Event Specific Multimodal Pattern Mining for Knowledge Base Construction

Hongzhi Li [*†], Joseph G. Ellis [*†], Heng Ji [‡] and Shih-Fu Chang [†]

[†]Columbia Univeristy, New York, NY 10027, USA
[‡]Rensselaer Polytechnic Institute,Troy, NY 12180, USA
{hongzhi.li, jge2105, shih.fu.chang}@columbia.edu, jih@rpi.edu

ABSTRACT

Knowledge bases, which consist of a collection of entities, attributes, and the relations between them are widely used and important for many information retrieval tasks. Knowledge base schemas are often constructed manually using experts with specific domain knowledge for the field of interest. Once the knowledge base is generated then many tasks such as automatic content extraction and knowledge base population can be performed, which have so far been robustly studied by the Natural Language Processing community. However, the current approaches ignore visual information that could be used to build or populate these structured ontologies. Preliminary work on visual knowledge base construction only explores limited basic objects and scene relations. In this paper, we propose a novel multimodal pattern mining approach towards constructing a high-level "event" schema semi-automatically, which has the capability to extend text only methods for schema construction. We utilize a large unconstrained corpus of weakly-supervised image-caption pairs related to high-level events such as "attack" and "demonstration" to both discover visual aspects of an event, and name these visual components automatically. We compare our method with several state-of-the-art visual pattern mining approaches and demonstrate that our proposed method can achieve dramatic improvements in terms of the number of concepts discovered (33% gain), semantic consistence of visual patterns (52% gain), and correctness of pattern naming (150% gain).

CCS Concepts

•**Information systems** → *Clustering;* •**Computing methodologies** → *Machine learning algorithms;*

Keywords

Multimodal, Pattern Mining, Convolutional Neural Network, Event Schema, Multimodal Knowledge Base

[*]Denotes equal contribution.

ACMMM '16 Amsterdam, The Netherlands

ISBN 978-1-4503-3603-1/16/10. . . $15.00

DOI: http://dx.doi.org/10.1145/2964284.2964287

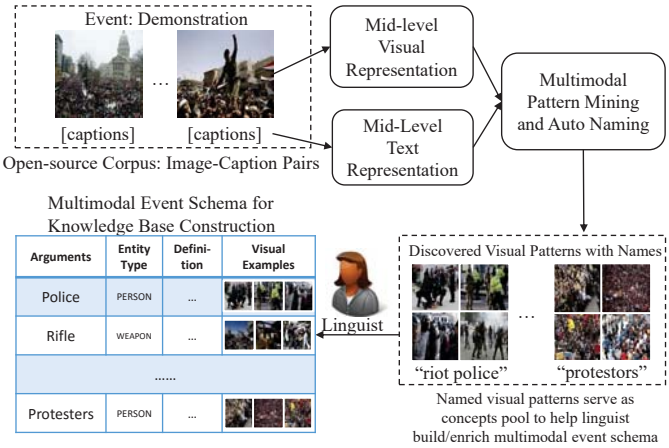

Figure 1: We propose a novel system to automatically discover and name visual patterns specific to each high-level event which helps experts construct event schema and related knowledge bases.

1. INTRODUCTION

With recent advances in computer vision, researchers have been able to demonstrate impressive performance at near-human level capabilities in difficult tasks such as image recognition. For example, computer vision systems now have the ability to recognize if a dog, cat, or car appears in an image. These advances are made possible by utilizing the massive amount of image datasets and label annotations, sometimes also with bounding boxes around the objects of interest within the image. However, the difficulty with the current systems is that every time when we want to learn a new visual concept, users must be looped in to manually define the target class and label the training data. Also, it is unclear how we can know apriori for a particular domain what the most important concepts we want to focus on and train corresponding detectors. For example, if an ontology expert is asked to construct a schema to describe high-level events like "attack", what visual classes or concepts are most relevant and can be automatically detected? Can we use data mining to discover such relevant classes directly from data available from the domain and use the discovered concepts to help the expert to build the new schema? Our intuition tells us that "gun", "knife", or "explosion", could be important things to try to detect from the images. However without inspecting the content of the images our preconceived notions of what is important may only tell part of the story of a given event. We will show later that other concepts such as "smoke", "air strike", and "police" might actu-

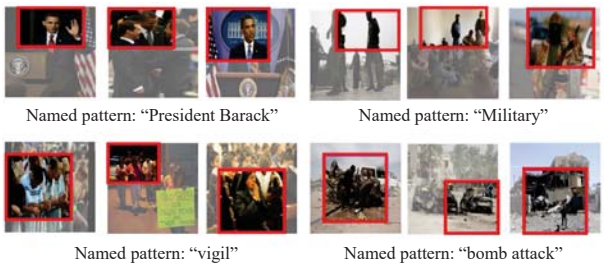

Figure 2: Examples of the localized patterns automatically discovered and named by the proposed multimodal approach.

ally appear much more frequently in the specific data corpus in the domain, but it would be difficult to know these concepts appear frequently without first inspecting a portion of the images. Situations like this can arise in many application domains (like sports, security, and open source content). Thus, there are great needs for automatically discovering concepts relevant to events of interest from data corpora in various specific domains.

In response to the challenge mentioned above, in this work we attempt to answer the following question: given a specific target domain and associated data corpora, how do we develop automatic methods to rapidly discover salient multimodal patterns that are semantically coherent, visually consistent, and can be automatically named with semantic concepts relevant to the high-level events in the target domains. The proposed architecture is described in Fig. 1. We use artifical neural network models to generate visual and text representation for image-caption pairs, and apply the association rule mining algorithm to discover multimodal visual patterns. In this work, we use news as an example domain and image caption pairs as example multimodal corpora to demonstrate the effectiveness of the proposed methods. But the overall proposed framework is general and can be easily extended to other domains.

