2-2020

# Image enhanced event detection in news articles

Meihan TONG

Shuai WANG

Yixin CAO
*Singapore Management University*, yxcao@smu.edu.sg

Bin XU

Juaizi LI

*See next page for additional authors*

# Image Enhanced Event Detection in News Articles

**Meihan Tong,**[*1] **Shuai Wang,**[*1]
**Yixin Cao,**[2] **Bin Xu,**[†1] **Juaizi Li,**[1] **Lei Hou,**[1] **Tat-Seng Chua**[2]
[1]Tsinghua University, [2]National University of Singapore
{tongmeihan, caoyixin2011, greener2009}@gmail.com, 18813129752@163.com,
{xubin, lijuanzi}@tsinghua.edu.cn, dcscts@nus.edu.sg

## Abstract

Event detection is a crucial and challenging sub-task of event extraction, which suffers from a severe ambiguity issue of trigger words. Existing works mainly focus on using textual context information, while there naturally exist many images accompanied by news articles that are yet to be explored. We believe that images not only reflect the core events of the text, but are also helpful for the disambiguation of trigger words. In this paper, we first contribute an image dataset supplement to ED benchmarks (i.e., ACE2005) for training and evaluation. We then propose a novel **D**ual **R**ecurrent **M**ultimodal **M**odel, DRMM, to conduct deep interactions between images and sentences for modality features aggregation. DRMM utilizes pre-trained BERT and ResNet to encode sentences and images, and employs an alternating dual attention to select informative features for mutual enhancements. Our superior performance compared to six state-of-art baselines as well as further ablation studies demonstrate the significance of image modality and effectiveness of the proposed architecture. The code and image dataset are avaliable at https://github.com/shuaiwa16/image-enhanced-event-extraction.

## Introduction

In Automatic Content Extraction (ACE), Event Detection (ED) aims to identify *event triggers* from sentences. *Event trigger* is the word that most clearly expresses the occurrence of an event (Doddington et al. 2004). In the left example in Figure 1, since *confront* indicates the occurrence of event *Meet*, it should be labeled as the event trigger of *Meet*. Detecting events in natural languages is significant for a variety of NLP tasks, such as information retrieval and question answering.

ED is a challenging task because the trigger words must be representative, which unfortunately usually are ambiguous. A single word can trigger different events, and the surrounding contexts are often not informative enough to disambiguate them. For example, in Figure 1, the trigger word *confront* evokes different events: *Meet* and *Attack*, as they

*Meet*
Ford was in his rush to **confront** members in Toronto's council chamber.

*Attack*
Police **confronted** protesters hurling stones

Figure 1: An example of ED. The word *confront*, noted in bold, is the event trigger, which triggers *Meet* and *Attack* event respectively. Images play a vital role in disambiguating the event trigger.

expresses distinct meanings. Existing methods cope with this problem by introducing global contexts in the whole passage (Duan, He, and Zhao 2017; Chen et al. 2018), or incorporating some extra linguistic resources (Liu et al. 2018b; Lu and Nguyen 2018).

Actually, the source data of ED task, such as news articles, are naturally accompanied by images on the web, but they are completely neglected by most existing methods. Images have been proved to be very effective to handle textual ambiguity (Zhang et al. 2018; Moon, Neves, and Carvalho 2018; Elliott, Frank, and Hasler 2015), and images are very suitable in ED scenario from two aspects. (1) The accompanied images usually reflect the core events of the texts. As shown in Figure 1, the first example contains two candidate verbs of the event trigger: *was* and *confront*, where *confront* is more representative in texts and is also the main content of the image. (2) Images are helpful for the disambiguation of trigger words as they provide complementary information, which is difficult to be depicted by words, such as dressing styles, facial expressions or motions. Zhang et al. (2017) also show the effect of images in ED, and by utilizing images on the disambiguation of entities, they obviously improve the performance on ED.

In this paper, we incorporate the original images of news articles into ED. It is a non-trivial task due to the following challenges. First, there is no image available in the ex-

[*]Authors contributed equally to this work, ordered alphabetically by surname.
[†]Corresponding author.

isting benchmarks, such as ACE2005. It is difficult to find appropriate images for these news articles, which has to be done manually. Second, despite multimodality tasks are increasingly researched, there is still no well-acknowledged method for merging image modality into NLP tasks. The semantic level at which the image should match also needs to be carefully considered. In ED scenario, images should help model recognize specific events, so these images should map to events rather than specific words, sentences or entities like in (Zhang et al. 2017), so the shallow connections in existing approaches unable to deal such a relation.

