Web-Based Semantic Fragment Discovery for Online Lingual-Visual Similarity

Xiaoshuai Sun,[1,2] **Jiewei Cao,**[1] **Chao Li,**[1] **Lei Zhu,**[1] **Heng Tao Shen**[1,3]
[1]The University of Queensland, Brisbane 4067, Australia. [2]Harbin Institute of Tehcnology, Heilongjiang 150001, China.
[3]University of Electronic Science and Technology of China, Chengdu 611731, China.
xiaoshuaisun.hit@gmail.com, {j.cao3, c.li1}@uq.edu.au, leizhu0608@gmail.com, shenht@itee.uq.edu.au

Abstract

In this paper, we present an automatic approach for on-line discovery of visual-lingual semantic fragments from weakly labeled Internet images. Instead of learning region-entity correspondences from well-labeled image-sentence pairs, our approach directly collects and enhances the weakly labeled visual contents from the Web and constructs an adaptive visual representation which automatically links generic lingual phrases to their related visual contents. To ensure reliable and efficient semantic discovery, we adopt non-parametric density estimation to re-rank the related visual instances and proposed a fast self-similarity-based quality assessment method to identify the high-quality semantic fragments. The discovered semantic fragments provide an adaptive joint representation for texts and images, based on which lingual-visual similarity can be defined for further co-analysis of heterogeneous multimedia data. Experimental results on semantic fragment quality assessment, sentence-based image retrieval, automatic multimedia insertion and ordering demonstrated the effectiveness of the proposed framework.The experiments show that the proposed methods can make effective use of the Web knowledge, and are able to generate competitive results compared to state-of-the-art approaches in various tasks.

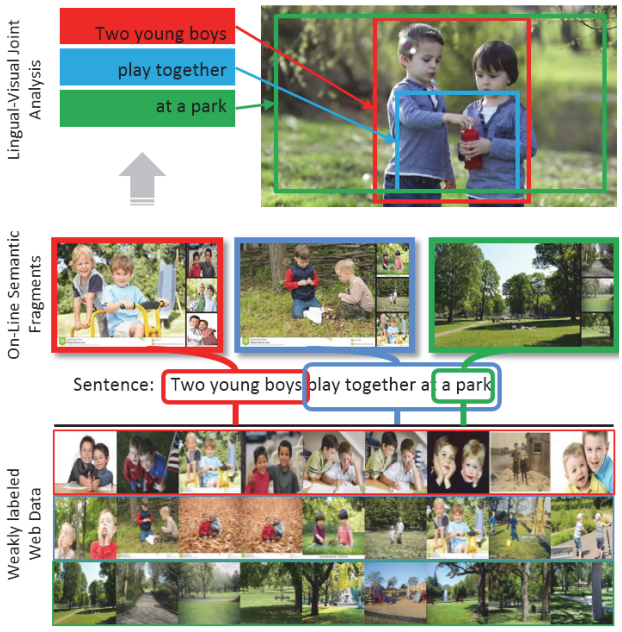

Figure 1: Concept Illustration. This work captures the meaning of a sentence based on a group of visual-linguistic semantic fragments. Unlike traditional methods that learn such latent representation from human labeled data, our method discovers such fragments directly from the Web which is highly efficient and free of human interactions.

Introduction

Every day, trillions of multimedia contents including texts, images and videos as well as the attached contextual data are generated and shared on the Internet, recording almost every aspect of human society, from the worldwide economic events to a person's personal feeling. For the last few decades, researchers from both academia and industry have developed a variety of information processing techniques trying to make computers understand the semantics in both texts and visual media. Despite the great progress made by the community, the visual intelligence of our computers is still at its early stage, which can only recognize simple lingual-visual links, and speak about shallow facts such as *what* and *where*. For example, image knowledge bases such as ImageNet(Deng et al. 2009) enable the possibility for the visualization of semantic entities defined as words or simple phrases. However, even as large as ImageNet, the current knowledge bases are still unable to characterize complex phrases like "the most dangerous dog in Asia", not to mention longer sentences and paragraphs. The huge gap between visual and lingual data still exists and becomes larger when the spatio-temporal scale goes up.

