall that in Table 1, we observe the severe label imbalance on our task. To deal with the issue, we adopt weighted cross-entropy loss, whose weights are set by the proportions of labels in training data.

5 Experimental Setup

Model Settings. The length of tweet texts (L) and captions (N) are both capped at 20 by truncation. The batch size is set to 100, the learning rate to 5×10^{-5} . The head number of all multi-head attention layers are set to 6. For image encoding,

Method	Insertion	Concretization	Projection	Restatement	Extension	F1
Baselines						
Qin et al. (2016)	41.13	69.91	26.13	39.67	41.15	61.67
Rutherford and Xue (2016)	43.17	70.78	32.62	42.31	40.82	62.73
Nam et al. (2017)	46.49	74.83	33.33	39.33	42.39	65.76
Text+Image						
CONCATFUSE	52.86	81.62	34.78	39.19	42.93	71.09
ATTENTION	54.30	82.64	33.71	39.23	39.41	71.48
CO-ATTENTION	51.90	83.31	36.36	42.57	40.59	72.37
MULTIHEADATT	53.69	84.33	36.96	42.11	42.01	73.33
Text+Caption						
CONCATFUSE	52.00	81.11	33.33	41.18	43.02	70.82
ATTENTION	54.79	81.26	36.78	39.72	42.55	70.97
CO-ATTENTION	53.73	82.13	37.20	41.78	39.16	71.38
MULTIHEADATT	53.79	82.27	34.55	43.96	43.46	72.08
Img+Text+Caption						
CONCATFUSE	52.48	82.41	32.97	43.01	42.39	71.88
ATTENTION	53.24	83.01	34.95	43.45	43.65	72.58
CO-ATTENTION	54.81	83.98	36.96	45.24	39.76	73.15
MULTIHEADATT(<i>full model</i>)	57.75*	84.88*	37.36	46.15*	44.19*	74.51*

Table 2: Comparison results of the baselines and our variants. Scores with * represent the significance tests of our full model over the baseline models with p-value<0.05.

image feature map size M is set to 14. For text and comment encoding, the representations are extracted from the bottom 6-layers of the Bertweet model, which are further fine-tuned in training. In the setup, we randomly split 80%, 10% and 10% for training, validation, and test. For evaluation, we report F1 scores in the prediction of each label and the weighted F1 to measure the overall results.

Baselines and Comparisons. We first consider two text-level discourse parsers proposed in Qin et al. (2016) and Rutherford and Xue (2016), where we extend their text encoders into multimodal encoders to fit the image-text pairs. Then, we compare with a popular multimodal classifier (Nam et al., 2017) that employs a dual attention network to fuse the visual and textual features.

Besides, we evaluate varying sets of feature combinations in our model *Text + Image*, *Text + Caption*, and *Text + Image + Caption* (the full set). Recall that our framework employs multi-head attention to integrate features learned from different modalities. In experiments, we also test the performance of other modality fusion alternatives based on simple feature concatenation (CONCATFUSE), the conventional attention mechanism (ATTENTION), the co-attention mechanism (CO-ATTENTION).

6 Experimental Discussions

This section first presents the main comparison results (§6.1). Then, we discuss model sensitivity to varying modalities and text length in §6.2. Finally,

§6.3 presents a case study to provide more insights.

6.1 Main Comparison Results

Table 2 shows the main comparison results of various multimodal encoders. The following observations can be drawn.

First, all models do not exhibit good F1. This indicates that cross-modality discourse prediction is a challenging task. A good understanding for that cannot be gained by trivially adapting discourse parsers to the multimodal settings or applying the existing vision-language encoders. Second, results on the two entity-level discourse labels (i.e., insertion and concretization) are relatively better than scene-level, indicating that local objects are easier to be captured than global scenes. Among all the labels, models perform the best in concretization, probably attributed to its richer data samples for feature learning (as shown in Table 1). And models obtain worst results in projection. The reasons might be that additional knowledge are needed for models to learn the implicit relation between the object and the entity.

Last, images, texts, and captions all contribute to building automatic discourse understanding. Joint modeling of the three modalities enables the corresponding models to outperform their text+image and text+caption counterparts.

