Improving Event Extraction via Multimodal Integration

Tongtao Zhang, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph Ellis, Lifu Huang, Wei Liu, Heng Ji, Shih-Fu Chang

1 Rensselaer Polytechnic Institute, USA 2 Columbia University, USA 3 Microsoft Research, USA 4 Tencent AI Lab, China

{zhangt13,whites5,huangl7,jih}@rpi.edu, {hz2471,hl2642,jge2105}@columbia.edu, {wliu,sfchang}@ee.columbia.edu

ABSTRACT

In this paper, we focus on improving Event Extraction (EE) by incorporating visual knowledge with words and phrases from text documents. We first discover visual patterns from large-scale text-image pairs in a weakly-supervised manner and then propose a multimodal event extraction algorithm where the event extractor is jointly trained with textual features and visual patterns. Extensive experimental results on benchmark data sets demonstrate that the proposed multimodal EE method can achieve significantly better performance on event extraction: absolute 7.1% F-score gain on event trigger labeling and 8.5% F-score gain on event argument labeling.

CCS CONCEPTS

• Computing methodologies → Information extraction; Image representations;

KEYWORDS

Event Extraction; Visual Pattern Discovery; Natural Language Processing; Multimodal Approach

1 INTRODUCTION

We are immersed in an ever-growing ocean of noisy, unstructured data of various modalities, such as text and images. For example, over 5 million articles on Wikipedia and 700 new articles are created per day1. In order to acquire a deeper understanding of the content, Event Extraction (EE) techniques have been developed to automatically extract information units, such as "what is happening" and "who, or what, is involved" [19, 24, 26], in a precise, clear, and structured form. EE can facilitate various downstream Web-scale applications such as automatic chronicle generation [9] and Wikipedia article generation [34].

To perform EE, programs and schema are created to define the EE task. By the definitions and terms in the Automatic Content Extraction (ACE) program2 [30], the aim of EE is to extract the following elements from a large corpus of text documents such as news articles, web blogs, discussion forum posts, and tweets (Figure 1a):  

- **Event**: An event denotes the dynamic interaction among arguments. An event includes a trigger and several arguments. /n

- **Trigger**: A word or phrase that clearly indicates the occurrence of an event. For example, in Figure 1a we have a Meet event triggered by "confront".

- **Argument**: An object involved in an event is an argument. Arguments can be people, organizations, weapons, vehicles, facilities, and locations. Each argument is assigned a role, which reveals the relation between the argument and event. For example, in Figure 1b, the Meet event has three arguments: "Mayor Ford", "enemies", and "council chamber".

A detailed list of 33 event types is presented in supplementary materials.

2https://www.ldc.upenn.edu/collaborations/past-projects/ace

Figure 1: Motivating example of our proposed multimodal approach, where external visual knowledge improves EE from text documents. (a) Ground truth of the EE output from the input sentence. (b) Incorrect extraction output using text-only features. (c) Corrected result using multimodal features.
Figure 2: The pipeline of the proposed multimodal approach: Given the text document as input, we retrieve visual patterns from a background visual repository constructed via Visual Argument Discovery (Section 4.1) using external image-caption pairs. We extract visual features from the retrieved patterns and integrate them with text features (Section 4.2). We train a structured perceptron classifier using the integrated features. The final output consists of “event-with-argument” structures.

Ford” and “members” as Entity arguments and “Toronto’s council chamber” as a Place argument.

A number of EE models have brought forth encouraging results by retrieving additional related text documents [19, 22, 37], extracting salient textual features [15, 24, 26], and adopting more advanced learning frameworks [5, 7, 16, 29]. However, it is well-known that the above text-based EE models are generally limited due to the ambiguity of natural language. For example, multiple meanings of the same word (i.e., polysemy) causes errors: the word “confront” is a trigger word of a Meet event as shown in Figure 1, however, in the sentence “Police confronted protesters hurling stones”, the word “confront” is a trigger word for an Attack event. Most methods [24, 29] predict the labels based on the probabilistic distribution of confront being an Attack trigger in the training set and external gazetteers or dictionaries. As a result, confront is mistakenly labeled as an Attack trigger and, consequently, "Mayor Ford" and “members” are treated as the Attacker and Target, respectively. Correcting such errors requires clues beyond the text domain.

