Multimedia Event Extraction From News With a Unified Contrastive Learning Framework

Jian Liu
Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University
Beijing, China
jianliu@bjtu.edu.cn

Yufeng Chen
Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University
Beijing, China
chenyf@bjtu.edu.cn

Jinan Xu
Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University
Beijing, China
jaxu@bjtu.edu.cn

ABSTRACT

Extracting events from news have seen many benefits in downstream applications. Today's event extraction (EE) systems, however, usually focus on a single modality — either for text or image, and such methods suffer from incomplete information because a news document is typically presented in a multimedia format. In this paper, we propose a new method for multimedia EE by bridging the textual and visual modalities with a unified contrastive learning framework. Our central idea is to create a shared space for texts and images in order to improve their similar representation. This is accomplished by training on text-image pairs in general, and we demonstrate that it is possible to use this framework to boost learning for one modality by investigating the complementary of the other modality. On the benchmark dataset, our approach establishes a new state-of-the-art performance and shows a 3 percent improvement in F1. Furthermore, we demonstrate that it can achieve cutting-edge performance for visual EE even in a zero-shot scenario with no annotated data in the visual modality.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Image representations**.

KEYWORDS

Multimedia event extraction, Contrastive learning, Image representation learning

ACM Reference Format:
Jian Liu, Yufeng Chen, and Jinan Xu. 2022. Multimedia Event Extraction From News With a Unified Contrastive Learning Framework. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548132

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

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548132

Figure 1: An example of multimedia EE.

1 INTRODUCTION

Extracting event information (or event extraction, EE) from news has seen many applications including news timeline construction [23], text summarization [18], and others [12, 33]. Despite many advances, today's EE methods are primarily for single modality — either text or image [16], and they run the risk of obtaining incomplete information because a news article is usually presented in a multimedia format. Consider the news article shown in Figure 1. This article describes an Attack event (expressed by a textual word "fires"); however, one of the event's arguments, namely [machine gun], appears only in the image. According to previous research, more than 30% of news images contain visual event arguments that are not present in texts [16], making multimedia EE a crucial topic.

There are two major impediments to the development of multimedia EE. The first is a data issue: because annotation is expensive, there are limited resources labeled with parallel textual-visual events. Existing datasets, such as the ACE 2005 corpora [20] in the text modality and the imSitu corpora [31] in the visual modality, have completely distinct event definition/schema, making cross-modality data sharing challenging. The second issue is related to modeling — due to a lack of parallel data, previous methods for multimedia EE often employ pipeline approaches and heuristics rules to align data [28, 33] and learn modality-invariant patterns (e.g., utilizing tied classifiers [16]). However, such techniques suffer

from error propagation in the pipeline stream and lack a holistic perspective to capture the interdependencies of different modalities.

In this study, we provide a new approach to multimedia EE, demonstrating advantages for addressing the above challenges efficiently. In comparison to previous efforts, our method instead creates a shared representation space for images, texts, and event schema, through a unified contrastive learning framework. We demonstrate that, using this framework, any paired text-image resource, regardless of schema, can be used for model training, which considerably alleviates the issue of a lack of parallel labeled data. This framework, on the other hand, avoids the complexity of the pipeline approach and enables a holistic perspective to model the interdependencies of different modalities. Particularly, given a multimedia document, 1) we can project images into the joint representation space and use them as additional evidence to boost textual EE, and 2) similarly, we can project texts into the joint representation space to find complementary clues to boost visual EE. Furthermore, by assessing the similarity of sentences and images, this joint space naturally enables cross-modality event co-reference.

We evaluate our approach on the M2E2 benchmark [16]. According to the results, our method significantly outperforms previous methods (including uni-modal methods and multi-modal methods) and achieves an improvement of 2.6 percent and 3.4 percent in F1 for event extraction and event argument extraction, respectively — this clearly justifies its effectiveness. Interesting, by utilizing this contrastive learning framework, we show even in a zero-shot scenario with no training data used for training, our approach achieves competitive performance to state-of-the-art methods. Additionally, we conduct a series of qualitative and quantitative studies to investigate the benefits and drawbacks of our approach.

To summarize, we have three contributions:

- We provide a new method for multimedia EE that uses a unified contrastive learning framework to address the data and model challenges. As a seminal study examining contrastive learning for multimedia EE, our work may inspire more research in this area.
- We show that, using our unified framework, it is possible to leverage resources in different modalities for learning, regardless of their annotation schema. Furthermore, by taking a holistic approach to modeling, this unified solution avoids the complexity of pipeline approaches.
- We established new state-of-the-art performance on standard benchmarks. Furthermore, we show that our approach performs competitively to previous methods for visual EE even in a zero-shot scenario.

2 RELATED WORK

2.1 EE With Cross-Modality Learning

Event extraction (EE) from news is an important topic for many applications, despite the fact that different domains/modalities have different event definitions and datasets. In the textual domain, EE is a classic information extraction task that seeks to extract event instances in texts [1, 17], with each event represented by an event trigger and several parameters (if any). Traditional methods to textual EE are based on lexical and syntactic features [17, 30], whereas modern models have relied on neural networks [19, 21]. EE in the visual domain, in contrast to textual EE, uses a different definition of event, and the task is generally referred to as *situation recognition* [24, 31]. Particularly, given an image, the goal of visual EE is to generate a brief summary of the scenario, including the main activity (i.e., an event verb), participating objects (i.e., event arguments), and the roles these participants play in the activity.

As previously stated, existing approaches for EE focus on only one modality, and directly using them for event extraction may result in the omission of essential information. Until now, multimedia EE is still in its early stage, and there has been very limited study on this topic. One of the seminal works [33] investigates boosting textual EE by incorporating image information acquired through a weakly supervised method. In a similar line, another work [28] improves the performance by manually obtaining images relevant to a news document. A recent work [16] builds more fine-grained alignment rules and strategies to learn cross-modality patterns for multimedia EE; Another work [27], on the other hand, conducts a study on multi-modal classification of urban events from an application perspective. Nonetheless, the existing approaches for multimedia EE generally suffer from error propagation in the pipeline stream and lack a comprehensive way to integrate various modalities. In this work, we propose a new joint contrastive learning framework for aligning images and texts, which can effectively addressing the aforementioned issues.

