Multimodal Pattern Mining from Twitter data for Image Concept Extraction

Brian Chen (陳柏穎), Chiou-Shann Fuh (傅楸善), Shi-Wei Wang (王熹偉)
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
brian271828@gmail.com, fuh@csie.ntu.edu.tw, swang200207@yahoo.com

Abstract—**In information retrieval problem, domain specific knowledge and related knowledge have been widely used. Most of the knowledge was generated by expert labeling. Based on the knowledge, we can perform various information extractions by neural net, which is very popular in Natural Language Processing. However, current system neglects information that can be used from images to build the structure data. Some previous works deal with extracting relations from images. In this paper, we try to generate a more high-level concept by considering both image caption pairs from Twitter data. We can comprehend the high-level concept of the image and classify into the certain category by convolution neural network. We use the image caption pair data to discover high-level concept such as "religion" and "disease". We will also name these patterns by our model. As a result, we can generate high-level concepts automatically, which can be incorporated with other knowledge for further usage.**

Keywords: Pattern Mining, Information Extraction, Convolutional Neural Network

I. INTRODUCTION

Nowadays computer vision researchers are able to demonstrate excellent performance in difficult image recognition and even surpass human vision. For instance, deep neural network models can easily classify animals shown in an image. With the great success of these works, we can perform image annotation, sometimes with bounding boxes easily. However, when we want to learn new concepts from pictures, we have to label the pictures manually. Moreover, we cannot extract high-level concepts like "religion" from an image. As a result, we can combine data mining techniques to find the pattern of religion which might usually contain "temple", "church", and "Monk". We also need to classify some misleading images such as castles, which may be very similar to churches. As a consequence, we need to automatically detect concepts from different domains.

In order to solve the problem mentioned above, our problem statement becomes: By given a set of images and its data corpora, how can we find the patterns and the high-level concepts in these images, and name these concepts in the images automatically? We use convolution neural network to capture the representation of image and use word embeddings to acquire the information from text. As a result, we can comprehend the image-caption pairs. In addition, we use association rules to discover patterns from the image-caption pairs to form high-level concepts. In this paper, we use Twitter data to show the performance of our proposed model. Our method can also be applied to other scenarios.

In many information retrieval tasks, we often use additional knowledge to improve our performance. Noted that generating additional knowledge has been a popular research topic. In this paper, we aim at generating high-level concepts shown in Fig. 1 instead of small objects e.g. "cats" and "dogs" to help information retrieval systems to acquire knowledge from different domains. We can extend the knowledge labeled by experts to construct new knowledge bases for further usage.

We can discover some patterns from image caption pairs. The concept behind the pattern should exist both in the image and the text. We also aim at discovering and naming these patterns automatically. The concept should be visually consistent and have relevant sematic between the images which have the same concept.

The evaluation part in our work is very challenging for two reasons. First, it will encounter the problem of novelty of the concept extracts. Second, it will encounter the problem of whether the quality of concept it captured is good or bad is very subjective. Unlike those in previous work, our final goal is not to enhance the performance of classifying these images more accurately. Rather, we focus on the procedure of finding and naming these patterns we extract, and also the quality and quantity of these concepts. In the experiment, we evaluate our method based on the similarity of our result with the ground truth, and the number of high-level concepts we can extract.

The main contributions of this work are as follows:

- We construct a multimodal for discovering patterns to form image-caption pairs for generating external knowledge. We can name these patterns automatically and make the procedure of generating additional knowledge be more convenient. In addition, the knowledge can be used in different information retrieval tasks.

- Our model is able to capture high-level concepts which are not only the items existing in the images such as identifying object as ImageNet. We can discover more meanings of the image at different domains.

- We consider both information from image and text to discover high-level concepts, which can comprehend the concept more precisely. The model can be applied to different scenarios that contains both image and text. Our model outperforms most of the existing models which only consider image or text.

