Distributional semantics from text and images

Elia Bruni
CIMeC, University of Trento
elia.bruni@unitn.it

Giang Binh Tran
EMLCT, Free University of Bolzano &
CIMeC, University of Trento
Giang.Tran@stud-inf.unibz.it

Marco Baroni
CIMeC, University of Trento
marco.baroni@unitn.it

Abstract

We present a distributional semantic model combining text- and image-based features. We evaluate this multimodal semantic model on simulating similarity judgments, concept clustering and the BLESS benchmark. When integrated with the same core text-based model, image-based features are at least as good as further text-based features, and they capture different qualitative aspects of the tasks, suggesting that the two sources of information are complementary.

1 Introduction

Distributional semantic models use large text corpora to derive estimates of semantic similarities between words. The basis of these procedures lies in the hypothesis that semantically similar words tend to appear in similar contexts (Miller and Charles, 1991; Wittgenstein, 1953). For example, the meaning of *spinach* (primarily) becomes the result of statistical computations based on the association between *spinach* and words like *plant*, *green*, *iron*, *Popeye*, *muscles*. Alongside their applications in NLP areas such as information retrieval or word sense disambiguation (Turney and Pantel, 2010), a strong debate has arisen on whether distributional semantic models are also reflecting human cognitive processes (Griffiths et al., 2007; Baroni et al., 2010). Many cognitive scientists have however observed that these techniques relegate the process of meaning extraction solely to linguistic regularities, forgetting that humans can also rely on non-verbal experience, and comprehension also involves the activation of non-linguistic representations (Barsalou et al., 2008; Glenberg, 1997; Zwaan, 2004). They argue that, without grounding words to bodily actions and perceptions in the environment, we can never get past defining a symbol by simply pointing to covariation of amodal symbolic patterns (Harnad, 1990). Going back to our example, the meaning of *spinach* should come (at least partially) from our experience with spinach, its colors, smell and the occasions in which we tend to encounter it.

We can thus distinguish two different views of how meaning emerges, one stating that it emerges from association between linguistic units reflected by statistical computations on large bodies of text, the other stating that meaning is still the result of an association process, but one that concerns the association between words and perceptual information.

In our work, we try to make these two apparently mutually exclusive accounts communicate, to construct a richer and more human-like notion of meaning. In particular, we concentrate on perceptual information coming from images, and we create a multimodal distributional semantic model extracted from texts and images, putting side by side techniques from NLP and computer vision. In a nutshell, our technique is based on using a collection of labeled pictures to build vectors recording the co-occurrences of words with image-based features, exactly as we would do with textual co-occurrences. We then concatenate the image-based vector with a standard text-based distributional vector, to obtain our multimodal representation. The preliminary results reported in this paper indicate that en-

riching a text-based model with image-based features is at least not damaging, with respect to enlarging the purely textual component, and it leads to qualitatively different results, indicating that the two sources of information are not redundant.

The rest of the paper is structured as follows. Section 2 reviews relevant work including distributional semantic models, computer vision techniques suitable to our purpose and systems combining text and image information, including the only work we are aware of that attempts something similar to what we try here. We introduce our multimodal distributional semantic model in Section 3, and our experimental setup and procedure in Section 4. Our experiments' results are discussed in Section 5. Section 6 concludes summarizing current achievements and discussing next directions.

2 Related Work

2.1 Text-based distributional semantic models

Traditional corpus-based models of semantic representation base their analysis on textual input alone (Turney and Pantel, 2010). Assuming the distributional hypothesis (Miller and Charles, 1991), they represent semantic similarity between words as a function of the degree of overlap among their linguistic contexts. Similarity is computed in a semantic space represented as a matrix, with words as rows and contextual elements as columns/dimensions. Thanks to the geometrical nature of the representation, words are compared using a distance metric, such as the cosine of the angle between vectors (Landauer and Dumais, 1997).

2.2 Bag of visual words

In NLP, "bag of words" (BoW) is a dictionary-based method in which a document is represented as a "bag" (i.e., order is not considered), which contains words from the dictionary. In computer vision, "bag of visual words" (**BoVW**) is a similar idea for image representation (Sivic and Zisserman, 2003; Csurka et al., 2004; Nister and Stewenius, 2006; Bosch et al., 2007; Yang et al., 2007).