In many applications like personalized news or social media tracking, it's often necessary to process information at the high-level event level (such as "Baltimore Riot" or "Syria Bombing"). Using knowledge bases to improve event-level processing has been a major research topic. In this work, we specifically focus on discovery of named multimodal patterns related to events in a specific domain. We use such patterns as candidate concepts that can be used to assist knowledge experts to construct new event knowledge bases.

As shown in Fig. 1 and 2, a pattern is formed by a collection of localized instances within images (referred to as image patches later in the paper) and their associated text captions. We specifically focus on patterns that are represented by both text and image modalities. We also emphasize the importance of the automatic naming process in order to assign semantic concepts relevant to high-level events (such as "explosion", "vigil" concepts for "attack" event). Such automatic discovery and naming processes will produce a pool of high-quality multimodal concepts that are semantically coherent, visually consistent, and semantically relevant to specific events of interest.

The proposed discovery and naming methods for multimodal patterns are distinct from the existing work in mining of visual patterns. Recently, [21] and [12] proposed discovery of representative and discriminative visual patterns as mid-level image representation in order to improve image classification performance, compared to the traditional representations using low-level features (such as visual bags of words). Visual patterns have also been used to summarize image collections by [31] [18]. Although such visual patterns capture unique visual characteristics that can be used to separate image classes, they often do not correspond to clear semantic concepts that are at a level suitable for defining entities included in high-level events. For example, visual patterns discovered in [12] from the MIT scene dataset are often at the lower level like chairs, windows, or furniture parts that do not meet the needs of high-level knowledge base construction.

Evaluation of pattern mining and naming is challenging due to the novelty of the problem and the subjectiveness of the task. It's important to note the goal is not to improve image classification tasks like what has been done in past works [12] [7]. Instead, our focus is on semantic naming of the discovered patterns and their relatedness to high-level events as mentioned earlier. In this work, we propose expert evaluation methods to compare the proposed approaches with several state-of-the-art baselines for visual discovery and naming procedures. The contributions of this paper can be summarized as follows:

- We develop a novel multimodal mining framework for discovering visual patterns from a collection of image-caption pairs and automatically naming the discovered patterns, producing a large pool of semantic concepts specifically relevant to a high-level event. The named visual patterns can be used to construct event schema needed in the knowledge base construction process.

- The proposed system can discover many novel semantic concepts not covered by existing visual ontologies such as ImageNet. This helps break the aforementioned bottleneck in extending existing visual classifiers to new domains.

- Our system exploits the joint multimodal representations in discovering unique patterns, which are shown to be more visually coherent and semantically correct, compared to baselines using separate processing of individual media modalities.

2. RELATED WORK

Low-level image features such as SIFT [15] and Bag-of-Words methods were widely used as a representation for image retrieval and classification. However, researchers have proven that these low-level features are insufficient for representing the semantic meaning of images. Mid-level image feature representations are often used to achieve better performance in a variety of computer vision tasks. Some frameworks for using middle level feature representations, such as [11, 28, 29, 30], have achieved excellent performance in object recognition and image retrieval. ImageNet [5], was introduced and has led to breakthroughs in tasks such as object recognition and image classification due to the availability of a massive amount of well-labeled data. Each of the images within ImageNet is manually labeled. Thus, it is a very expensive and time-consuming task. Other similar datasets, including SUN[25], MSCOCO [13], and UCF101 [22] are created for object/scene/concept classification tasks. However, the manually defined ontologies are quite limited and oftentimes do not extend to real-world data, and therefore may not cover the concepts needed to build an event schema for high-level events. This work looks beyond a manually defined ontology, and instead focuses on mining multimodal patterns automatically from weakly supervised data to attempt to unbind researchers from the need for costly supervised datasets. We approach this problem from a multi-modal perspective (using the image and caption together), which allows us to name and discover higher-level image concepts.

Visual pattern mining is an important task since it is the foundation of many middle-level feature representation frameworks. [9] and [31] use low-level features and a hashing approach to mine visual patterns from image collections. [26] utilizes a spatial random partition to develop a fast image matching approach to discover visual patterns. All of these methods obtain image patches from the original image collection either by random sampling or salient object proposal and utilize image matching or clustering to discover similar patches to create visual patterns. These methods are computationally intense, because they have to examine possibly hundreds or thousands of image patches from each image. These methods rely heavily on low-level image features, and therefore do not often produce image patches that exhibit high-level semantic meaning. The generated image patterns are typically visually duplicated or near-duplicated image patches.

Convolutional neural networks (CNN) have achieved great success in many research areas [20], [10]. Recently, [12] combined the image representation from a CNN and the association rule mining technique to effectively mine visual patterns. They first uniformly sampled image patches from the original image and extracted the fully connected layer response as features for each image patch utilized in an association rule mining framework. This approach is able to find consistent visual patterns, but cannot guarantee the discovered visual patterns are relevant to specific events and can be used in event schema construction. Most of the existing visual pattern mining work focuses on how to find visually consistent patterns.

Other related works focus on visual knowledge mining. [6] propose a visual instance mining system to find the unique visual elements that are the most distinctive for a certain geo-spatial area. They attempt to answer the question "What makes Paris looks like Paris?", which is answered by the visual patterns discovered by their system. They first collect images from Google Street View. Then the images are sampled to obtain image patches at different scales. A discriminative clustering approach is proposed to take into account the weak geographic supervision and discover the visual patterns with the unique character of a specific city. This idea is very inspiring, as the visual patterns discovered are used not only as a mid-level representation, but as valuable knowledge by themselves.

The NEIL system [3] is another example of visual knowledge mining work. NEIL automatically discovers common sense relationships and labels instances of given visual categories. In NEIL, the discovered knowledge consists of relationships between predefined objects and concepts. The NEIL system is also able to find new visual instances of given categories. However, while NEIL is limited to a predefined ontology, our proposed multimodal pattern mining method aims to discover and name new objects/concepts from an unstructured set of image-caption pairs.