Recent research has pushed forward with a new challenge of bridging the semantic gap between textual and visual media in a long-term sentential level. Some pioneering models have been proposed to generate textual descriptions for images and videos (Farhadi et al. 2010; Yao et al. 2010; Rohrbach et al. 2013; Kulkarni et al. 2013; Socher et al. 2014; Donahue et al. 2015; Vinyals et al. 2015; Chen and Lawrence Zitnick 2015) or building image-semantic correspondences (Lin et al. 2014; Kong et al. 2014; Zhu et al. 2015; Karpathy and Fei-Fei 2015).These mod-

els can build discrete links between large textual fragments and visual contents by learning from human labeled data, which become the foundation for a variety of novel applications. However, the current models often brought in unnecessary restrictions on both the lingual complexity and the application domain. Specifically, most models were trained in a supervised manner based on human labeled sentence-image pairs, which naturally results in an inevitable scalability issue for real-world applications. Take Karpathy and Li's framework(Karpathy and Fei-Fei 2015) for instance, new sentences can only be processed if it contains lingual snippets pre-existed in the training set, which results in a restricted application capacity depending on the number of training samples and the scalability of the pre-trained model. Furthermore, most methods take snippets as the minimum computation unit, thus, cannot support fuzzy inputs which are quite common in real-world applications.

In this paper, we exploit an automatic Web-based semantic fragment discovery method to support on-line similarity computation between long lingual description (sentences) and visual contents (image or video). Inspired by (Chen, Shrivastava, and Gupta 2013; Chen et al. 2014; Singh et al. 2015), our key insight is to take Web image collections as a universal knowledge database, from which we extract and evaluate the links between the lingual fragments and their related visual contents. We aim to collect such high-quality links to form a reliable and dynamic joint representation and to further support deep lingual-visual analysis of multimedia contents. Unlike traditional methods that learn such joint representation from human labeled data, our method discovers the semantic fragments directly from the Web in a fully automatic manner. We make code, datasets and annotations publicly available on our project page. Our contribution consists of three aspects:

- We proposed a novel on-line methodology for generic lingual-visual joint analysis. Unlike traditional supervised methods that learn visual knowledge or priors from human labeled data, we propose to discover such knowledge directly from the Web without human interactions.
- We proposed an effective quality measure for the discovery of semantic fragments from the Web and provided an on-line similarity measure for long texts and visual contents. Compared with the training-based visual-semantic alignment methods, our approach has better scalability and achieves state-of-art performance without additional human interaction and domain restrictions.
- We created two new datasets for the research and evaluation of Web-based semantic discovery, and also explore potential applications such as automatic multimedia intersection and ordering enabled by the proposed method.

Related Work

Joint modeling of lingual and visual signals has been extensively investigated in computer vision community. The research effort has been devoted mostly to the problem of image captioning and text-image alignment, where explicit and latent correspondences between visual segmentations and text fragments are explored.

Image and Video Captioning

Early works (Farhadi et al. 2010; Jia, Salzmann, and Darrell 2011) learned fixed visual representations from a set of labeled data and translate them into short lingual descriptions. Later methods (Ordonez, Kulkarni, and Berg 2011; Hodosh, Young, and Hockenmaier 2013; Socher et al. 2014) followed a similar methodology which pose the captioning task as a retrieval problem, where the most compatible annotation in the training set is retrieved and transferred to the test image. Some other approaches generate captions by filling in fixed templates according the the visual contents presented in the given image (Yao et al. 2010; Gupta and Mannem 2012; Kulkarni et al. 2013). Recent models adopt recurrent neural networks (RNN) to learn joint image-text embedding for the generation of complex captions for both images (Karpathy, Joulin, and Li 2014; Kiros, Salakhutdinov, and Zemel 2014; Mao et al. 2014; Karpathy and Fei-Fei 2015) and short video clips (Venugopalan et al. 2014).

