Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations

Guillem Collell
Computer Science Department
KU Leuven
3001 Heverlee, Belgium
gcollell@kuleuven.be

Marie-Francine Moens
Computer Science Department
KU Leuven
3001 Heverlee, Belgium
sien.moens@cs.kuleuven.be

Abstract

Human concept representations are often grounded with visual information, yet some aspects of meaning cannot be visually represented or are better described with language. Thus, vision and language provide complementary information that, properly combined, can potentially yield more complete concept representations. Recently, state-of-the-art distributional semantic models and convolutional neural networks have achieved great success in representing linguistic and visual knowledge respectively. In this paper, we compare both, visual and linguistic representations in their ability to capture different types of fine-grain semantic knowledge—or attributes—of concepts. Humans often describe objects using attributes, that is, properties such as shape, color or functionality, which often transcend the linguistic and visual modalities. In our setting, we evaluate how well attributes can be predicted by using the unimodal representations as inputs. We are interested in first, finding out whether attributes are generally better captured by either the vision or by the language modality; and second, if none of them is clearly superior (as we hypothesize), what type of attributes or semantic knowledge are better encoded from each modality. Ultimately, our study sheds light on the potential of combining visual and textual representations.

1 Introduction

Vision and language capture complementary information that humans automatically integrate in order to build mental representations of concepts. Certain concepts or properties of objects cannot be explicitly visually represented while, at the same time, not all the properties are easily expressible with language. For example, there are clearly visual differences between cats and dogs although these are not easy to describe with language. Recent advances in deep learning had led to breakthroughs in learning unimodal representations (a.k.a. embeddings) in both, computer vision (CV) and natural language processing (NLP) (LeCun et al., 2015). However, the automatic integration of visual and linguistic modalities is still a challenging—and usually task-dependent—problem that has gained increasing popularity within the NLP and CV communities. Lately, several studies have achieved reasonable success in integrating visual and linguistic representations, showing improvement over the unimodal baselines in simple linguistic tasks such as concept similarity (Lazaridou et al., 2015; Kiela and Bottou, 2014; Silberer and Lapata, 2014)—which is only possible if vision and language encode complementary knowledge.

In this paper we do not tackle the problem of *how* to integrate both modalities, but instead, we systematically study *what* type of fine-grain semantic knowledge is encoded in each modality, shedding light on the potential benefit of combining vision and language. By fine-grain semantics we refer to the recognition of different types of *attributes* or properties (e.g., shape, function, sound, etc.) that concrete nouns might exhibit. A recent study by Rubinstein et al. (2015) evidenced that state-of-the-art linguistic-only representations do not succeed at capturing all types of attributes equally well. Here, we extend their work into the multimodal domain by comparing the performance between visual and linguistic representations at encoding different types of attributes. In contrast with Rubinstein et al. (2015)'s unimodal research, here we aim to answer two different research questions. First, whether either vision or

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2807–2817, Osaka, Japan, December 11-17 2016.

language provide a superior ground for capturing fine-grain attributes; and second, if none of them is clearly superior—as we hypothesize—what type of attributes are better captured by each modality. To the best of our knowledge, no studies have systematically compared vision and language in these terms. Ultimately, this work provides insight on building better representations of concepts, which in turn is essential towards improving automatic language understanding.

The rest of the paper is organized as follows. In the next section we review and discuss related work. In Section 3 we describe our design for evaluating the success of visual and text embeddings in encoding attributes. Next, we present and discuss our experimental results. Finally, in conclusions and future work, we summarize our findings and suggest future lines of research.

2 Related Work

2.1 Unimodal Representations

Convolutional neural networks (CNN) have rapidly become the state-of-the-art approach in computer vision (Krizhevsky et al., 2012). To some extent, CNN algorithms emulate human visual perception, in which the learning occurs at different levels of abstraction—or layers of a network. On the language side, distributional models (DMs) have been employed for learning semantic representations a long time ago. These are based on the distributional hypothesis: *Words which are similar in meaning occur in similar contexts* (Rubenstein and Goodenough, 1965). Recently, neural-based distributional models or word embeddings (Mikolov et al., 2013; Pennington et al., 2014) have achieved great success, rapidly replacing the old DMs (Turney et al., 2010) based on word co-occurrence counts (Baroni et al., 2014). Instead of counting words, neural-based DMs capture words co-occurrences by trying to predict the context given a word (skip-gram) or by trying to predict a word given its context (CBOW). Alternative approaches such as generative probabilistic models that learn the probability distribution of a vocabulary word in a context window as a latent variable have also been proposed (Deschacht et al., 2012; Deschacht and Moens, 2009).