To address the issues, we manually supplement images dataset for benchmark ACE2005, and propose a novel **D**ual **R**ecurrent **M**ultimodal **Model** (DRMM) to conduct deep interactions between images and sentences for modality feature aggregation. We manually recover visual contexts for articles in ACE2005 by searching the original website, and expand our dataset by searching images from other four authoritative websites. The extension allows our datasets to contain rich images depicting events in different angles. Our proposed model DRMM adopts a recurrent network to sequentially encode multiple images and employs a novel alternating dual attention at each step to pick up informative textual information and filter out irrelevant noise for feature abstraction. The novel alternation dual attention has a two-round structure for deep interaction between text and image modalities, capable of repeatedly merging useful event-related images and texts.

We conduct a variety of experiments on our image-enhanced ACE dataset. The overall result strikingly outperforms the current SOTA approaches in ED. The subsequent ablation experiments demonstrate the significance of introducing image modality and the superiority of the proposed DRMM in ED. The experiments also show that the image modality is especially helpful for low-frequency triggers, which also alleviate data sparsity problem in ED.

Our contributions can be summarized as follows:

- We manually construct image datasets for Event Detection benchmark ACE2005, which may also benefit other related tasks in event extraction.
- We propose a novel dual recurrent multi-modal model (DRMM) to integrate two types of modality features via an alternating dual attention mechanism. It thus conducts deep interactions between images and sentences.
- For evaluation, we have verified the quality of the constructed image enhanced ED datasets based on language model. We conducted a series of experiments on the benchmark ACE2005, and compared with six state-of-the-art baseline models. The results as well as further ablation studies demonstrate the effectiveness of our model.

## Related Work

### Event Detection (ED)

In Automatic Content Extraction (ACE), event detection (ED) aims to detect event triggers (usually verbs or nouns) from unstructured news reports, which has a long history of research (Ahn 2006; Nguyen and Grishman 2018). ED serves as the fundamental task in information extraction, same as NER (Cao et al. 2019) and entity linking (Cao et al. 2017; 2018). Due to the flexibility and diversity of natural language, event triggers can be very ambiguous (Hogenboom et al. 2011). The same event trigger can trigger different events in various contexts. Previous methods prove lexical and sentence-level information quite helpful for event detection (Ahn 2006; Nguyen and Grishman 2015).

Several researchers further incorporate document-level information to disambiguate the event (Duan, He, and Zhao 2017; Chen et al. 2018; Liu et al. 2018b). Other researchers use multiple linguistic resources to enhance event semantic understanding. Liu et al. (2018a) proposes a gated attention to dynamically integrate parallel training corpus from different languages. In addition, open-domain lexical database (WordNet, FrameNet) is adopted as extra auxiliary resources (Lu and Nguyen 2018; Liu et al. 2016) or extra training datasets (Liu et al. 2016; Wang et al. 2019) to improve event detection performance.

However, Event does not solely exist in textual modality (Zhang et al. 2017). All the above methods totally ignore information from different heterogeneous sources like image. We propose a novel dual recurrent multimodal model to leverage visual context in the news article to improve event detection.

### Multimodal Learning

Multimodal learning aims to build models that can integrate information from heterogeneous moralities, such as image, video and audio. Recently, multimodal learning has been widely adopted to handle NLP issues, such as NER (Moon, Neves, and Carvalho 2018) and machine translation (Heo, Kang, and Yoo 2019). These approaches enhance short and coarse text understanding from the perspective of visual context, and propose various modality attentions to integrate information from different heterogeneous sources.

Zhang et al. (2017) integrate image modality into ED by visualizing entities in sentences, but event typically scatters all over the article and is unsuitable to disambiguate at the entity level. Our work manually recovers the original images that directly reflect event semantics and proposes a novel alternating dual attention to squeeze multiple images into the disambiguation process to ensure panoramic observation of image modality. Extensive experiments on benchmark demonstrate the effectiveness of this design.

## Methodology

Figure 2 illustrates our **D**ual **R**ecurrent **M**ultiModal **M**odel (DRMM). DRMM has three components. First, **Feature Extraction** extracts text and image features from large-scale pre-trained BERT and ResNet network. Next, **Multimodal Integration** performs two round for deep interaction between text and image modalities with a novel alternating dual attention (ADA). Finally, **Event Prediction** employs a fully connected layer to map the final multimodal representation to the event-type semantic space to complete event detection.