Text-Image Alignment

Compared to image captioning, our work is more related to text-image alignment models (Lin et al. 2014; Kong et al. 2014; Karpathy, Joulin, and Li 2014; Zhu et al. 2015; Karpathy and Fei-Fei 2015), since the focus of our work is on sentence understanding which is more like a reverse problem of image captioning. Our objective is different with traditional text-to-image alignment methods, which mainly focused on discovering the coherence between texts and visual contents in specified domains, e.g retrieving driving videos via complex textual queries(Lin et al. 2014), aligning books and movies(Zhu et al. 2015), or parsing indoor scenes using lingual descriptions (Kong et al. 2014). Most related to our work, (Karpathy and Fei-Fei 2015) proposed a visual-semantic alignment model which learns the inner-modal correspondences by combining convolutional neural networks over image regions and bidirectional recurrent neural networks over sentences. All the above-mentioned approaches are domain specific and training-based, which is the main obstacle against large-scale applications and commercialization.

Web Knowledge Discovery

Web data has been widely used to discover both textual and visual knowledge. (Berg, Berg, and Shih 2010) proposed a framework to automatically mine attributes from Internet images and their associated textual descriptions. (Chen, Shrivastava, and Gupta 2013) use web data to exploit common sense relationships and thus generate weak labels for images. Free from low-level features and pre-specified attributes, (Habibian, Mensink, and Snoek 2014) obtain textual descriptions of videos from the web and learn a multimedia embedding for few-example event recognition. (Singh et al. 2015) proposed an iterative framework based on Google Image Search and action centric part of speech model to discover descriptive concepts and predict complex events. The success of these methods inspires us to explore and utilize Web knowledge for lingual-visual joint analysis of multimedia data.

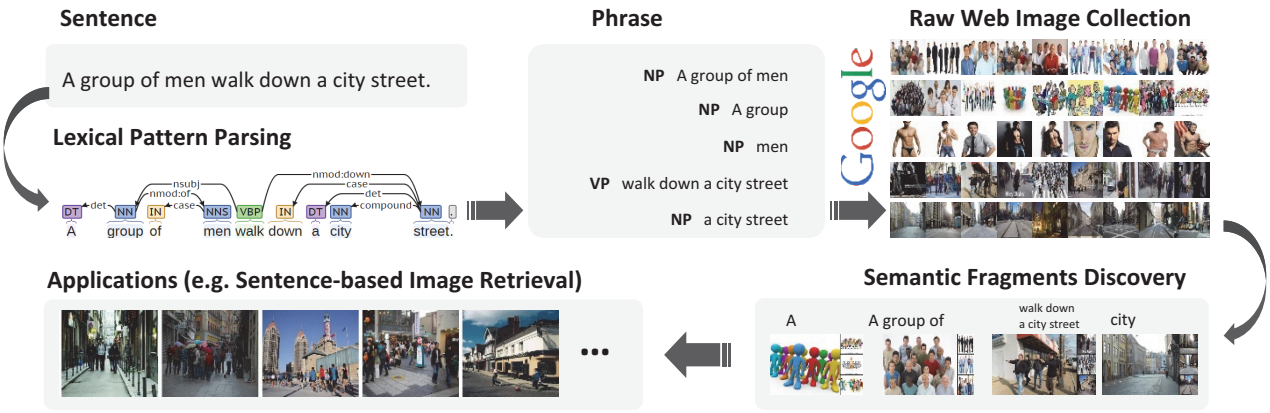

Figure 2: Overview of our approach. Given a sentence as input, we first parse the lexical patterns and achieve a set of linguistic fragments in the form of noun and verb phrases. Then, we retrieve Web images using search engine for every phrase to construct a raw fragment set. Based on the statistical analysis of the visual contents, we re-rank the images and filter out low quality fragments. The discovered semantic fragments can be used in many applications that require lingual-visual joint analysis.

The Model

Problem Definition

Given a sentence $\mathbf{L}$ and an image $\mathbf{F}$, our major objective is to construct an on-line function $S(\mathbf{L}, \mathbf{F})$, which measures the semantic similarity between the two heterogeneous item.

