

Measuring the semantic relatedness between words and images

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

Measures of similarity have traditionally focused on computing the semantic relatedness between pairs of words and texts. In this paper, we construct an evaluation framework to quantify cross-modal semantic relationships that exist between arbitrary pairs of words and images. We study the effectiveness of a corpus-based approach to automatically derive the semantic relatedness between words and images, and perform empirical evaluations by measuring its correlation with human annotators.

1 Introduction

Traditionally, a large body of research in natural language processing has focused on formalizing word meanings. Several resources developed to date (e.g., WordNet (Miller, 1995)) have enabled a systematic encoding of the semantics of words and exemplify their usage in different linguistic frameworks. As a result of this formalization, computing semantic relatedness between words has been possible and has been used in applications such as information extraction and retrieval, query reformulation, word sense disambiguation, plagiarism detection and textual entailment.

In contrast, while research has 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 *meaning* of an image in isolation is not well-defined and it is mostly task-specific. A given image, for instance, may be simultaneously labeled by a set of words using an automatic image annotation algorithm, or classified under a different set of semantic tags in the image classification task, or simply draw its meaning from a few representative regions following image segmentation performed in an object localization framework.

Given that word meanings can be acquired and disambiguated using dictionaries, we can perhaps express the meaning of an image in terms of the words that can be suitably used to describe it. Specifically, we are interested to bridge the *semantic gap* (Smeulders et al., 2000) between words and images by exploring ways to harvest the information extracted from visual data in a general framework. While a large body of work has focused on measuring the semantic similarity of words (e.g., (Miller and Charles, 1998)), or the similarity between images based on image content (e.g., (Goldberger et al., 2003)), very few researchers have considered the measure of semantic relatedness¹ between words and images.

But, how exactly is an image related to a given word? In reality, quantification of such a cross-modal semantic relation is impossible without supplying it with a proper definition. Our work seeks to address this challenge by constructing a standard evaluation framework to derive a semantic relatedness metric for arbitrary pairs of words and images. In our work, we explore methods to build a representation model consisting of a joint semantic space of images and words by combining techniques widely adopted in computer vision and natural language processing, and we evaluate the hypothesis that we can automatically derive a semantic relatedness score using this joint semantic space.

Importantly, we acknowledge that it is significantly harder to decode the semantics of an image, as its interpretation relies on a subjective and perceptual understanding of its visual components (Biederman,

¹In our paper, we are concerned with semantic *relatedness*, which is a more general concept than semantic *similarity*. Similarity is concerned with entities related by virtues of their likeness, e.g., *bank-trust company*, but dissimilar entities may also be related, e.g., *hot-cold*. A full treatment of the topic can be found in Budanitsky and Hirst (2005).

1987). Despite this challenge, we believe this is a worthy research direction, as many important problems can benefit from the association of image content in relation to word meanings, such as automatic image annotation, image retrieval and classification (e.g., (Leong et al., 2010)) as well as tasks in the domains of text-to-image synthesis, image harvesting and augmentative and alternative communication.

2 Related Work

Despite the large amount of work in computing semantic relatedness between words or similarity between images, there are only a few studies in the literature that associate the meaning of words and pictures in a joint semantic space. The work most similar to ours was done by Westerveld (2000), who employed LSA to combine textual words with simple visual features extracted from news images using colors and textures. Although it was concluded that such a joint textual-visual representation model was promising for image retrieval, no intensive evaluation was performed on datasets on a large scale, or datasets other than the news domain. Similarly, Hare et al. (2008) compared different methods such as LSA and probabilistic LSA to construct joint semantic spaces in order to study their effects on automatic image annotation and semantic image retrieval, but their evaluation was restricted exclusively to the Corel dataset, which is somewhat idealistic and not reflective of the challenges presented by real-world, noisy images.

Another related line of work by Barnard and Forsyth (2001) used a generative hierarchical model to learn the associative semantics of words and images for improving information retrieval tasks. Their approach was supervised and evaluated again only on the Corel dataset.

More recently, Feng and Lapata (2010) showed that it is possible to combine visual representations of word meanings into a joint bimodal representation constructed by using latent topics. While their work focused on unifying meanings from visual and textual data via supervised techniques, no effort was made to compare the semantic relatedness between arbitrary pairs of word and image.