Another category of related works is image captioning. In recent years, many researchers have focused on teaching machines to understand images and captions jointly. Image caption generation focuses on automatically generating a caption that directly describes the content in an image using a language model. Multimodal CNN [8] is often used to generate sentences for the images. All the existing works use supervised approaches to learn a language generation model based on carefully constructed image captions created for this task. The datasets used in caption generation, such as the MSCoco dataset [13] consist of much simpler sentences than appear in news image-caption pairs. We differ from these approaches in that we do not try to generate a caption for images, but instead use them jointly to mine and name the patterns that appear throughout the images.

3. MULTIMODAL PATTERN MINING

In this section, we discuss our multimodal pattern mining (MMPM) framework. In particular, we will describe how we collect a large-scale dataset, generate feature-based transactions from the images and captions, and discover and name semantic visual patterns.

3.1 Weakly Supervised Event Dataset Collection

We believe that by using weakly supervised image data from target categories that are sufficiently broad, we can automatically discover meaningful and easily nameable image patch patterns for structured ontology generation. To accomplish this task, we collect a set of image caption pairs from a variety of types of news event categories.

We begin by crawling the complete Twitter feeds of four prominent news agencies, the Associated Press, Al Jazeera, Reuters, and CNN. Each of these agencies has a prominent Twitter presence, and tweet links to their articles multiple times a day. We collect the links to the articles and then download the HTML file from each extracted link. The articles span the time frame from 2007-2015, and cover a variety of different topics. We then parse the raw HTML files and find the image and caption pairs from the downloaded news articles. Through this process, we were able to collect approximately 280k image-caption pairs.

Once we have collected the dataset, we want to find image-caption pairs that are related to different events covered in news. We utilized the event ontology that was defined for the Knowledge Base Population (KBP) task in the National Institute for Standards and Technology Text Analysis Conference in 2014 to provide supervision to our dataset. Within this task, there is an event track with the stated goal to "extract information such that the information would be suitable as input to a knowledge base." This track goal closely models the goals of learning patterns that are easily nameable with semantic concepts and hence could be used in knowledge base population. This makes this ontology a perfect fit for our task.

The KBP event task utilizes the ontology defined by the Linguistic Data Consortium in 2005 [24]. This event ontology contains 34 distinct event types; the events are broad actions that appear commonly throughout news documents, such as *demonstrate*, *divorce*, and *convict*. Provided in the training data with these event ontologies is a list of trigger words for each of the events that are used to detect when an event appears in text data. An example of some of the trigger words used for the *demonstrate* event are: protest, riot, insurrection, and rally. We search each of the captions for a trigger word from the event category, and if an image caption contains that trigger word, we assign that image caption pair to the given event category. An example of the number of images for some representative events can be seen in Table 1.

Table 1: Number of images per event category for some of the most popular event categories in our dataset.

Event	# of Images	Event	# of Image
Attack	52649	Injure	5853
Demonstrate	20933	Transport	51187
Elect	9265	Convict	1473
Die	26475	Meet	32787

3.1.1 Review of Pattern Mining

In this section, we will review the basic ideas and definitions necessary for pattern mining. Assume that we are given a set of n

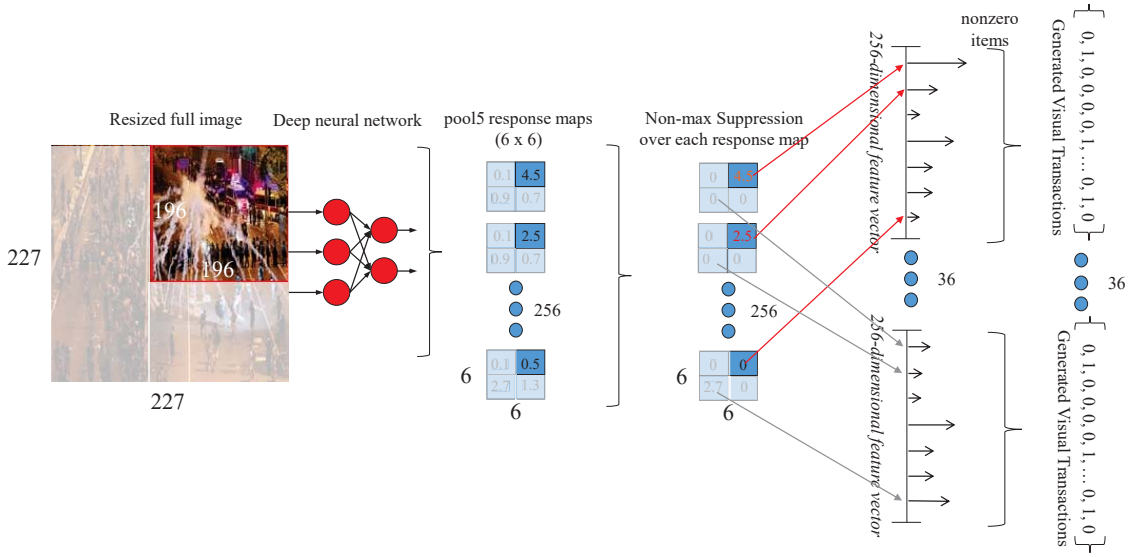

Figure 3: The visual transaction generation pipeline utilizing the last convolutional layer of a convolutional neural network. Using this pipeline, we are able to obtain representations of each image that can localize the presence of a pattern within the image. We use convolutional layers from AlexNet as an example.