Events do not solely exist in the single modality of text – similar event types, participants, or contexts may co-exist in rich multimedia content (e.g., news articles usually come with textual documents and images/videos referring to the same or similar events). As a human reader who attempts to tackle ambiguities, such as the example shown in Figure 1b, one may draw on visual clues to provide clarity. For example, we can use “mayor”, “member”, and/or “council chamber” as keywords in a large multimedia repository and retrieve images, as in Figure 1c, from which we can observe visual concepts involving “well-dressed people” or “tables and chairs in a parliament setting” indicating non-violent events (e.g., Meet) instead of violent ones (e.g., Attack). Analogously, we can conceive of an automatic EE approach which leverages visual information.

We propose a multimodal approach which integrates explicit visual information to improve EE performance on text documents. As shown in Figure 2, visual information serves as auxiliary external knowledge to resolve ambiguities of the text-only modality and enhance EE performance. In order to acquire such knowledge, we use external multimedia resources, such as images and captions crawled from a large source of news articles, to construct an adaptive and scalable background repository of visual patterns which depict arguments like “police”, “protesters” or “soldiers” in an Attack event) from reference or training documents, and assigns the proper label to the event (e.g., assign Meet to the event in as in Figure 1c).

We use the multimodal model to improve EE results on test sentences, compared to a text-only approach. We conduct experiments on two standard benchmark data sets – ACE2005 [39] and ERE (Entities, Relations and Events) [37] – and empirically validate that our proposed multimodal approach successfully improves EE performance, with up to an absolute 7.1% gain in F1 score on trigger labeling and 8.5% gain on argument labeling.

Our contributions are as follows:

(1) To the best of our knowledge, we are the first to propose a multimodal framework for EE from text documents by utilizing visual background knowledge.

(2) We adopt a visual pattern discovery approach to generate a background visual repository of entities for each specific event, which provides additional background knowledge augmenting textual arguments with visual representations. We also improve
the visual pattern discovery approach by introducing more information from textual captions.

3) Our work proposes a framework that tightly integrates multimodal evidences, features and information, instead of simple post-processed or re-ranked results from separate detection systems solely relying on data of individual modalities.

2 RELATED WORK

2.1 Event Extraction

Besides the text-only EE methods mentioned in Section 1, which detect and extract structured events from text documents, there are some visual approaches, such as [13, 27, 43], which manage to detect and generate similar event-argument structures from visual data and represent them as a tuple of subject, predicate and object. The contents of the extracted tuples are equivalent to the event structure in our work, with the predicate in each tuple as the trigger and the subject/object as the arguments. The frameworks in [2, 42] detect and extract event triggers from a series of images accompanied with textual descriptions. [18] present approaches to combine object and action detection results to form descriptive phrases and assign them to segmented images. While these methods leverage textual information for a better understanding of images, our method goes in the opposite direction: we utilize vivid and explicit visual information to improve Event Extraction on purely textual documents.

2.2 Visual Pattern Mining/Discovery

Most visual pattern mining approaches, such as [25], focus on data sets of a single modality, such as images. The approach in [18] can be viewed as multimodal pattern discovery since it assigns textual information to segmented patches. However, it still requires fine-grained natural language labels as prior knowledge. In our work, we construct our background visual repository with an unsupervised multimodal visual pattern discovery framework from [21], which is more scalable and generalizable.

2.3 Multimodal Approaches

There are also many NLP tasks where information, resources, and features of multiple modalities are utilized. [12, 31, 38] propose approaches of image/video captioning. Visual question answering is tackled in [8, 32]. [20] extends Word2Vec to the visual domain. [14] presents caption translation with multimodal information. [35] takes visual features to identify metaphors. [4, 41] introduce summarization with visual information. Most of these approaches require parallel and well-aligned multimodal data to ensure one-to-one mapping on each data instance. Our work is the first to demonstrate a new approach that transfers visual knowledge from rich external multimodal resources to documents lacking visual information.