Here, an image is treated as a document, and features from a dictionary of visual elements extracted from the image are considered as the "words" representing the image. The following pipeline is typically adopted in order to group the local interest points into types (**visual words**) within and across images, so that then an image can be represented by the number of occurrences of each visual word type in it, analogously to BoW. From every image of a data set, keypoints are automatically detected and represented as vectors of various descriptors. Keypoint vectors are then projected into a common space and grouped into a number of clusters. Each cluster is treated as a discrete visual word (this technique is generally known as vector quantization). With its keypoints mapped onto visual words, each image can then be represented as a BoVW feature vector according to the count of each visual word. In this way, we move from representing the image by a varying number of high-dimensional keypoint descriptor vectors to a representation in terms of a single sparse vector of fixed dimensionality across all images. What kind of image content a visual word captures exactly depends on a number of factors, including the descriptors used to identify and represent local interest points, the quantization algorithm and the number of target visual words selected. In general, local interest points assigned to the same visual word tend to be patches with similar low-level appearance; but these common types of local patterns need not be correlated with object-level parts present in the images. Figure 1 illustrates the procedure to form bags of visual words. Importantly for our purposes, the BoVW representation, despite its unrelated origin in computer vision, is entirely analogous to the BoW representation, making the integration of text- and image-based features very straightforward.

2.3 Integrating textual and perceptual information

Louwerse (2011), contributing to the debate on symbol grounding in cognitive science, theorizes the *interdependency account*, which suggests a convergence of symbolic theories (such as distributional semantics) and perceptual theories of meaning, but lacks of a concrete way to harvest perceptual information computationally. Andrews et al. (2009) complement text-based models with experiential information, by combining corpus-based statistics with speaker-generated feature norms as a proxy of perceptual experience. However, the latter are an unsatisfactory proxy, since they are still verbally pro-

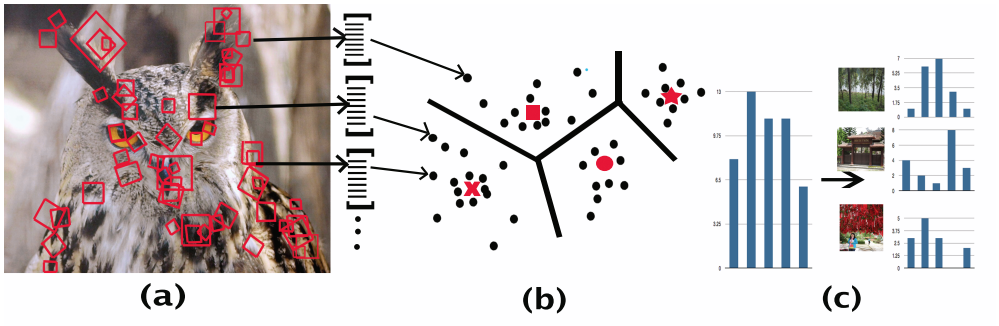

Figure 1: Illustration of *bag of visual words* procedure: (a) detect and represent local interest points as descriptor vectors (b) quantize vectors (c) histogram computation to form BoVW vector for the image

duced descriptions, and they are expensive to collect from subjects via elicitation techniques.

Taking inspiration from methods originally used in text processing, algorithms for image labeling, search and retrieval have been built upon the connection between text and visual features. Such models learn the statistical models which characterize the joint statistical distribution of observed visual features and verbal image tags (Hofmann, 2001; Hare et al., 2008). This line of research is pursuing the reverse of what we are interested in: using text to improve the semantic description of images, whereas we want to exploit images to improve our approximation to word meaning.

Feng and Lapata are the first trying to integrate authentic visual information in a text-based distributional model (Feng and Lapata, 2010). Using a collection of BBC news with pictures as corpus, they train a Topic model where text and visual words are represented in terms of the same shared latent dimensions (topics). In this framework, word meaning is modeled as a probability distribution over a set of latent multimodal topics and the similarity between two words can be estimated by measuring the topics they have in common. A better correlation with semantic intuitions is obtainable when visual modality is taken into account, in comparison to estimating the topic structure from text only.