2.2 Multimodal Representations

There exist certain properties of perceptible objects that are poorly captured by language. For example, everyone can easily tell from an image whether a face is attractive, yet if one has to describe with language what properties make a face attractive will certainly struggle. Recently, CNN-based computer vision models have achieved reasonable success in the task of recognizing attractiveness (Rothe et al., 2015). Furthermore, psychological research evidences that human concept formation is strongly grounded in visual perception (Barsalou, 2008). All this suggests that linguistic representations can benefit from visual grounding. In this direction, recent studies have shown that multimodal embeddings are often able to outperform text-only embeddings in simple semantic tasks such as concept similarity (Lazaridou et al., 2015; Silberer and Lapata, 2014; Kiela and Bottou, 2014) or categorization (i.e., grouping objects into categories such as "fruit", "furniture", etc.) (Silberer and Lapata, 2014). Several ways of combining representations from both modalities have been devised so far—yet an exhaustive review of them would deviate from the target of this work. To name a few, Kiela and Bottou (2014) proposed the simple concatenation of visual and text representations, although more sophisticated methods such as the extension of skip-gram models into the multimodal domain (Lazaridou et al., 2015) or to apply stacked autoencoders to the unimodal representations (Silberer and Lapata, 2014) have also been considered.

2.3 Attribute Representations

Previous multimodal research often evaluates representations in word similarity tasks (Lazaridou et al., 2015; Silberer and Lapata, 2014; Kiela and Bottou, 2014) which offer rather a coarse-grain indicative of the quality of the embeddings and are not very informative about fine-grain aspects of meaning—or attributes. This gap is thus an important motivation for the present study. Attributes are often used by humans to describe objects (McRae et al., 2005), providing a powerful way to represent knowledge in terms of shape, color, taxonomic information, etc. Several studies have leveraged attributes to build representations that can be used in linguistic and visual tasks. For example, Silberer and Lapata (2014)

used an attribute-based hidden representation from stacked autoencoders as multimodal embeddings. The attributes were learned from both, visual and textual input. In contrast with Silberer and Lapata (2014), here we aim at spotting differences on the fine-grain semantic knowledge encoded by vision and language instead of building multimodal representations and to use them in a task. Furthermore, attribute representations exhibit the advantage that they can be learned transcending the task or the modality at hand, e.g., from either linguistic or visual input. Leveraging this transcendence, Lampert et al. (2009) showed that it is possible to classify objects from unseen classes (i.e., zero-shot learning) by training with attributes specified by humans—such as shape or color—instead of using images. Afterwards, new classes can be identified provided that one has their list of attributes at hand. Further, attributes transcend class boundaries. For instance, the attribute "stripped" can be learned from zebras, bees or tigers (Lampert et al., 2009). In this direction, Farhadi et al. (2009) proposed that the goal of image recognition should be describing rather than naming, that is, for instance, labeling "spotty dog" instead of just "dog" or replacing "unknown class" by "hairy and four-legged."

2.4 Attribute Prediction

The categorization of attributes proposed by McRae et al. (2005) has been widely used in NLP and CV studies (Baroni and Lenci, 2008; Silberer and Lapata, 2014; Rubinstein et al., 2015). Their attribute taxonomy includes individual attributes that belong to more general attribute types (e.g., *tactile* or *taxonomic*). For example, *has_legs* is a *form_and_surface* attribute, while *is_a_bird* or *is_a_fruit* are instances of *taxonomic* attributes.