3 EVENT EXTRACTION VIA MULTIMODAL INTEGRATION

3.1 Baseline Text-only Approach

In this paper, we use JointIE [24] as our baseline approach to EE and we briefly introduce the approach in this subsection. As shown in Figure 3, given a sentence \( S \) (e.g., “Police officers confronted protesters hurling stones”), we construct several hypothesis graphs \( Y \) via Beam Search as in [24]. For example, in Figure 3a, “police officers”, “confronted” and “protesters” are nodes in the graph, and the edges connecting them demonstrate the argument roles.

We have an assignment score function, \( F(\cdot, \cdot) \), given by

\[
F(S, Y_i) = \sum_j w_j f(x_j, y_{ij}),
\]

where \( i \) denotes the index of hypothesis graph, \( j \) denotes the index of a feature, and \( w_j \) denotes a weight of feature \( j \).

\[
f(x_j, y) = \log p(y|x),
\]

which can be estimated from training data.

The tuple of \((x, y)\) is a nominal feature, where \( x \) denotes an attribute of a node in the graph (e.g., uni-gram of “hurling”, bi-gram of “police officers”, or “confronted” as past form) and \( y \) represents a substructure of a hypothesis graph (e.g., “police officers” being an argument, “police officers” involved in an Attack event triggered by “confronted”, or “protesters” being an Attacker argument in an Attack event). The features used by JointIE include: local trigger/argument features, which mainly focus on the triggers/arguments themselves and interactions (e.g., dependency parsing [3] results) among other arguments or within the same event; and global trigger/argument features, which focus on the interactions (e.g., co-existence) among triggers/arguments across different events in the same sentences.

During training, a structured perceptron model estimates the weight coefficients \( w_j \) based on features extracted from the ground-truth graph as well as other generated hypothesis graphs and ensures that the ground-truth graph’s assignment score is the highest ranked.

In the testing phase, given a sentence, JointIE also heuristically generates multiple hypothesis graphs with Beam Search, and pursues the highest assignment score among these graphs, which is given by:

\[
\hat{y} = \arg \max_{y_i} \sum_j w_j \log p(y_{ij}|x_j)
\]

and decodes them as the EE results.

In all, the JointIE approach simultaneously captures all EE results – including event triggers, event types, arguments, and argument roles – from target documents. It aims to determine the most feasible structure from multiple hypothesis graphs, where all nodes and edges can contribute their own weights or counterbalance the impact of other units.

---

3 Terminology varies between domains. For work in the vision domain, such as [27, 43], the word “tuple” refers relations between objects. While in EE work (including ours), “tuple” means event.

4 A detailed list of features is presented in the supplementary documents.
Hypothesis with Wrong Argument Detection

We alter Equation 1 as follows:

$$\mathcal{F}(S, Y_i) = \sum_j w_j f(x_j, y_{ij}) + \sum_k w_k f(z_k, y_{ik}).$$  \hspace{1cm} (4)

where \((z_k, y_k)\) denotes a visual feature, and \(z_k\) is a visual-based attribute. The objective function after integrating visual features becomes:

$$\hat{Y} = \arg \max_{Y_i} (\mathcal{F}(S, Y_i))$$

$$= \arg \max_i \left( \sum_j w_j \log p(y_{ij} | x_j) + \sum_k w_k \log p(y_{ik} | z_k) \right)$$  \hspace{1cm} (5)

The JointIE approach considers a comprehensive and global view of the context of text documents. Visual features further expand its scope with deeper real world knowledge. For example, we can expect that the visual features from “mayor”, “congress”, “chamber council”, and “council members” often appear with event types such as Meet, Start-Position, and End-Position, while they are less likely to appear in, or co-exist with, events such as Attack, Die, or Injury. Therefore, the trigger word “confront” is considered as a Meet event instead of an Attack event. Such background information is often uniquely and directly inferred from visual features, and goes far beyond the reach of simple textual dictionaries or gazetteers, which merely reveal superficial knowledge of local structures in the graphs.