Although Feng and Lapata's work is very promising and the main inspiration for our own, their method requires the extraction of a single distributional model from the same mixed-media corpus. This has two important drawbacks: First, the textual model must be extracted from the same corpus images are taken from, and the text context extraction methods must be compatible with the overall multimodal approach. Thus, image features cannot be added to a state-of-the-art text-based distributional model – e.g., a model computed on the whole Wikipedia or larger corpora using syntactic dependency information – to assess whether visual information is helping even when purely textual features are already very good. Second, by training a joint model with latent dimensions that mix textual and visual information, it becomes hard to assess, quantitatively and qualitatively, the separate effect of image-based features on the overall performance. In order to overcome these issues, we propose a somewhat simpler approach, in which the text- and image-based models are independently constructed from different sources, and then concatenated.

3 Proposed method

Figure 2 presents a diagram of our overall system. The main idea is to construct text-based and image-based co-occurrence models separately and then combine them. We first describe our procedure to build both text-based and image-based models. However, we stress the latter since it is the more novel part of the procedure. Then, we describe our simple combination technique to integrate both models and create a multimodal distributional semantic space. Our implementation of the proposed method is open-source[1].

[1] https://github.com/s2m

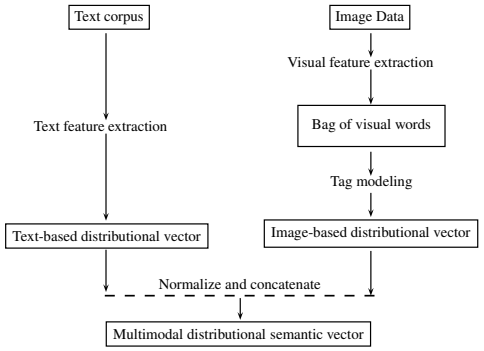

Figure 2: Overview of our system architecture

3.1 Text-based distributional model

Instead of proposing yet another model, we pick one that is publicly available off-the-shelf and has been shown to be at the state of the art on a number of benchmarks. The picked model (**DM**)[2] is encoded in a matrix in which each target word is represented by a row vector of weights representing its association with collocates in a corpus. See Section 4.1 for details about the text-based model.

3.2 Image-based distributional model

We assume image data where each image is associated with word labels (somehow related to the image) that we call **tags**.

The primary approach to form the image-based vector space is to use the BoVW method to represent images. Having represented each image in our data set in terms of the frequency of occurrence of each visual word in it, we construct the **image-based distributional vector** of each tag as follows. Each tag (textual word) is associated to the list of images which are tagged with it; we then sum visual word occurrences across that list of images to obtain the co-occurrence counts associated with each tag. For uniformity with the treatment of textual co-occurrences (see Section 4.1), the raw counts are transformed into Local Mutual Information scores computed between each tag and visual word. Local Mutual Information is an association measure that closely approximates the commonly used Log-Likelihood Ratio while being simpler to compute (Evert, 2005).

In this way, we obtain an image-based distributional semantic model, that is, a matrix where each row corresponds to a tag vector, summarizing the distributional history of the tag in the image collection in terms of its association with the visual words.

3.3 Integrating distributional models

We assemble the two distributional vectors to construct the multimodal semantic space. Given a word that is present both in the text-based model and (as a tag) in the image-based model, we separately normalize the two vectors representing the word to length 1 (so that the text and image components will have equal weight), and we concatenate them to obtain the multimodal distributional semantic vector representing the word. The matrix of concatenated text- and image-based vectors is our multimodal distributional semantic model. We leave it to future work to consider more sophisticated combination techniques (preliminary experiments on differential weighting of the text and image components did not lead to promising results).

4 Experimental setup

4.1 The DM text-based model

DM has been shown to be near or at the state of the art in a great variety of semantic tasks, ranging from modeling similarity judgments to concept categorization, predicting selectional preferences, relation classification and more.

The DM model is described in detail by Baroni and Lenci (2010), where it is referred to as TypeDM. In brief, the model is trained on a large corpus of about 2.8 billion tokens that include Web documents, the Wikipedia and the BNC. DM is a structured model, where the collocates are labeled with the link that connect them to the target words. The links are determined by a mixture of dependency parse information and lexico-syntactic patterns, resulting in distributional features (the dimensions of the semantic space) such as *subject_kill*, *with_gun* or *as_sharp_as*. The score of a target word with a feature is not based on the absolute number of times they co-occur in the corpus, but on the variety of different surface realizations of the feature the word co-occurs with. For example, for the word *fat* and the feature *of_animal*, the raw score is 9 because *fat* co-occurs with 9 different forms of the feature (*a*

[2] http://clic.cimec.unitn.it/dm

fat of the animal, *the fat of the animal*, *fats of animal*...). Refer to Baroni and Lenci (2010) for how the surface realizations of a feature are determined. Raw scores are then transformed into Local Mutual Information values.