3 Bag of Visual Codewords

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, characterizing its appearance. 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 representation in all images, or *local* – specific to a given image type and task requirement. Global features are often described using a continuous feature space, such as color histogram in three different color spaces (RGB, HSV and LAB), or textures using Gabor and Haar wavelets (Makadia et al., 2008). In comparison, local features such as key points (Fei-Fei and Perona, 2005) are often distinct across different objects or scenes. 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. Typically, we can employ normalized cuts to divide an image into irregular regions, or apply uniform segmentation 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 that are being extracted from the image. Typically, feature descriptors (global or local) 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.

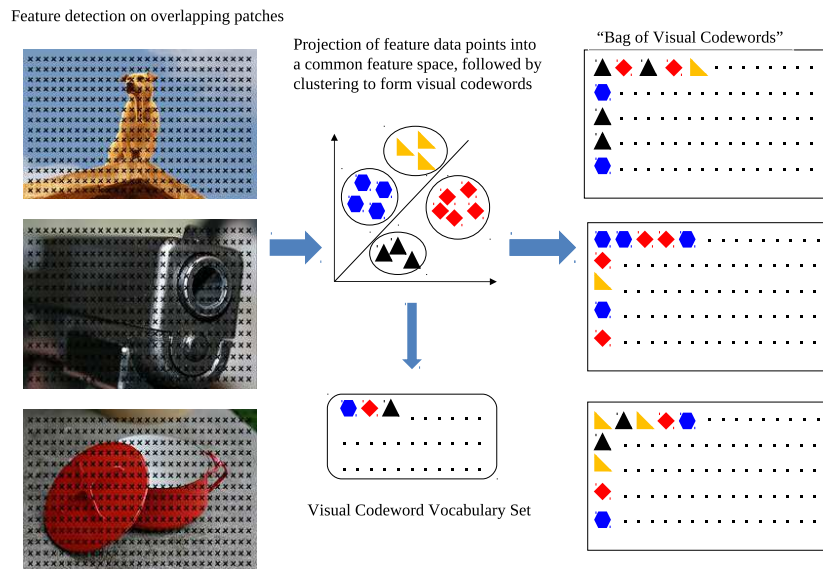


Figure 1: An illustration of the process of generating “Bag of Visual Codewords”

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 collection of clusters defines the visual vocabulary for that image collection. Each image region or patch abstracted in feature detection is now represented by the visual codeword mapped from its corresponding feature vector.

The process of visual codeword generation is illustrated in Figure 1. Fei-Fei and Perona (2005) has shown that, unlike most previous work on object or scene classification that focused on adopting global features, local features 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.

4 Semantic Vector Models

The underlying idea behind semantic vector models is that concepts can be represented as points in a mathematical space, and this representation is learned from a collection of documents such that concepts related in their meanings are near to one another in that space. In the past, semantic vector models have been widely adopted by natural language processing researchers for tasks ranging from information retrieval and lexical acquisition, to word sense disambiguation and document segmentation. Several variants have been proposed, including the original vector space model (Salton et al., 1997) and the Latent Semantic Analysis (Landauer and Dumais, 1997). Generally, vector models are attractive because they can be constructed using unsupervised methods of distributional corpus analysis and assume little language-specific requirements as long as texts can be reliably tokenized. Furthermore, various studies (Kanerva, 1998) have shown that by using collaborative, distributive memory units to represent semantic vectors, a closer correspondence to human cognition can be achieved.

While vector-space models typically require nontrivial algebraic machinery, reducing dimensions is often key to uncover the hidden (latent) features of the terms distribution in the corpus, and to circumvent the sparseness issue. There are a number of methods that have been developed to reduce dimensions – see e.g., Widdows and Ferraro (2008) for an overview. Here, we briefly describe one commonly used

technique, namely the Latent Semantic Analysis (LSA), noted for its effectiveness in previous works for reducing dimensions.

In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a **Singular Value Decomposition (SVD)** on the term-by-document matrix \mathbf{T} representing the corpus. SVD is a well-known operation in linear algebra, which can be applied to any rectangular matrix in order to find correlations among its rows and columns. SVD decomposes the term-by-document matrix \mathbf{T} into three matrices $\mathbf{T} = \mathbf{U}\Sigma_k\mathbf{V}^T$ where Σ_k is the diagonal $k \times k$ matrix containing the singular k values of \mathbf{T} , $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k$ and \mathbf{U} and \mathbf{V} are column-orthogonal matrices. When the three matrices are multiplied together the original term-by-document matrix is re-composed. Typically we can choose $k' \ll k$ obtaining the approximation $\mathbf{T} \simeq \mathbf{U}\Sigma_{k'}\mathbf{V}^T$.