3.2 Multimodal Integration

In this work, we improve the framework described in Section 3.1 by integrating visual features, which will be elaborated in Section 4.2. We alter Equation 1 as follows:

$$\mathcal{F}(S, Y_i) = \sum_j w_j f(x_j, y_{ij}) + \sum_k w_k f(z_k, y_{ik}).$$  \hspace{1cm} (4)

where \((z_k, y_k)\) denotes a visual feature, and \(z_k\) is a visual-based attribute. The objective function after integrating visual features becomes:

$$\hat{Y} = \arg \max_{Y_i} (\mathcal{F}(S, Y_i))$$

$$= \arg \max_i \left( \sum_j w_j \log p(y_{ij} | x_j) + \sum_k w_k \log p(y_{ik} | z_k) \right)$$  \hspace{1cm} (5)

The JointIE approach considers a comprehensive and global view of the context of text documents. Visual features further expand its scope with deeper real world knowledge. For example, we can expect that the visual features from “mayor”, “congress”, “chamber council”, and “council members” often appear with event types such as Meet, Start-Position, and End-Position, while they are less likely to appear in, or co-exist with, events such as Attack, Die, or Injury. Therefore, the trigger word “confront” is considered as a Meet event instead of an Attack event. Such background information is often uniquely and directly inferred from visual features, and goes far beyond the reach of simple textual dictionaries or gazetteers, which merely reveal superficial knowledge of local structures in the graphs.

4 IN-DOMAIN VISUAL FEATURE DISCOVERY

In our work, we require a background visual repository as a source of visual features. It should provide auxiliary background knowledge of explicit and vivid visual patterns, to fill the gap between visual materials and text documents, and facilitate our multimodal approach to EE.

An ideal background visual repository should contain a large amount of visual clusters, each of which may consist of multiple visual patterns depicting a world object – or an argument, which names the cluster.

Image data sets designed for image classification and object detection tasks, such as ImageNet [6], can be a candidate resource for a visual repository. Methods such as Region proposals with CNN features (R-CNN) [10, 11, 33] are also capable of generating a visual repository by providing bounding boxes labeled with object names on input images. However, the pattern names generated from these data sets and methods cover a limited, fixed, and closed subset of real world objects. For example, the ImageNet data set includes the annotations of bounding boxes for 3,627 labels\(^7\). Although it has attempted to cover as many real world objects as possible, and the community continues expanding the annotation, there are still lots of concepts that remain missing. Common concepts such as “police”, “politician”, and “businessman” are not included. Expensive annotation costs also obstruct deployment to open domains.

4.1 Visual Argument Discovery

To tackle the aforementioned problems, we adopt Visual Argument Discovery (VAD) to cluster, mine, and name visual patterns automatically. Based upon Visual Pattern Discovery (VPD) [21], this unsupervised framework consumes a large external corpus of images which are well aligned with their descriptive captions and generates a rich repository of visual patterns, which are assigned to various clusters according to the objects they share.

This pattern discovery framework is adaptive and scalable. It can accept text documents and accompanying images from open domains and the output will not be confined to fixed topics (animal, vehicle etc.). Rather, it will generate the list of pattern clusters covering all topics mentioned in the input documents. Figure 4 demonstrates some sample results from 285,900 image-caption pairs mentioned in [21].

However, there are two major issues for the original visual pattern discovery approach in [21].

First, the arbitrary determination of the maximum length of n-grams (unigram and bi-gram) makes the approach unable to handle longer phrases, such as “law enforcement officials”.

\(^7\)The full data set contains images with 21,841 labels, while only 3,627 of them are clearly annotated with bounding boxes. In this paper, we use the subset with the 3,627 label annotations.
we can prevent trigger words from being processed in the cluster-formation. As a consequence, VPD generates some visual pattern clusters with names of potential trigger words of events such as “burn” and “air strike”, which are confused with patterns of “fire” and “smoke”, respectively. Moreover, the clustering algorithm does not tackle polysemy or word sense disambiguation. The output of [21] provides biased discovery results (e.g., “fire” can be the object “flame”, a trigger word of Attack, or a trigger word of End-Position, whereas the Figure 4e merely demonstrates the first meaning). Per our empirical observation, if we impose a constraint where we filter out the cluster names which can only serve as triggers, such as “burn” and “air strike”, and concentrate on the arguments, there could be less confusion and fewer biases; although we may encounter other dubieties such as “apple” as a fruit or “Apple” as a company.