The DM semantic space is a matrix with 30K rows (target words) represented in a space of more than 700M dimensions. Since our visual dimension extraction algorithms are maximally producing 32K dimensions (see Section 4.2 below), we make the impact of text features on the combined model directly comparable to the one of visual features by selecting only the top n DM dimensions (with n varying as explained below). The top dimensions are picked based on their cumulative Local Mutual Information mass. We show in the experiments below that trimming DM in this way does not have a negative impact on its performance, so that we are justified in claiming that we are adding visual information to a state-of-the-art text-based semantic space.

4.2 Visual Information Extraction

For our experiments, we use the ESP-Game data set.[3] It contains 50K images, labeled through the famous "game with a purpose" developed by Louis von Ahn (von Ahn and Dabbish, 2004). The tags of images in the data set form a vocabulary of 11K distinct word types. Image labels contain 6.686 tags on average (2.357 s.d.). The ESP-Game corpus is an interesting data set from our point of view since, on the one hand, it is rather large and we know that the tags it contains are related to the images. On the other hand, it is not the product of experts labelling representative images, but of a noisy annotation process of often poor-quality or uninteresting images (e.g., logos) randomly downloaded from the Web. Thus, analogously to the characteristics of a textual corpus, our algorithms must be able to exploit large-scale statistical information, while being robust to noise.

Following what has become an increasingly standard procedure in computer vision, we use the Difference of Gaussian (DoG) detector to automatically detect keypoints from images and consequently map them to visual words (Lowe, 1999; Lowe, 2004). We use the Scale-Invariant Feature Transform (SIFT) to depict the keypoints in terms of a 128-dimensional real-valued descriptor vector. Color version SIFT descriptors are extracted on a regular grid with five pixels spacing, at four multiple scales (10, 15, 20, 25 pixel radii), zeroing the low contrast ones. We chose SIFT for its invariance to image scale, orientation, noise, distortion and partial invariance to illumination changes. To map the descriptors to visual words, we cluster the keypoints in their 128-dimensional space using the K-means clustering algorithm, and encode each keypoint by the index of the cluster (visual word) to which it belongs. We varied the number of visual words between 250 and 2000 in steps of 250. We then computed a one-level 4x4 pyramid of spatial histograms (Grauman and Darrell, 2005), consequently increasing the features dimensions 16 times, for a number that varies between 4K and 32K, in steps of 4K. From the point of view of our distributional semantic model construction, the important point to keep in mind is that standard parameter choices such as the ones we adopted lead to distributional vectors with 4K, 8K, ..., 32K dimensions, where a higher number of features corresponds, roughly, to a more granular analysis of an image. We used the VLFeat implementation for the entire pipeline (Vedaldi and Fulkerson, 2008). See the references in Section 2.2 above for technical details.

4.3 Model integration

We remarked above that the visual word extraction procedure naturally leads to 8 kinds of image-based vectors of dimensionalities from 4K to 32K in steps of 4K. To balance text and image information, we use DM vectors made of *top n* features ranging from 4K to 32K in the same 4K steps. By combining, we obtain 64 combined models (4K text and 4K image dimensions, 4K text and 8K image dimensions, etc.). Since in the experiments on WordSim (Section 5.1 below) we observe best performance with 32K text-based features, we report here only experiments with (at least) 32K dimensions. Similar patterns to the ones we report are observed when adding image-based dimensions to text-based vectors of different dimensionalities.

For a thoroughly fair comparison, if we add n visual features to the text-based model and we notice

[3] http://www.espgame.org

an improvement, we must ask whether the same improvement could also be obtained by adding more text-based features. To control for this possibility, we also consider a set of purely text-based models that have the same number of dimensions of the combined models, that is, the top 32K DM features plus 8K, ..., 32K further DM features (the next top features in the cumulative Local Mutual Information score ranking). In the experiments below, we refer to the purely textual model as **text** (always 32K dimensions), to the purely image-based model as **image**, to the combined models as **combined**, and to the control in which further text dimensions are added for comparability with *combined* as **text+**.