Framework Overview

An illustration of our method is presented in Figure 2. Given a sentence waiting to be processed, our first step is to discover the potential visualizable semantic fragments from the sentence. By parsing the lexical patterns of the sentence with pre-defined structural constraints, we are able to achieve a set of linguistic semantic fragments in the form of either single words or long phrases. We then propose a data-driven approach to examine the quality (visualizability) of every semantic fragment based on the statistical analysis of online-retrieved images from Google. We obtain visualizable semantic fragments by thresholding the quality score. Finally, we use the discovered semantic fragments to define fragment-based lingual-visual similarity.

Semantic Fragment Discovery

On-Line Semantic Fragment Parsing Following the proposal of Lin et al.(Lin et al. 2014), we use a similar graph representation to capture the semantic structure of a sentence. As illustrated in Figure 2, we first adopt Stanford Lexical Parser (Le 2013) to obtain the initial semantic graph from the sentence, and then extract noun and verb-phrases from the graph to further construct a candidate set of semantic fragments. The fragment candidates are then utilized as queries to retrieve related images from Google Image Search. For noun phrases, we directly use the phrase as the query. While for verb phrases, we search the semantic graph to find its subject noun phrase, and then combine them together as the final query.

Deep Visual Representation Our image representation is based on the network architecture of **VGG-16** (Simonyan and Zisserman 2014). The **VGG-16** network contains 16 weight layers, 13 of which are convolutional layers and the rest 3 are fully connected layers. We follow the notions in (Girshick et al. 2014; Simonyan and Zisserman 2014) to define different types of CNN features: fc_6 and fc_7 refer to the activation of the first and the second fully-connected layers; fc_6_relu and fc_7_relu denote the activations after Rectified Linear Unites of fc_6 and fc_7 respectively. Practically, we adopt MatConvNet Toolkit (Vedaldi and Lenc 2015) with pre-trained model of VGG-16 (Simonyan and Zisserman 2014) for image feature extraction. The activations of fc_6, fc_7, fc_6_relu and fc_7_relu are extracted and normalized using L2 norm. Despite the type of the activation, each image is represented as a 4096-D normalized CNN vector in the rest of our paper.

Discovering High Quality Semantic Fragments The semantic fragments, represented as lingual phrases and their corresponding ranked images, form a small customized image knowledge base for the given textual inputs. Some semantic fragments contain images with high semantical relevance, while some others consist of images with inconsistent or even irrelevant visual contents. This is reasonable since the images only carry weak labels given by Google and such label becomes even weaker when the query phrase gets longer. Thus, we should filter out those low quality fragments with labels that are not strong enough to reflect the visual semantics of the corresponding phrase. Our intuition for the estimation of fragment quality is to compare the visual appearance of images within each fragment candidate set. Higher self-similarity means better within set consistency and thus indicates better label accuracy.

Specifically, for a given fragment, we first extract visual features from the top-N attached images and then compute the self-similarity to measure the overall quality of the fragment. Finally, we use a fixed threshold to filter out the low-

quality candidates and get the true visualizable semantic fragment set. Given a fragment candidate g represented as a group of CNN features $\mathbf{F}_g = \{f_1^g, f_2^g, ..., f_N^g\}$, we measure the quality of g by the self-similarity of $\mathbf{F}_g$:

$$Q(g) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} f_i^{g\mathrm{T}} f_j^g, \quad (1)$$

where the value of $Q(.)$ is ranging from 0 to 1 (1 is the best). We adopt a straight-forward thresholding approach to filter out low quality fragment candidates.

Density-based Data Reranking. We can use the original Google rank to organize the images of each semantic fragment. However, since we are focusing on the visual semantics, the ranking order can be further improved using unsupervised Parzen density estimation (Mezuman and Weiss 2012) over the CNN feature space. To this end, a typical Parzen density estimater can be formulized as:

$$\hat{f}_{pdf}^g(f) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(f - f_i^g), \quad (2)$$

where $\{f_i^g\}_{i=1}^N$ are image features of fragment g and $K_\sigma(x) = e^{-\|x\|^2/2\sigma^2}$ is a Gaussian kernel. To boost the speed, we construct a new discrete kernel K_Q based on $Q(.)$ to replace K_σ:

$$K_Q(x) = \begin{cases} 1 & \text{if } x > Q(g) \\ 0 & \text{else} \end{cases} \quad (3)$$

Intuitively, we use the estimated probability density to rerank images and thus improve the visualization quality.