To address the issues above, we require a dynamic approach to generate candidate text strings of variable length instead of exhaustively searching through all fixed-length n-grams. We also need to further disambiguate the candidates even if they are limited to arguments. Accordingly, we use the Abstract Meaning Representation (AMR) [1] parser to parse the captions [40]. AMR provides a graph of a clear semantic representation of a sentence as shown Figure 5. This semantic graph provides stemmed information that the verbs “fire” and “arrest” in the first sentence and the noun “arrest” in the second sentence indicate actions and we can consider them as potential event triggers. Using these disambiguated results, we can prevent trigger words from being processed in the clustering algorithm. Moreover, from the AMR structure, we are able to propose text segments of variable length, e.g., “police officer” and “law enforcement officials”, as candidate arguments.

Additionally, as stated in [17], arguments can be disambiguated when using additional representations from their context, especially the actions in which they are involved. We introduce additional dimensions of representations (i.e., entity typing as proposed in [17]). As shown in Figure 5, we append the encoded embedding of “arrest” and the ARG0 (subject) relation with the original Word2Vec embedding of “law enforcement officials”. Similarly, the representation of “police officers” includes entity typing embeddings generated from “fire”, “arrest” and their ARG0 relations. We use these candidates and their word/phrase embeddings in place of their counterparts in the original VPD in [21]. After we adopt these procedures, in Figure 4, the cluster “burn” and “air strike” are removed, while the cluster “fire” is still retained because all instances of “fire” eligible to be candidate arguments focus on the “flame” concept.

Finally, with the improved argument embeddings and the visual response of the images, we do clustering and mine the named clusters using association mining rules, and we achieve an argument-centric visual repository.

### 4.2 Visual Feature Extraction

After we construct a visual repository, we can provide each argument in the hypothesis graphs of the multimodal JointIE with visual features if the argument string matches the pattern name in the visual repository.

Given a visual repository, \( V \), and a query (or hypothesis argument), \( Q \), we retrieve a set of visual patterns, \( I = \{I_1, I_2, \ldots, I_n\} \),
where \( n \) is the number of visual patterns in a specific cluster, by finding the cluster whose name exactly matches \( Q \) and collecting all the visual patterns that belong to the matched cluster. Next, for each visual pattern, \( I_j \), we extract a visual feature vector, \( z_j, j = 1, \ldots, n \). We do so by providing \( I_j \) as input to a pre-trained VGGNet [36] and using the response of the penultimate FullyConnected layer, known as the fc7 layer, as \( z_j \). The response of the fc7 layer provides a representation of the input visual pattern that can be used to distinguish between similar and dissimilar visual patterns [28]. The result of this process is a set of visual features vectors \( Z = \{z_1, \ldots, z_n\} \), each of which corresponds to a single visual pattern in \( I \).

It is important to note that, for different exactly matched queries \( Q \), the numbers of visual patterns often vary. Moreover, we are not able to provide any visual information for a target sentence whose entities do not match any visual cluster names. Last but not least, features used in JointIE[24] are nominal features, which are expressed in terms of conditional probabilities of labels given existing attributes, while the ones extracted from images are vectors of numerical features. We need further steps to handle such heterogeneous input.

We notice that, after we rank the entry values of each visual feature vectors from the largest to smallest, the ranks within each visual cluster are quite similar. For example, the 3,704th, 1,292th and 1,175th dimension values are always among the top-20 largest features for patterns in the “police” cluster, but none of them appear in the top-20 largest features for visual pattern feature vectors of the “smoke” cluster. Although the aforementioned visual features were extracted from a hidden layer in pre-trained neural network, meaning that we have not semantically defined each of the dimensions in the 4,096 feature vectors, these visual features still provide sufficiently enriched information to the original text-only approach.

We posit that similar input (visual patterns in the same cluster) can provide similar output whose vector entry value rankings generally remain stable. Therefore, for a query argument, \( Q \), we consider the average feature vector given by

\[
\bar{z} = \frac{1}{n} \sum_{j=1}^{n} z_j,
\]

where \( z_j \in Z \).

We determine the indexes of the \( l \) largest values in the average feature vector \( \bar{z} \) and treat them as visual attributes of the argument query \( Q \), then we encode them with the correspond substructure in the graphs. Finally, we can use Equation 4 and 5 to train and test the multimodal event extraction model.