6.2 Sensitivity to Modalities and Text Length

Varying Modalities. To further examine the effects of varying modalities, we compare the F1 scores of our full model with its caption-only,

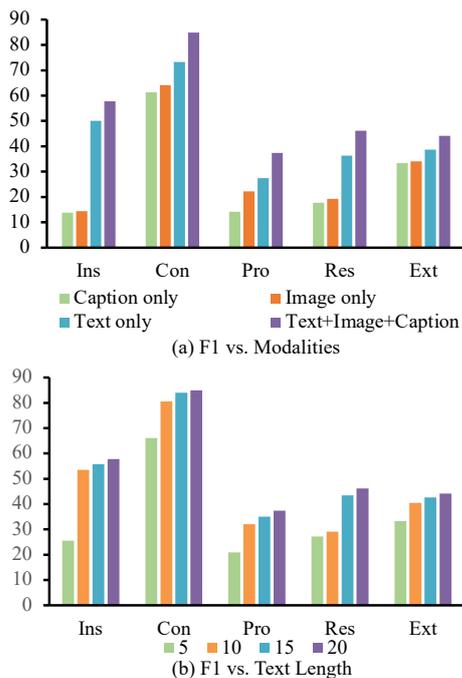


Figure 5: Full model performance compared with varying modality ablations in (a) and its results over varying text length (b). X-axis: insertion, concretization, projection, restatement, and extension; Y-axis: F1 scores. For each label, bars from left to right show the caption only, image only, text only ablations, and the full model in (a) and the tweet texts capped at 5, 10, 15, and 20 in (b).

image-only, and text-only ablations in Fig. 5(a). It is seen that text modality contributes relatively more to discourse modeling observed from all labels, especially for insertion, where Name Entities are omitted and makes the text style easy to recognize. Nevertheless, the joint effects of images, texts, and captions together present the best performance over all labels.

Varying Text Length. As discussed above, text features are crucial to predict cross-modality discourse. Here we further examine the effects of text length on model performance and the results of our full model are shown in Fig. 5(b). Better scores are observed for longer texts as richer contents can be captured. This again demonstrates the essential signals provided by texts to infer cross-modality discourse.

6.3 Qualitative Analysis

Discussions above mostly concern caption and text modalities. Here we present a case study to probe into how the model reflects discourse indicators over vision signals.

Case Study. Visual features are analyzed by the heatmap (in Fig. 6) visualizing the text-aware atten-

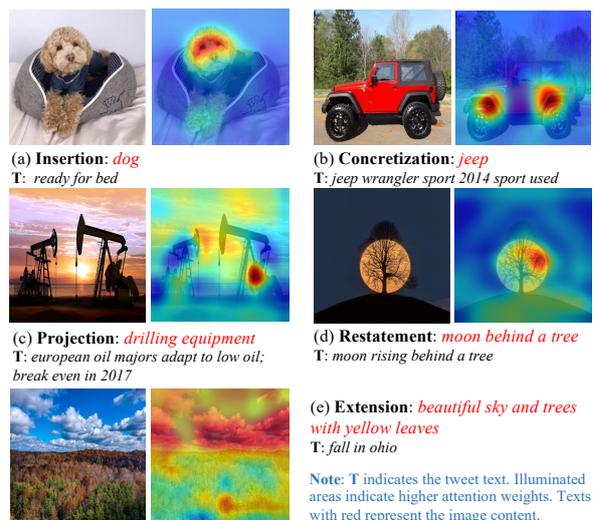


Figure 6: Visualization of multi-head attention heatmaps over sample images.

tion weights over images (Eq. 3), which is captured from image-text interactions. As can be seen, attentions are able to highlight salient regions that signal the essential semantic links with the texts, e.g., the entities (dog and jeep) in (a) and (b). It is also observed that the attention would vary in their focus in regions: for entity level discourse, it tends to concentrate on the some parts of a salient object (entity), while for scene level, attention also examines the background to capture the global view.

7 Conclusion

We have presented a novel task to learn cross-modality discourse that advances models to gain social media language understanding capability in multimedia contexts. To handle the intricate image-text interactions, the visual semantics are first converted into text-formed subtitles and then compared with post texts to explore deep syntactic relations in linguistic space. For empirical studies, we further contribute the first dataset presenting 16K human-annotated tweets with discourse labels for image-text pairs. The main comparison results on our dataset have shown the effectiveness of multi-head attentions in exploring interactions among text, image, and caption modalities. Further discussions demonstrate our potential to produce meaningful representations indicating implicit image-text structure. These discourse features, conveying essential linguistic clues consistent with human senses, may largely benefit the future advances of automatic cross-modality understanding on social media.

Limitations

Class imbalance is one of the main limitations of this work. As illustrated in Table 1, Concretization is the majority category which occupies 66.0% of the dataset, while the minority categories, e.g. Projection and Insertion only account for 4.3% and 5.2% respectively. Although such uneven distribution reflects the real scenario of image-text relationships among tweets, future work should acquire a larger amount of minority categories for better interpretation of image-text relationships.

Cross-lingual and multi-platform studies should also be considered in later studies. It would be interesting and insightful to investigate the cross-modality discourse categories distribution among different languages. Are there any cultural traits that affect the use of image and text? Meanwhile, social media platforms can also exhibit preference for image and text usage. For example, will users on Instagram prefer to omit the Name Entities (Insertion category) than Twitter users?