2.2 Contrastive Learning

Contrastive learning [5] is a machine learning technique for learning general features without labels by instructing the model on which example pairs are similar or dissimilar; it has seen wide use in applications such as data augmentation [6, 11], feature clustering [2], unsupervised sentence embedding learning [10], and others [3, 7]. As for connecting texts and images, Radford et al. [25] offer a method that jointly trains a text encoder and an image encoder using contrastive learning over a pre-training task. Our approach is motivated by [25] to learn the joint representation space of images and texts, but we extend it to include label information and design a method for capturing fine-trained patterns for multimedia EE that cannot be addressed simply by aligning the two modalities. To the best of our knowledge, this is the first work to introduce contrastive learning for multimedia EE; additionally, while our approach uses image and text modalities as the case, it may also apply to scenarios with more than two modalities.

3 PROBLEM SETUP

We follow Li et al. [16] to define the multimedia EE task: Each news document is assumed to have a collection of sentences $\mathcal{S} = \{s_1, s_2, \cdots, s_{|\mathcal{S}|}\}$ and a set of images $\mathcal{M} = \{m_1, m_2, \cdots, m_{|\mathcal{M}|}\}$. Each sentence can be further expressed as $s = \{w_1, w_2, \cdots\}$, with w_j being the j^{th} word. We further suppose that each sentence is associated with a set of entities $\mathcal{T} = \{t_1, t_2, \cdots\}$, and each entity is an individually unique object that refers to a real-world object (i.e., a person, an organization, a facility, and an location). Given this, we can characterize the multimedia EE problem by establishing the following two subtasks:

Event Mention Extraction. Given a multimedia news document, the purpose of event mention extraction is to extract a set of *event*

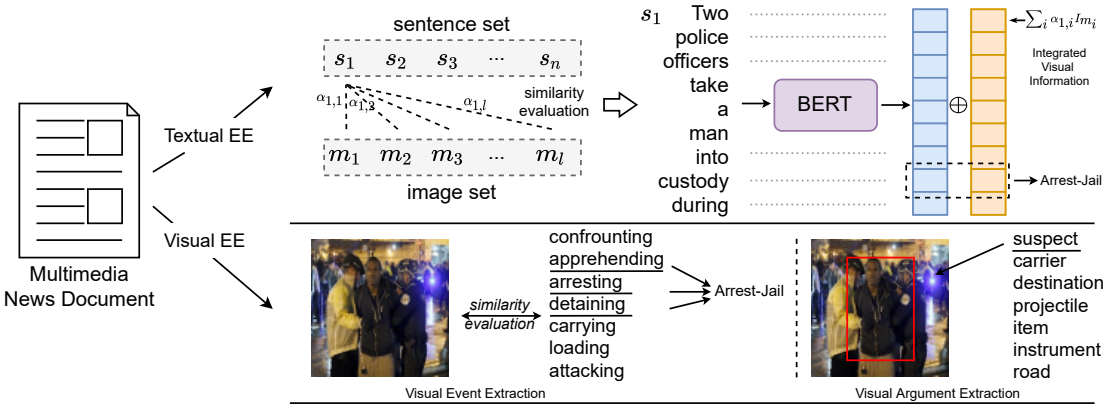

Figure 2: The overview of our approach. Using the shared representation space obtained by our contrastive learning framework (§ 4.1), we augment both textual EE and visual EE by incorporating the complimentary information in the other modality.

mentions: $\mathcal{E} = \{e_i\}_{i=1}^{|\mathcal{E}|}$, each with the following representation:

$$e = \{(w, m),\ y_e\} \tag{1}$$

where w is the word that most clearly expresses the presence of the event (i.e., an event trigger [1]), m is an image that best matches the event, and y_e denotes the event type. When both w and m exist, indicating the textual trigger and the visual image referring to the same event, the event is referred to as a *multimedia event*. Alternatively, if only w or m exists, the event is referred to as a *text-only* or an *image-only* event.

Event Argument Extraction. The purpose of event argument extraction, given an identified event mention e, is to find a set of arguments (i.e., participants) that each plays a specified role in the event. In the same way as we represent an event mention, we represent an event argument as:

$$a = \{(t, o),\ y_a\} \tag{2}$$

where t identifies a specific entity in a sentence, o a visual object represented as a bounding box in an image, and y_a the role it plays. Similarly to the case of event mention extraction, t and o can both exist or only one of them exists.

Note that there are no large-scale datasets labeled with parallel text-image events, making it difficult to train a model for multimedia EE directly. As a workaround, and similarly to Li et al. [16], we employ the following resources for training: 1) ACE 2005 [20], which contains only textual events, 2) imSitu [31], which contains only visual events (notice that this dataset does not give bounding boxes of arguments), and 3) VOA Image-Caption dataset [16], which contains parallel image-caption pairs (with no annotation of events). We adapt the M2E2 dataset [16] for evaluation, which contains 245 documents labeled with both textual and visual events. More detailed statistics of the datasets are shown in § 5.1.

4 APPROACH

Figure 2 depicts a high-level overview of our method. Particularly, we first introduce a contrastive learning framework to learn the shared representation space for images, texts, and event ontology (e.g., event types and semantic roles). Then, using this shared space, we conduct both textual and visual EE to incorporate the complimentary information in the other modality. Finally, we execute a cross-modality event co-reference process with similarity measurement to combine events from multiple modalities. Our approach's technical specifics are provided below.

4.1 Learning a Shared Representation Space via Contrastive Learning

The first stage of our approach is to learn a shared representation space in which images, texts, and event ontology all have a unified representation. To accomplish this, we use contrastive learning [25] to encourage matched image-text pairs to have higher scores than unmatched image-text pairs, in order to learn the shared space that allows for further cross-modality matching and reasoning.