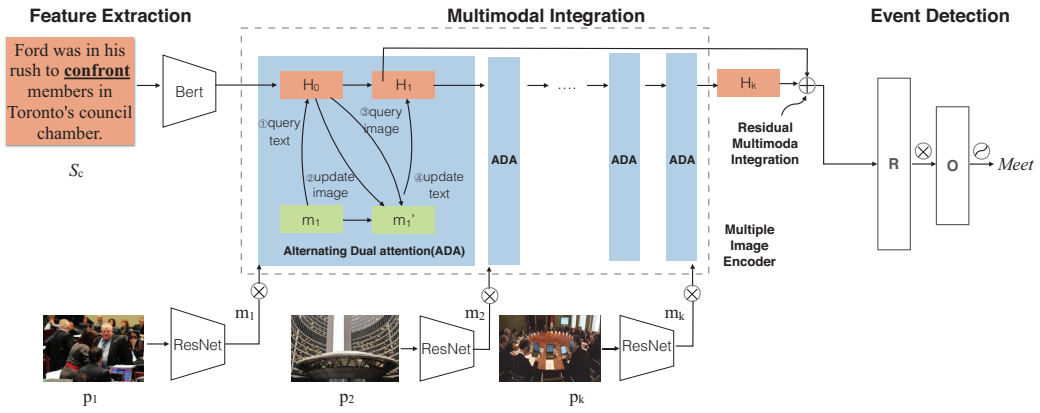


Figure 2: The architecture of the proposed **Dual Recurrent MultiModal Model** (DRMM). From left to right, DRMM first extracts text and image features from pre-trained BERT and ResNet respectively. Next it enhances text representation $H$ with image modality knowledge $p_1, p_2, p_3$ via a novel Alternating Dual Attention (ADA). Finally, DRMM detects the event via a fully connected layer. As indicated in the dotted box, DRMM processes image modality information step by step via a recurrent structure, with ADA as its basic unit. At each step, ADA first refines image representation from the text side, and then reversely updates text representation from the image side.

## Notation

Following Feng, Qin, and Liu (2018), we regard event detection as a sequence tagging task. Formally, given a sentence $S = \langle w_1, w_2, \ldots, w_n \rangle$ and its related multiple images $P = \{p_1, p_2, \ldots, p_k\}$, event detection aims to identify the event type $Y = \{y_1, y_2, \ldots, y_n\}$, where $Y$ has 34 categories in ACE. If the word $w_i$ is not an event trigger, which is the most common case, $y_i$ will turn out to be *Negative*.

## Image Dataset Construction

We manually recover illustrations of news articles in ACE2005 from the original website [1]. However, the original news websites usually provide no or very few images. Based on the fact that the same event is often reported by many different websites and these websites sometimes provide their own images, we expand the image dataset by searching for news from four more news websites: CNN, Fox News, NPR and The Guardian, which are authoritative and able to ensure the quality of the images. The 'same event' is defined as the events sharing the same event arguments: subject, object and place. For instance, '*Wildfires Rip Through Southern California*' reported by NPR is the same event as *Massive wildfires rage in California* reported by CNN, since they both report a fire event and share the same event arguments: *Massive wildfires* and *California*. Event arguments are obtained by parsing the title of the news with AMR parser (Banarescu et al. 2013). We try to include more images of the same event, even they are in different years, and find the date issue has no negative impacts on the detection of events in ACE. We employ 3 students and adopt the union of their searched images as the final collection of images. Finally, we acquire 2815 images altogether for ACE2005.

[1] https://www.nytimes.com

## Feature Extraction

In the section, we illustrate the details of Feature Extraction layer. Since event exists not only in text modality but also in image modality, we simultaneously extract features from text and image modalities.

**Text Feature Extraction** BERT (Devlin et al. 2018) is a pre-trained language representation model and has achieved great success on a wide range of down-streaming natural language tasks, like conversational systems, question answering and event detection. The powerful capability of BERT is applicable to event detection, and we conduct extensive experiments to show that the fine-tuned BERT model has achieved superior performance.

We adopt BERT as our text feature extractor. Formally, we feed the input sentence $S = \langle w1, w2, \ldots, w_n \rangle$ into BERT and use the sequential output as the sentence representation $H_0 = \langle h_1, h_2, \ldots, h_n \rangle$.

$$H_0 = BERT(S) \tag{1}$$

**Image Feature Extraction** ResNet has been found to be an effective image representation (He et al. 2016). Given multiple images $P = \{p_1, p_2, \ldots, p_k\}$ in the news article, we feed each image $p_i$ into ResNet, and then adopt the last residual block output as the image hidden representation $u_i$.

$$u_i = ResNet(p_i) \tag{2}$$

To map images into the same latitude space as text (from 2048 to 768), we adopt a sigmoid function to generate the final image representation $m_i$:

$$m_i = \sigma(W_u u_i + b_u) \tag{3}$$

## Multimodal Integration

In the section, we illustrate the procedures in multimodal integration. We first obtain an image-enhanced text representation via a recurrent multiple images encoder. At each

step, we propose a novel Alternating Dual Attention (ADA) to first refine the image representation with textual information and then reversely for deep interaction. After that, we aggregate image-enhanced text representation and the original outputs of BERT with a residual network to get the final multimodal representation.