5 Semantic Relatedness between Words and Images

Although the bag of visual codewords has been extensively used in image classification and retrieval tasks, and vector-space models are well explored in natural language processing, there has been little connection between the two streams of research. Specifically, to our knowledge, there is no research work that combines the two techniques to model multimodal meaning relatedness. Since we are exploring new grounds, it is important to clarify what we mean by computing the semantic relatedness between a word and an image, and how the nature of this task impacts our hypothesis. The assumptions below are necessary to validate our findings:

1. Computing semantic relatedness between a word and an image involves comparing the concepts invoked by the word and the salient objects in the image as well as their interaction. This goes beyond simply identifying the presence or absence of specific objects indicated by a given word. For instance, we expect a degree of relatedness between an image showing a soccer ball and the word “jersey,” since both invoke concepts like {sports, soccer, teamwork} and so on.
2. The semantics of an image is dependent on the focus, size and position of distinct objects identified through image segmentation. During labeling, we expect this segmentation to be performed implicitly by the annotators. Although it is possible to focus one’s attention on specific objects via bounding boxes, we are interested to harvest the meaning of an image using a holistic approach.
3. In the case of measuring the relatedness of a word that has multiple senses with a given image, humans are naturally inclined to choose the sense that provides the highest relatedness inside the pair. For example, an image of a river bank expectedly calls upon the “river bank” sense of the word “bank” (and not “financial bank” or other alternative word senses).
4. A degree of semantic relatedness can exist between any arbitrary word and image, on a scale ranging from being totally unrelated to perfectly synonymous with each other. This is trivially true, as the same property holds when measuring similarity between words and texts.

Next, we evaluate our hypothesis that we can measure the relatedness between a word and an image empirically, using a parallel corpus of words and images as our dataset.

5.1 ImageNet

We use the ImageNet 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 three reasons. First, by leveraging on reliable keyword annotations in WordNet (i.e., words in the synset and their gloss naturally serve as annotations for the corresponding images), 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 building a large-scale joint semantic space of images and words. Finally, third, although we can search for relevant images using keywords in ImageNet,² there is currently no method to query it in the reverse direction. Given a test image, we must search through millions of images in the database to find the most similar image and its corresponding synset. A joint semantic model can hopefully augment this shortcoming by allowing queries to be made in both directions. Figure 2 shows an example of a synset and the corresponding images in ImageNet.

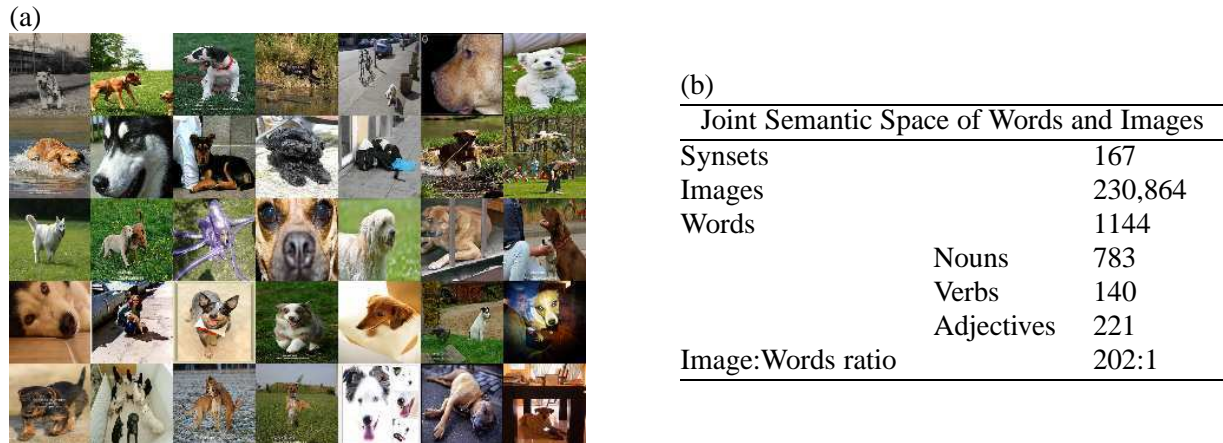


Figure 2: (a) A subset of images associated with a node in ImageNet. The WordNet synset illustrated here is $\{Dog, domestic\ dog, Canis\ familiaris\}$ with the gloss: *A member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; “the dog barked all night”* (b) A table showing statistical information on our joint semantic space model

5.2 Dataset

For our experiments, we randomly select 167 synsets³ from ImageNet, covering a wide range of concepts such as plants, mammals, fish, tools, vehicles etc. We perform a simple pre-processing step using Tree Tagger (Schmid, 1994) and extract only the nouns. Multiwords are explicitly recognized as collocations or named entities in the synset. Not considering part-of-speech distinctions, the vocabulary for synset words is 352. The vocabulary for gloss words is 777. The shared vocabulary between them is 251.