Rubinstein et al. (2015) distinguished between two types of attributes: *taxonomic* and *attributive* properties. An *attributive* property is any of the remaining attribute types from McRae et al. (2005) categorization that are not taxonomic (Tab. 1). These include attribute types such as *tactile* (e.g., *is soft*), *form_and_surface* (e.g., *is made of metal*) or *encyclopedic* (e.g., *is_dangerous*). By trying to predict attributes using word embeddings as input, Rubinstein et al. (2015) concluded that DMs are significantly better at capturing *taxonomic* attributes rather than *attributive* properties—showing thus a limitation of the distributional hypothesis for certain attributes. Their findings align with previous research that showed that *taxonomic* attributes are generally more abundant in text compared to *attributive* properties (Baroni and Lenci, 2008). While Rubinstein et al. (2015) investigated whether DMs are equally good at capturing each type of attribute, our research questions are different. First, we want to answer whether there are differences between textual and visual representations in the type of attributes that they encode; and second, where these differences are. In other words, we present an inter-modality analysis while Rubinstein et al. (2015) performed only intra-modality comparisons.

In addition to the survey of Rubinstein et al. (2015), the closest work to ours is a study by Bruni et al. (2012) who showed that a very particular type of attribute, namely *color*, is better captured by visual representations than by DMs. Here, we go one step further and compare performance between visual and text embeddings for a larger number of visual attributes, as well as for other non-visual attributes such as *taxonomic*, *functional* or *encyclopedic*.

3 Approach and Experimental Settings

In this section we describe the procedure that we follow in order answer our research questions. An explanatory diagram is shown in Fig. 1.

3.1 Visual Representations

We use ImageNet (Russakovsky et al., 2015) as our source of visual information. ImageNet is currently the largest labeled image bank, with a coverage of 21,841 WordNet synsets (or meanings) (Fellbaum, 1998) and 14,197,122 images. Our choice of ImageNet is motivated by: (i) large word coverage; (ii) images are generally clean and with the relevant object at the foreground; and (iii) replicability of our experiments. Here, we only keep synsets with more than 50 images, and we set an upper bound of 500 images per synset for computational reasons. After this selection, 11,928 synsets are kept.

We extract a 4096-dimensional vector of features for each image using the output of the last layer

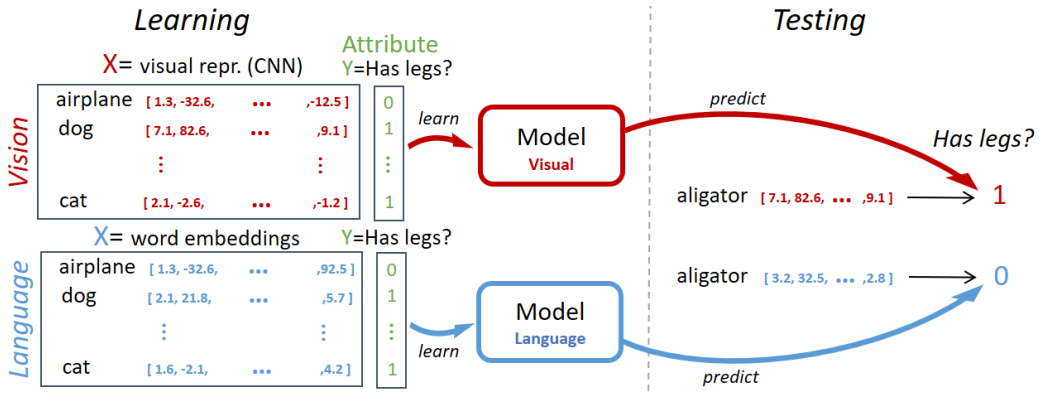

Figure 1: Overview of our experimental setting. Attributes are learned from the embeddings of each modality (left side), and afterwards new concepts are classified on whether the attribute is present or not (classification) or to which degree the attribute is present (regression). For clarity, we omitted the regression problem since its setting is identical to classification except for a continuous output $\mathcal{Y}$ instead of 0/1.

