Going Beyond Text: A Hybrid Image-Text Approach for Measuring Word Relatedness

Chee Wee Leong and Rada Mihalcea
Department of Computer Science and Engineering
University of North Texas
cheeweeleong@my.unt.edu, rada@cs.unt.edu

Abstract

Traditional approaches to semantic relatedness are often restricted to text-based methods, which typically disregard other multimodal knowledge sources. In this paper, we propose a novel image-based metric to estimate the relatedness of words, and demonstrate the promise of this method through comparative evaluations on three standard datasets. We also show that a hybrid image-text approach can lead to improvements in word relatedness, confirming the applicability of visual cues as a possible orthogonal information source.

1 Introduction

Measuring the semantic relatedness of words is an important task with applications in information extraction and retrieval, query reformulation, word sense disambiguation, plagiarism detection and textual entailment. Owing mainly to the nature of this task, research efforts in the past have typically centered around methodologies employing the use of knowledge-based or corpus-based textual resources, with only little (if any) work paying attention to evidence provided by other multimodal sources, such as visual cues presented by the images that are associated with a given word. While it can be shown that the human cognitive system is sensitive to visual information, and incorporating a dual linguistic-and-pictorial representation of information can actually enhance knowledge acquisition (Potter and Faulconer, 1975), the use of visual information to improve tasks in natural language processing has been largely unexplored.

In this paper, we hypothesize that the relatedness between the visual representations of a pair of words can be effectively used to gauge their similarity. We first discuss a technique widely used in computer vision termed as "bag of visual words" to show how distinctive features of an image can be harvested. We next introduce the main resource, ImageNet, used in our work to bridge the semantic gap between words and images. Finally, we show how a new relatedness metric based exclusively on visual information can be constructed for the semantic relatedness task. We evaluate this metric alongside existing corpus-based (Turney and Pantel, 2010) and knowledge-based metrics (Pedersen et al., 2004) either in a standalone or combined setting and present our findings.

2 Bag of Visual Words

Inspired by the bag-of-words approach employed in information retrieval, the "bag of visual codewords" is a similar technique used mainly for scene classification (Yang et al., 2007). Starting with an image collection, visual features are first extracted as data points from each image. By projecting data points from all the images into a common space and grouping them into a large number of clusters such that similar data points are assigned to the same cluster, we can treat each cluster as a "visual codeword" and express every image in the collection as a "bag of visual codewords." This representation enables the application of methods used in text retrieval to tasks in image processing and computer vision.

Typically, the type of visual features selected can be *global* – suitable for representing an entire image, or *local* – specific to a given region in the image, depending on task requirement. For a global representation, features are often described using a continuous feature space, such as a color histogram in three different color spaces (RGB, HSV and LAB), or textures using Gabor and Haar wavelets (Makadia et al., 2008). Likewise, local descriptors such as key points (Fei-Fei and Perona, 2005) can also adopt such a representation. Regardless of the features used, visual codeword generation involves the following three important phases.

1. Feature Detection: The image is divided into partitions of varying degrees of granularity from which features can be extracted and represented. We can employ normalized cuts to divide an image into irregular regions, or apply uniform seg-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1403–1407,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

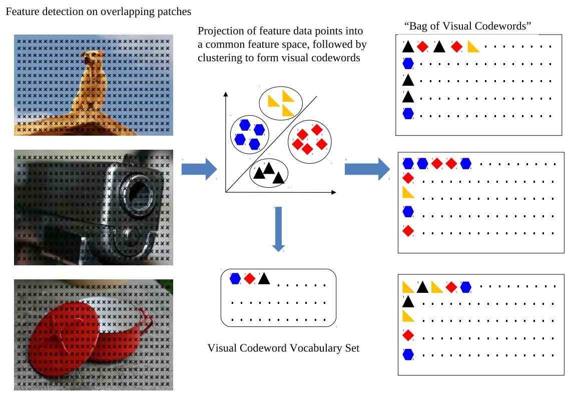
Feature detection on overlapping patches
Projection of feature data points into a common feature space, followed by clustering to form visual codewords
“Bag of Visual Codewords”
Visual Codeword Vocabulary Set

Figure 1: An illustration of the process of generating “Bag of Visual Codewords”

mentation to break it into smaller but fixed grids, or simply locate information-rich local patches on the image using interest point detectors.