**Multiple Images Encoder** In news articles, multiple images tend to portray an event from a different perspective. For instance, when we talk about an *Earthquake* event, we may talk about the damage situation with an image of a road collapse. We may also refer to the reconstruction situation with an image of workers carrying the sheets. Different from previous approaches (Zhang et al. 2018; Heo, Kang, and Yoo 2019) which only consider single image, our method is able to dynamically aggregate information from multiple images information to disambiguate the event.

The structure of the multiple images encoder is shown in the dotted box in Figure 2. Multiple images encoder recurrently updates text representation by reading multiple images sequentially. Specifically, at the $t$-th step, multiple images encoder relies on image representation $m_t$ to update the previous image-enhanced text representation $H_t$ into the new image-enhanced text representation $H_{t+1}$. The updating procedure is carried out by a novel block called Alternating Dual Attention (ADA).

We will first give a formulated description of ADA, which serves as the basic unit of multiple images encoder, and then illustrate the whole procedure of multiple images encoder.

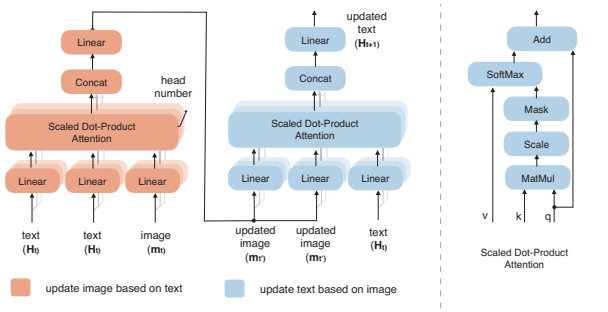


Figure 3: The illustration of Alternating Dual Attention (ADA). ADA has a dual structure for deep interaction, which first updates image representation based on textual information (indicated by the red part), and then reversely (indicated by the blue part). Two of the three inputs for each part are the same, one for attention calculate and the other for residual integration.

**Basic Unit: Alternating Dual Attention (ADA)** As shown in Figure 3, ADA has a dual structure, in the sense that textual information is used to guide the image attention, and then image information is adopted to guide the textual attention. We employ the dual structure since the image and text information affect each other. In different text backgrounds, the focus areas of the same image are different. Also, the same word can describe different events in different visual contexts.

Specifically, ADA is a two round multi-head attention module. We first introduce the first round, and then the second round.

For the first round (indicated in red part in Figure 3), ADA aims to update image representation by textual information. Formally, we employ three fully connected layers to map text representation $H_t$ into the first two inputs and image representation $m_t$ into the third input of the scaled dot-product attention module, noted as $v$, $k$ and $q$ respectively.

We then calculate the attention $\alpha$ by querying $k$ with $q$. We re-scale the attention value by dividing the dimension of $k$ to avoid vanishing gradients (Vaswani et al. 2017). Next, we dot-product the learned attention $\alpha$ with the third input $v$ to obtain the weighted image representation $z$.

$$\begin{aligned} \mathbf{s} &= \frac{\mathbf{q} \cdot \mathbf{k}}{\sqrt{d_k}} \\ \alpha_i &= \frac{s_i}{\sum_{i=1}^{L} s_i} \\ z &= \alpha \mathbf{v}^{\mathrm{T}} \end{aligned} \tag{4}$$

We repeat the above procedure $u$ times and adopt a linear transformation to obtain the final attention-revised image representation $h$

$$\begin{aligned} Z &= [z_1; z_2; \ldots; z_u] \\ h &= W_h Z + b_o \end{aligned} \tag{5}$$

where ";" indicates concatenation operation.

Finally, we adopt a residual block to directly send the query signal $q_p$ to the attention-revised output $h_p$ to obtain the refined image representation $m_t^{'}$ at $t$-th step.

$$m_t^{'} = h + q \tag{6}$$

We denote the operations from Formula 4 to 6 as $\Omega$. Then, the first round procedure can be summarized as

$$m_t^{'} = \Omega(m_t, H_t) \tag{7}$$

For the second round (indicated in blue part in Figure 3), ADA aims to update text representation by image information. The middle operations are the same as the first round, but the input is different. We exchange the input by mapping $H_t$ into the third input and $m_t^{'}$ into the first and second inputs in the scaled dot-product attention module. We formulate the second round procedure as:

$$H_{t+1} = \Omega(m_t^{'}, H_t) \tag{8}$$

**Recurrent Structure** We adopt the recurrent structure to consider multiple images in the article to disambiguate each event trigger. Formula 7 and 8 probe into the recurrent structure at step $t$, when ADA exploits the $t$th image $m_t$ to update text representation from $H_t$ to $H_{t+1}$. Denoting the images of the news article as $M = \langle m_0, m_1, \ldots, m_k \rangle$, we update text representation by exploiting images in $M$ sequentially. The output of the last step of recurrent procedure $H_k$ is adopted as the final image-enhanced text representation.