(before the softmax layer) of a pre-trained AlexNet CNN model implemented with Caffe toolkit (Jia et al., 2014). Other than CNN, there exist a variety of methods for obtaining visual features such as SIFT (Lowe, 1999), HOG (Dalal and Triggs, 2005) or SURF (Bay et al., 2006); to name a few. An exhaustive comparison will deviate from the goal of this paper, which is to show that at least some visual embeddings are able to better represent certain attributes than state-of-the-art DMs. Thus, we employ an off-the-shelf CNN model, as CNNs generally outperform the old approaches such as SIFT, HOG or SURF (LeCun et al., 2015). Additionally, we have repeated our experiments with ResNet (He et al., 2015), a more recent CNN model known to outperform AlexNet in image classification. Similar results are obtained with both models, suggesting thus that our vision-language comparisons are relatively independent of the CNN choice.

For each concept, several ways of integrating the representations from its individual images into a single vector could be devised. Here, we apply the following two common approaches (Kiela and Bottou, 2014):

(i) Averaging: Computes the component-wise average of the CNN feature vectors of individual images. This is equivalent to the cluster center of the individual representations.

(ii) Maxpool: Computes the component-wise maximum of the CNN feature vectors of individual images. This approach makes sense intuitively because CNN vector components can be interpreted as "visual properties."

For simplicity of notation we henceforth refer to the averaged and maxpooled visual representations as VIS_{avg} and VIS_{max} respectively.

3.2 Word Embeddings

We employ 300-dimensional GloVe vectors (Pennington et al., 2014) pre-trained in the largest available corpus (840B tokens and a 2.2M words vocabulary from Common Crawl corpus) from the author's website[1]. For completeness, we have repeated our experiments with word2vec embeddings (Mikolov et al., 2013) and we have found GloVe to perform slightly better. Thus, we report results with GloVe as it provides a stronger baseline to compare visual representations with.

3.3 McRae et al. Dataset

The data set collected by McRae et al. (2005) consists of data gathered from human participants that were asked to list properties—attributes—of concrete nouns. For each noun, 30 participants listed its attributes. For example, for "airplane", the attribute *has_wings* (i.e., a *form_and_surface* attribute) was

[1] http://nlp.stanford.edu/projects/glove

listed by 20 subjects, while the attribute *used_for_travel* (i.e., a *function* attribute) was listed by 7. The McRae et al. (2005) data contains 541 concepts, 2,526 different attributes, and 10 attribute types.

3.4 Binary Classification Setup

To evaluate fine-grain semantic understanding of the different embeddings, we evaluate how well the attributes from McRae et al. (2005) can be predicted by using the embeddings as input (Fig. 1). We use both a classification and a regression setting.

For each attribute a, we build a data set with the concepts to which this attribute applies as the positive class instances and the rest of concepts form the negative class. For example, a "beetle" is a negative instance and "airplane" a positive instance for the attribute a = *is_large*. We consider that an attribute applies to a noun if a minimum of 5 people have listed it[2]. Table 1 shows a summary of the number of positive class concepts per attribute type. For each attribute a we learn a predictor:

$$f_a : \mathcal{X} \rightarrow \mathcal{Y}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is the input space of (d-dimensional) concept representations and $\mathcal{Y} = \{0, 1\}$ the binary output space.

To guarantee that the number of positive instances is enough to actually learn the attributes, only attributes with at least 25 positive instances are kept. This leads to a total of 43 attributes (Fig. 3), which can be seen as a total of 43 data sets. The concept selection in ImageNet described in Sect. 3.1 results in a visual coverage of 400 concepts (out of 541 from McRae et al. (2005) data), and, for a fair vision-language comparison, only the word embeddings (from GloVe) of these nouns are employed. Hence, our training data $\{(\overrightarrow{x_i}, y)\}_{i=1}^{400}$ consists of 400 instances. Since we have three types of representations (GloVe, VIS_{avg} and VIS_{max}), three different models f_a^{GloVe}, f_a^{avg} and f_a^{max} are learned from the three different input data $X_{GloVe} \in \mathbb{R}^{400\times300}$, $X_{avg} \in \mathbb{R}^{400\times4096}$ and $X_{max} \in \mathbb{R}^{400\times4096}$ respectively, where $X = \{\overrightarrow{x_i}\}_{i=1}^{400}$ and $\overrightarrow{x_i} \in \mathcal{X}$, $y \in \mathcal{Y}$.

Classification performance is evaluated with the F1 measure of the positive class—that is, the harmonic mean of precision and recall—since F1 is insensitive to class imbalance.