There are a total of 230,864 images associated with the 167 synsets, with an average of 1383 images per synset. We randomly select an image for each synset, thus obtaining a set of 167 test images in total. The technique explained in Section 3 is used to generate visual codewords for each image in this dataset.⁴ Each image is first pre-processed to have a maximum side length of 300 pixels. Next, SIFT descriptors are obtained by densely sampling the image on 20x20 overlapping patches spaced 10 pixels apart. 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.

To create the gold-standard relatedness annotation, for each test image, six nouns are randomly selected from its associated synset and gloss words, and six other nouns are again randomly selected from the shared vocabulary words.⁵ In all, we have $167 \times 12 = 2004$ word-image pairs as our test dataset. Similar to previous word similarity evaluations (Miller and Charles, 1998), we ask human annotators to rate each pair on a scale of 0 to 10 to indicate their degree of semantic relatedness using the evaluation framework outlined below, with 0 being totally unrelated and 10 being perfectly synonymous with each other. To ensure quality ratings, for each word-image pair we used 15 annotators from Amazon Mechanical

²<http://www.image-net.org/>

³Not all synsets in ImageNet are annotated with images. We obtain our dataset from the Spring 2010 version of ImageNet built around Wordnet 3.0.

⁴For our experiments, we obtained the visual codewords computed a priori from ImageNet. Test images are not used to construct the model

⁵12 data points are generally considered sufficient for reliable correlation measures (Vania Kovic, p.c.).




		
Synset {sunflower, helianthus}	Synset {oxygen-mask}	Synset {submarine , pigboat , sub , U-boat}
Gloss any plant of the genus <i>Helianthus</i> having large flower heads with dark disk florets and showy yellow rays	Gloss a breathing device that is placed over the mouth and nose; supplies oxygen from an attached storage tank	Gloss a submersible warship usually armed with torpedoes
Relatedness Scores color (5.13) dog (0.53) florete (6.53) flower (9.67) freshwater (2.40) hair (1.00) garden (6.60) head (3.80) plant (8.47) ray (3.67) sunflower (9.80) reed (2.27)	Relatedness Scores basketball (0.20) central (1.53) device (5.47) family (0.80) iron-tree (0.47) mouth (5.13) oxygen-mask (7.73) tank (4.47) storage (3.07) supply (5.20) nose (6.20) time (1.13)	Relatedness Scores africa (0.80) brass (1.73) door (1.67) good (2.40) pacific (2.40) pigboat (6.47) sub (8.20) submarine (9.67) tail (0.93) torpedo (7.60) u-boat (7.47) warship (8.73)

Table 1: A sample of test images with their synset words and glosses : The number in parenthesis represents the numerical association of the word with the image (0-10). Human annotations reveal different degree of semantic relatedness between the image and words in the synset or gloss.

Turk.⁶ Finally, the average of all 15 annotations for each word-image pair is taken as its gold-standard relatedness score⁷. Note that only the pairs of images and words are provided to the annotators, and not their synsets and gloss definitions.

The set of standard criteria underlying the cross-modal similarity evaluation framework shown here is inspired by the semantic relations defined in Wordnet. These criteria were provided to the human annotators, to help them decide whether a word and an image are related to each other.

1. **Instance of itself:** Does the image contain an entity that is represented by the word itself (e.g. an image of “Obama” vs the word “Obama”) ?
2. **Member-of Relation:** Does the image contain an entity that is a member of the class suggested by the word or vice versa (e.g. an image of an “apple” vs the word “fruits”) ?
3. **Part-of Relation:** Does the image contain an entity that is a part of a larger entity represented by the word or vice versa (e.g. an image of a “tree” vs the word “forest”) ?
4. **Semantically Related:** Do both the word and the image suggest concepts that are related (e.g. an image of troops at war vs the word “peace”) ?
5. **Semantically Close:** Do both the word and the image suggest concepts that are not only related but also close in meaning? (e.g. an image of troops at war vs the word “gun”) ?