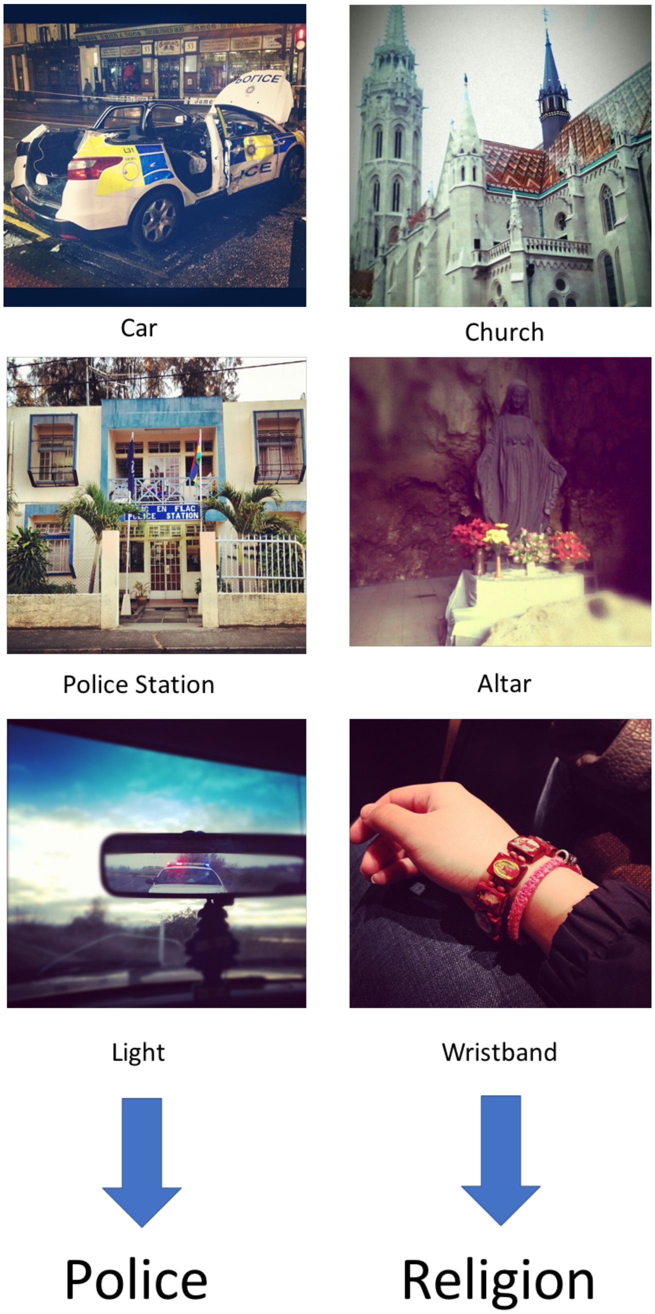


Fig. 1: example of concept generated.

II. Related work

Some of the related work of image retrieval and classification includes using images features such as Bag-of-words and SIFT [4]. However, previous studies have shown that using these low-level features is not enough for representing the concept of the image. As a result, using mid-level image features has become very popular in recent computer vision works. One of the most famous work is ImageNet [3], which performs the image classification and is widely used in several areas. It can accurately classify image by its massive training data. Hence, to reach the performance of their work is very time consuming. Moreover, the category they define is not always suitable for different tasks in the real world. It might encounter such scenarios that the image does not belong to any categories. Moreover, it is unable to capture high-level concepts since it focuses on object recognition. In our work, we aim at discovering the concept that might not belong to any categories predefined in ImageNet. We also work in a multi-modal approach and consider both image and its caption, in order to discover more high-level concept behind the image.

Pattern mining in visual areas is also explored to our work. Previous works such as frequent pattern mining [5] were used to mine visual patterns from images. Most of the work apply low-level features and combine with hashing methods. However, directly apply frequent pattern mining is neither efficient nor effective. The features it used are enormous, which require abundant computational time. Moreover, as we mentioned above, those low-level features are not representative for those concepts in the images. As a result, the performance of the pattern it discovers was not meaningful in most of the cases. With these kind of result, it is hard for human to comprehend or even apply to the knowledge bases.

Recently, artificial neural network has achieved great success in various areas. Convolutional neural networks (CNN) [6] in particular has the state of the art performance in image recognition. We can also use CNN to extract features in other tasks. In our problem, we combine CNN and frequent pattern mining for discovering high-level concepts in images. However, directly using CNN cannot guarantee we can capture the concept that is relevant to its captions. Since most images contain several different objects, we have to consider image features and text features jointly for the pattern mining.