possible observations $X = \{x_1, x_2, ...x_n\}$; a *transaction*, T, is a set of observations such that $T \subseteq X$. Given a set of transactions $S = \{T_1, T_2, ...T_m\}$ containing m transactions, our goal is to find a particular subset of X, say t^*, which can accurately predict the presence of some target element $y \in T_a$, given that $t^* \subset T_a$ and $y \cap t^* = \emptyset$. t^* is referred to as a *frequent itemset* in the pattern mining literature. The relationship from $t^* \rightarrow y$ is known as an *association rule*. The support of t^* reflects how often t^* appears in S and is defined as,

$$s(t^*) = \frac{|\{T_a | t^* \subseteq T_a, T_a \in S\}|}{m} \quad (1)$$

Our goal is to find association rules that accurately predict the correct event category for the image-caption pairs. Therefore, we want to find patterns such that if t^* appears in a transaction there is a high likelihood that y, which represents an event category, appears in that transaction as well. We define the *confidence* as the likelihood that if $t* \subseteq T$ then $y \in T$, or,

$$c(t^* \rightarrow y) = \frac{s(t^* \cup y)}{s(t^*)}. \quad (2)$$

3.1.2 Transaction Generation from Images

Certain portions of a CNN are only activated by a smaller region of interest (ROI) within the original image. Throughout this paper, we will utilize the CNN defined in [10], which is a common CNN structure that is often used for computer vision tasks. The last layer in which the neurons in that layer do not correspond to the entire image is the output of the final convolutional and pooling layer. Based on this observation, for each image we find the maximum magnitude response from a particular feature map from this layer of the CNN. The last pooling layer of [10] is commonly known as "pool5". This layer consists of 256 filters; and the response of each of the filters over a 6×6 mapping of the image. The corresponding ROI from the original image in which all the pixels in that region contribute to the response of a particular neuron in the pool5 layer is a 196×196 image patch from the 227×227 resized image. These 196×196 image patches come from a zero-padded representation of the image with zero-padding around the image edges of 64 pixels and a stride of 32. Namely, from a 227×227 scaled input image, a total of 6×6 (36) patch areas are covered from all the stride positions, resulting in a 6×6 feature map for each filter in this layer. Using this approach, we are able to leverage the existing architecture to compute the filter responses for all patches at once without actually changing the network structure. This idea allows us to extract image patch patterns in a way that is much more efficient than many current pattern mining methods, which utilize a sampling approach.

We use the pre-trained CNN model from [10], trained on the ImageNet dataset for extracting the pool5 features for the news event images. For each image, we keep the maximum response over the 6×6 feature map and set other non-maximal values to zero for all 256 filters, which is similar to non-maximum suppression that appears throughout the CNN literature. This operation finds the patch triggering the highest response in each filter and helps avoid redundant patches in the surrounding neighborhood in the image that may also generate high responses. The above process results in a 256-dimensional feature vector representation for each image patch. We then set the nonzero items in this feature vector to 1. This creates a binary representation of which filters are activated for each image patch. We use these sparse binarized features to build *transactions* for each image patch as discussed in Sec. 3.1.1, where the nonzero dimensions are the items in our transaction.

By utilizing the architecture of the CNN directly, we are able to efficiently extract image features that come from specific ROI that are suitable for pattern mining. The current state-of-the-art pattern mining technique proposed by [12] requires a sampling of the image patches within each image and then operating the entire CNN over this sampled and resized image patch. This procedure is very costly, because the CNN must be used to extract features from a number of sampled images that can be orders of magnitude larger than the dataset size. For example, for the MIT indoor

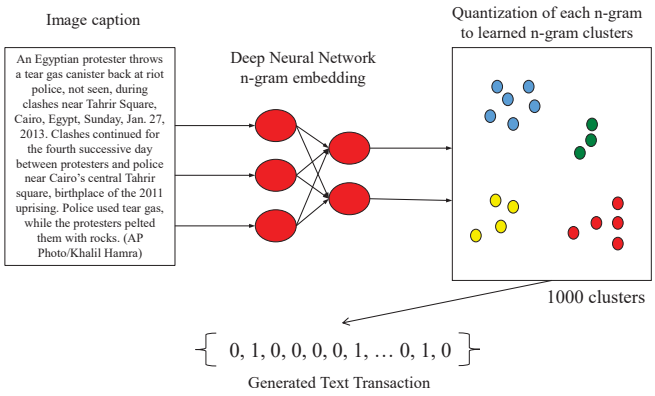

Figure 4: The text transaction generation pipeline.

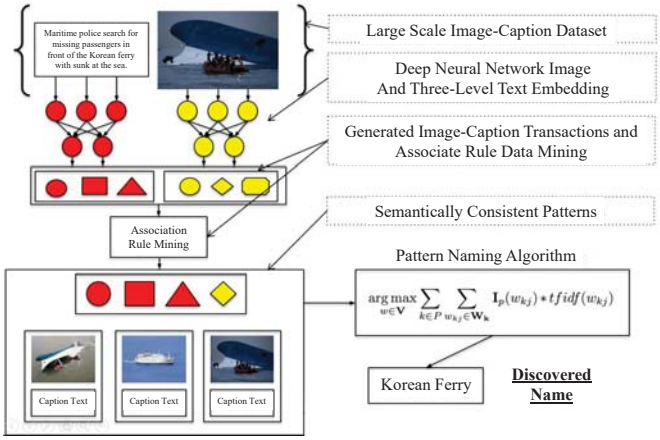

Figure 5: Our full multimodal pattern discovery and naming pipeline

dataset, the authors of [12] sample 128×128 size image patches with a stride of 32 from images that have been resized such that their smallest dimension is of size 256. Thus, the number of image samples that are taken for each image is greater than or equal to $(\frac{256-128}{32}+1)^2 = 25$. The full CNN must operate on all of these sampled images. In contrast, our method works directly on the images themselves without any sampling, but still has the ability to localize images within the dataset. We are able to extract representations for 36 image patches from an image while only having the CNN operate on the image once. By leveraging the structure of the CNN during test or deployment, our method is *at least* 25 times less computationally intensive than the current state of the art. We will discuss the speed-up we have in deploying our event-specific pattern mining compared with other approaches in our experiment section. Fig. 3 shows our visual transaction generation pipeline in detail.