2. Feature Description: A descriptor is selected to represent the features extracted from the image. Typically, feature descriptors are represented as numerical vectors, with each vector describing the feature extracted in each region. This way, an image is represented by a set of vectors from its constituent regions.

3. Visual Codeword Generation: Clustering methods are applied to group vectors into clusters, where the center of each cluster is defined as a visual codeword, and the entire set of clusters defines the visual vocabulary for that image collection. Each image region or patch abstracted in feature detection is now represented by the codeword mapped from its corresponding feature vector.

The process of visual codeword generation is illustrated in Figure 1. Fei-Fei and Perona (2005) have shown that, unlike most previous work on object or scene classification that focused on adopting global features, local regions are in fact extremely powerful cues. In our work, we use the Scale-Invariant Feature Transform (SIFT) introduced by Lowe (2004) to describe distinctive local features of an image in the feature description phase. SIFT descriptors are selected for their invariance to image scale, rotation, differences in 3D viewpoints, addition of noise, and change in illumination. They are also robust across affine distortions.

3 ImageNet

Given the maturity of techniques used to extract visual content from images, it is possible to study the synergistic relationships between semantic representations of words and images given the availability of a large lexical resource with associated relevant images. For such a resource, we turn to the ImageNet[1] database (Deng et al., 2009), which is a large-scale ontology of images developed for advancing content-based image search algorithms, and serving as a benchmarking standard for various image processing and computer vision tasks. ImageNet exploits the hierarchical structure of WordNet by attaching relevant images to each synonym set (known as “synset”), hence providing pictorial illustrations of the concept associated with the synset. On average, each synset contains 500-1000 images that are carefully audited through a stringent quality control mechanism. Compared to other image databases with keyword annotations, we believe that ImageNet is suitable for evaluating our hypothesis for two important reasons. First, by leveraging on reliable semantic annotations in WordNet (i.e., words in the synset), we can effectively circumvent the propagation of errors caused by unreliable annotations, and consequently hope to reach more conclusive results for this study. Second, unlike other image databases, ImageNet consists of millions of images, and it is a growing resource with more images added on a regular basis. This aligns with our long-term goal of extending our image-based similarity metric to cover more words in the lexicon. Figure 2 shows an example of a synset and the corresponding images in ImageNet.

Figure 2: A subset of images associated with a node in ImageNet. The WordNet synset illustrated here is {*Dog, domestic dog, Canis familiaris*}

4 Datasets

To evaluate the effectiveness of our image-based model for measuring word-to-word relatedness, we selected three datasets widely used in the past:

[1] http://image-net.org/. ImageNet currently hosts 12,184,113 images in 17624 synsets, each of which is classified under a high level category such as animal, fish, plant, structure etc

Rubenstein and Goodenough (RG65) consists of 65 word pairs ranging from synonymy pairs (e.g., car - automobile) to completely unrelated terms (e.g., noon - string). The 65 noun pairs were annotated by 51 human subjects. All the nouns pairs are non-technical words scored using a scale from 0 (not-related) to 4 (perfect synonymy).

Miller-Charles (MC30) is a subset of the Rubenstein and Goodenough dataset, consisting of 30 word pairs, whose relatedness was rated by 38 human subjects, using a scale from 0 to 4.

WordSimilarity-353 (WS353), also known as Finkelstein-353, consists of 353 word pairs annotated by 13 human experts, on a scale from 0 (unrelated) to 10 (very closely related). The dataset also includes proper names and technical terms, therefore posing an additional degree of difficulty for any relatedness metric.

5 Experiments

In our experiments, we seek answers to the following questions. First, what is the effectiveness of our image-based method in measuring word-to-word relatedness, as compared to existing text-based methods? Second, can our image-based method complement these text-based methods via a combination of their outputs ?