Lingual-Visual Similarity

Based on semantic fragment discovery, we can represent a sentence by a group of visual semantic fragment $\mathbf{L} = \{g_1, g_2, ..., g_n\}$. As discussed in (Karpathy and Fei-Fei 2015), the visual semantics can appear at any location of a related image, thus using global matching would not be accurate enough to pull out fragment-level similarity. Consequently, we follow the proposal of (Karpathy and Fei-Fei 2015), and parse the input images into semantic regions to allow precise fragment-level similarly estimation. To avoid unnecessary restriction, we utilize Objectness(Alexe, Deselaers, and Ferrari 2012) to get object proposals. Objectness is designed for generic object detection which is based on pure bottom-up cues (e.g. saliency) and thus has no domain restrictions. Practically, we use the top 19 object proposals in addition to the whole image to get a CNN feature set $\mathbf{F}_t = \{f_1^t, f_2^t, ..., f_{20}^t\}$, which forms a similar fragment representation to the query sentence. The similarity between an input image t and the sentence $\mathbf{L}$ can then be formulized as:

$$S(\mathbf{L}, \mathbf{F}_t) = \sum_{i=1}^{n} \max_{f_j^t, f_k^{g_i}} f_j^{t\mathrm{T}} f_k^{g_i}, f_j^t \in \mathbf{F}_t \text{ and } f_k^{g_i} \in \mathbf{F}_{g_i}. \quad (4)$$

Eqn. 4 uses max pooling over all image regions and all instances within the fragment to compute the similarity. Here,

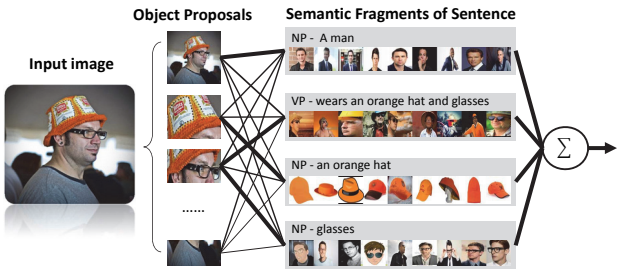

Figure 3: Illustration of Sentence-Image similarity computation. We use max pooling over all object proposals and all instances within the fragment to compute fragment-level similarity. The final output is a straight summation of all fragment similarity.

we introduce an alternative reformulation which achieved better results while costing much less computations:

$$S'(\mathbf{L}, \mathbf{F}_t) = \sum_{i=1}^{n} \max_{f_j^t \in \mathbf{F}_t} f_j^{t\mathrm{T}} f_r^{g_i}, \quad (5)$$

where $f_r^{g_i}$ is the representative vector of g_i. In the experiments, we compared three definitions of $f_r^{g_i}$, including:

$$f_r^{g_i} = \begin{cases} f_1^{g_i} & \textbf{T} \text{ - Rank Top 1} \\ \frac{1}{N}\sum_{m=1}^{N} f_m^{g_i} & \textbf{G} \text{ - Geometric Center} \\ \frac{1}{N}\sum_{m=1}^{N} \hat{f}_{pdf}^{g_i}(f_m^{g_i}) \cdot f_m^{g_i} & \textbf{P} \text{ - Probabilistic Center} \end{cases} \quad (6)$$

An illustration of sentence-image similarity computation is presented in Figure 3. It clearly shows how image and sentence are connected via the discovered semantic fragments.

The Experiments

We first present the quality assessment results of our self-similarity metric with different visual features on labeled semantic fragment dataset. Then, we test our lingual-visual similarity in sentence-based image retrieval task. We also show qualitative results of some other applications, such as automatic multimedia insertion and ordering.

Datasets

We release two new datasets as extensions of Flickr30K (Young et al. 2014) to enable research and comparisons on Web-based unsupervised sentence understanding tasks.