In this paper, we choose to use the response map of the pooling layer after the last convolutional layer to build the image transactions. This choice is inspired by the research on visualizing the convolutional neural network [27] and the recent success of using convolutional layers for object semantic segmentation [14] [19]. Those works have proved that the filters in the last convolutional layers have strong capability to capture the semantic objects in the input images. Thus, we use this particular setup in our model. However, it is important to note that our proposed multimodal framework is very flexible so that the image transaction generation model or the text transaction generation model can be replaced by other advanced models to achieve even better results.

3.1.3 Transaction Generation from Captions

We have discussed how we generated transactions by binarizing and thresholding the CNN features that are extracted from the images. Similarly, we require an analogous algorithm for generating transactions from image captions.

We begin by cleaning each of the image captions by removing stopwords and other ancillary words that are not relevant (HTML tags or URLs). We then tokenize each of the captions and find all of the words that appear in at least 10 captions in our dataset. Once we find these words, we use the skip-gram model proposed in [16] that was trained on a corpus of Google News articles to map each word to a 300-dimensional embedded space. This model works well in our setting, because the structure of image captions is similar to that of news article text, which this model was trained on. The skip-gram model works well in our context because words with similar uses end up being embedded close to each other in the feature space. Words such as "religious clergy", "priest", and "pastor" all end up close in Euclidean distance after embedding and far away from words that are not similar. We cluster the words using K-means clustering to generate 1,000 word clusters.

To generate transactions for each caption, we map each word back to its corresponding cluster, then include that cluster index in the transaction set. We remove patterns that contain cluster indices that are associated with commonly used words by having a high confidence score threshold as defined in Eq. 2. The cluster indices that frequently appear in captions from a particular event category but rarely for other categories are found through our association rule mining framework.

We require our discovered patterns to contain items from both the visual and text transactions. By requiring words with similar meaning to appear in all captions of a visual pattern, we are able to discard patterns that may be visually similar but semantically incoherent. The skip-gram-based algorithm is able to handle differences in word choice and structure between captions to effectively encode meaning into our multimodal transactions.

3.2 Mining the Patterns

We add the event category of each image as an additional item in the generated transaction for each of the image caption pairs. Inspired by [12], we use the popular apriori algorithm [1] to find patterns within the transactions that predict which event category the image belongs to. We only find the association rules that have a confidence higher than 0.8, and calculate the support threshold that ensures that at least 30 image patches exhibit a found pattern. Finally, we also remove any rules that only contain items generated from the image or caption transactions, ensuring that we only retain truly multimodal patterns. Therefore, our pattern requirements can be described mathematically as,

$$\begin{aligned} &c(t^* \rightarrow y) \geq c_{min} \\ &s(t^*) \geq s_{min} \\ &t^* \cap \mathbf{I} \neq \emptyset \\ &t^* \cap \mathbf{C} \neq \emptyset. \end{aligned} \tag{3}$$

where as defined in Eq 1 and Eq 2, y is the event category of

the image-caption pair, c_{min} is the minimum confidence threshold, s_{min} is our minimum support threshold, $\mathbf{I}$ represents the items generated from the image transaction pipeline, and $\mathbf{C}$ are those generated from the caption pipeline. At the end, each multimodal pattern t* contains a set of visual items (fired filter responses in pool5 in CNN model) and a set of text patterns (clusters in the skip-gram embedded space).

3.3 Naming the Patterns

We name the generated patterns so that they can be used for higher-level information extraction tasks, such as event schema generation. We leverage the fact that we have captions associated with each of the images to generate names for each pattern.

We begin the process of name generation by removing the words that are not generally useful for naming but appear often in captions. The words that are removed include standard English language stop words (or, I, etc.), the name of each month, day, and directional words such as "left" and "right". After cleaning the caption words, we then encode both unigram and bigrams into a vector using tf-idf encoding. We ignore any unigrams or bigram that does not appear at least 10 times across our caption dataset.

Once these words are removed we then sum the TF-IDF vector representations of each word in all of the captions associated with a particular pattern. We then take the argument max over the summed TF-IDF representations to obtain the name for this pattern. The word embedding described in Sec. 3.1.3 ensures that words with semantically similar usages and meanings will be clustered together, and the TF-IDF naming algorithm chooses the most appropriate word from the associated clusters for a particular pattern. This procedure is explained mathematically in the following way: Let p be a found multimodal itemset (pattern), and T_k is the multimodal transaction for the k'th generated transaction in our dataset. We define the set P as all the indices of transactions that contain p, or $P = \{i | p \subseteq T_i, \forall i\}$. In Eq. 4, $\mathbf{V}$ is our vocabulary, W_k is the set of words from the k'th caption, $\mathbf{I}_p(w)$ is an indicator function on whether w corresponds to a word cluster in the itemset of p, and w_{kj} is the j'th word in the k'th caption,

$$w_{name} = \arg\max_{w \in \mathbf{V}} \sum_{k \in P} \sum_{w_{kj} \in W_k} \mathbf{I}_p(w_{kj}) * tfidf(w_{kj}) \quad (4)$$

Once the names are found, we remove any name that appears in more than 10% of the captions of a particular event. This is important because for particular events like "injure", words such as *injure* and *wounded* appear across many captions, and may lead to poor naming. Some examples of discovered patterns and the names that we have assigned to them can be seen in Fig 6. Our full pattern naming algorithm and pipeline can be seen in Fig 5.

4. EVALUATION

4.1 Baseline Methods

In this paper, we propose a complete end-to-end pipeline for discovering and naming visual patterns from large-scale image-text datasets, in particular news event-related content. The discovered visual patterns with names are used to help linguists build the multimodal event schema. To the best of our knowledge, there is no existing research that attempts to solve the specific problem that we have addressed. However, some existing techniques can be modified and then used to address this problem. We propose 3 different baseline approaches based on some state-of-the-art pattern mining techniques and then compare our approach with those baselines.