Assume D is a collection of paired images and texts[1]. First, we develop two Transformer-based encoders [29], for images and texts respectively. Then, we sample N image-text pairs $\{(m_i, s_i)\}_{i=1}^{N}$ from D and encode the images as $\boldsymbol{I} = [I_1, I_2, \cdots, I_N]$ and the texts as $\boldsymbol{T} = [T_1, T_2, \cdots, T_N]$, respectively. Based on $\boldsymbol{I}$ and $\boldsymbol{T}$, we can construct an matrix $\boldsymbol{U} \in \mathcal{R}^{N \times N}$, with the element $\boldsymbol{U}_{i,j}$ denoting the matching score of I_i and T_j. In this matrix view, the diagonal elements represent matched image-text pairs, whereas the rest indicate unmatched ones, as seen in Figure 3. Our goal is to have the matched ones have a higher score than the unmatched ones, and we train the model using the following loss function:

$$\mathcal{L} = -\sum_i \log \text{softmax}(\boldsymbol{U}_{i,i}, \boldsymbol{U}_{i,*}) - \sum_j \log \text{softmax}(\boldsymbol{U}_{j,j}, \boldsymbol{U}_{*,j}) \tag{3}$$

where $\boldsymbol{U}_{i,*}$ and $\boldsymbol{U}_{*,j}$ signify the i^{th} row and j^{th} column of $\boldsymbol{U}$, respectively. Here softmax denotes a function that normalizes a scalar value x over a vector Y (notice $x \in Y$) as follows: $\text{softmax}(x, Y) = \frac{\exp(x)}{\sum_{y \in Y} \exp(y)}$. Given this, for a matched image-text pair (m, s), the first term in Eq. (3) encourages their score to be greater than the score of the image m and any other text, and the second term encourages their score to be greater than the score of the text s and

[1]Each image-text pair can be (image, caption) from the VOA Image-Caption dataset [16], or (image, activity verb/semantic role) from the imSitu dataset [31].

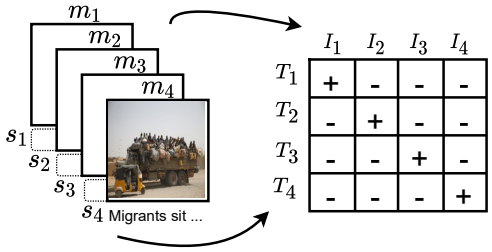

Figure 3: Illustration of positive pairs (+) and negative pairs (-) in our contrastive learning framework.

any other image, forming a contrastive learning framework. After learning, we end up with a shared representation space for images and texts (by training on the VOA Image-Caption dataset), and also event ontology (by training on the imSitu dataset).

We train the model in three steps: 1) Initialization with CLIP settings [25], which aims to directly exploit supervisions from general image-caption pairs for learning. 2) Pre-training on the VOA Image-Caption dataset [16], which aims to acquire more domain-specific knowledge about news documents. 3) Fine-tuning on the imSitu dataset [31], which aims to learn more domain-specific knowledge about news events. Given that the imSitu dataset defines both activity verbs and semantic roles, we design two strategies for fine-tuning: 1) a joint training strategy in which a single model is trained on both activity verbs and semantic roles simultaneously, and 2) a separate training strategy in which two independent models are trained on activity verbs and semantic roles separately. The ablations of the above strategies are investigated in § 6.5.

4.2 Textual EE With Visual Clues

The shared representation space allows for cross-modality information integration. We begin by introducing our method of using visual clues to improve textual EE, which consists of three steps: image-fused representation learning, textual event trigger extraction, and textual event argument extraction.

4.2.1 Image-Fused Representation Learning. Given a news document represented as a set of sentences $\mathcal{S}$ and images $\mathcal{M}$, for each sentence $s = \{w_1, w_2, \cdots\} \in \mathcal{S}$, we first construct a BERT based encoder [8] to learn the contextualized representation of each word. Assume the representation is $\boldsymbol{H}_s = \{H_1, \cdots, H_{|s|}\}$. Then, we propose to augment the representation using the image information, by computing an integrated image representation as:

$$H_{\mathcal{M}} = \sum_{j=1}^{|\mathcal{M}|} \alpha_j I_j; \qquad \alpha_j = \frac{\exp\left(\cos(I_j, T_s)\right)}{\sum_{\hat{j}} \exp\left(\cos(I_{\hat{j}}, T_s)\right)} \tag{4}$$

where I_j denotes the representation of the j^{th} image $m_j \in \mathcal{M}$ in the joint space; α_j is the weight of m_j, computed as a normalized value of the cosine similarity to T_s (T_s is the representation of s in the joint space). Finally, we concatenate $H_{\mathcal{M}}$ and each word's original representation to create an image-fused representation: $\boldsymbol{H}'_{s,\mathcal{M}} = \{H'_1, H'_2, \cdots, H'_{|s|}\}$, where $H'_i = H_i \oplus H_{\mathcal{M}} \oplus T_s$ with $\oplus$ being the concatenation operator. It is worth noting that to enrich the representation, we additionally concatenate T_s, the entire sentence's representation to the word's original representation.

4.2.2 Textual Event Trigger Extraction. Based on the image-fused representation $\boldsymbol{H}'_{s,\mathcal{M}} = \{H'_1, H'_2, \cdots, H'_{|s|}\}$, we predict an event label[2] for each word to indicate whether it is an event trigger or not:

$$O_i = \text{softmax}(\boldsymbol{W}_t H'_i + B_t) \tag{5}$$

where O_i is an output vector containing the probability of each event type for w_i (i.e., the i^{th} word in s); $\boldsymbol{W}_t$ and B_t are model parameters to be trained. The final predicted event type for w_i is the event type whose index has the highest value in O_i.