Flickr30K-Phrase This dataset consists of 3.2 million images with 32,486 weak phrase labels. More specifically, We parsed 5K sentences from the Flickr30K test image set into 32,486 noun and verb phrases, each attached by 100 urls retrieved from Google Image Search. The dataset is quite challenging, it contains large amount of noises since no human interaction is involved. The dataset can be regarded as a fragment-level Web reference set for Flick30K's 5K testing captions, based on which new Web-driven retrieval and captioning methods can be developed and evaluated.

Flickr30K-Quality We sample 20K images with 1K phrase labels from Flickr30K-Phrase, and ask in-house annotators to score the quality of each phrase. In each round of annotation, one phrase are presented with all the related images, the score of the phrase are defined according to the semantical correctness and visual appearance consistency of its attached images, including 1(High), 0.5(Normal), and 0(Low) respectively. Each phrase are labeled at least by three annotators and the final score is obtained by averaging all annotated scores. The dataset can be used for the training and evaluation of both semantic discovery and label quality assessment methods.

Semantic Fragment Discovery Evaluation

We evaluate our semantic fragment discovery method by taking it as a binary classier for the classification of high quality and low quality phrase fragments. Specifically, We compute the area under ROC curve to measure the overall performance. Table 1 shows the results of our method on Flickr30K-Quality dataset using GIST (Oliva and Torralba 2001) and 4 kinds of CNN features. The results demonstrate the effectiveness of the self-similarity measure. Our method with fc_7 features achieved an AUC score of 0.8537, which means we are able to filter out most of the low quality fragments by a simple thresholding operation.

Feature	GIST	fc_6	fc_6_relu	fc_7	fc_7_relu
AUC	0.7376	0.8512	0.8512	**0.8537**	0.8525

Table 1: Evaluation results of semantic fragment quality assessment. We show AUC scores of the self-similarity metric with GIST and 4 CNN features.

Sentence-based Image Retrieval Evaluation

We adopt the 1K test images in Flickr30K for quantitative evaluation. We take the 5K sentence annotations as queries, and use Eqn. 4 and Eqn. 5 as our similarity measure to sort the test images. We set the size of each fragment $N = 20$, and fix the self-similarity threshold to 0.2 in all the tests. Recall@K (Socher et al. 2014) is introduced as the main evaluation metric, which measures the fraction of times a correct item was found among the top K results.

We compared our method with several state-of-art approaches. Note that our method runs without training, while others are pre-trained based on the Flickr30K training data. As shown in Table 2, our full model achieved competitive performance with **DeFrag** (Karpathy, Joulin, and Li 2014), and it also succeeded in beating some of the supervised models like **DeViSE** (Frome et al. 2013) and **SDT-RNN** (Socher et al. 2014). Note that, among all the tested methods, our approach is the only one that can run on-line. With such property, the model is able to be deployed to real-world search engines with no additional restrictions on either application domain or human supervision.

We show the influences of parameters including the type of features (Table 3) and the number of object proposals (Table 4). The fc_7_relu CNN layer leads to a relatively better result and the full model performance stops increasing when Top-20 object proposal were used.

Model	**R@1**	**R@5**	**R@10**
Supervised Model			
DeViSE	6.7	21.9	32.7
SDT-RNN	8.9	29.8	41.1
DeFrag	**10.2**	**30.8**	**44.2**
Our Label-Free Model			
Ours (**T**+ 1 OP)	7.3	19.1	26.7
Ours (**P**+ 1 OP)	10.3	24.8	33.5
Ours (**G**+ 1 OP)	10.3	25.0	33.7
Ours (**T** + 20 OP)	7.4	19.8	28.3
Ours (**P** + 20 OP)	10.4	27.6	37.1
Ours (**G** + 20 OP)	10.5	**27.8**	37.4
Ours (Eqn. 4 + 20 OP)	**10.7**	27.4	**37.7**

Table 2: Flickr30K experiments. **R@K** is Recall@K (high is good). **T**, **G**, **P** denote the option in Eqn. 5. **OP** means object proposal. Compared to the state-of-art models, our model can generate competitive results without pre-training on human labels.

CNN layer	fc_6	fc_6_relu	fc_7	fc_7_relu
Recall@1	7.8	9.2	9.2	**10.5**
Recall@5	20.7	24.5	25.4	**27.8**
Recall@10	28.4	33.6	33.9	**37.4**

Table 3: Influence of CNN features on Flickr30K. All features are normalized by L2 norm.