4.1.1 *Baseline 1: Object Proposal and Clustering (OPC)*

We follow the commonly used pipeline that is widely used in the visual pattern mining literature [6] [4] for this baseline approach. We first sample the images at multiple scales to obtain image patches, and then image features are extracted from each patch. The patches are then clustered using K-Means clustering to group the image patches into visual patterns. We use existing state-of-the-art methods in each component to implement this baseline. Selective Search [23] is used to propose multi-scale image patches. The response from second to last fully connected layer of the CNN, VGG19 [20], is used as the feature representation of each image patch. We attempted different cluster numbers for the K-Means algorithm and report the best performance in the following section. The selection of parameters is discussed in Sec. 4.4.

To name the discovered visual patterns, we utilize topics discovered by Latent Dirichlet Allocation (LDA) [2] to find pattern names from the associated caption text for each visual pattern. We first collect all the image captions from one visual pattern, and this collection of captions represents a document in our LDA framework. Then we discover n topics over all the documents, in this work, n was set to be 1,000. Each document is then assigned to be generated from some distribution over the topics. We first take the intersection of the set of all the words from its document and the set of the topic words from the topic. The words in this intersection are then ranked based on the corresponding topic score for each word. The word ranking score can be seen in Eq. 5 where w_i is the word being ranked, W_k is the set of words that appear in the captions of a pattern, T_j is the set of words from the j^{th} topic, and t_{ij} is topic score for word w_i in relation to topic j. In this equation, T_j represents the topic that is chosen by LDA to be most representative of this document.

$$w_{score}(w_i) = \sum_{w_i \in W_k \cap T_j} t_{ij} \quad (5)$$

We choose the top 3 words based on our above ranking algorithm to name each visual pattern.

4.1.2 *Baseline 2: Mid-level Deep Pattern Mining (MDPM)*

MDPM is proposed by Li et al. in [12]. They use the response of a fully connected layer of a pre-trained CNN to build transactions and apply association rule mining to find visual patterns. Their method is the most similar current approach to ours, but they only use visual information in their approach, while we use both visual and text modalities to build the multimodal transactions. This method is a state-of-the-art approach for visual pattern mining, and we will demonstrate the performance gains that can be achieved using multimodal information as opposed to visual only. We use the code provided by the authors of [12] in our experiments. As their method is designed for finding visual patterns only, we use the same LDA naming algorithm introduced in section 4.1.1 to name the discovered visual patterns.

4.1.3 *Baseline 3: Object Detection to Find Concepts from Image Dataset (OD)*

With the recent rapid development of artificial neural networks for image classification, researchers can accurately detect objects and concepts from images. We use a pre-trained convolutional neural network model [20] to detect concepts and objects from our dataset. We select the top k frequently detected objects/concepts in each event as k visual patterns. We set k to be 50 in our experiment. Naturally, all the images with the same detected concept are

Figure 6: Some examples of named visual patterns discovered by our MMPM. These examples are discovered for the "attack" event. Although each pattern may contain more images, we only show the first six images for each pattern.

Figure 7: Some examples of visual patterns discovered by baseline 1: object proposal and clustering (OPC) approach. It is important to note that OPC is also capable of finding meaningful concepts. We selectively show these examples to demonstrate the common problem with OPC. For example, the partial object in "visual pattern 2" and the solid color pattern in "visual pattern 3". These types of patterns are not useful for multimodal event schema construction.

grouped as a visual pattern, and the name of the synset corresponding to the detected visual pattern is used as the pattern name.

4.2 Subjective Evaluation

We first show some examples of visual patterns discovered by our method and the baseline approaches in Fig. 6 and 7. Some subjective conclusions can be drawn from the illustration. These conclusions are representative of the algorithms and the patterns that are generated by all of the methods.

Image patch sampling-based approaches (OPC and MDPM) have a tendency to find low-level visual patterns. Examples of this are solid color patches and partial objects (Visual Pattern 2 and 3 in Fig. 7). This is due to the fact that these methods sample image patches at the beginning of their pipeline, which is necessary for discovering visual patterns at different scales. However, the solid color patches and partial object patches are often produced using these approaches. Unfortunately, it is very difficult and expensive to avoid generating such patches by either a random/uniform image patch sampling approach or using object proposals. It is, therefore, not a surprise to find meaningless or very low-level conceptual image patterns, such as color similarity or "bars", as can be seen in "Visual Pattern 2" in Fig. 7. In our MMPM method, we use the response map of the last convolutional layer of a CNN to generate visual transactions and localize the image patches. It has been shown in [27] and [32] that the filters in the last convolutional layer of a CNN are particularly useful in capturing high-level semantic concepts/objects from local regions in an image. This high-level semantic nature of the filters we use in conjunction with the text information that we extract from the captions naturally leads to higher-level semantically meaningful patterns than the baseline approaches. Another issue with visual-only approaches is that the visual clustering based approach (OPC) fails to distinguish different semantic concepts if they have the similar visual appearances. For example, "Visual Pattern 1" in Fig. 7 shows a mix of patches that were generated from images containing "rocket", "smoke", and "air strike" concepts. It is an obvious drawback of a visual-only approach and particularly difficult to overcome since those concepts are indeed visually similar. Our MMPM approach can overcome this problem by involving the semantically relevant text information from image captions to distinguish between visually similar patterns. Examples of "air strike", "burns", "smoke", and "rockets" and the differences can be seen in Fig. 6.

4.3 Quantitative Evaluation

We evaluate the quality of multimodal visual patterns by evaluating whether the named visual patterns generated by each method

can be used to construct an event schema. The names of visual patterns are usually used as entities or arguments in the multimodal event schema. The images from the visual pattern are used as visual examples and can be further used to train visual classifiers to expand the visual example set. To obtain this annotation, we have hired linguists and others who have experience in designing structured event schemas and ontologies. We give the annotators a set of named visual patterns for each event. The task they complete is to build a multimodal event schema using the concepts provided by the pattern mining methods. The linguist is asked to determine if the pattern name, which is a unigram or bigram, and the associated images can be used to construct an event schema based on the following three criteria:

1. **Event relatedness:** The discovered concepts (pattern names and related images) are relevant to the event, and would be useful entities or attributes in an event schema for that event.