Criterion (1) basically tests for synonym relation. Criteria (2) and (3) are modeled after the hyponym-hypernym and meronym-holonym relations in WordNet, which are prevalent among nouns. Note that none of the criteria is preemptive over the others. Rather, we provide these criteria as guidelines in a *subjective* evaluation framework, similar to the word semantic similarity task in Miller and Charles (1998). Importantly, criterion (4) models dissimilar but related concepts, or any other relation that indicates frequent association, while criterion (5) serves to provide additional distinction for pairs of words and images on a higher level of relatedness toward similarity. In Table 1, we show sample images from our test dataset, along with the annotations provided by the human annotators.

⁶We only allowed annotators with an approval rating of 97% or higher. Here, we expect some variance in the degree of relatedness between the candidate words and images, hence annotations marked with all 10s or 0s are discarded due to lack of distinctions in similarity relatedness

⁷Annotation guidelines and dataset can be downloaded at <http://lit.csci.unt.edu/index.php/Downloads>

5.3 Experiments

Following Erk and McCarthy (2009), who argued that word meanings are graded over their senses, we believe that the meaning of an image is not limited to a set of “best fitting” tags, but rather it exists as a distribution over arbitrary words with varying degrees of association. Specifically, the focus of our experiments is to investigate the correlation between automatic measures of such relatedness scores with respect to human judgments.

To construct the joint semantic space of words and images, we use the SVD described in Section 4 to reduce the number of dimensions. To build each model, we use the 167 synsets from ImageNet and their associated images (minus the held out test data), hence accounting for 167 latent dimensions. We first represent the synsets as a collection of documents D , each document containing visual codewords used to describe their associated images as well as textual words extracted from their gloss and synset words. Thus, computing a cross-modal relatedness distance amounts to comparing the cosine similarity of vectors representing an image to the vector representing a word in the term-document vector space. Note that, unlike textual words, an image is represented by multiple visual codewords. Prior to computing the actual cosine distance, we perform a weighted addition of vectors representing each visual codeword for that image.

To illustrate, consider a single document d_i , representing the synset “snail,” which consists of $\{cw_0, cw_{555}, cw_{23}, cw_{124}, cw_{876}, \text{snail}, \text{freshwater}, \text{mollusk}, \text{spiral}, \text{shell}\}$, where cw_X represents a particular visual codeword indexed from 0-999⁸, and the textual words are nouns extracted from the associated synset and gloss. Given a test image I , it can be expressed as a bag of visual codewords $\{cw_1, \dots, cw_k\}$. We first represent each visual codeword in I as a vector of length $|D|$ using term-frequency inverse-document-frequency (*tfidf*) weighting, e.g., $cw_k = \langle 0.4*d_1, 0.2*d_2, \dots, 0.9*d_m \rangle$, where $m=167$, and perform an addition of k such vectors to form a final vector v_i . To measure the semantic relatedness between image I and a word w , e.g., “snail,” we simply compute the cosine similarity between v_i and v_w , where v_w is also a vector of length $|D|$ calculated using *tfidf*.

This paper seeks answers to the following questions. First, what is the relation between the discriminability of the visual codewords and their ability to capture semantic relatedness between a word and an image, as compared to the gold-standard annotation by humans? Second, given the unbalanced dataset of images and words, can we use a relatively small number of visual codewords to derive such semantic relatedness measures reliably? Third, what is the efficiency of an unsupervised vector semantic model in measuring such relatedness, and is it applicable to large datasets?

Analogous to text-retrieval methods, we measure the discriminability of the visual codewords using two weighting factors. The first is *term-frequency (tf)*, which measures the number of times a codeword appears in all images for a particular synset, while the second, *image-term-frequency (itf)*, captures the number of images using the codeword in a synset. For the two weighting schemes, we apply normalization by using the total number of codewords for a synset (for *tf* weighting) and the total number of images in a synset (for *itf* weighting).

We are interested to quantify the relatedness for pairs of words and images under two scenarios. By ranking the 12 words associated with an image in reverse order of their relatedness to the image, we can determine the ability of our models to identify the most related words for a given image (**image-centered**). In the second scenario, we measure the relatedness of words and images regardless of the synset they belong to, thus evaluating the ability of our methods to capture the relatedness between any word and any image. This allows us to capture the correlation in an (**arbitrary-image**) scenario. For the evaluations, we use the Spearman’s Rank correlation.

To place our results in perspective, we implemented two baselines and an upper bound for each of the two scenarios above. The *Random* baseline randomly assigns ratings to each word-image pair on the same 0 to 10 scale, and then measures the correlation to the human gold-standard. The *Vector-Based (VB)* method is a stronger baseline aimed to study the correlation performance in the absence of dimensionality reduction. As an upper bound, the *Inter-Human-Agreement (IHA)* measures the correlation of the rating by each annotator against the average of the ratings of the rest of the annotators, averaged over the 167 synsets (for the image-centered scenario) and over the 2004 word-image pairs (for the arbitrary-image scenario).