A more concrete model, e.g. vision-language Transformers, could also be employed to encode the text, caption, and image jointly. Current model runs efficiently on single NVIDIA RTX3080Ti GPU, while the training consumption of vision-language Transformers could be costly and requires larger dataset. Future studies could explore the trade-off between computation cost and classification performance.

Ethical Considerations

We declare our dataset will cause no ethics problem. First, we follow the standard data acquisition process regularized by Twitter API. We downloaded the data for a purpose of academic research and is consistent with the Twitter terms of use. Then, we thoroughly navigated the data and ensured that no content will rise any ethics concerns, e.g. toxic languages, human face images, and censored images. Next, we perform the data anonymization to protect the user privacy. For the language use, we only keep the posts with English text. For the human annotations, we recruited the annotators as part-time research assistants with 16 USD/hour payment.

Acknowledgements

This paper is substantially supported by NSFC Young Scientists Fund (No.62006203, 62106105), 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).

References

- Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020a. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689.
- Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel R. Tetreault, and Alejandro Jaimes. 2020b. Multimodal categorization of crisis events in social media. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14667–14677. IEEE.
- Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. 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 570–575. Association for Computational Linguistics.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2212–2218. The Association for Computational Linguistics.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019a. Multimodal 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.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019b. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.
- Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco. 2019. Extracting possessions from social media: Images complement language. 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 663–672. Association for Computational Linguistics.
- Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5374–5386. Association for Computational Linguistics.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Lun Huang, Wenmin Wang, Yaxian Xia, and Jie Chen. 2019. Adaptively aligned image captioning via adaptive attention time. In *Advances in Neural Information Processing Systems*, pages 8942–8951.
- Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 996–1005. Association for Computational Linguistics.
- Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. 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 4395–4405. Association for Computational Linguistics.
- Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4621–4631. Association for Computational Linguistics.
- Jing Li, Yan Song, Zhongyu Wei, and Kam-Fai Wong. 2018. A joint model of conversational discourse and latent topics on microblogs. *Comput. Linguistics*, 44(4).
- Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2557–2568. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014a. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014b. 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.
- Linlin Liu, Xiang Lin, Shafiq R. Joty, Simeng Han, and Lidong Bing. 2019. Hierarchical pointer net parsing. 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 1007–1017. Association for Computational Linguistics.
- 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.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 289–297.
- William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
- Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.
- Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.
- Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307.

- 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.
- Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695.
- Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2019. Towards personalized image captioning via multimodal memory networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(4):999–1012.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.
- Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2263–2270, Austin, Texas. Association for Computational Linguistics.
- Keith Rayner. 2009. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506.
- Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.
- Attapol Rutherford and Nianwen Xue. 2016. Robust non-explicit neural discourse parser in English and Chinese. In *Proceedings of the CoNLL-16 shared task*, pages 55–59, Berlin, Germany. Association for Computational Linguistics.
- Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Improved semantic-aware network embedding with fine-grained word alignment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1829–1838. Association for Computational Linguistics.
- Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 269–277, Online. Association for Computational Linguistics.
- Qingyi Si, Zheng Lin, Ming yu Zheng, Peng Fu, and Weiping Wang. 2021. Check it again: progressive visual question answering via visual entailment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4101–4110, Online. Association for Computational Linguistics.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based BERT model for multimodal NER. 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 13860–13868. AAAI Press.
- Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. 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 5099–5110. Association for Computational Linguistics.
- Quoc-Tuan Truong and Hady W Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 305–312.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- 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.
- Yue Wang, Shafiq R. Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C. H. Hoi. 2020a. VD-BERT: A unified vision and dialog transformer with BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3325–3338. 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.
- Yue Wang, Jing Li, Michael R. Lyu, and Irwin King. 2020b. Cross-media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3311–3324. Association for Computational Linguistics.
- Yue Wang, Jing Li, Michael R Lyu, and Irwin King. 2020c. Cross-media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings. *arXiv preprint arXiv:2011.01565*.
- Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pages 451–466. Springer.
- Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Discourse-aware neural extractive text summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5021–5031. Association for Computational Linguistics.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.
- Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou. 2018. Using active learning to expand training data for implicit discourse relation recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 725–731. Association for Computational Linguistics.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.
- Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. [Vision guided generative pre-trained language models for multimodal abstractive summarization](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6281–6290. Computer Vision Foundation / IEEE.
- Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. [Vinvl: Revisiting visual representations in vision-language models](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE.
- Suwei Zhang, Yuan Yao, Feng Xu, Hanghang Tong, Xiaohui Yan, and Jian Lu. 2019. Hashtag recommendation for photo sharing services. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5805–5812. AAAI Press.
- Luwei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. 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 13041–13049. AAAI Press.