Recently, with the great success of CNN, image captioning has become popular. People want to train the model to comprehend image-caption pairs. Most of the work aimed at generating captions from images automatically by combining with language models like Long Short-Term memory (LSTM) [7]. As a result, it can use this multimodal to generate a sentence describing the image by these two neural architectures: LSTM and CNN. However, all of the models mentioned above are by supervised learning which needs enormous data labeling such as MSCoco dataset [8] before the work. In contrast, our work aims at generating label data automatically without previous labeled data, which is unsupervised learning. We are able to mine the patterns from unlabeled data and name the patterns we discover for further knowledge base usage.

There are related works that also deal with image-caption pairs to extract concepts such as MMPM [16]. The work extracts event specific concepts from news related data. Our work discovers various types of concepts that not only focus on events e.g. religion and gifts. In consequence, our work covers more aspects of the concept in the image-caption pair data.

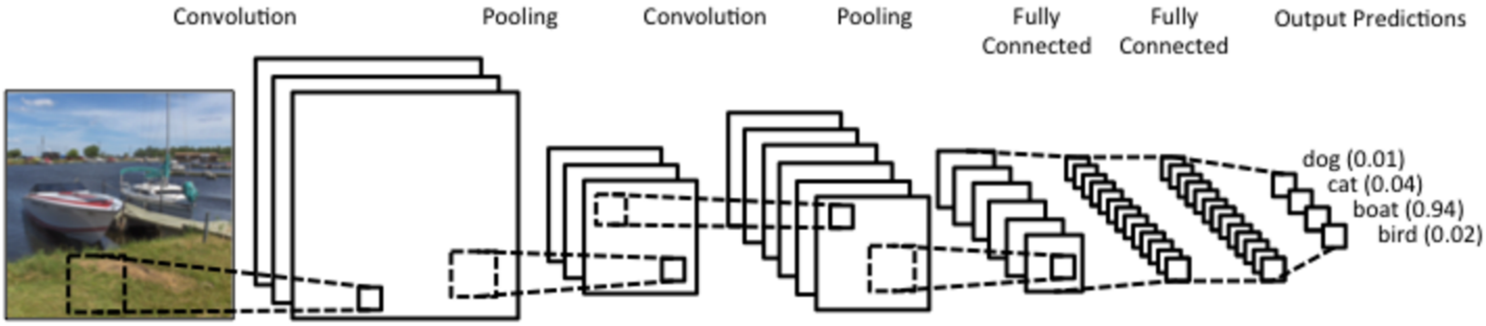


Fig. 2: Convolutional Neural Network

III. Methodology

We describe each component in our model, including feature extraction from images and text. We also explain the method of discovering patterns and generate high-level concepts.

1) *Pattern Mining*

We first explain our main goal: pattern mining. Pattern mining considers a set of transactions $S = \{T_1, T_2, \ldots, T_m\}$ for each T may include some observation $X = \{x_1, x_2, \ldots, x_n\}$, where $T \subseteq X$. Our goal is to discover a subset X, say t^*, which is able to predict the target element $y \in T_a$, given that $t^* \subset T_a$ and $y \cap t^* = \varnothing$, where t^* is called the frequent item set. In addition, the rule of $t^* \rightarrow y$ is called association rule. How often t^* appears in S is called the support of t^* which is defined as,

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

The final goal of our model is to discover the association rules from the image-caption pairs and find out some high-level concepts. As a result, we should find the patterns that t^* exists in a transaction which has high likelihood that y, which represents a concept category, also appears in the transaction. We set the confidence as the likelihood that if $t^* \subseteq T$ then $y \in T$.

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

Later on, we are going to explain how to generate those transactions for pattern mining, including image transactions and text transactions.

2) *Generating Image Transactions*

Throughout this paper, we will utilize the CNN (Convolutional Neural Network) Fig. 2. [12] ResNet (residual networks.) defined in [1], which is a common CNN structure that is often used for computer vision tasks. ResNet has a characteristic of it is easier to train deeper network than those methods before,

It reconstructs the layers to residual functions with reference to the layer inputs, replace of learning unreferenced functions. In addition, the ResNet is easier to optimize and enhance its accuracy by adding the depth of the network. Its special building block is shown if Fig. 3.