4.2.3 Textual Event Argument Extraction. For event argument extraction, we use an approach similar to event trigger extraction, but predict the semantic role (instead of event type) for each entity (instead of word) using the following calculation:

$$O_j = \text{softmax}(\boldsymbol{W}_a [H'_{\text{trigger}} \oplus H'_{t_j}] + B_a) \tag{6}$$

where H'_{trigger} indicates the representation of a predicted event trigger; H'_{t_j} is the representation of the j^{th} entity in the sentence; $\boldsymbol{W}_a$ and B_a are model parameters to be trained. When an entity has multiple words, we utilize the mean as the representation.

4.2.4 Training and Optimization. For optimization, we use cross-entropy loss and Adam [13] with default hyper-parameters. Notice that there is a gap between the training and testing stages — because the original ACE 2005 dataset lacks images, learning the image-fused representation is challenging. We address this issue with a weak supervised method: for each ACE 2005 document, we enumerate each sentence and select up to five best-matched images from the VOA Image-Caption dataset, which we then utilize in conjunction with the sentences for learning.

4.3 Visual EE With Textual Clues

Visual EE is defined differently than textual EE: visual event mention extraction seeks to predict an event type (represented by an activity verb [31]) for each image, whereas visual argument extraction seeks to locate an argument (represented by a box) in the image that corresponds to a given semantic role. Instead of using a classification based approach, we adopt a unified query-based strategy for the two subtasks.

4.3.1 Visual Event Mention Extraction. Given an input news document represented as a set of sentences $\mathcal{S}$ and images $\mathcal{M}$, for each image $m \in \mathcal{M}$, we first construct a list of activity verbs $\mathcal{V} = \{v_1, v_2, \cdots, v_{|\mathcal{V}|}\}$ defined in the imSitu dataset [31]. Then, for each verb, we compute its similarity to the image m and normalize the score as a probability value. For example, the normalized similarity score of v_i and m is calculated as follows:

$$o_{v_i} = \frac{\exp\left(\text{score}(v_i, m)\right)}{\sum_{i'=1}^{|\mathcal{V}|} \exp\left(\text{score}(v_{i'}, m)\right)} \tag{7}$$

One simple method is to define the score as the cosine similarity of the verb and image in the joint representation space, i.e., letting $\text{score}(v_i, m) = \cos(T_{v_i}, I_m)$. This strategy, however, disregards the text information and may foster false patterns such as "snowing"

[2]Following most previous studies on textual EE [1, 12, 16, 32, 33], we introduce a None class to indicate non-trigger word.

that have no event semantic. As a solution, we take into account the textual clues and devise the following score function:

$$\text{score}(v_i, m, \mathcal{S}) = \lambda \cos(T_{v_i}, I_m) + (1 - \lambda) \cos(T_{v_i}, T_{\mathcal{S}}) \quad (8)$$

where $T_{\mathcal{S}}$ is the mean of all sentence representations in the joint space; λ is a trade-off coefficient balancing image similarity and text similarity. In this way, only verbs that simultaneously matches the image and the texts would receive a high score. Finally, to get an event type for the image, we consider Top-K verbs and use a major voting approach to map them into imSitu event types (the mapping between imSitu activity verbs to M2E2 event types is based on [16]).

4.3.2 Visual Event Argument Extraction. Visual event argument extraction is more difficult than visual event mention extraction because we should: 1) determine which semantic role is realized in the image, and 2) locate the argument in the image using a box.

4.3.3 Semantic Role Identification. To determine which semantic role is realized in the image, we apply a similar method to visual event mention extraction, but use a set of pre-defined semantic roles, denoted by $\mathcal{R} = \{r_1, r_2, \cdots, r_{|\mathcal{R}|}\}$ (instead of activity verbs) to calculate similarity. Because semantic role identification occurs after event mention extraction, we should additionally filter out irrelevant roles in terms of the predicted event type. The text clues are also incorporated to produce more accurate predictions.

4.3.4 Event Argument Locating. For each role in the prediction set, we should locate a box in the image. Because the existing training datasets do not include argument box annotations, however, we are unable to directly train a model to achieve this. Inspired by a recent study on model interpretability [4], we locate an object box of a semantic role following three steps: i) Forward passing, in which we use the role and image as input to calculate their score. ii) Gradient-based attention integration, in which we conduct a simple average across attention heads to generate an gradient-augmented attention map $\bar{A}$, along the forward passing procedure:

$$\bar{A} = E_{hd}\left(\frac{\partial c}{\partial A} \odot A\right) \quad (9)$$

where *hd* enumerates each attention head in the image encoder, c is the output of the model (i.e., the similarity score), and $\odot$ denotes the Hadamard product. iii) Relevance map calculation, which aims to covert the attention map to a relevance map. Particularly, we first introduce four relevance maps: $\boldsymbol{R}^{tt}$, $\boldsymbol{R}^{ii}$, $\boldsymbol{R}^{ti}$, $\boldsymbol{R}^{it}$ defining the text-to-text, image-to-image, text-to-image, and image-to-text relevance matrices. Then 1) for self-attention, we adopt the following update rules for the relevance maps:

$$R^{tt} = R^{tt} + \bar{A} \cdot R^{tt} \quad (10)$$

$$R^{ti} = R^{ti} + \bar{A} \cdot R^{ti} \quad (11)$$

which can be seen as adding contexts to the relevance maps along with the self-attention computation procedure. 2) For bi-model attention, we adopt the following update rules, which enforces a normalization over the self-attention maps:

$$R^{ti} = R^{ti} + \tilde{R}^{tt} \cdot \bar{A} \cdot \tilde{R}^{ii} \quad (12)$$

$$R^{tt} = R^{tt} + \bar{A} \cdot R^{it} \quad (13)$$

where $\tilde{R}^{tt}$ and $\tilde{R}^{ii}$ represent the row-normalized R^{tt} and R^{ii}, respectively. The final R^{ti} is used as the relevance map to predict a box.