5 EXPERIMENTS

5.1 Data sets

In order to evaluate the EE performance with our proposed multimodal approach, we use two standard evaluation corpora for EE: **ACE2005**: (Automatic Content Extraction) ACE2005 [39] is a text-only corpus consisting of 600 documents including news wires, web logs, and discussion forum posts. 4,700 events covering 33 types of events and 9,700 arguments are labeled within the documents. The documents in the ACE2005 data set were generated between 2003 and 2005.

**ERE**: The LDC Entities, Relations, and Events (ERE) corpus [37] contains 336 text documents of news articles and discussion forums. 1,068 events and 2,448 arguments are labeled. The documents in the ERE data set were generated between 2010 and 2013.

We use the following data sets to generate our visual repositories:

- **ImageNet** [6]: We utilize a subset of the ImageNet images including 3,627 objects, which are annotated with bounding boxes in the images.

- **Image-Caption Pairs**: We take the 285,900 image-caption pairs used in [21] to generate our visual repository. Per [21], the 285,900 image-caption pairs are crawled and generated from all tweets of four major news agency’s accounts (the Associated Press, Al Jazeera, Reuters, and CNN) between 2007 and 2015. This data set is crawled in an indifferent manner and it covers most of the daily topics and events (including the 33 ACE event types). Since the text documents in ACE2005 and ERE data do not contain any images or captions, this image-caption data is considered as the external resource to the text-only documents.

5.2 Experiment Setup

5.2.1 Evaluation Metrics. The criteria of the evaluation follow the previous ACE event extraction work [19, 23]:

- A trigger is correct if its event type and offsets match a trigger in the ground truth.
- An argument is correctly labeled if its event type, offsets, and role match any of the argument mentions in the ground truth.

The training, validation, and test data set splits are identical to the previous work as well.

5.2.2 Baseline and Visual Repositories. We use JointIE[24] with textual features as our baseline. For our multimodal approach, we integrate the visual features with textual features and retrain the multimodal models using JointIE’s structured perceptron and make predictions on test data with the retrained models.

We generate four visual repositories in our experiments as our resources for visual features:

- **ImageNet**: The ImageNet images form a repository of 3,627 clusters named after the object names. This is the only human-generated visual repository in our experiments.

- **Faster R-CNN**: We trained a Faster R-CNN [33] model from the ImageNet subset we used. Due to hardware performance and capacity limitations, we randomly sample 50 images for each object that has more than 50 annotated images in ImageNet. We trained 10 epochs on those sampled images. The 285,900 crawled captioned images are then passed through the trained neural network and we obtain a visual repository of 617 clusters.

- **VPD**: The visual pattern discovery approach is applied on the 285,900 images. For word embeddings, we train Word2Vec on the August 11, 2014 Wikipedia dump to obtain 200-d word embeddings. The initial number of visual and textual clusters for X-means is set to 3,000. We obtain a visual repository containing 2,730 clusters.

- **VAD**: The parameters used in VAD are identical to the ones in VPD, the only difference is that we use the additional entity typing representations, which consist of 200-d embeddings. We obtain a visual repository with 1,921 clusters. We extract visual features from the retrieved visual patterns (i.e., the patches within the red bounding
Table 1: The performance (%) of event extraction. ImageNet denotes the human constructed repository. VPD and VAD are the original frameworks in [21] and our argument-centric variance, respectively.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>ACE2005 [39]</th>
<th>ERE [37]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tasks</td>
<td>Trigger Labeling</td>
<td>Argument Labeling</td>
</tr>
<tr>
<td>Metric</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>JointIE [24]</td>
<td>73.7</td>
<td>62.3</td>
</tr>
<tr>
<td>ImageNet [6]</td>
<td>72.1</td>
<td>63.1</td>
</tr>
<tr>
<td>Faster R-CNN [33]</td>
<td>72.7</td>
<td>62.9</td>
</tr>
<tr>
<td>VPD [21]</td>
<td>72.9</td>
<td>57.6</td>
</tr>
<tr>
<td>VAD (our work)</td>
<td>75.1</td>
<td>64.3</td>
</tr>
</tbody>
</table>

5.2.3 Parameter Tuning. We tune the parameters based on the F-scores of argument labeling on development sets. The tunable parameters in our experiments include the aforementioned ones, such as the initial X-means cluster numbers and epoch number of Faster R-CNN. Using a validation set, we determine that the best scores can be achieved when we include top-6 visual feature indexes.