**Residual Integration** Instead of directly adopting the last step output of ADA $H_k$ as the final multimodal representation, we employ a residual block to integrate text-only representation $H_0$ back to image-enhanced text representation $H_k$. We want the final multimodal representation $R$ still preserves the original text semantics as much as possible. We also consider from the perspective of optimization. By bridging BERT output $H_0$ into the final multimodal representation $R$, we prevent the parameters in BERT from gradient vanishing during the training procedure.

$$R = H_0 + H_k \tag{9}$$

## Event Prediction

In this section, we aim to illustrate the event detector module. As illustrated in Figure 2, given the output of Multimodal Integration $R$, we employ a non-linear layer to transform the dimension of $R$ to the number of event types.

Let $x_i = \langle S, P \rangle$ and $y_i = Y$ denote the $i$-th training sample, where $S$, $P$ ,$Y$ respectively represent the sentence, multiple images and event label from the same news article. Event Prediction will output a result vector $O$, where $O_{ijc}$ represents the probability that the $j$-th word in $x_i$ belongs to the $c$-th event class. The conditional probability is normalized by the softmax function.

$$p(y_{(i)}|x_{(i)}, \theta) = \sum_{j=1}^{n} \frac{exp(o_{ijc})}{\sum_{c=1}^{C} exp(o_{ijc})} / n \tag{10}$$

Given the input corpus $D = \{x_i, y_i\}|_{i=1}^{I}$, the negative loss function is defined as:

$$J(\theta) = -\sum_{i=1}^{I} \log p(y_{(i)}|x_{(i)}, \theta) \tag{11}$$

We use Adam as the gradient descent optimizer.

# Experiment

In this section, we evaluate the proposed dataset and approach by extensive experiments. We first give a description of dataset and hyperparameters in the experiment. We then will compare our results with several existing SOTA approaches on the same benchmarks to show the effectiveness of our image dataset and the superiority of the proposed approach. Next, we conduct experiments to answer three questions: 1) the quality of images, 2) whether to use images and 3) how to use images. Finally, we analyze when and how the images are helpful in ED by a case study.

## Experiment Setup

**Datasets** We employ the publicly available dataset in Event Detection ACE2005. ACE corpus includes 6 news areas, a total of 8 event types and 33 subtypes[2]. We directly classify the subtypes of event. The size of train/dev/test for ACE2005 is 529/30/40 (Chen et al. 2015). Each article in ACE2005 corresponds to several images (uncertain number). The details of our image dataset is shown in Figure 2. The images are human-searched on news websites mentioned above. 2815 images are collected altogether with 4.7 images for each article on average.

**Hyperparameters** We encode sentences by pre-trained BERT and images by pre-trained ResNet50. We expand the final pooling layer of Resnet50 from 7*7*1024 feature map to a 49 * 1024 sequence. In the integration module, we employ a multi-head attention with 8 heads and 768 hidden units. Additionally, We add an identity connection from the output of BERT to the final output. Our batch size is 32, learning rate being 2e-5, and epoch is 4. Our codes are implemented by tensorflow and all models can be fit into a single GPU with the help of Tensorflow Large Model Support[3]. We will make all our datasets and source code publicly available once the paper is published.

**Baselines.** We denote the proposed method as DRMM. To validate its effectiveness, we compare DRMM with the following baselines. **VAD**: an image-enhanced event detection model that incorporates visual knowledge at word and phrase level (Zhang et al. 2017). **DLRNN**: a LSTM-based model extracting cross-sentence clues to improve the sentence-level event detection (Duan, He, and Zhao 2017). **ANN-FN**: ANN-FN aligns the taxonomy of FrameNet with ACE to obtain more training corpus. (Liu et al. 2016). **GMLATT**: a gated multilingual attention approach. It is the best reported sentence-level attention approaches (Liu et al. 2018a). **HBTNGMA**: a hierarchical and bias tagging networks to detect multiple events and gated to fuse the sentence-level and document-level information with multi-level attention (Chen et al. 2018). **AD-DMBERT**: an adversarial imitation based event detection model which adopts BERT as the basic feature extractor (Wang et al. 2019).