Table 1: Statistics of different datasets.

Dataset	# Type	# Role	# Event	# Argument
ACE 2005 [20]	33	36	5,349	61,321
imSitu [31]	504	1,788	205,095	1,481,851
VOA Caption [16]	123,078 image-caption pairs			
M2E2 [16]	8	15	1,688	2,357

Particularly, to generate a bounding box for a role, we threshold R^{ti} using the adaptive value of 0.75 * p, where p is the peak value of the map's local max. Then we compute the tightest bounding box encloses all the region as the prediction box.

4.4 Cross-Modality Event Co-reference

Finally, we should combine textual and visual events in order to conduct a cross-modality event co-reference. Using the joint representation space for cross-modality event co-reference is simple and direct: given an input news document, we first compute the cosine similarity between each sentence and image, and by linking them, we end up with a bipartite graph, with each edge connecting a sentence and an image. Then, we utilize the Hungarian algorithm [14] to find the best matches between the sentences and the images, which formulates the task as a maximum-weight matching problem. We finally combine the best-matched textual and visual events into a single event set. In practice, we find that a greedy algorithm that iteratively chooses the best matching sentence for each image functions well. The influence of various solutions for cross-modality event co-reference is investigated in § 6.4.

5 EXPERIMENTAL SETUPS

5.1 Datasets

In our experiments, we use the following resources for model training: 1) ACE 2005 [20], which annotates textual events with 33 event types and 36 semantic roles, 2) imSitu [31], which annotates visual events with 504 activity verbs and 1,788 semantic roles (note that this dataset does not provide bounding boxes of arguments), and 3) VOA Image-Caption dataset [16], which contains 123,078 parallel image caption pairs (Note that this dataset has no annotation of events). For evaluation, we use the M2E2 dataset [16], which contains 245 documents labeled with parallel textual and visual events. The M2E2 event schema is aligned with 8 ACE types and 98 imSitu types, and the event instances are divided into 1,105 text-only events, 188 image-only events, and 385 multimedia events. The detailed data statistics are summarized in Table 1. Following most studies on EE [1, 16, 22, 32, 33], we use precision (P), recall (R), and F1 score (F1) are evaluation metrics.

5.2 Implementations

In our contrastive learning framework, the text encoder is implemented as a transformer architecture [26], with 12 layers, 512 hidden units, and 8 attention heads, and the image encoder is implemented as a visual transformer (ViT) architecture [9], with a 16x16 patch size, 12 layers, 768 hidden units, and 12 heads. We set the initialized parameters according to [25]. In the following

pre-training/fine-tuning stage, the learning rate is set to 1e-7, and the batch size is set to 20. In textual EE, we use a BERT-based architecture [8] to learn the contextualized word representation. For both event trigger and event argument extraction, the maximum input length is set to 200, the learning rate is set to 1e-5, and the batch size is set to 10. For visual EE, the value of K is set to 10, implying that we consider the 10 verbs with the highest scores for predicting the event type, and the trade-off coefficient λ is set to 0.5. The threshold p for visual argument extraction is set to 0.7, which is consistent with [16], except that we compute a salience map (other than using the original attention map) to locate all of the arguments. To enable reproducibility, we have made our code public at https://github.com/jianliu-ml/Multimedia-EE.

5.3 Baselines

We compare our approach to the following baseline models: 1) JMEE [22], a state-of-the-art method that combines syntactic information with graph representation learning for textual EE. 2) GAIL [32], a method that uses a reinforcement learning strategy to jointly identify entities and events for textual EE. 3) VAD [33], which uses a weakly-supervised technique to collect images and jointly trains a model with textual features and visual patterns for textual EE. 4) WASE, a model based on the basic architectures given in [16]. Note that this model only examines uni-modal features for learning. 5) Flat [16], a model concatenating textual and visual features for learning. This model is used for both textual EE and visual EE. 6) WASEatt and WASEobj [16], the state-of-the-art approaches for multimedia EE; they employ either an attention map or an object detection model to locate the argument box. 7) CLIPEvent [15], which introduces event structure to align multimedia events. We use T (text-only), V (image-only), and T+V (text plus image) to specify which information is utilized to train the above model. Finally, we refer to our approach as UniCL (Multimedia Event Extraction with a Unified Contrastive Learning Framework).

6 EXPERIMENT RESULTS

6.1 Quantitative Performance

Table 2 investigates the quantitative performance of various approaches. According to the results, our approach achieves the best performance in various settings and outperforms previous methods by a wide margin, for example, +3.1% and +4.3% for textual EE on event mention and argument extraction, +6.7% and +4.3% for visual EE, +2.6% and +4.2% for multimedia EE. The significant improvement clearly demonstrates the efficacy of our approach. The results also suggest that our approach can exploit complimentary information from the other modality for learning — for textual EE, adding complementary information from the visual modality can increase performance by +1.1% for trigger extraction and +4.3% for argument extraction. The results also suggest that our approach can leverage the complementary information in the other modality for learning — for textual EE, by adding the complementary information from the visual modality, the performance can be boosted by +1.1% for trigger extraction and +4.3% for argument extraction. The same is true for visual EE; by incorporating complementing information from the textual modality, the performance can be increased by +1.3% for visual event extraction and +0.7% for visual event argument extraction.

Table 2: Results on textual EE (top), visual EE (middle), and multimedia EE (bottom). T, V, and T+V indicate using text-only, image-only, and multi-modal information.