2. **Visual semantic coherence:** The visual pattern associated with the name is semantically consistent. Namely, the images shown under the pattern depict a coherent semantic concept, and not a mix of many different concepts. If the majority of images in a pattern are consistent, with few outliers, the pattern is considered to exhibit this property.

3. **Text-visual matching:** The pattern name correctly describes the semantic concept of the visual pattern.

In our experiment, three experts [1] examine the event relatedness of approximately 2,500 unique pattern names discovered from the dataset we described in Sec. 3.1 by our method and the baselines. We also check the visual semantic coherence and name quality (text-visual matching) of 10,000 visual patterns. Each pattern name is checked by three experts to determine the event relatedness independently. To avoid bias, the discovered pattern names from all the methods are mixed and then presented to the experts without the information of which method discovers the pattern name. Similarly, the semantic coherence and name quality of each pattern are also checked by three experts independently without bias. The final results reported in the following sections are determined using majority vote of all the examiners.

4.3.1 Coverage of the Discovered Concepts

We want to measure how many concepts are correctly discovered by each method and how many unique concepts each method is responsible for. This experiment does not aim to prove which method is better, but we make the point that the proposed MMPM can find additional concepts that cannot be discovered by the baseline visual pattern mining plus a naming algorithm and do not currently exist in the current concept/object ontologies defined by computer vision researchers, such as ImageNet[5], Places[33], MIT Indoor[17], and others. Table 2 shows the coverage of discovered concepts for each method. The metrics shown are averaged across each of the events tested for each method. We show two metrics for each method: 1. the number of detected concepts by each method relevant to a particular event, and 2. The number of unique concepts that a method detects relevant to the event and that are not found by other methods.

The MMPM method finds many more event relevant concepts than other approaches, with at least 33% increase. Among the discovered concepts in each event, there are approximately 58 unique concepts that cannot be found by any other baseline methods. Compared with the commonly used visual ontology ImageNet, over 95% of the concepts discovered by MMPM are not covered by the ImageNet ontology. We find that there are only around 5 related concepts discovered by the visual ontology ImageNet for each event. This demonstrates the narrowness of current state-of-the-art visual ontologies. Those visual concepts do not meet the needs of high-level knowledge base construction, particularly for events. It is therefore important to adopt a multimodal approach to concept discovery.

[1] The evaluators include a linguist and two CS PhD students.

Table 2: Average number of relevant concepts discovered by each method (per event), and number of unique concepts that are only discovered by each method

Method	# concepts discovered (per event)	# unique concepts
OPC	73	29
OD	5	3
MDPM	50	25
MMPM	97	58

Table 3: Number of found patterns in news events.

Event	# of Patterns	Event	# of Patterns
Attack	573	Convict	0
Demonstrate	1247	Die	146
Elect	42	Injure	45
Meet	5159	Transport	509

4.3.2 Visual Semantic Coherence

A pattern is judged to be semantically coherent if the majority of annotators judge patches associated with the patter have consistent semantic concepts. As shown in Table 5, all the methods achieve reasonable results in finding visual patterns that exhibit semantic coherence. We can see that our method outperforms the competing baselines by a large margin, and exhibits a 52% performance improvement over the current state-of-the-art approach. The ability to leverage the text information in our algorithm for discovering patterns tends to lead to patterns that exhibit this semantic coherence, because we ensure that not only is the visual content similar in appearance, but the corresponding text content also has similar meaning. The other approaches that utilize only visual information can be misled by image patches that are visually similar but share little to no semantic similarity, as shown in Fig. 7. The added visual semantic coherence of our patterns allows them to be leveraged in event schema construction, which demonstrates why we are able to find more patterns suitable for this task as shown in Fig. 6.

4.3.3 Correctness of Visual Pattern Names

We show the accuracy of attempting to name the visual patterns in Table 6. MMPM significantly outperforms the other methods for visual pattern naming, with approximately 150% improvement in naming accuracy. This is because MMPM combines information from both visual and textual modalities to build the image caption representations, and therefore the discovered patterns tend to be more semantically consistent, and are discovered based on the appearance of semantically similar words in the captions. MMPM can distinguish different semantic concepts even if they are visually similar, as shown in the naming results. The baseline approaches first find the visual patterns and then use the LDA naming model to name the visual patterns. In this approach, text is not leveraged in the pattern discovery process. In such cases, the naming

Table 4: Concepts (visual pattern names) discovered in "attack" event using our multimodal pattern mining method

air strike	damage	rockets
aggravated assault	dead	security forces
army	destroyed	shot
arrested	explosion	strike
bomb	fighting	tank
bomb attack	forces	tear gas
bomber	gun	terrorist
burns	helicopter	troops
confrontation	riot police	

Table 5: Semantic coherence of the visual patterns. We show the percent of semantically coherent visual patterns among all the patterns discovered by each method.

Method	Semantically Coherent Visual Patterns (%)
OPC	33.76
OD	31.25
MDPM	50.15
MMPM	**76.32**

algorithm cannot find correct names for these patterns, because the patterns may be visually similar, but semantically different. We noticed that the performance of baseline 3, object detection, does not achieve good performance in pattern semantic coherence and name quality evaluation. The cross-domain issue is the reason why baseline using object detectors trained in a different domain could not achieve high performance. However, the comparison is still appropriate since our method and other baselines also use the same network architecture and pretrained model without fine-tuning using the target dataset (although we only use the top five convolutional layers). This demonstrates again the deficiency of existing visual ontology and object detectors in extracting knowledge in a new target domain, and importantly why we require a multimodal approach to find visual patterns for event schema construction, particularly when naming them is important.