Table 1: Overall Performance on ACE2005 dataset (%)

| Method | Precision | Recall | F1 |
|---|---|---|---|
| VAD | 75.1 | 64.3 | 69.3 |
| DLRNN | 77.2 | 64.9 | 70.5 |
| ANN-FN | 77.6 | 65.2 | 70.7 |
| GMLATT | **78.9** | 66.9 | 72.4 |
| HBTNGMA | 77.9 | 69.1 | 73.3 |
| AD-DMBERT | 77.9 | 72.5 | 75.1 |
| DRMM(Our) | 77.9 | **74.8** | **76.3** |

Table 2: Statistics of our image dataset.

| Measure | number |
|---|---|
| total number | 2815 |
| average per article | 4.7 |
| max number per article | 6 |
| min number per article | 3 |

## Overall Performance

We present the overall performance of the proposed approach on ACE2005 in Table 1. As shown in Table 1,

[2]https://catalog.ldc.upenn.edu/LDC2006T06

[3]https://github.com/IBM/tensorflow-large-model-support

Table 3: The performance of the language model with and without integration of images.

| Dataset | LM | LM-image |
|---|---|---|
| language model | 75.8 | 79.4 |

DRMM (our) outperforms previous state-of-the-art models, showing the effectiveness of introducing image modality into ED and the superiority of the proposed alternative dual attention. Compared with VAD, which also incorporates image dataset, our method improves F score by over 7%. This indicates that the choice of images and the method of fusion are of vital importance on the performance. VAD introduces entity level images and fuses multi-modality features by naive fully connected layer, unable to make a deeper connection between the two modalities and makes an event level interaction. AD-DMBERT and DRMM (our) both employs BERT as feature extractor. The difference is that AD-DMBERT uses text while we incorporate images as extra resources. Our approach obviously outperforms AD-DMBERT, which implies the superiority of multi-modality resources. Another interesting phenomenon is that the proposed method principally achieves the highest recall. Due to the small scale of ACE, many triggers suffer from zero-shot and few-shot issue, so it is difficult to retrieve events based solely on textual information. Knowledge from image modality provides similarities to the event trigger's distributional semantics with other training examples, and thus our model successfully retrieves more events.

## Evaluation of Image Dataset

Since the image dataset is one of the most principal contribution of the paper, we evaluate the quality of images by a series of experiments. Firstly the statistics of our image dataset is given in Table 2.

To validate the effectiveness of the image dataset, two questions need to be answered. The first is to what extent news articles are related to the images. It is necessary that images are closely related to their articles. Otherwise, the images are noises that may harm the understanding of texts. Secondly, how much extra information images can provide to the understanding of texts.

We answer the first question by an image caption task. Specifically, we pretrain an image caption model (Wang, Li, and Lazebnik 2016) by replacing the text and image representation by BERT and ResNet50. Then, we search images based on articles in by the model. If the resulting image is the illustration of the text, we treat it as correct otherwise wrong. The top3 accuracy is 75%. It is obvious that images are closely related to their according articles.

To answer the second question, a language model is adopted. As the fundamental task in NLP, language model reflects the understanding of text, as demonstrated in BERT (Devlin et al. 2018). Hence, if image information helps to train a better language model, then images are considered to provide extra information for the understanding of texts. We train a Masked Language Model (MLM) as in BERT (Devlin et al. 2018) with and without image incorporation. The metrics is F score. As ACE2005 is a small dataset, we only mask words with top 50 frequency to control the vocabulary size to an appropriate number. The result is shown in Table 3. We can see the obvious improvement of image-enhanced language model, which shows the significance of image information for the understanding of texts.

## Effectiveness of Image Modality Knowledge

In the section, we discuss how much improvement the image modality brings to Event Detection. We employ the same models with and without image modality on ACE2005 dataset. Different from the overall part, we employ three different base encoders, including CNN, RNN and BERT, in order to show the improvement of image modality on ED is omnipresent rather than a model-related situation. We employ two layers of Bi-LSTM model with hidden units 384 for each direction. For CNN model, we employ three layers of convolution with kernel size of 3,4,5 respectively and a final fully connected layer with 768 output units.

As shown in Table 5, the incorporation of image modality improves the performance of Event Detection on Precision, Recall and F score independent to specific models. The three models are the most commonly used neural network models, so the results validate the significance of image modality in Event Detection. Note that the improvement for CNN and LSTM encoders is obviously bigger than that on BERT, which reflects the complementary role of images in Event Detection. When the capacity of text encoder is small, images can bring in larger improvement.

Dataset distribution is skewed in ACE2005 and most events have scarce annotated data. We are curious about whether image modality knowledge would gain more benefits on the data scarcity situation. We analyze the performance of BERT+image in zero-shot, few-shot and high-frequency situations in Figure 4. Our taxonomy is the frequency of occurrence of event triggers in the training dataset. For instance, we regard triggers that do not appear in the training corpus as the zero-shot event trigger.

Precision results show that BERT+image obviously surpasses text-only method (BERT) on zero-shot (+3.9%) and few-shot (+2.7%) situations and has comparable performance on high-frequency case (-0.7%). The results indicate that image modality knowledge is particularly effective for low-resource settings. The failure of text-only method (BERT) in low-resource settings may be due to the flexibility of the sentence representation. With less training data, even the same event can have very different expressions on the text modality. Therefore, it is better to consider the similarity on the image modality simultaneously to correctly distinguish the types of events.