		Event Ext.			Argument Ext.		
Eval.	**Method**	**P**	**R**	**F1**	**P**	**R**	**F1**
Textual EE	JMEE (T)	42.5	58.2	48.7	22.9	28.3	25.3
	GAIL (T)	43.4	53.5	47.9	23.6	29.2	26.1
	WASE (T)	42.3	58.4	48.2	21.4	30.1	24.9
	Flat (T+V)	34.2	63.2	44.4	20.1	27.1	23.1
	VAD (T+V)	34.8	64.4	45.2	23.1	27.5	25.1
	WASEatt (T+V)	37.6	66.8	48.1	27.5	33.2	30.1
	WASEobj (T+V)	42.8	61.9	50.6	23.5	30.3	26.4
	UniCL (T)	42.8	**68.3**	52.6	26.5	32.9	29.4
	UniCL (T+V)	**49.1**	59.2	**53.7**	**27.8**	**34.3**	**30.7**
Visual EE	WASE (V)	29.7	61.9	40.1	9.1	10.2	9.6
	Flat (T+V)	27.1	57.3	36.7	4.3	8.9	5.8
	WASEatt (T+V)	32.3	63.4	42.8	9.7	11.1	10.3
	WASEobj (T+V)	43.1	59.2	49.9	14.5	10.1	11.9
	UniCL (V)	50.8	**63.2**	56.3	14.6	**14.5**	14.5
	UniCL (T+V)	**54.6**	60.9	**57.6**	**16.9**	13.8	**15.2**
Mul. EE	Flat (T+V)	33.9	59.8	42.2	12.9	17.6	14.9
	CLIPEvent (T+V)	41.3	**72.8**	52.7	21.1	13.1	17.1
	WASEatt (T+V)	38.2	67.1	49.1	18.6	21.6	19.9
	WASEobj (T+V)	43.0	62.1	50.8	19.5	18.9	19.2
	UniCL (T+V)	**44.1**	67.7	**53.4**	**24.3**	**22.6**	**23.4**

Table 3: Qualitative analysis on textual EE. Arg. (GT) indicates the use of golden event triggers to omit their impact.

	Model					
Eval.	**CNN**	**CNN (V)**	**RNN**	**RNN (V)**	**JMEE**	**JMEE (V)**
Event	47.9	49.5 (+1.6)	48.1	49.6 (+1.5)	48.7	50.3 (+1.6)
Arg.	25.1	27.4 (+2.3)	25.0	28.1 (+2.9)	25.3	27.9 (+2.4)
Arg. (GT)	38.6	39.9 (+1.3)	40.4	42.3 (+1.9)	40.9	42.7 (+1.8)

After further examination, we discover that the main improvement is from the precision score. One explanation is that the data distribution in the training set ACE 2005 differs from that in the evaluation set M2E2, therefore directly adopting a model for assessment outputs more false patterns. For example, for ACE 2005, 83% of the "shootings" are annotated with an Attack event, whereas only 30% in M2E2 (such as "Not all police *shootings* fit the pattern"). By including the image's complementary information, the model becomes more cautious in its prediction, lowering the possibility of incorrect predictions and increasing precision score.

6.2 Qualitative Analysis on Textual EE

To further understand the efficacy of our approach, we conduct a qualitative analysis on textual EE, exploring the usage of several architectures for integrating visual information, such as CNN, RNN,

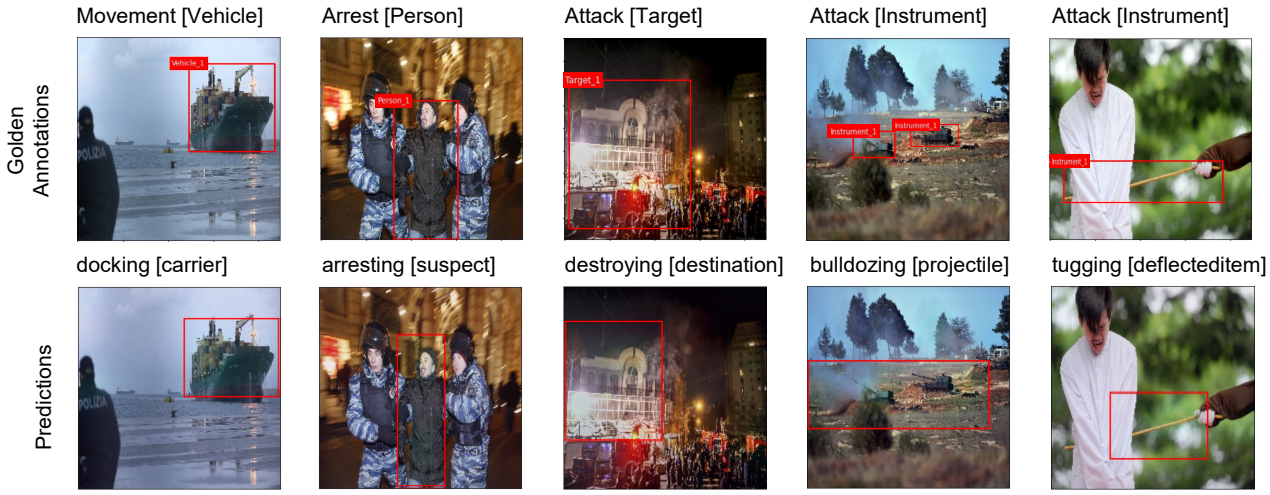

Figure 4: Typical cases on visual argument grounding, with predictions at the bottom.

Table 4: Qualitative analysis on visual EE.

	Arg. Detection			Arg. Grounding		
Method	**P**	**R**	**F1**	**P**	**R**	**F1**
WASE (V)	32.3	35.7	33.9	9.1	10.2	9.6
Flat (T+V)	34.6	37.7	36.1	4.3	8.9	5.8
WASEatt (T+V)	37.6	38.6	38.1	9.7	11.1	10.3
WASEobj (T+V)	36.3	38.4	37.3	14.5	10.1	11.9
UniCL (V)	39.2	42.0	40.5	14.6	**14.5**	14.5
UniCL (T+V)	**40.6**	**42.9**	**41.7**	**16.9**	13.8	**15.2**

Table 5: Impact on cross-modality event co-reference.