5.3 Performance Analysis

5.3.1 General Discussion. Table 1 demonstrates that the performance after the introduction of visual repositories is significantly boosted. We present a qualitative analysis of the performance boosts.

In the sentence in Figure 6a, the trigger “cart away” does not frequently appear in the whole corpus, and external text-only dictionaries do not include either “cart” or “cart away” as Transport triggers. Therefore, although the baseline approach using text-only features provides a few hypothesis graphs which contain detections of a Transport event, because “away” is a potential indicator, the final output still fails to promote the confidence of the correct graph. However, our multimodal approach leverages the visual features from the potential arguments “medical team” and “wounded victim”, which have similar visual features to those extracted from patterns like “rescue team” and “injured man” that frequently exist in Transport events, to correctly detect the event triggered by “cart away”.

Moreover, our proposed method also improves argument labeling. For example, in Figure 6b, the traditional text-only approach missed “Ukrainian armored personnel carrier” as Instrument in Attack event triggered by “battle”, because the sentence lacks explicit textual clues to capture the relation between “battle” and “armored personnel carrier”. After we incorporate visual information, our multimodal approach can acquire from the training documents that patterns of “artillery/tank” serve as Instrument in Attack events. Since these patterns resemble those of “armored personnel carrier”, our method successfully recovers the missed Instrument argument.

However, we also observe some errors with our multimodal approach due to the joint impacts of visual and textual features. For example, in Figure 6c, our proposed method misses the Attack
event triggered by “raided” while the baseline does not. The reason for this is "headquarters" in the visual repository is primarily represented with people in business suits, which is more likely to appear in business events, so the Attack event is removed from the result.

5.3.2 Intermediate Results. Table 2 shows intermediate results on argument identification (before assigning roles within any events). From the numbers, we can conclude that the identification of arguments (including the offsets) is largely impacted by the visual repository.

We also notice that with the visual repository generated by the original VPD approach, the performance is lower than with our argument-centric repository as well as the baseline using ACE2005 documents. As discussed in earlier sections, some visual clusters are assigned names which are actually trigger words of events. This will inevitably introduce mistakes and lower the performance since the visual features are often ambiguous. For example, “air strike” patterns have similar visual features with “smoke” patterns, and will be mistakenly considered as an argument in an Attack event triggered by another word, such as "launch".

The curves in Figure 7 demonstrate that, in the early iterations, the introduction of visual features yields relatively lower performance, and convergence comes later than the text-only model. However, from 10 iterations on, the multimodal performance exceeds that of the single-modal approach. The performance gaps become stable after 15 iterations because the updates in weights of both visual and textual features tend to cease.

5.3.3 Coverage. From Table 3, we can conclude that a visual repository from a list of pre-defined cluster names can only provide limited performance boosts. We notice that fewer patterns can be retrieved from ImageNet and FRCNN during the training and testing phases. These two repositories do not provide as many clusters as the VPD and VAD repositories. Hence, the performance boost is less significant with the ImageNet and FRCNN repositories.

5.3.4 Comparison with Text-Only Approaches. Table 4 provides a comparison of our approach, the baseline, and the state-of-the-art text-only EE approach LiberalIE [16]. [16] utilizes text clustering and AMR parsing to determine the event triggers and their arguments.

Our approach has better performance on ACE data, but is not the top performer on ERE. In [16], both text document data sets (ACE and ERE) are parsed by the AMR parser, which is trained on perfect, human annotation on ERE data. The quality parsing results (where results on ERE data are far better than on ACE data) are crucial and heavily impact the performance on ACE data. Our multimodal approach (where no AMR parsing was used directly on the text documents) still provides steady improvement.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a multimodal approach to improve the performance of event extraction on text-only documents by integrating visual features from an external visual repository with conventional textual counterparts. We demonstrate a successful transfer of visual background knowledge from an established multimodal repository to target data of a single modality and observe a significant boost in the performance. In the future, we are seeking more advanced approaches to comprehensively extract information from both visual contents and text documents and to expand the schemas by discovering new event types and roles.

7 ACKNOWLEDGEMENT

This work was supported by the U.S. DARPA Multimedia Seedling grant, DARPA DEFT No. FA8750-13-2-0041 and FA8750-13-2-0045. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.