## Effectiveness of Multimodal Fusion

As mentioned above, it is still an open problem to integrate image modality into textual tasks because of the complication of scenarios. In order to show the superiority of DRMM (our), we compare DRMM with three common multimodal fusion approaches, including the traditional concatenation, modality attention (Moon, Neves, and Carvalho

Table 4: Error analysis: When does the image modality knowledge improve ED? GT is the ground truth and event triggers are marked by underlined. For interpretability, we describe images from the perspective of people, background and action instead of showing the actual figure vector.

| | Sentence | Image Tags | GT | Prediction | |
|---|---|---|---|---|---|
| | | | | Text | Text+Image |
| + | *S1: We do not think that America **won**, said Dmitry Rogozin.* | armed soldier, battlefield, explosion | **O** | **Elect** | **O** |
| | *S2:Thousands of Iraq's majority Shiite Muslims **marched** to their main mosque.* | protest crowd, chaotic street, shouting | **Demonstrate** | **Transport** | **Demonstrate** |
| | *S3: Palestinian forces returned before the outbreak of the 33-month palestinian **uprising**.* | armed soldier, bloody bus, conflicting | **Attack** | **O** | **Attack** |
| - | *S4: The EU is set to **release** 20 million eurosin immediate humanitarian aid for iraq* | wounded people, refuge tents, rescue | **Transfer -Money** | **Transfer -Money** | **O** |

Table 5: The evaluation of image modality.

| Method | Precision | Recall | F1 |
|---|---|---|---|
| CNN | 72.3 | 51.2 | 59.9 |
| CNN+image | 74.9 | 56.1 | 63.3 |
| improvement | **+2.6** | **+4.9** | **+3.4** |
| LSTM | 71.2 | 52.2 | 60.2 |
| LSTM+image | 74.3 | 58.3 | 64.8 |
| improvement | **+3.1** | **+5.1** | **+4.6** |
| BERT | 76.4 | 73.8 | 75.1 |
| BERT+image | 77.9 | 74.8 | 76.3 |
| improvement | **+1.5** | **+1.0** | **+1.2** |

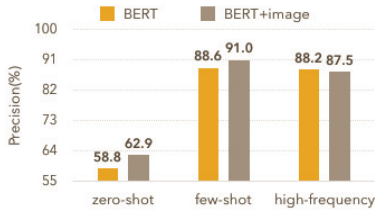


Figure 4: Precision of BERT+image in zero-shot, few-shot and high-frequency situations.

2018) and co-attention (Qian et al. 2017). The knowledge in co-attention model refers to images in our setting.

Results from Table 6 indicate that DRMM outperforms all of the common fusion approaches by over 1.5%. The failure of concatenation is inevitable due to equal treatment of multimodality information, unsuitable in ED scenarios in which textual and image modalities playing leading and supporting roles respectively. Modality attention and co-attention also are inferior to DRMM by ignoring the importance of contextual information, which emphasized by several approaches in the fusion process (Atrey, Kankanhalli, and Jain 2006).

## Case Study and Error Analysis

Table 4 gives example cases about how image modality knowledge affects predictions of ED. In S1, as '*won*' always meaning election victory in the training corpus, text-only method turns to overfitting, and thus mistakenly thinks '*won*' triggers an '*Elect*' event. The image modality knowledge "*soldier, battlefield, explosion*" helps disambiguate the event trigger, making the model correctly predict it as a non-trigger word. In S2, the event trigger '*marched*' itself refers to walk in a military manner, making text-only method mistakenly classifies it as a '*Transport*' event. However, by considering the image modality knowledge '*protest crowd, chaotic street, shouting*', '*marched*' is more suitable to recognize as the event trigger of '*Demonstrate*' in this context. In S3, with few descriptions about riots in the surrounding context, text-only method becomes confused and conservative, erroneously thinking '*uprising*' does not trigger an event. However, with extra knowledge from image modality '*soldier, blood stain wall, conflicting*', our model successfully recognizes that '*uprising*' is the event trigger of '*Attack*'. In a few cases, image modality knowledge harms the performance of ED, primarily because images are unrelated to the event trigger or the surrounding textual contexts. For instance, "release" triggers a "Transfer-Money" event in S4, but the mainly content of the article describes the war in Iraq, and so do the images, making it impossible to disambiguate the "Transfer-Money" event. In the future, we will try to remove unrelated or low-quality images before the model.

Table 6: Effectiveness of multimodal fusion in DRMM

| Fusion Methods | Precision | Recall | F |
|---|---|---|---|
| Concatenation | 71.2 | 67.3 | 69.2 |
| Modality Attention | 78.9 | 69.4 | 73.8 |
| Co-Attention | 75.3 | 74.0 | 74.6 |
| DRMM(our) | 77.9 | **74.8** | **76.3** |

## Conclusion

In this paper, we propose to utilize accompanied images in news articles to enhance Event Detection. We contribute a supplement image dataset for ED benchmark ACE2005, which can be further analyzed in related tasks such as event extraction. For image enhanced ED, we propose a novel fusion method, DRMM, which conducts a deeper connection between the two modalities and makes an event level interaction. For evaluation, not only we verify the quality of the image datasets supplement to ACE2005, but also conduct a series of experiments on it. The results are compared with

six baseline methods demonstrate effectiveness of DRMM.