	Event Ext.			Argument Ext.		
Method	**P**	**R**	**F1**	**P**	**R**	**F1**
Bipartite Matching	44.1	**67.7**	**53.4**	**24.3**	**22.6**	**23.4**
Greedy Matching	**44.4**	65.1	52.8	**24.3**	21.5	22.8
Sequential Matching	**44.4**	49.1	46.1	16.2	18.1	17.1

and JMEE. Table 3 displays the results, where we also investigate the scenario of employing golden triggers for argument extraction to eliminate their impact on prediction (denoted as Arg. (GT)). According to the results, our approach is universal for each architecture, resulting in an F1 improvement ranging from 1.3% to 2.9%. This highlights the efficacy of our contrastive learning framework for multimodal information integration in multimedia EE.

6.3 Qualitative Analysis on Visual EE

In Table 4, we present a qualitative analysis of visual EE, comparing the performance of different models on argument detection, where we only identify which role is realized, and argument grounding, where we additionally should identify the bounding box. According to the results, the most difficult step is argument grounding, and our model can reach above 40% for argument detection, which merely determines which semantic role is realized in the image. Figure 4 gives several cases of argument grounding. The first two cases, in particular, are near-perfect examples. The error cases include: 1) Partial match, like in Case 3, where only a portion of the object is recognized. 2) Multiple objects, such as in Case 4, where two tanks are labeled in the original annotation, but our method identifies a union box containing both. 3) Shading, as in Case 5, in which our approach only annotates the visible portion of the object. One key cause is that there is no annotation of the grounding box in the visual EE training set. Adding box annotations for training is a promising direction for improvement, which, however, requires a significant amount of human effort.

6.4 Results of Cross-Modality Co-reference

We investigate the impact of cross-modality event co-reference in Table 5, where we compare our bipartite matching strategy with a greedy matching strategy, which iteratively chooses the best matching sentence for each image, and a sequential matching strategy, which selects the first sentence containing the same event type as the co-reference result for each image (without similarity measurement). The results reveal that our bipartite matching strategy is the most effective method since it targets global optimization. The greedy matching technique also performs well since it finds the best matched sentence for each image, and the sentences rarely overlap. The sequential matching approach performs poorly as it ignores similarity measurement in the matching process.

6.5 Impact on Contrastive Learning

Because learning the joint representation space is crucial in our approach, we investigate the influence of various contrastive learning strategies. The results are shown in Figure 6, where CLIP indicates only using CLIP parameters for initialization, CLIP+VOA indicates pre-training on VOA Image-Caption dataset as well, CLIP+VOA+S indicates using a shared strategy to train the model on activity verbs and semantic roles of the imSitue dataset, and CLIP+VOA+D indicates training separate models on activity verbs and semantic

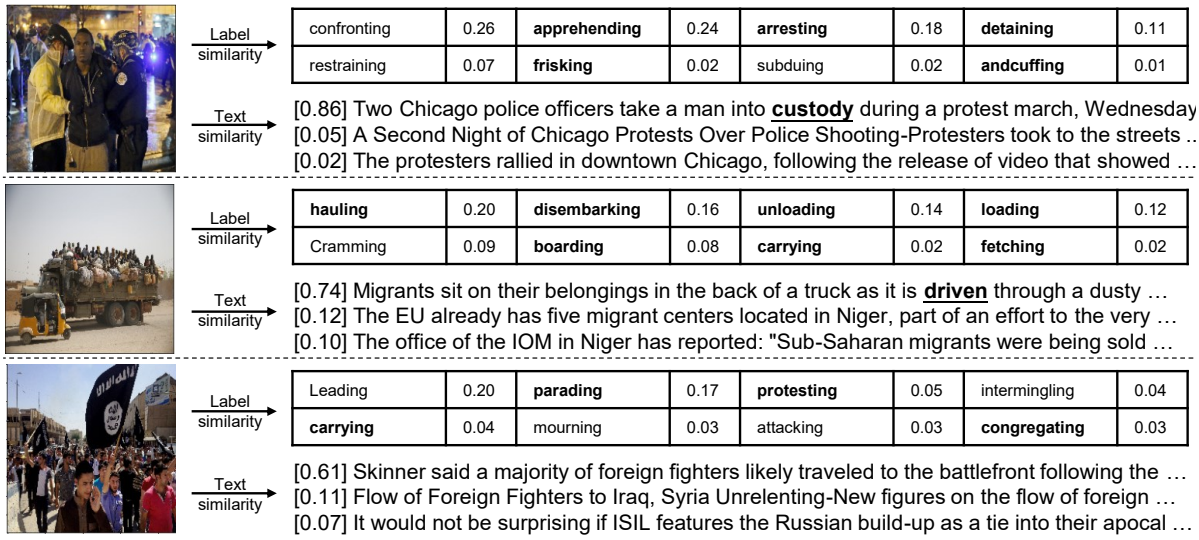

Figure 5: A finer-grained case study on the joint representation space, exploring label and text similarities for each image.

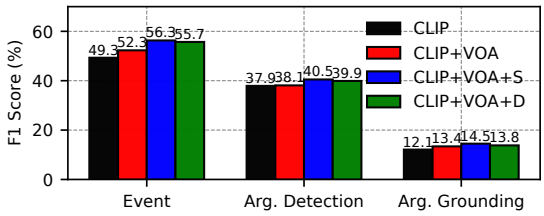

Figure 6: Impact of different contrastive learning strategies.

roles. According to the results, adding the VOA dataset and the imSitu dataset greatly improves learning, and CLIP+VOA+S gets the best performance across all settings, implying that training a single model is the optimal solution.

One interesting finding is that our approach can get good results even without any training on visual event resources. To demonstrate this, in Table 6, we show the performance of our approach in a zero-shot case with merely initialized CLIP parameters, where the results suggest that our approach achieves competitive performance to state-of-the-art methods even without any training.