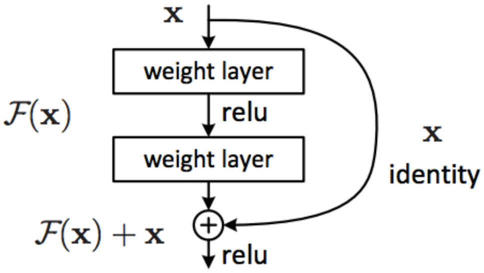
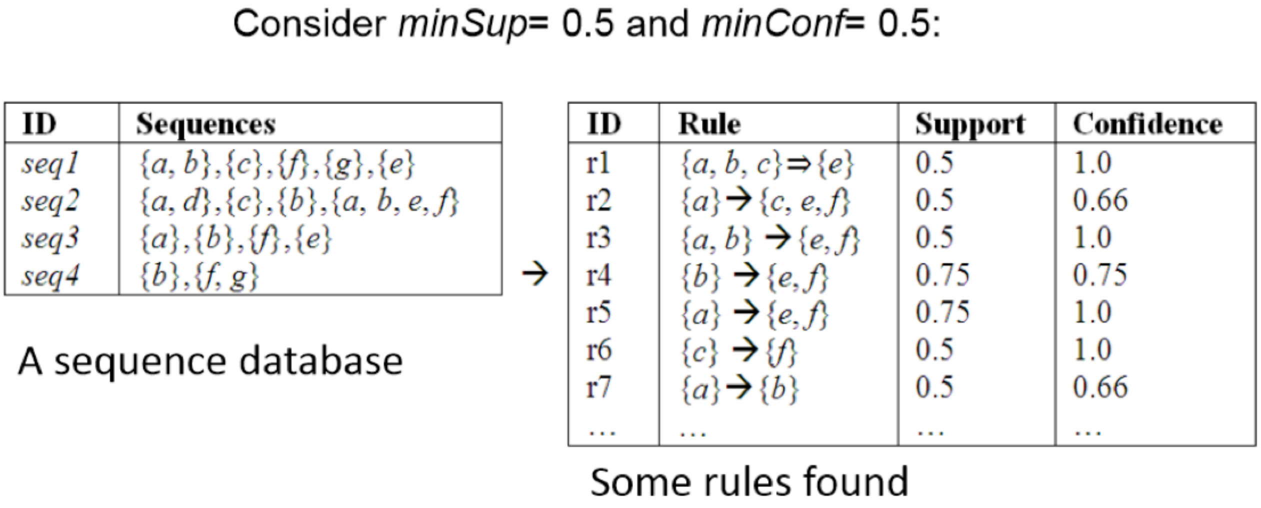

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 3: Residual learning: a building block.

We use the pre-trained CNN model from [3], trained on the ImageNet dataset for extracting the pool5 features for the news event images. We use these classification results to build transactions for each image patch as discussed in Sec. 3.1, where the predicted labels are the items in our transaction. We are able to efficiently extract image features that come from images that are suitable for pattern mining.

3) *Generating Captions Transactions*

Similar to the generation of image transactions mentioned above. We also need to generate image captions transactions for further pattern mining.



Consider *minSup*= 0.5 and *minConf*= 0.5:

ID	Sequences
seq1	$\{a, b\},\{c\},\{f\},\{g\},\{e\}$
seq2	$\{a, d\},\{c\},\{b\},\{a, b, e, f\}$
seq3	$\{a\},\{b\},\{f\},\{e\}$
seq4	$\{b\},\{f, g\}$

A sequence database

→

ID	Rule	Support	Confidence
r1	$\{a, b, c\}\Rightarrow\{e\}$	0.5	1.0
r2	$\{a\}\rightarrow\{c, e, f\}$	0.5	0.66
r3	$\{a, b\}\rightarrow\{e, f\}$	0.5	1.0
r4	$\{b\}\rightarrow\{e, f\}$	0.75	0.75
r5	$\{a\}\rightarrow\{e, f\}$	0.75	1.0
r6	$\{c\}\rightarrow\{f\}$	0.5	1.0
r7	$\{a\}\rightarrow\{b\}$	0.5	0.66
…	…	…	…

Some rules found

Fig. 5: Example of frequent pattern mining

We first remove the stop words from the captions by begin by NLTK[9] libraries and delete URLs by parsers .We tokenize the sentences by NLTK and set a threshold of 5 to collect words that appears more than the threshold in our data. In order to vectorize our data, we use pre-trained word embeddings Stanford's Glove[10], which represents a global word-word co-occurrence statistics from the huge corpus containing various domains, and gain the representation of each word in word vector space. The vector represents semantic meanings between words which shown in Fig. 4. At the end, we gain the representation of each word and use clustering algorithm K-means to cluster the words with similar semantic meanings.