Note that as ImageNet is still a resource under development, not all word pairs in the datasets presented in section 4 are covered. To level the playing field, in our experiments we only select those pairs of words of which both words would appear as surface forms in the synsets of ImageNet with validated images. Moreover, due to coverage issues, an anomaly exists in situations such as $monk - slave$, where both words may appear in single-candidate synsets, i.e., {monk,monastic} and {slave ant} respectively, but are represented using fundamentally different images (person vs animal). To prevent this, we further constrain the selection of word pairs of which at least a pair of candidate synsets each representing a word in the pair belong to the same high level category. Note that both selection steps are performed automatically, and thus the identification of the word pairs that can be used in conjunction with the image-based approach can be effectively applied to any dataset, regardless of size. Our trimmed dataset consists of 10 word-pairs from the Miller-Charles dataset (**MC10**), 18 word-pairs from the Rubenstein-Goodenough dataset (**RG18**) and 56 word-pairs from the Word Similarity dataset (**WS56**).

For each word in a pair, we randomly select 50 images from the validated image pool of its associated synset[2], and extract all the visual codewords from all such images, using the technique explained in section 2. Each image is first preprocessed to have a maximum side length of 300 pixels. Next, SIFT gray-scale descriptors are obtained by densely sampling the image on 20x20 overlapping patches spaced 10 pixels apart using a publicly available image-processing toolkit.[3] K-means clustering is applied on a random subset of 10 million SIFT descriptors to derive a visual vocabulary of 1,000 codewords. Each descriptor is then quantized into a visual codeword by assigning it to the nearest cluster. As such, each image J can now be expressed as a vector $< tf_i.w_i >$, where i=1:1000 and tf_i is the frequency of occurrence of visual codeword w_i in image J. For each synset, we sum the vectors of all 50 images and normalize each w_i by its total frequency in the synset.

Image Metric: Given a word pair w_i and w_j, let $S_i = \{v_k^i\}$ and $S_j = \{v_m^j\}$ be their set of candidate visual vectors respectively. Then, computing the semantic relatedness of two words amounts to finding the maximum visual relatedness between all the possible pairings of synsets representing both words, using the cosine similarity between the visual vectors of the synsets, given below. The dimensionality of the vector, n, is set to 1000, which is the size of the visual codeword vocabulary.

$$Sim_{img}(w_i, w_j) =$$

$$\max_{v_k \in S_i, v_m \in S_j} \frac{\sum_{p=1}^{n} v_k^p v_m^p}{\sqrt{\sum_{p=1}^{n} (v_k^p)^2} \sqrt{\sum_{p=1}^{n} (v_m^p)^2}}$$

Text Metric: For a comparative study, we evaluate several knowledge-based methods, including Roget and WordNet Edges (Jarmasz, 2003), H&S (Hirst and St-Onge, 1998), L&C (Leacock and Chodorow, 1998), J&C (Jiang and Conrath, 1997), LIN (Lin, 1998), RES (Resnik, 1995), and two corpus-based methods Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) and Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007).

Combined Metric: In the combined setting, we attempt to integrate the output of our image-based metric with that of existing text-based metrics in a pairwise manner via two combination functions, which were previously noted for

[2] Note that a word may appear as surface forms across multiple synsets. In such cases, we randomly sample 50 images from each of the synsets

[3] http://www.image-net.org/challenges/LSVRC/2011

	Text-based measures								Image
	WNE	**H&S**	**J&C**	**L&C**	**LIN**	**RES**	**LSA**	**ESA**	**metric**
MC10									
STANDALONE	0.846	0.883	0.685	0.846	0.685	0.328	0.867	0.515	0.851
SUM	**0.879**	**0.927**	**0.830**	**0.855**	0.806	0.842	0.915	**0.842**	
F1	0.855	0.855	0.806	**0.855**	**0.842**	**0.891**	**0.927**	0.782	
RG18									
STANDALONE	0.867	0.775	0.828	0.867	0.820	0.580	0.546	0.611	0.820
SUM	0.887	**0.826**	**0.867**	0.887	0.863	**0.813**	**0.728**	**0.827**	
F1	**0.893**	0.796	0.833	**0.907**	**0.869**	0.793	0.607	0.627	
WS56									
STANDALONE	**0.482**	0.453	0.454	0.515	0.496	0.469	0.520	0.453	0.404
SUM	0.457	0.474	0.471	0.507	**0.523**	0.524	0.538	0.440	
F1	0.453	**0.583**	**0.513**	**0.546**	0.520	**0.570**	**0.588**	**0.475**	