⁸For simplicity, we only show the top 5 visual codewords

	Spearman’s Rank Coefficient (image-centered)									
Top K codewords	100	200	300	400	500	600	700	800	900	1000
<i>LSA tf</i>	0.228	0.325	0.273	0.242	0.185	<u>0.181</u>	0.107	0.043	-0.018	0.000
<i>LSA tf (norm)</i>	0.233	0.339	<u>0.293</u>	<u>0.254</u>	0.202	0.180	<u>0.124</u>	<u>0.047</u>	<u>-0.012</u>	0.000
<i>LSA tf*itf</i>	<u>0.268</u>	0.317	0.256	0.248	<u>0.219</u>	0.166	0.081	-0.004	-0.037	0.000
<i>LSA tf*itf (norm)</i>	0.252	0.327	0.257	0.246	0.211	0.153	0.097	0.002	-0.042	0.000
<i>VB tf</i>	0.243	0.168	0.101	0.055	-0.021	-0.084	-0.157	-0.210	-0.236	-0.332
<i>VB tf (norm)</i>	0.240	0.181	0.110	0.062	-0.010	-0.082	-0.152	-0.204	-0.235	-0.332
<i>VB tf*itf</i>	0.262	0.181	0.107	0.065	-0.019	-0.081	-0.156	-0.211	-0.241	-0.332
<i>VB tf*itf (norm)</i>	0.257	0.180	0.116	0.068	-0.014	-0.079	-0.150	-0.250	-0.237	-0.332
Random	0.001	0.018	0.016	-0.008	0.008	0.005	-0.001	0.014	-0.035	0.012
IHA	0.687									
	Spearman’s Rank Coefficient (arbitrary-image)									
Top K codewords	100	200	300	400	500	600	700	800	900	1000
<i>LSA tf</i>	0.236	0.341	0.291	0.249	0.208	0.183	0.106	<u>0.033</u>	-0.039	0.000
<i>LSA tf (norm)</i>	0.230	0.353	<u>0.301</u>	<u>0.271</u>	0.220	<u>0.186</u>	<u>0.115</u>	0.032	<u>-0.029</u>	0.000
<i>LSA tf*itf</i>	<u>0.291</u>	0.332	0.289	0.262	<u>0.235</u>	0.172	0.092	0.008	-0.041	0.000
<i>LSA tf*itf (norm)</i>	0.277	0.345	0.292	0.269	0.234	0.164	0.098	0.015	-0.046	0.000
<i>VB tf</i>	0.272	0.195	0.119	0.059	-0.012	-0.088	-0.164	-0.218	-0.240	-0.339
<i>VB tf (norm)</i>	0.277	0.207	0.130	0.069	-0.003	-0.083	-0.160	-0.215	-0.242	-0.339
<i>VB tf*itf</i>	0.287	0.206	0.127	0.062	-0.008	-0.085	-0.161	-0.214	-0.241	-0.339
<i>VB tf*itf (norm)</i>	0.286	0.212	0.132	0.071	-0.005	-0.081	-0.158	-0.214	-0.241	-0.339
Random	-0.024	-0.014	0.015	-0.015	-0.004	-0.014	0.024	-0.009	-0.007	0.007
IHA	0.764									

Table 2: Correlation of automatically generated scores with human annotations on cross-modal semantic relatedness, as performed on the ImageNet test dataset of 2004 pairs of word and image. Correlation figures scoring the highest within a weighting scheme are marked in bold, while those scoring the highest across weighting schemes and within a visual vocabulary size are underlined.

6 Discussion

Our experimental results are shown in Table 2. A somewhat surprising observation is the consistency of correlation figures between the two scenarios. In both scenarios, a representative set of 200 visual codewords is sufficient to consistently score the highest correlation ratings across the 8 weighting schemes. Intuitively, based on the experimental results, automatically choosing the top 10% or 20% of the visual codewords seems to suffice and gives optimal correlation figures, but requires further justification. Conversely, the relatively simple weighting scheme using *tf (normalized)* produces the highest correlation in six visual codeword sizes (K=200,300,400,700,800,900) for the image-centered scenario, as well as in another six visual codeword sizes (K=200,300,400,600,700,900) for the arbitrary-image scenario. Unlike stopwords in text retrieval accounting for most of the highest *tf* scores, visual codewords weighted by the same scheme *tf* and a similar *tf (normalized)* scheme seem to be the most discriminative. The correlation for including the entire visual vocabulary set (1000) produces identical results for all vector-based and LSA weighting schemes, as images across synsets are now encoded by the same set of visual codewords without discrimination between them.