6.6 Case Study

Finally, Figure 5 depicts a finer-grained case study on the joint representation space, in which we show label and text similarity for each image. For each image, our method can find appropriate activity verbs, such as "apprehending" and "arresting" for case 1, "disembarking" and "loading" for case 2, and "parading" and "protesting" for case 3, which clearly indicate the event type. Such information provides strong clues for textual EE; for example, in case 1, the word "custody" does not appear in the ACE 2005 training set, which normally fails a regular text event detector, but our approach successfully identifies it as a trigger of Arrest-Jail by integrating image information. We also investigate text similarity to see if the images and texts match for cross-modality co-reference.

Table 6: Results of our approach for zero-shot adaption.

	Event Ext.			Argument Ext.		
Method	**P**	**R**	**F1**	**P**	**R**	**F1**
WASE (V)	29.7	61.9	40.1	9.1	10.2	9.6
Flat (T+V)	27.1	57.3	36.7	4.3	8.9	5.8
WASEatt (T+V)	32.3	63.4	42.8	9.7	11.1	10.3
WASEobj (T+V)	43.1	59.2	49.9	14.5	10.1	11.9
UniCL Zero (V+T)	44.1	55.9	49.3	14.9	10.2	12.1
UniCL Full (V+T)	54.6	60.9	57.6	16.9	13.8	15.2

In cases 1 and 2, for example, the retrieved sentences perfectly match the image. This demonstrates the efficacy of our approach to learning the joint representation space.

7 CONCLUSION

In this paper, we propose a new method for multimedia EE by proposing a joint contrastive learning framework to bridge the textual and visual modalities. This is accomplished by training on text-image pairs in general, and we demonstrate that it is possible to use this framework to boost learning for each modality by investigating the complementary of the other modality. Our approach achieves state-of-the-art performance and outperforms previous methods by a wide margin. In the future, we will strive for more precise argument box identification, as well as considering more modalities (e.g., video) for multimedia EE.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No.62106016). This work is also supported by Fundamental Research Funds for the Central Universities (No. 2021RC234) and the Open Projects Program of National Laboratory of Pattern Recognition.

REFERENCES

[1] David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Association for Computational Linguistics, Sydney, Australia, 1–8. https://aclanthology.org/W06-0901

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep Clustering for Unsupervised Learning of Visual Features. *ArXiv preprint* abs/1807.05520 (2018). https://arxiv.org/abs/1807.05520

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html

[4] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 397–406.

[5] S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 539–546 vol. 1. https://doi.org/10.1109/CVPR.2005.202

[6] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. 2018. AutoAugment: Learning Augmentation Policies from Data. *ArXiv preprint* abs/1805.09501. https://arxiv.org/abs/1805.09501

[7] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. 2020. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[11] Daniel Ho, Eric Liang, Xi Chen, Ion Stoica, and Pieter Abbeel. 2019. Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 2731–2741. http://proceedings.mlr.press/v97/ho19b.html

[12] Heng Ji and Ralph Grishman. 2011. Knowledge Base Population: Successful Approaches and Challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 1148–1158. https://aclanthology.org/P11-1115

[13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[14] Harold W. Kuhn. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly* 2, 1–2 (March 1955), 83–97. https://doi.org/10.1002/nav.3800020109

[15] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and S. Chang. 2022. CLIP-Event: Connecting Text and Images with Event Structures.

[16] 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*. Association for Computational Linguistics, Online, 2557–2568. https://doi.org/10.18653/v1/2020.acl-main.230

[17] Qi Li, Heng Ji, and Liang Huang. 2013. Joint Event Extraction via Structured Prediction with Global Features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 73–82. https://aclanthology.org/P13-1008

[18] Quanzhi Li and Qiong Zhang. 2020. *Abstractive Event Summarization on Twitter*. Association for Computing Machinery, New York, NY, USA, 22–23. https://doi.org/10.1145/3366424.3382678

[19] Sha Li, Heng Ji, and Jiawei Han. 2021. Document-Level Event Argument Extraction by Conditional Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 894–908. https://doi.org/10.18653/v1/2021.naacl-main.69

[20] Linguistic Data Consortium (Ed.). 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events* (5.4.3 2005.07.01 ed.).

[21] Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event Extraction as Machine Reading Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1641–1651. https://doi.org/10.18653/v1/2020.emnlp-main.128

[22] Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1247–1256. https://doi.org/10.18653/v1/D18-1156

[23] Jakub Piskorski, Vanni Zavarella, Martin Atkinson, and Marco Verile. 2020. Timelines: Entity-centric Event Extraction from Online News. In *Text2Story@ECIR*.

[24] Sarah M. Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded Situation Recognition. *ArXiv preprint* abs/2003.12058 (2020). https://arxiv.org/abs/2003.12058

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research)*, Marina Meila and Tong Zhang (Eds.), Vol. 139. PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners. (2018). https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[27] Maarten Sukel, Stevan Rudinac, and Marcel Worring. 2019. Multimodal Classification of Urban Micro-Events. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. 1455–1463. https://doi.org/10.1145/3343031.3350967

[28] Meihan Tong, Shuai Wang, Yixin Cao, Bin Xu, Juanzi Li, Lei Hou, and Tat-Seng Chua. 2020. Image Enhanced Event Detection in News Articles. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 9040–9047. https://doi.org/10.1609/aaai.v34i05.6437

[29] 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*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[30] Bishan Yang and Tom M. Mitchell. 2016. Joint Extraction of Events and Entities within a Document Context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 289–299. https://doi.org/10.18653/v1/N16-1033

[31] Mark Yatskar, Luke S. Zettlemoyer, and Ali Farhadi. 2016. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 5534–5542. https://doi.org/10.1109/CVPR.2016.597

[32] Tongtao Zhang and Heng Ji. 2018. Event Extraction with Generative Adversarial Imitation Learning. *ArXiv preprint* abs/1804.07881 (2018). https://arxiv.org/abs/1804.07881

[33] Tongtao Zhang, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph G. Ellis, Lifu Huang, Wei Liu, Heng Ji, and Shih-Fu Chang. 2017. Improving Event Extraction via Multimodal Integration. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. 270–278. https://doi.org/10.1145/3123266.3123294