Figure 4 presents some typical retrieval results of our model. We notice that our approach performs well on images with a relatively small number of salient, well-defined objects. The performance of our method can be further improved when 1) the media data on Web gets denser or 2) the weak phrase labels get stronger. The first condition increases the probability to retrieve high consistent data, while the second improves the inferencing confidence.

Extensions and Limitations

Based on our on-line lingual-visual similarity, we can fulfill joint analysis of weakly-correlated texts and visual contents. Given a text document and a few weakly-correlated photos (e.g. travel blog and the related photos), the goal of automatic multimedia insertion is to insert the photos (videos) into proper locations of the text document according to their visual semantics. Figure 5 shows an example of automatic multimedia insertion. Photos are inserted at the locations where the visual-lingual similarity are maximized. We can also use the same algorithm to recover the order of the photos based on the corresponding textual descriptions.

The proposed visual-lingual similarity can also be used to transform textual descriptions into multimedia data flow which contains not only texts but also images and videos. Such kind of media flow can further be used to generate personal albums, video stories, or documentary videos.

Although our method generates promising results, the proposed framework still suffers from serval limitations.

Figure 4: Detailed retrieval results on Flickr30K dataset. In each case, we show the ground truth image (left) and the top 5 images (right) retrieved using our lingual-visual similarity. As shown from the results, our method is quite effective for accurate descriptions but lack of robustness against unimportant/unrecognizable phrases. For example, in one of the failure case, the query mentioned about horse but the ground truth actually doesn' t contain any recognizable horse in the image.

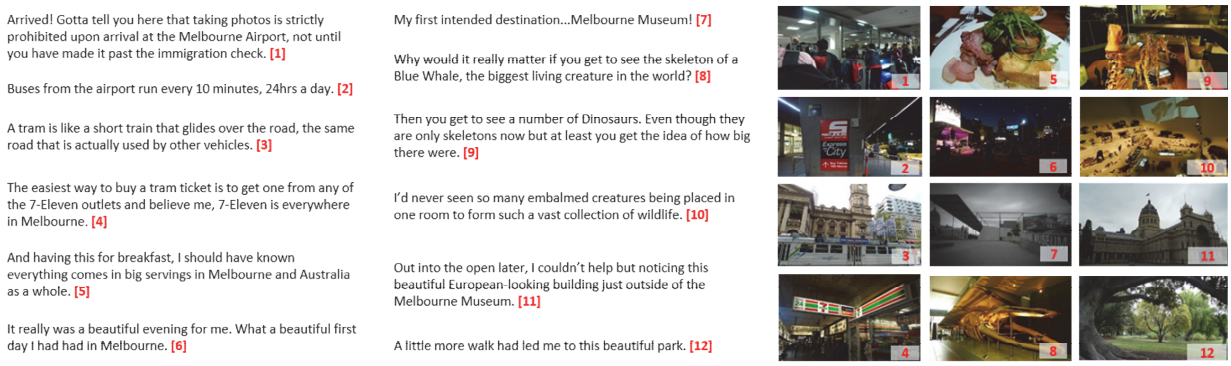

Figure 5: An example of automatic multimedia insertion and ordering. On the left is a travel blog without visual illustrations. By using our on-line lingual-visual similarity, the travel photos on the right can be automatically inserted into proper locations of the travel blog (indicated by red numbers). Similarly, the travel photos can also be ordered according the textual descriptions of the blog, instead of using their original time stamp.

Box Number	Top 1	Top 5	Top 10	Top 20
Recall@1	9.0	10.3	**10.7**	10.5
Recall@5	22.5	26.5	27.1	**27.8**
Recall@10	31.0	35.3	36.7	**37.4**

Table 4: Influence of the number of object proposals (bounding boxes) on Flickr30K. We use fc_7_relu as the feature. Object bounding boxes are ranked according to their objectness (Alexe, Deselaers, and Ferrari 2012).