Dimensionality reduction using SVD gains an advantage over the vector-based method for both scenarios, with the highest correlation rating in LSA (200 visual codeword, *tf(norm)*) achieving 0.077 points better than the corresponding highest correlation in Vector-based (100 visual codeword, *tf*itf*) for the image-centered scenario, representing a 29.3% improvement. Similarly, in the arbitrary-image scenario, the increase in correlation from 0.287 (VB *tf*itf* at 100 visual codeword) to 0.353 (LSA *tf(norm)* at 200 visual codeword) underlines a gain of approximately 23.0%. Overall, the arbitrary-image scenario also scores consistently higher than the image-centered scenario under similar experimental conditions. For instance, for the top 200 visual words, the same weighting schemes produce consistently lower correlation figures for the image-centered scenario. This is also true for the Inter-Human-Agreement score, which is higher in the arbitrary-image scenario (0.764) compared to the image-centered scenario (0.687). Note that for all the experiments, the semantic relatedness scores generated from the semantic vector space are significantly more correlated with the human gold-standard than the random baselines.

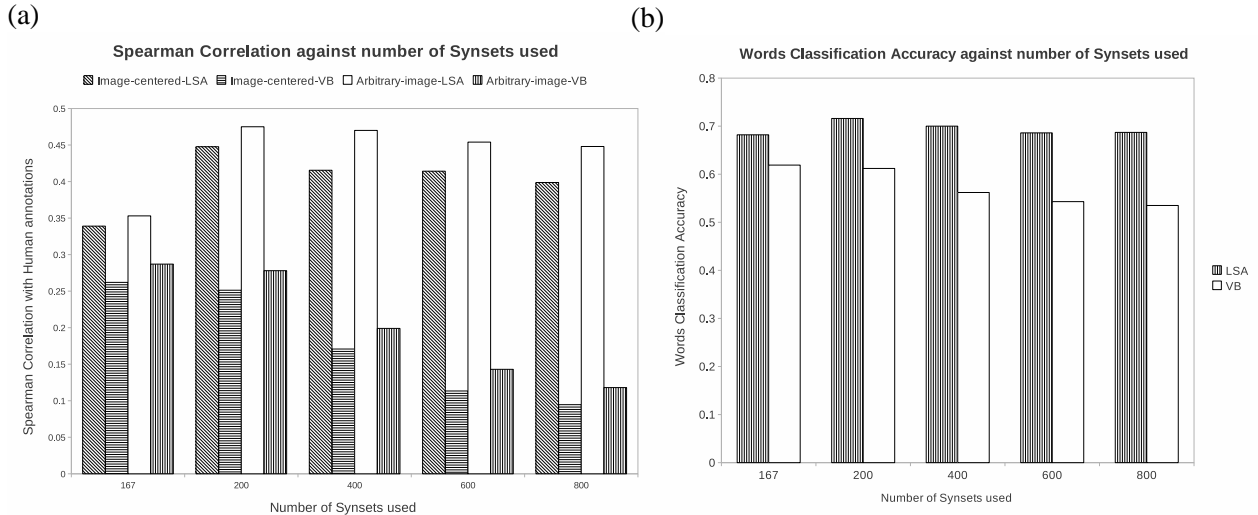


Figure 3: (a) Correlation performance, and (b) Classification accuracy, as more data is added to construct the semantic space model.

To investigate the effectiveness of the model when scaling up to large datasets, we employ the best combination of weighting scheme and vocabulary size shown in Table 2, i.e., a visual vocabulary size of 200 and *tf (normalized)* weighting for LSA, and vocabulary size of 100 and *tf*idf* weighting for the vector-based model, and incrementally construct models ranging from 167 synsets to 800 synsets (all randomly selected from ImageNet). We then measure the correlation of relatedness scores generated using the same test dataset with respect to human annotations. The dataset was randomly selected to increase by approximately five times, from a total of 230,864 images with 878 words to a total of 1,014,528 images with 3887 words. Furthermore, for each unseen test image taken from Synset S_i and the associated 12 candidate words, we evaluate the ability of the model to identify which of the candidate words actually appear in the gloss or the synset of S_i , in a task we term as word classification. Here, the top six words are predictably classified as those appearing in S_i while the last six are classified as outside of S_i , after all 12 words are ranked in reverse order of their relatedness to the test image. We measure the accuracy of the word classification task using $\frac{TP+TN}{2004}$, where TP is the number of words correctly classified as synset or gloss words, and TN is the number of words correctly classified as outside of synset or gloss, both summed over the 2004 pairs of words and images.