4.4 Evaluation benchmarks

We conduct our most extensive evaluation on the **WordSim353** data set (Finkelstein et al., 2002), a widely used benchmark constructed by asking 16 subjects to rate a set of word pairs on a 10-point similarity scale and averaging the ratings (*dollar/buck* receive a high 9.22 average rating, *professor/cucumber* a low 0.31). We cover 260 WordSim (mostly noun/noun) pairs. We evaluate models in terms of the Spearman correlation of the cosines they produce for the WordSim pairs with the average human ratings for the same pairs (here and below, we do not report comparisons with the state of the art in the literature, because we have reduced coverage of the data sets, making the comparison not meaningful).

To verify if the conclusions reached on WordSim extend to different semantic tasks, we use two **concept categorization** benchmarks, where the goal is to cluster a set of (nominal) concepts into broader categories. The Almuhareb-Poesio (**AP**) concept set (Almuhareb, 2006), in the version we cover, contains 230 concepts to be clustered into 21 classes such as *vehicle* (*airplane, car...*), *time* (*aeon, future...*) or *social unit* (*brigade, nation*). The **Battig** set (Baroni et al., 2010), in the version we cover, contains 72 concepts to be clustered into 10 classes. Unlike AP, Battig only contains concrete basic-level concepts belonging to categories such as *bird* (*eagle, owl...*), *kitchenware* (*bowl, spoon...*) or *vegetable* (*broccoli, potato...*). For both sets, following the original proponents and others, we cluster the words based on their pairwise cosines in the semantic space defined by a model using the CLUTO toolkit (Karypis, 2003). We use CLUTO's built-in *repeated bisections with global optimization* method, accepting all of CLUTO's default values. Cluster quality is evaluated by percentage *purity* (Zhao and Karypis, 2003). If n_r^i is the number of items from the *i*-th true (gold standard) class that were assigned to the *r*-th cluster, n is the total number of items and k the number of clusters, then: $\mathrm{Purity} = \frac{1}{n}\sum_{r=1}^{k}\max_i(n_r^i)$. In the best case (perfect clusters), purity is 100% and as cluster quality deteriorates, purity approaches 0.

Finally, we use the Baroni-Lenci Evaluation of Semantic Similarity (**BLESS**) data set made available by the GEMS 2011 organizers.[4] In the version we cover, the data set contains 174 concrete nominal concepts, each paired with a set of words that instantiate the following 6 relations: hypernymy (*spear/weapon*), coordination (*tiger/coyote*), meronymy (*castle/hall*), typical attribute (an adjective: *grapefruit/tart*) and typical event (a verb: *cat/hiss*). Concepts are moreover matched with 3 sets of randomly picked unrelated words (nouns, adjectives and verbs). For each true and random relation, the data set contains at least one word per concept, typically more. Following the GEMS guidelines, we apply a model to BLESS as follows. Given the similarity scores provided by the model for a concept with all associated words within a relation, we pick the term with the highest score. We then z-standardize the 8 scores we obtain for each concept (one per relation), and we produce a boxplot summarizing the distribution of z scores per relation across the concepts (i.e., each box of the plot summarizes the distribution of the 174 scores picked for each relation, standardized as we just described). Boxplots are produced accepting the default boxplotting option of the R statistical package[5] (boxes extend from first to third quartile, median is horizontal line inside the box).

[4] http://sites.google.com/site/geometricalmodels/shared-evaluation

[5] http://www.r-project.org/

5 Results

5.1 WordSim

The WordSim results for our models across dimensionalities as well as for the full *DM* are summarized in Figure 3.