After we gain the cluster of the word exist in image captions, we want to generate transactions that occurs in both text and image. We will match our transactions generated to the image transactions and eliminate those transactions that has to semantic similarity to any images. As a result, we can ensure our transaction is also talking about concepts in the images.

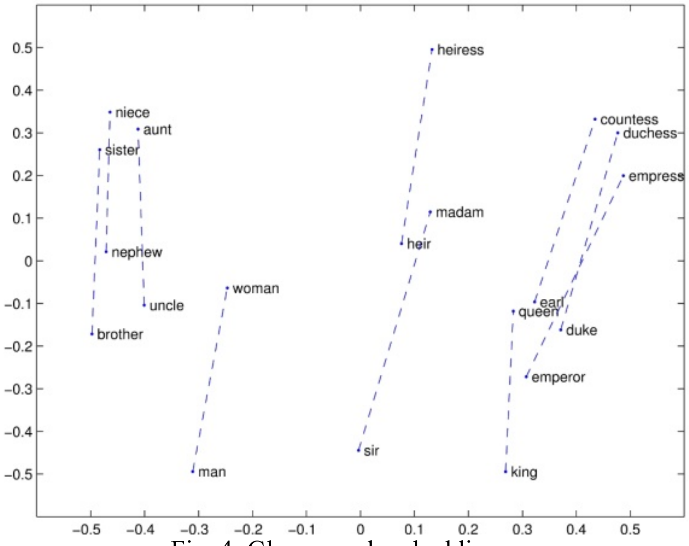


Fig. 4: Glove word embeddings

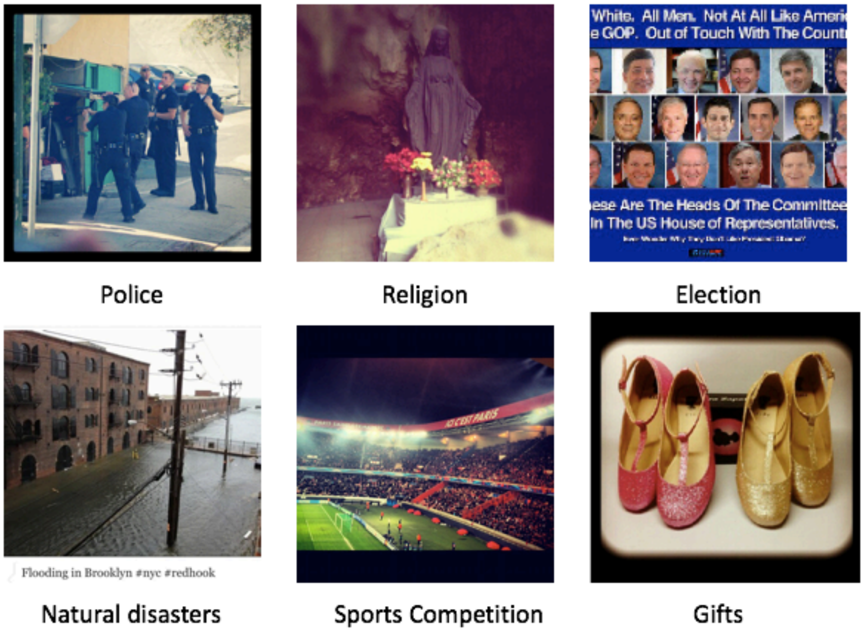
4) Mining the Patterns

As we mentioned in 3.1 about the pattern mining approach, we want to use the transactions from images and image captions to predict which concept behind the image caption pairs. We apply the apriori algorithm [11] to discover the frequent pattern exist in these transactions. We set the confidence of the association rule to be 0.6 and the support to be 10. We will also filter the results that deal with only either image object or image captions, to discover the patterns that consider both image and text. Fig. 5 [13] shows an example of frequent pattern mining.