Attribute type	# Attr.	Avg. # concepts	SD
encyclopedic	4	32.7	1.5
function	3	46	27.9
sound	1	34	-
tactile	1	26	-
taste	1	33	-
taxonomic	7	42	24.8
color	7	42.4	12.0
form_and_surface	14	63.7	29.9
motion	4	37.5	5.7

Table 1: Attribute types, number of attributes per attribute type (# Attr.), and average number of concepts (i.e., positive instances) per attribute type (Avg. # concepts) with their respective standard deviations (SD).

3.5 Regression Setup

Let us consider the same scenario as in the classification one above, yet with a continuous output space $\mathcal{Y} = [0, 1]$ instead of a binary. The proportion of participants (out of 30) who have listed the attribute a for a given concept is taken as ground truth $y \in \mathcal{Y}$. This can be interpreted as the saliency of attribute a for this concept. Regression performance is evaluated with the Spearman ρ correlation coefficient between the predicted outputs and the ground truth.

[2] This threshold was set by McRae et al. (2005)

3.6 Experimental Setup

In both classification and regression we perform 2 runs of 5-fold stratified cross validation. That is, we create 5 (stratified) disjoint splits, repeating it with two different seeds. The use of 5 folds is convenient since the data set is small and contains just a few instances of the minority class—which ranges from 6.25% to 30.5% of the data. Thus, the use of $4/5$-th of the data for learning (and $1/5$-th for testing) is more likely to yield well-learned attributes than smaller training proportions. To handle class imbalance in classification we set the training class weights inversely proportional to the class priors.

This work relies on the basic assumption that, if a given attribute can be predicted using some embeddings as input (e.g., by means of a classifier or a regressor), then these embeddings contain encoded information about this attribute. The inverse is not necessarily true, that is, if an attribute cannot be predicted from a given embedding, this does not necessarily mean that the information is not present. In this case, the bottleneck might be any sort of technical issue such as our classifier choice, regularizer choice, data scaling, etc. Thus, in order to validate our conclusions, we repeated the same experiments with different classifier choices (SVM, logistic regression, bagging ensemble of decision trees and AdaBoost); different regressors (SVM, neural networks, ensemble of regression trees and gradient boosting); and data scalings (max/min scaling and component-wise centering plus normalizing by the standard deviation). We found results to be notably stable across classifier and regressor choices. We report results with a linear SVM regressor and a linear SVM classifier, both implemented with the scikit machine learning toolkit (Pedregosa et al., 2011). Data scaling affected only the performance of VIS_{max}, conceivably because their values are extreme by definition.

4 Results and Discussion

4.1 Intra-Modality Performance per Attribute Type

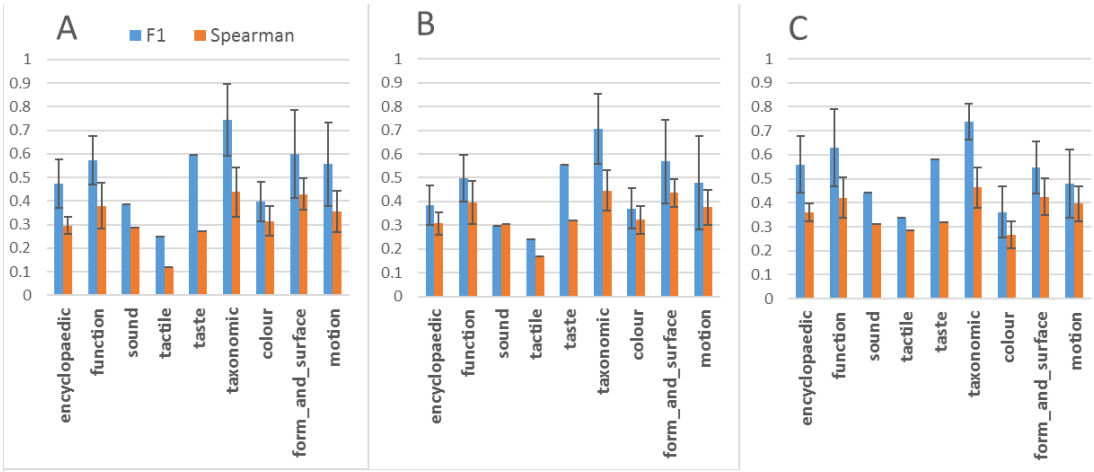

Figure 2: Averages of F1 (classification) and Spearman (regression) measures per attribute type (i.e., averaging individual attributes) for VIS_{avg} (A), VIS_{max} (B) and GloVe (C). Error bars show standard error.