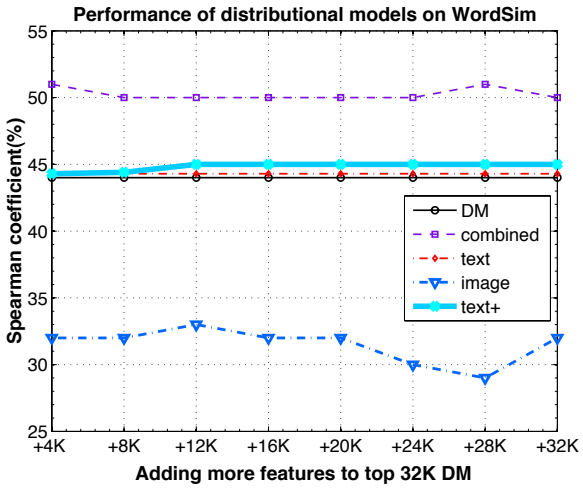

Figure 3: Performance of distributional models on WordSim

The purely image-based model is having the worst performance in all settings, although even the lowest image-based Spearman score (*0.29*) is significantly above chance ($p. < 0.05$), suggesting that the model does capture some semantic information. Contrarily, adding image-based dimensions to a textual model (*combined*) consistently reaches the best performance, also better – for all choices of dimensionality – than adding an equal number of text features (*text+*) or using the full *DM* matrix. Interestingly, the same overall result pattern is observed if we limit evaluation to the WordSim subsets that Agirre et al. (2009) have identified as *semantically similar* (e.g., synonyms or coordinate terms) and *semantically related* (e.g., meronyms or topically related concepts).

Based on the results reported in Figure 3, further analyses will focus on the *combined* model with +20K image-based features, since performance of *combined* does not seem to be greatly affected by the dimensionality parameter, and performance around this value looks quite stable (it is better only at the boundary +4K value, and with +28K, where, however, there is a dip for the *image* model). The *text+* performance is not essentially affected by the dimensionality parameter, and we pick the +20K version for maximum comparability with *combined*.

The difference between *combined* and *text+*, although consistent, is not statistically significant according to a two-tailed paired permutation test (Moore and McCabe, 2005) conducted on the results for the +20K versions of the models. Still, very interesting qualitative differences emerge. Table 1 reports those WordSim pairs (among the ones with above-median human-judged similarity) that have the highest and lowest *combined*-to-*text+* cosine ratios, i.e., pairs that are correctly treated as similar by *combined* but not by *text+*, and *vice versa*. Strikingly, the pairs characterizing the image-feature-enriched *combined* are all made of concrete, highly imageable concepts, whereas the *text+* pairs refer to very abstract notions. We thus see here the first evidence of the complementary nature of visual and textual information.

combined	*text+*
tennis/racket	physics/proton
planet/sun	championship/tournament
closet/clothes	profit/loss
king/rook	registration/arrangement
cell/phone	mile/kilometer

Table 1: WordSim pairs with highest (first column) and lowest (second column) *combined*-to-*text+* cosine ratios

5.2 Concept categorization

Table 2 reports percentage purities in the AP and Battig clustering tasks for full *DM* and the representative models discussed above.

model	*AP*	*Battig*
DM	81	96
text	79	83
text+	80	86
image	25	36
combined	78	96

Table 2: Percentage AP and Battig purities of distributional models

Once more, we see that the *image* model alone is not at the level of the text models, although both its AP and Battig purities are significantly above

chance ($p < 0.05$ based on simulated distributions for random cluster assignment). Thus, even alone, image-based vectors do capture aspects of meaning. For AP, adding image features does not improve performance, although it does not significantly worsen it either (a two-tailed paired permutation test confirms that the difference between *text+* and *combined* is far from significance). For Battig, adding visual features improves on the purely text-based models based on a comparable number of features (although the difference between *text+* and *combined* is not significant), reaching the same performance obtained with the full *DM* model (that in these categorization tests is slightly above that of the trimmed models). Intriguingly, the Battig test is entirely composed of concrete concepts, so the difference in performance for *combined* might be related to its preference for concrete things we already observed for WordSim.

5.3 BLESS

The BLESS distributions of text-based models (including *combined*) are very similar, so we use here the full *DM* model as representative of the text-based set – its histogram is compared to the one of the purely *image*-based model in Figure 4.

We see that purely text-based *DM* cosines capture a reasonable scale of taxonomic similarity among nominal neighbours (coordinates then hypernyms then meronyms then random nouns), whereas verbs and adjectives are uniformly very distant, whether they are related or not. This is not surprising because the *DM* links mostly reflect syntactic patterns, that will be disjoint across parts of speech (e.g., a feature like *subject_kill* will only apply to nouns, save for parsing errors). Looking at the *image*-only model, we first observe that it can capture differences between related attributes/events and random adjectives/verbs (according to a Tukey HSD test for all pairwise comparisons, these differences are highly significant, whereas *DM* only significantly distinguishes attributes from random verbs). In this respect, *image* is arguably the "best" model on BLESS. However, perhaps more interestingly, the *image* model also shows a bias for nouns, capturing the same taxonomic hierarchy found for *DM*. This suggests that image analysis is providing a decomposition of concepts into attributes shared by similar entities, that capture ontological similarity beyond mere syntagmatic co-occurrence in an image description.