5) Naming the Patterns

The last part of the model is naming the patterns automatically. We want to capture high level concepts and name the pattern for further knowledge base usage. From the previous section, we acquire the frequent pattern from images and image captions. We first remove words that are not suitable for naming concepts such as image object categories or person's name. After removing those words, we apply TF-IDF (Term Frequency - Inverse Document Frequency) encoding. We only consider unigrams or bigram that surpass the threshold we set in the caption dataset. We apply the TF-IDF representations to name those patterns.

In conclusion, the Stanford's Glove captures the sematic similarity of each words and TF-IDF picks the suitable words from cluster to name the pattern.


White. All Men. Not At All Like Ameri
e GOP. Out of Touch With The Countr
ese Are The Heads Of The Committee
In The US House of Representatives.
Flooding in Brooklyn #nyc #redhook

Police

Religion

Election

Natural disasters

Sports Competition

Gifts

Fig 6: Example concepts in our data set

IV. EXPERIMENTAL RESULTS

A. Data set

We use Twitter data [14] from DVMM Laboratory in Columbia University. The resulting dataset consists of 800 image-caption pairs with different topics. We also know the ground true of what concept the image delivers for us to evaluate. The high-level concepts can be seen in Table 1. Concept figures are shown in Fig. 6.

Table 1: The number of concepts in our data set.

Concept	count
Disease	69
Religion	46
Police	40
Election	36
Sports Competition	28
Natural disasters	20

B. Baseline Methods

In this paper, we compare with 2 simple baseline models which considers only image and considers only image captions to find the high-level concepts.

1) Baseline 1: Considers only image

As mentioned in generating image transactions, we use pre-trained CNN by ResNet to gain the representation of the images. In this baseline model, we simply use the classification results of the ResNet which shows the probability of the image belonging to different categories. We generate the transactions by these categories that ResNet predicts after classifying the images. We also use pre-trained word embeddings Glove to transform the category results into vectors, and perform K-means to cluster the words. After we get the transactions, we perform frequent pattern mining algorithm.

2) Baseline 2: Considers only image captions

Similar to the method mentioned above, we use pre-trained word embedding Globe to transform the image captions into vectors and perform clustering. In addition, apply frequent pattern mining algorithm after the transactions were generated.

C. Accuracy compare with flag Category accuracy

Since we have the ground truth of the concept of each image caption pairs, we can evaluate the unsupervised method by comparing the results by the clustering similarity. As a result, we compare them with the homogeneity score

$$H = \sum_{k} \sum_{i \in ck} \sum_{j \in ck, j \neq i} \| \mathrm{Xi} - \mathrm{Xj} \|_2$$

The measure HH reflects the pairwise similarity of the cluster elements. It is not based on the cluster mean, and therefore does not favor compact clusters. Table 2 shows different Homogeneity scores of different methods.

Table 2: Homogeneity score.

Methods	Homogeneity score
Baseline1(Image)	0.14
Baseline2(Text)	0.10
Our Method(Both)	0.22

We can find out that our method is more similar to the cluster of the concept ground truth. Moreover, we can find out that using only image to generate transactions performs better that using only text. Consequently, we can conclude that the image representation in our data set is more informative than image captions.

D. Compare number of concept generated

In this section, we evaluate the total number of patterns generated by different methods. Table 3 shows the number of concept each method generated.

Table 3: Number of concept generated.

Methods	Concept generated
Baseline1(Image)	32
Baseline2(Text)	20
Our Method(Both)	58

The result shows that our method generates most concepts among all.

E. Uniqueness of the concept

Since we have two different sources of data including image and text, we can evaluate the uniqueness of concepts and count how many concepts they have. We want to know how many patterns do each method generates and do not overlap with other methods. Table 4 shows the number of unique concepts generated by each method

Table 4: Number of unique concepts.

Methods	Unique Concepts
Baseline1(Image)	13
Baseline2(Text)	5
Our Method(Both)	20

We can see that image and text have their own unique concepts in their data, because not all of the image captions are about the picture. In addition, the image caption may have some different meanings beside the image shows, e.g. red balloons in the picture shown in Fig. 7. However, the picture is mainly about preventing AIDS. Hence, we have to consider the text to capture these kinds of hidden meaning behind the picture, since we can only detect balloons and people in the picture.