As a first noteworthy finding, one may observe from Fig. 2 the—perhaps unexpected—resemblance that visual and textual representations present at predicting different types of attributes, which suggests interesting commonalities between visual and textual representations. Further, this seems to indicate that some attribute types are genuinely more difficult to predict than others (e.g., *tactile*), conceivably because the nouns to which these attributes apply tend to have little in common in terms of both, visual resemblance and word co-occurrences in similar contexts. This can be further appreciated in Fig. 5 (bottom row) from the scattered pattern of the attribute *is_soft*, which proves to be a difficult target for both, vision and language. In turn, this suggests that neither vision nor language might be sufficiently informative about certain attribute types (e.g., *tactile* or *sound*).

From Fig. 2 (C) it can also be observed that our results align with Rubinstein et al. (2015)'s findings with DMs. That is, from text-only embeddings, *taxonomic* attributes can be generally more accurately predicted than most of the *attributive* properties (i.e., all attribute types from Tab. 1 except *taxonomic*). We additionally find a similar behavior for visual embeddings (Fig. 2 A and B).

4.2 Visual Vs. Text Performance

Fig. 3 provides an answer to our first research question, showing that, clearly, neither vision nor language absolutely dominates the other in grasping fine-grain semantic knowledge but they rather show preference for different attributes. In general, visual embeddings (especially VIS_{avg}) perform better than GloVe in three main attribute types: *motion*, *form_and_surface* and *color* (Fig. 3 and 4). On the other hand, GloVe clearly outperforms vision in *encyclopedic* and *function* attribute types (Fig. 3 and 4), which are seldom visual. For the *taxonomic* type, vision or language clearly dominate in different individual attributes (Fig. 3). The visual performance gains with respect to GloVe (in e.g., *is_a_bird*) are particularly interesting since previous research evidenced that *taxonomic* is the attribute type where text-only DMs are the strongest (Baroni and Lenci, 2008; Rubinstein et al., 2015). Hence, these results suggest that the representation of taxonomic knowledge can further benefit from visual grounding.