## Acknowledgments

This work is supported by the National Key Research and Development Program of China (2018YFB1005100 and 2018YFB1005101), NSFC key projects (U1736204, 61533018, 61661146007). It also got partial support from National Engineering Laboratory for Cyberlearning and Intelligent Technology, and Beijing Key Lab of Networked Multimedia. This research is part of NExT research which is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative. We would like to thank Fan Ding [4] for his contribution in data construction and eperimental assistance.

## References

Ahn, D. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, 1–8.

Atrey, P. K.; Kankanhalli, M. S.; and Jain, R. 2006. Information assimilation framework for event detection in multimedia surveillance systems. *Multimedia systems* 12(3):239–253.

Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; and Schneider, N. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186.

Cao, Y.; Huang, L.; Ji, H.; Chen, X.; and Li, J. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *ACL*, 1623–1633.

Cao, Y.; Hou, L.; Li, J.; and Liu, Z. 2018. Neural collective entity linking. In *COLING*, 675–686.

Cao, Y.; Hu, Z.; Chua, T.-s.; Liu, Z.; and Ji, H. 2019. Low-resource name tagging learned with weakly labeled data. In *(EMNLP-IJCNLP)*, 261–270.

Chen, Y.; Xu, L.; Liu, K.; Zeng, D.; and Zhao, J. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *IJCNLP*, volume 1, 167–176.

Chen, Y.; Yang, H.; Liu, K.; Zhao, J.; and Jia, Y. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *EMNLP*, 1267–1276.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S. M.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, 1.

Duan, S.; He, R.; and Zhao, W. 2017. Exploiting document level information to improve event detection via recurrent neural networks. In *IJCNLP*, 352–361.

[4] ding-fan@outlook.com

Elliott, D.; Frank, S.; and Hasler, E. 2015. Multi-language image description with neural sequence models. *CoRR, abs/1510.04709*.

Feng, X.; Qin, B.; and Liu, T. 2018. A language-independent neural network for event detection. *Science China Information Sciences* 61(9):092106.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Heo, Y.; Kang, S.; and Yoo, D. 2019. Multimodal neural machine translation with weakly labeled images. *IEEE Access*.

Hogenboom, F.; Frasincar, F.; Kaymak, U.; and De Jong, F. 2011. An overview of event extraction from text. In *ISWC*, volume 779, 48–57. Citeseer.

Liu, S.; Chen, Y.; He, S.; Liu, K.; and Zhao, J. 2016. Leveraging framenet to improve automatic event detection. In *ACL*, volume 1, 2134–2143.

Liu, J.; Chen, Y.; Liu, K.; and Zhao, J. 2018a. Event detection via gated multilingual attention mechanism. *Statistics* 1000:1250.

Liu, S.; Cheng, R.; Yu, X.; and Cheng, X. 2018b. Exploiting contextual information via dynamic memory network for event detection. *arXiv preprint arXiv:1810.03449*.

Lu, W., and Nguyen, T. H. 2018. Similar but not the same: Word sense disambiguation improves event detection via neural representation matching. In *EMNLP*, 4822–4828.

Moon, S.; Neves, L.; and Carvalho, V. 2018. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862*.

Nguyen, T. H., and Grishman, R. 2015. Event detection and domain adaptation with convolutional neural networks. In *IJCNLP*, volume 2, 365–371.

Nguyen, T. H., and Grishman, R. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *AAAI*.

Qian, C.; Zhu, X.; Ling, Z.-H.; Inkpen, D.; and Wei, S. 2017. Neural natural language inference models enhanced with external knowledge. *arXiv preprint arXiv:1711.04289*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*. 5998–6008.

Wang, X.; Han, X.; Liu, Z.; Sun, M.; and Li, P. 2019. Adversarial training for weakly supervised event detection. In *NAACL*.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In *CVPR*, 5005–5013.

Zhang, T.; Whitehead, S.; Zhang, H.; Li, H.; Ellis, J.; Huang, L.; Liu, W.; Ji, H.; and Chang, S.-F. 2017. Improving event extraction via multimodal integration. In *MM*, 270–278. ACM.

Zhang, K.; Lv, G.; Wu, L.; Chen, E.; Liu, Q.; Wu, H.; and Wu, F. 2018. Image-enhanced multi-level sentence representation net for natural language inference. In *ICDM*, 747–756. IEEE.