4.4 Discussion of Parameters

We discuss some parameters of MMPM and baseline methods in this section. As is natural in clustering methods, the number of clusters is an important tunable parameter. In baseline 1, OPC, we tried different numbers of clusters in the experiments. The clusters are formed over each event. Instead of setting a unique cluster number for all the events, we calculated the number of clusters as the number of patches in each event divided by the expected number of patches in each cluster. We tried $\{20, 30, 50\}$ as the expected number of patches in each cluster. We only report the best performance in the experiment section. Since we use the apriori algorithm [1] for association rule mining in our MMPM method and baseline 2 MDPM method, there are two tunable parameters within the algorithm: support rate and confidence rate; we discussed them in Section 3.2. In general, our performance is not sensitive to those two parameters. However, we do note that a small support rate may lead to finding many duplicate visual patterns for both methods. The number of selected concepts in baseline 3 (object detection) is not sensitive to parameters. In our experiment, the top 50 concepts are able to cover all the potential discovered concepts in the ImageNet ontology. We examined the top 100 concepts, but we found that there were no "relevant concepts" discovered after the top 50 concepts for each event.

Table 6: Evaluation of naming the discovered patterns. The accuracy of each method is shown.

Method	Pattern Naming Accuracy (%)
OPC	15.0
OD	21.4
MDPM	23.0
MMPM	**57.4**

4.5 Complexity and Efficiency

MMPM is not only effective for finding and naming multimodal visual patterns, but it is also quite efficient when compared with the baseline approaches. Most current state-of-the-art pattern mining methods work on the image patch level. The complexity of the algorithm can be roughly estimated by the number of patches. Since the value of the response map of the last convolutional layer is usually sparse, on average, we only select about $3-5$ patches per image after non-max suppression. (We ignore the transactions that only contain "0" in each dimension of the feature vector.) Compared with the other baselines, object proposal usually generates hundreds of image patches per image and MDMP generates approximately one hundred patches per image. Therefore, our method is faster by more than an order of magnitude than competing approaches, because as discussed in Sec. 3.1.2 our method works on an image level, unlike other approaches. It is important to note that although our method takes the whole image as input, it is still able to localize the visual patterns, similar to the other image patch based methods. The actual running time of our experiments demonstrates what we have described here. We run our experiments on the same workstation with two Intel Xeon E5 CPUs, 64GB memory, and Nvidia TITAN X GPU. The dataset used in our experiments has about 100,000 images. MMPM takes about 2 hours to finish the entire pipeline, including feature extraction, transaction generation, pattern mining, and naming. OPC takes about 18 hours, and MDPM requires approximately 36 hours. It is necessary to mention that MDPM is implemented in MATLAB by the original authors of [12]. The actual running time may be improved by using a more efficient implementation.

5. CONCLUSIONS

We have developed a novel dataset and algorithm for mining and naming multimodal visual patterns from a corpus of high-level news event image caption pairs. Our multimodal pattern mining method is able to discover patterns that are more informative than the state-of-the-art vision-only approaches, and accurately name those patterns. These patterns are then leveraged to build multimodal event schemas for each particular news event. We demonstrate that our method discovered patterns that greatly outperform other competing methods for this task. This work represents the first approach for using multimodal pattern mining to discover and name high-level semantically meaningful image patterns for event schema construction. The combination of our ability to find meaningful patterns and name them allows for many applications in high-level information extraction tasks, such as knowledge base population using multimodal documents and automatic event ontology creation. Our work can be leveraged as a bridge between structured information extraction tasks in the Computer Vision and Natural Language Processing communities.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-11-44155 and a multimedia seedling grant through the DARPA DEFT Program #FA8750-12-2-0347. 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.

7. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *20th International Conference on Very Large Data Bases*, pages 487–499, 1994.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.

[4] A. Cruz-Roa, J. C. Caicedo, and F. A. González. Visual pattern mining in histology image collections using bag of features. *Artificial intelligence in medicine*, 2011.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[6] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

[7] B. Fernando, E. Fromont, and T. Tuytelaars. Mining mid-level features for image classification. *International Journal of Computer Vision*, 108(3):186–203, 2014.

[8] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[9] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, 2000.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105. 2012.

[11] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[12] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Mid-level deep pattern mining. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 971–980, 2015.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[14] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Computing Research Repository (CoRR)*, abs/1310.4546, 2013.

[17] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.

[18] K. Rematas, B. Fernando, F. Dellaert, and T. Tuytelaars. Dataset fingerprints: Exploring image collections through data mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4867–4875, 2015.

[19] H. Schulz and S. Behnke. Learning object-class segmentation with convolutional neural networks. In *ESANN*, 2012.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository (CoRR)*, abs/1409.1556, 2014.

[21] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. *European Conference on Computer Vision (ECCV)*, pages 73–86, 2012.

[22] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[23] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 2013.

[24] C. Walker, S. Strassel, J. Medero, and K. Maeda. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 2006.

[25] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[26] J. Yuan and Y. Wu. Spatial random partition for common visual pattern discovery. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[27] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. 2014.

[28] H. Zhang, X. Shang, W. Yang, H. Xu, H. Luan, and T.-S. Chua. Online collaborative learning for open-vocabulary visual classifiers. In *CVPR*, 2016.

[29] H. Zhang, Y. Yang, H. Luan, S. Yang, and T.-S. Chua. Start from scratch: Towards automatically identifying, modeling, and naming visual attributes. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 187–196. ACM, 2014.

[30] T. Zhang, H. Li, H. Ji, and S.-F. Chang. Cross-document event coreference resolution based on cross-media features. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*, 2015.

[31] W. Zhang, H. Li, C.-W. Ngo, and S.-F. Chang. Scalable visual instance mining with threads of features. In *ACM International Conference on Multimedia*, pages 297–306, 2014.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[33] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.