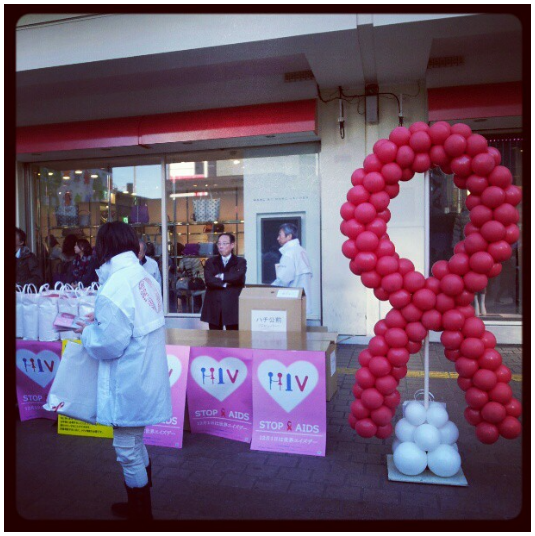


Fig. 7: Picture of AIDS prevention

F. Compare the number of concepts matching the ground truth.

In this section, we evaluate quality of the patterns generated by different method. We compare their results by the number of patterns generated in the previous section. We also match the ground truth to evaluate how many similar concepts generated by each method matched the ground truth concepts. Table 5 shows the number of concept each method generated are similar to the ground truth concept.

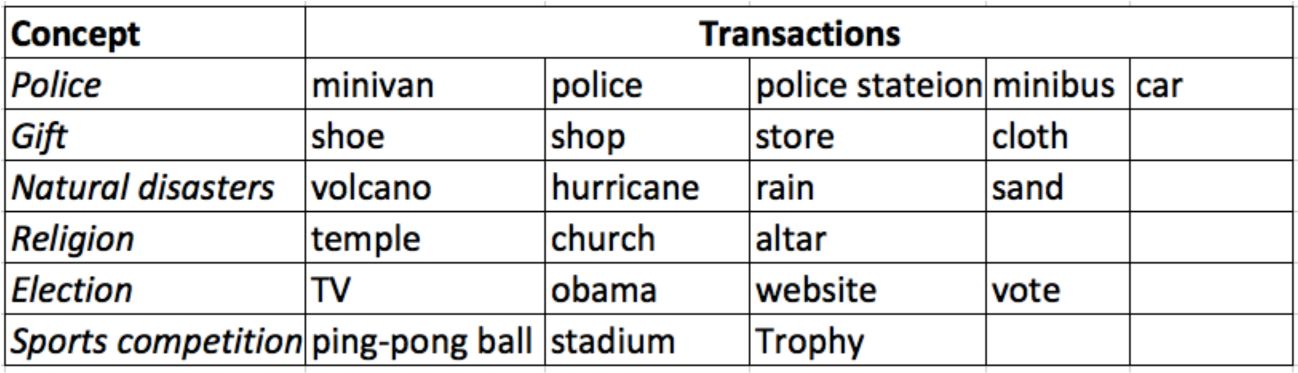
Table 5: Number of concept matched the ground truth.

Methods	Concept generated
Baseline1(Image)	11
Baseline2(Text)	7
Our Method(Both)	20

The result shows that our method generates more similar concepts to the ground truth than other baseline methods, meaning that the concept generated by our method is more likely to apply in the knowledge base.

G. Frequent Pattern generated by our methods

We show some examples of frequent patterns generated by our method and the concept it discovers. The results are shown in Fig. 8. We list the concept discovered and the transactions in the concepts. The transactions may include information from both image and text. In consequence, we utilize both information on gain more insight of the image-caption pair.



Concept	Transactions				
Police	minivan	police	police stateion	minibus	car
Gift	shoe	shop	store	cloth	
Natural disasters	volcano	hurricane	rain	sand	
Religion	temple	church	altar		
Election	TV	obama	website	vote	
Sports competition	ping-pong ball	stadium	Trophy		

Fig. 8: Example of frequent pattern mining

V. Discussion

In this section, we discuss some scenarios we encounter that causes our model to misunderstand the concepts or miss the concept. We will talk about them case by case.

A. *Image-caption is not informative*

In some data, the image-caption doesn't talk about the image. In consequence, we can only guess the concept based on image. For example, the image in Fig. 9 is about a police box, but it is not clear enough. Moreover, the author doesn't write anything about the police. As a result, we missed this concept.