To support this latter claim, we counted the average number of times that the related terms picked by the *image* model directly co-occur with the target concepts in an ESP-Game label. It turns out that this count is higher for both attributes (10.6) and hypernyms (7.5) than for coordinates (6.5). So, the higher similarity of coordinates in the image model demonstrates that its features do generalize across images, allowing us to capture "attributional" or "paradigmatic" similarity in visual space. More in general, we find that, among all the related terms picked by the *image* model that have an above-average cosine with the target concept, almost half (41%) *never* co-occur with the concept in the image set, again supporting the claim that, by our featural analysis, we are capturing visual properties of similar concepts beyond their co-occurrence as descriptions of the same image.

A final interesting point pertains to the specific instances of each (non-random) relation picked by the textual and visual models: of 870 related term pairs in total, almost half (418) differ between *DM* and *image*, suggesting that the boxplots in Figure 4 hide larger differences in what the models are doing. The randomly picked examples of mismatches in top attributes from Table 3 clearly illustrate the qualitative difference between the models, and, once more, the tendency of *image*-based representations to favour (not surprisingly!) highly visual properties such as colours and shapes, vs. the well-known tendency of text-based models to extract systemic or functional characteristics such as *powerful* or *elegant* (Baroni et al., 2010). By combining the two sources of information, we should be able to develop distributional models that come with more well-rounded characterizations of the concepts they describe.

6 Conclusion

We proposed a simple method to augment a state-of-the-art text-based distributional semantic model with information extracted from image analysis. The method is based on the standard bag-of-visual-words representation of images in computer vision. The image-based distributional profile of a word is

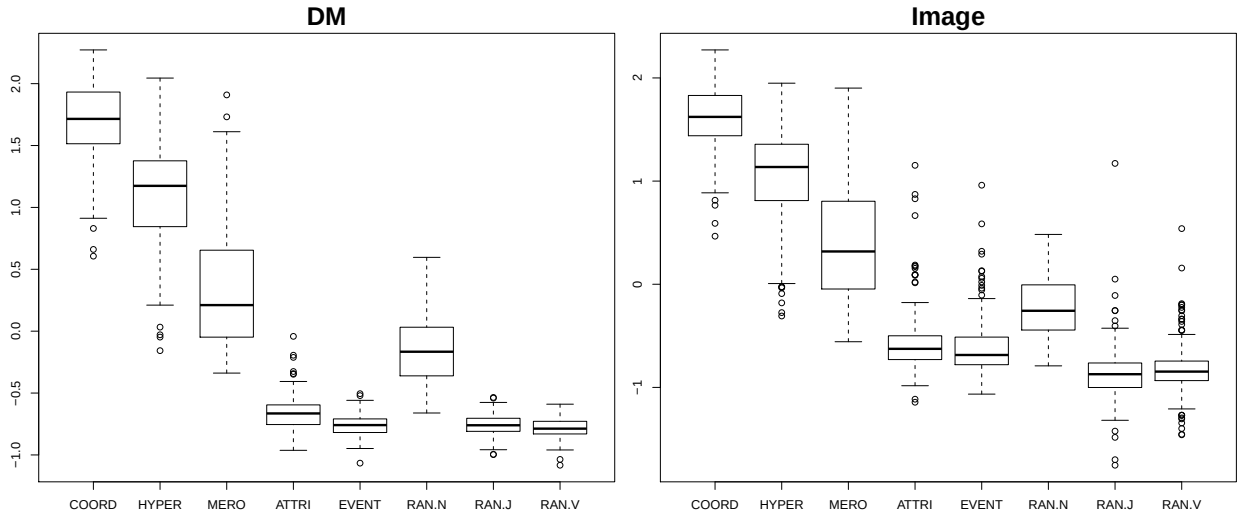

Figure 4: Distribution of z-normalized cosines of words instantiating various relations across BLESS concepts.