Table 1: Results obtained with individual knowledge-based and corpus-based text-based measures, with our image measure, and with two combination functions (SUM and F1). The bold correlation numbers represents the highest among all metrics per text-based measure per dataset.

their effectiveness in Information Retrieval systems (Fox and Shaw, 1994). Specifically, we combine the text-based and image-based metrics by summing their relatedness figures (SUM) and by calculating their F-measure (F1) defined as the harmonic mean of the two input metrics. Because the similarity scores are differently distributed across various methods, we apply a normalization step within each metric to assert the same lower and upper-bound prior to the combination: $Score_{norm} = (Score_{original} - Score_{min})/(Score_{max} - Score_{min})$.

For each dataset and metric, we obtain the Spearman rank correlation of the automatically generated similarity scores with the ground-truths by human subjects.

6 Discussion

The results in Table 1 show that our image-based method can be an effective metric on its own, scoring a competitive Spearman correlation of 0.851 on the MC10 dataset, and 0.820 on the RG18 dataset. Perhaps not surprisingly, these two datasets consists mainly of words such as car, $forest$, $bird$, $furnace$, which are *picturable*, concrete entities that possess distinctive and unambiguous visual representations. Its performance, however, degrades on the WS56 dataset with a somewhat low correlation rating of 0.404, possibly due to the presence of more broadly defined words lacking a visual identity (e.g., $equipment$ in the word pair $phone{-}equipment$),

Regardless of the performance of the individual image-based metric, the hybrid image-text approach improves over the standalone text-based metric in almost all cases, and this holds for both knowledge-based and corpus-based methods. These results are encouraging, as they suggest that image-based approaches can be effectively used to improve even basic tasks in natural language processing such as word relatedness.

While we are aware that the limited coverage of ImageNet restricts the applicability of this hybrid image-text method to word relatedness, the continued growth of this resource should provide alleviation. Future work will also consider a comparison of multi-way combinations between knowledge-based, corpus-based and image-based metrics for further advancement of the state-of-the-art.

7 Related Work

Recently, some attention has been given to modelling synergistic relationships between the semantics of words and images (Leong and Mihalcea, 2011; Bruni et al., 2011). The research that is most closely related to ours is the work of (Feng and Lapata, 2010), where it has been shown that it is possible to combine visual representations of word meanings into a joint bimodal representation constructed by using probabilistic generative latent topic models. Unlike our approach, however, (Feng and Lapata, 2010) relied on a news corpus where images and words in a document are assumed to be generated by a set of latent topics, rather than a lexical resource such as ImageNet. While they provided a proof-of-concept that using the visual modality leads to an improvement over their purely text-based model (an increase of Spearman correlation of 0.071 on a subset of WordSim353 dataset), no attempt has been made to evaluate the image-based models independently, or to combine image models with previously proposed knowledge-based and corpus-based measures of relatedness.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. *Proceedings of the EMNLP Geometrical Models for Natural Language Semantics Workshop*.

Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Li Fei-Fei and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Proceedings of the Annual Conference of the North American Chapter of the ACL*.

Edward A. Fox and Joseph A. Shaw. 1994. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *International Joint Conferences on Artificial Intelligence*.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press.

Mario Jarmasz. 2003. Rogets thesaurus as a lexical resource for natural language processing. In *Ph.D. Dissertation*, Ottawa-Carleton Institute for Computer Science, School of Information Technology and Engineering, University of Ottawa.

Jay J. Jiang and David A. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonom. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*.

Thomas Landauer and Susan Dumais. 1997. A solution to platos problem: The latent semantic analysis theory of acquisition. In *Psychological Review*, volume 104, pages 211–240.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In *The MIT Press*, pages 265–283.

Chee Wee Leong and Rada Mihalcea. 2011. Measuring the semantic relatedness between words and images. In *Proceedings of International Conference on Computational Semantics*.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*.

Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. 2008. A new baseline for image annotation. In *Proceedings of European Conference on Computer Vision*.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, pages 1024–1025.

Mary C. Potter and Babara A. Faulconer. 1975. Time to understand pictures and words. In *Nature*, volume 253, pages 437–438.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. In *Journal of Artificial Intelligence Research*, volume 37, pages 141–188.

Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. In *ACM Multimedia Information Retrieval Workshop*.