As shown in Figure 3, when a small number of synsets (33) was added to the original semantic space, correlation with human ratings increased steeply to around 0.45 and higher for LSA in both scenarios, while the vector-based method suffers a slight decrease in correlation ratings from 0.262 to 0.251 (image-centered) and from 0.287 to 0.278 (arbitrary-image). As more images and words are added, correlation for the vector-based model continues to decrease markedly. Comparatively, LSA is less sensitive to data scaling, as correlation figures for both scenarios decreases slightly but stays within a 0.40 to 0.45 range. Additionally, we infer that LSA is consistently more effective than the vector-based model in the words classification task (as also seen in Figure 3). Even with more data added to the semantic space, word classification accuracy stays consistently at 0.7 for LSA, while it drops to 0.535 for the vector-based model at a synset size of 800.

7 Conclusion

In this paper, we provided a proof of concept in quantifying the semantic relatedness between words and images through the use of visual codewords and textual words in constructing a joint semantic vector space. Our experiments showed that the relatedness scores have a positive correlation to human gold-standards, as measured using a standard evaluation framework.

We believe many aspects of this work can be explored further. For instance, other visual codeword attributes, such as pixel coordinates, can be employed in a structured vector space along with the existing model for improving vector similarity measures. To improve textual words coverage, a potentially effec-

tive way would be to create mappings from WordNet synsets to Wikipedia entries, where the concepts represented by the synsets are discussed in detail. We also plan to study the applicability of the joint semantic representation model to tasks such as automatic image annotation and image classification.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340 and IIS award #1018613. 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

- Barnard, K. and D. Forsyth (2001). Learning the semantics of words and pictures. In *Proceedings of International Conference on Computer Vision*.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. In *Psychological Review*, Volume 94, pp. 115–147.
- Budanitsky, A. and G. Hirst (2005). Evaluating wordnet-based measures of lexical semantic relatedness. In *Computational Linguistics*, Volume 32.
- Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Erk, K. and D. McCarthy (2009). Graded word sense assignment. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Fei-Fei, L. and P. Perona (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Feng, Y. and M. Lapata (2010). Visual information in semantic representation. In *Proceedings of the Annual Conference of the North American Chapter of the ACL*.
- Goldberger, J., S. Gordon, and H. Greenspan (2003). An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Proceedings of IEEE International Conference on Computer Vision*.
- Hare, J. S., S. Samangooei, P. H. Lewis, and M. S. Nixon (2008). Investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In *Proceedings of the international conference on content-based image and video retrieval*.
- Kanerva, P. (1998). Sparse distributed memory. In *MIT Press*.
- Landauer, T. and S. Dumais (1997). A solution to platos problem: The latent semantic analysis theory of acquisition. In *Psychological Review*, Volume 104, pp. 211–240.
- Leong, C. W., R. Mihalcea, and S. Hassan (2010). Text mining for automatic image tagging. In *Proceedings of the International Conference on Computational Linguistics*.
- Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*.
- Makadia, A., V. Pavlovic, and S. Kumar (2008). A new baseline for image annotation. In *Proceedings of European Conference on Computer Vision*.
- Miller, G. (1995). Wordnet: A lexical database for english. In *Communications of the ACM*, Volume 38, pp. 39–41.
- Miller, G. and W. Charles (1998). Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1).
- Potter, M. C. and B. A. Faulconer (1975). Time to understand pictures and words. In *Nature*, Volume 253, pp. 437–438.
- Salton, G., A. Wong, and C. Yang (1997). A vector space model for automatic indexing. In *Readings in Information Retrieval*, pp. 273–280. San Francisco, CA: Morgan Kaufmann Publishers.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.
- Smeulders, A. W., M. Worring, S. Santini, A. Gupta, and R. Jain (2000). Content-based image retrieval at the end of the early years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 22, pp. 1349–1380.
- Westerveld, T. (2000). Image retrieval: Context versus context. In *Content-Based Multimedia Information Access*.
- Widdows, D. and K. Ferraro (2008). Semantic vectors: a scalable open source package and online technology management application. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.
- Yang, J., Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo (2007). Evaluating bag-of-visual-words representations in scene classification. In *ACM Multimedia Information Retrieval Workshop*.