concept	*DM*	*image*	*concept*	*DM*	*image*
ant	small	black	potato	edible	red
axe	powerful	old	rifle	short	black
cathedral	ancient	dark	scooter	cheap	white
cottage	little	old	shirt	fancy	black
dresser	new	square	sparrow	wild	brown
fighter	fast	old	squirrel	fluffy	brown
fork	dangerous	shiny	sweater	elegant	old
goose	white	old	truck	new	heavy
jet	fast	old	villa	new	cosy
pistol	dangerous	black	whale	large	gray

Table 3: Randomly selected cases where nearest attributes picked by DM and *image* differ.

encoded in a vector of co-occurrences with "visual words", that we concatenate with a text-based co-occurrence vector. A cautious interpretation of our results is that adding image-based features is at least not damaging, when compared to adding further text-based features, and possibly beneficial. Importantly, in all experiments we find that image-based features lead to interesting qualitative differences in performance: Models including image-based information are more oriented towards capturing similarities between concrete concepts, and focus on their more imageable properties, whereas the text-based features are more geared towards abstract concepts and properties. Coming back to the discussion of symbol grounding at the beginning of the paper, we consider this (very!) preliminary evidence for an integrated view of semantics where the more concrete aspects of meaning derive from perceptual experience, whereas verbal associations mostly account for abstraction.

In future work, we plan first of all to improve performance, by focusing on visual word extraction and on how the text- and image-based vectors are combined (possibly using supervision to optimize both feature extraction and integration with respect to semantic tasks). However, the most exciting direction we intend to follow next will concern evaluation, and in particular devising new benchmarks that address the special properties of image-enhanced models directly. For example, Baroni and Lenci (2008) observe that text-based distributional models are seriously lacking when it comes to characterize physical properties of concepts such as their colors or parts. These are exactly the aspects of conceptual knowledge where image-based information should help most, and we will devise new test sets that will focus specifically on verifying this hypothesis.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasça, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of HLT-NAACL*, pages 19–27, Boulder, CO.

Abdulrahman Almuhareb. 2006. *Attributes in Lexical Acquisition*. Phd thesis, University of Essex.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

Marco Baroni and Alessandro Lenci. 2008. Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Eduard Barbu, Brian Murphy, and Massimo Poesio. 2010. Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Lawrence Barsalou, Ava Santos, Kyle Simmons, and Christine Wilson, 2008. *Language and Simulation in Conceptual Processing*, chapter 13, pages 245–283. Oxford University Press, USA, 1 edition.

Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Image Classification using Random Forests and Ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8.

Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99, Los Angeles, California. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Arthur Glenberg. 1997. What memory is for. *Behav Brain Sci*, 20(1), March.

Kristen Grauman and Trevor Darrell. 2005. The pyramid match kernel: Discriminative classification with sets of image features. In *In ICCV*, pages 1458–1465.

Tom Griffiths, Mark Steyvers, and Josh Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114:211–244.

Jonathon Hare, Sina Samangooei, Paul Lewis, and Mark Nixon. 2008. Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR '08, pages 359–368, New York, NY, USA. ACM.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, June.

Thomas Hofmann. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196, January.

George Karypis. 2003. CLUTO: A clustering toolkit. Technical Report 02-017, University of Minnesota Department of Computer Science.

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Max Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3:273–302.

David Lowe. 1999. Object Recognition from Local Scale-Invariant Features. *Computer Vision, IEEE International Conference on*, 2:1150–1157 vol.2, August.

David Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), November.

George Miller and Walter Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

David Moore and George McCabe. 2005. *Introduction to the Practice of Statistics*. Freeman, New York, 5 edition.

David Nister and Henrik Stewenius. 2006. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2161–2168.

Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October.

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

Andrea Vedaldi and Brian Fulkerson. 2008. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, New York, NY, USA. ACM.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell, Oxford. Translated by G.E.M. Anscombe.

Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. In James Ze Wang, Nozha Boujemaa, Alberto Del Bimbo, and Jia Li, editors, *Multimedia Information Retrieval*, pages 197–206. ACM.

Ying Zhao and George Karypis. 2003. Criterion functions for document clustering: Experiments and analysis. Technical Report 01-40, University of Minnesota Department of Computer Science.

Rolf Zwaan. 2004. The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation: Advances in Research and Theory, Vol 44*, 44.