First, the discovery module of our method is effective in identifying most of the low quality phrases, yet it also wrongly filtered out many high quality fragments because of the gap between feature-based similarity and the true semantic similarity. Second, the current method is not able to identify ambiguous phrases, which probably leads to poor retrieval performance and confused visualizations. Lastly, the proposed method must be supported or implemented within the architecture of a search engine to ensure acceptable processing speed for real-world on-line services.

Conclusions

We introduced a fully automatic framework for semantic fragment discovery and on-line lingual-visual similarity measurement without domain restrictions and pre-training requirement. Experimental results demonstrate the promising potential of our method on Flick30K and our new datasets. In our future work, object co-segmentation and contextual constraints within texts will be introduced to overcome the limitations on both fragment quality and ambiguity. Although far from perfect, the proposed framework shows good potential for real-world applications since its intelligence is originated from the Web, which is not restricted by the scale limitations of the human labels.

Acknowledgement

This research was partially supported by the Discovery Project of Australian Research Council, DP130103252, FT120100718 and FT130101530.

References

Alexe, B.; Deselaers, T.; and Ferrari, V. 2012. Measuring the objectness of image windows. *TPAMI* 34(11):2189–2202.

Berg, T. L.; Berg, A. C.; and Shih, J. 2010. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010*. Springer. 663–676.

Chen, X., and Lawrence Zitnick, C. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *IEEE CVPR'15*.

Chen, J.; Cui, Y.; Ye, G.; Liu, D.; and Chang, S.-F. 2014. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR'14*,

Chen, X.; Shrivastava, A.; and Gupta, A. 2013. Neil: Extracting visual knowledge from web data. In *IEEE ICCV'13*, 1409–1416.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR'09*, 248–255.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE CVPR'15*.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*. Springer. 15–29.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *NIPS'13*, 2121–2129.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR'14*, 580–587.

Gupta, A., and Mannem, P. 2012. From image annotation to image description. In *NIPS'12*, 196–204. Springer.

Habibian, A.; Mensink, T.; and Snoek, C. G. 2014. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*, 17–26.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR,* 853–899.

Jia, Y.; Salzmann, M.; and Darrell, T. 2011. Learning cross-modality similarity for multinomial data. In *IEEE ICCV'11*, 2407–2414.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE CVPR'15*, 3128–3137.

Karpathy, A.; Joulin, A.; and Li, F. F. F. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS'14*, 1889–1897.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Kong, C.; Lin, D.; Bansal, M.; Urtasun, R.; and Fidler, S. 2014. What are you talking about? text-to-image coreference. In *IEEE CVPR'14*, 3558–3565.

Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE TPAMI* 35(12):2891–2903.

Le, Q. V. 2013. Building high-level features using large scale unsupervised learning. In *IEEE ICASSP'13*, 8595–8598.

Lin, D.; Fidler, S.; Kong, C.; and Urtasun, R. 2014. Visual semantic search: Retrieving videos via complex textual queries. In *IEEE CVPR'14*, 2657–2664.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; and Yuille, A. L. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Mezuman, E., and Weiss, Y. 2012. Learning about canonical views from internet image collections. *NIPS'12*, 719–727.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Iournal of Computer Vision* 42(3):145–175.

Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2text: Describing images using 1 million captioned photographs. In *NIPS'11*, 1143–1151.

Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; and Schiele, B. 2013. Translating video content to natural language descriptions. In *IEEE ICCV'13*, 433–440.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, B.; Han, X.; Wu, Z.; Morariu, V. I.; and Davis, L. S. 2015. Selecting relevant web trained concepts for automated event retrieval. *IEEE ICCV'15*.

Socher, R.; Karpathy, A.; Le, Q. V.; Manning, C. D.; and Ng, A. Y. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2:207–218.

Vedaldi, A., and Lenc, K. 2015. Matconvnet-convolutional neural networks for matlab.

Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; and Saenko, K. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *IEEE CVPR'15*.

Yao, B. Z.; Yang, X.; Lin, L.; Lee, M. W.; and Zhu, S.-C. 2010. I2t: Image parsing to text description. *Proceedings of the IEEE* 98(8):1485–1508.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.

Zhu, Y.; Kiros, R.; Zemel, R.; salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *IEEE ICCV'15*.