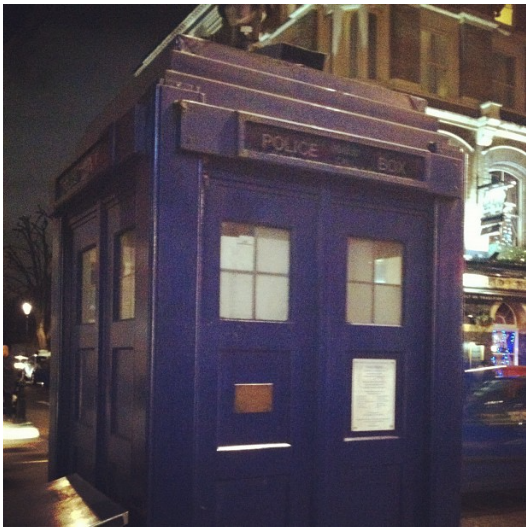


Fig. 9: Police Box

B. *Image-caption is misleading*

In this case, the image-caption talks about something that may be confusing. For example, the picture in Fig. 10 has an image-caption “Religion”. However, the “religion” that the author means is about its dressing style that he believes it is awesome, not the same as our concept definition of religion. Our definition of religion is the belief in a god or in a group of gods [15]. As a result, we have to avoid this kind of misleading captions in our model.

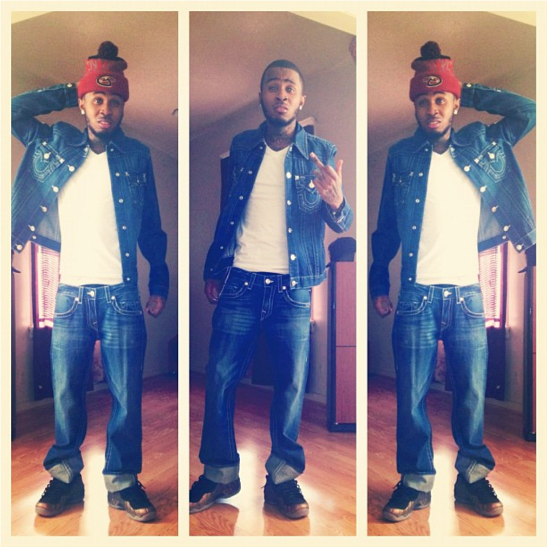


Fig. 10: Religion

VI. Conclusion and Future work

We have constructed a model for discovering meaningful concepts from image-caption pairs and name these patterns automatically. We consider both image and text data to capture more informative concepts from data, which can comprehend deeper insights behind the image. We also compare with other baseline methods. Our model generates more concepts than others and the concepts are more accurate. With the help of our model, we can apply these concepts for further usage in tasks which need additional knowledge bases.

For the future work, we aim at generating more precise concepts and understand the image-caption pair more clearly. We found out some sarcasm in our image-caption pairs which talks about total different concepts in the image and text. We can design new method to comprehend this kind of difficult situations by combining two neural architecture including

Convolution Neural Network for image and Recurrent Neural Network architecture Long Short-Term Memory for text to discover more knowledge in our data.

REFERENCE

[1] Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition" *arXiv preprint arXiv:1512.03385. 2015*

[2] Distributed representations of words and phrases and their compositionality. Computing Research Repository (CoRR), abs/1310.4546, 2013.

[3] Simon, Marcel and Rodner, Erik and Denzler, Joachim, "ImageNet pre-trained models with batch normalization" arXiv preprint arXiv:1612.01452, 2016

[4] 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.

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

[6] 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.

[7] Sepp Hochreiter. “Long Short-Term Memory”. Neural Comput. November 15, 1997.

[8] 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.

[9] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

[11] 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.

[12] http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

[13] http://data-mining.philippe-fournier-viger.com/introduction-frequent-pattern-mining/

[14] DVMM Data http://www.ee.columbia.edu/ln/dvmm/vso/download/twitter_dataset.html#download

[15] https://www.merriam-webster.com/dictionary/religion

[16] Li, H, Ellis Joseph G, Heng J, Chang S-F (2016), Event specific multimodal pattern mining for knowledge base construction. In: Proceedings of the 2016 ACM on multimedia conference, pp 821–830. ACM