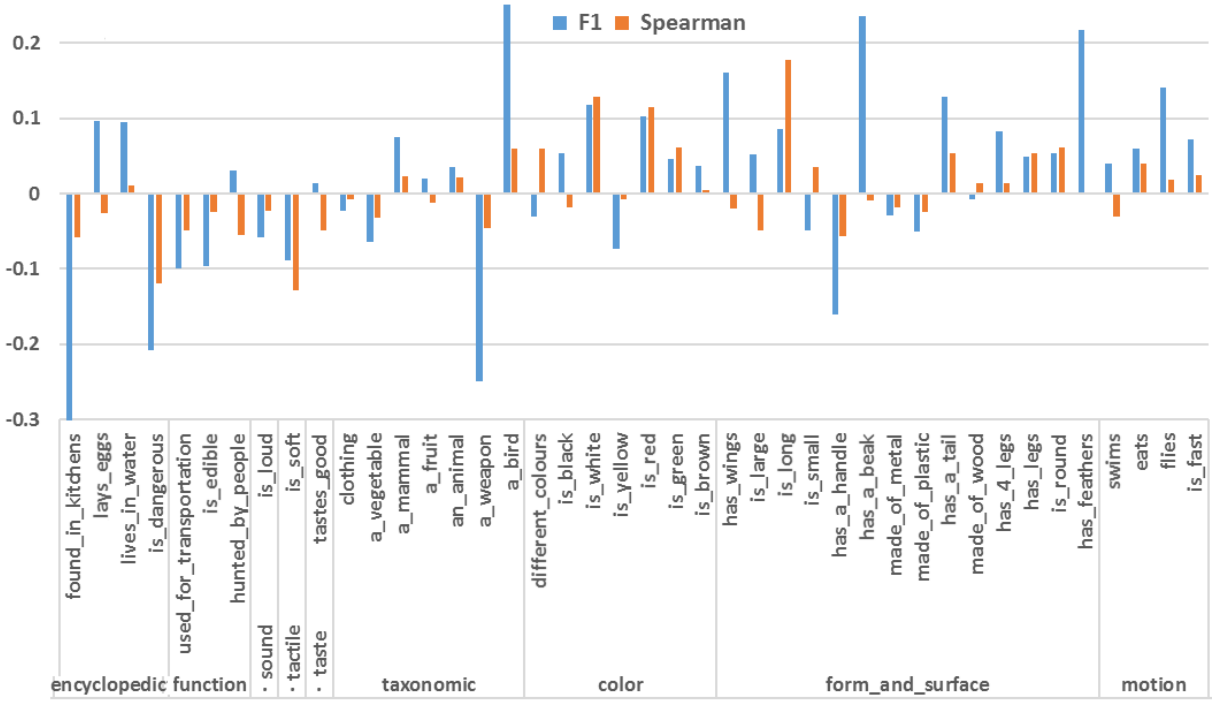

Figure 3: Difference of performance between VIS_{avg} minus GloVe. Attributes are shown on the horizontal axis and grouped by their type. Positive bars indicate better performance of visual embeddings and negative bars otherwise. Results with VIS_{max} are omitted as they exhibit almost identical patterns as VIS_{avg}, yet slightly worse.

Interestingly, even in the attribute types where either vision or language generally dominate, there are exceptions. For example, VIS_{avg} seems to outperform GloVe in classifying *lays_eggs* (i.e., an *encyclopedic* attribute), while GloVe seems to capture better *has_a_handle* (a *form_and_surface* attribute) which is predominantly visual. It is important to notice that visual attributes do not equal "less abstract" knowledge. For example, the visual attribute *has_a_handle* clearly requires more abstract semantic understanding than purely sensory visual attributes such as *is_green* since the definition of "handle" is clearly functionally-motivated rather than visual. For instance, the ball-shaped handle of a door has virtually no visual resemblance with the handle of a bag, yet they both have the same function. All this suggests that not only the attribute type is important but there are other factors to be taken into account. More concretely, the visual resemblance among objects to which the same attribute applies

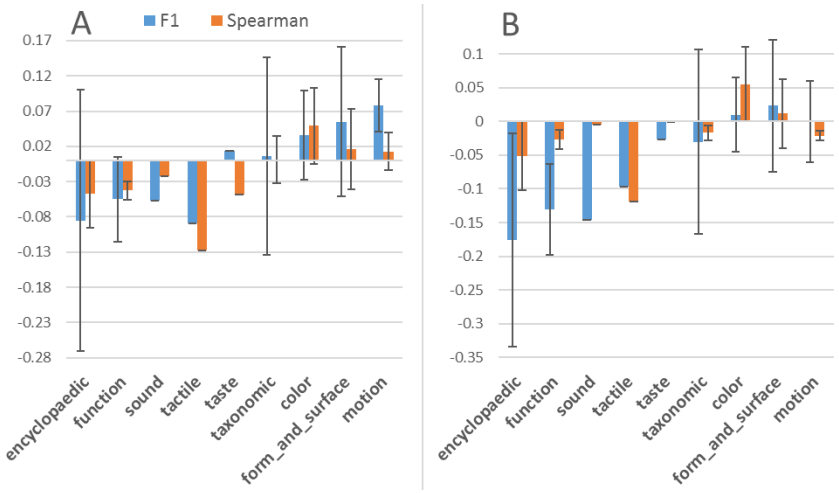

Figure 4: Averages of performance difference per attribute type. For each attribute type (e.g., taxonomic, taste, etc.), the bar indicates the average performance difference of its set of attributes. Plot A shows performance difference between VIS_{avg} and GloVe and B between VIS_{max} and GloVe. As in Fig. 3, positive bars indicate better performance of visual embeddings and negative bars otherwise. Error bars show standard error.

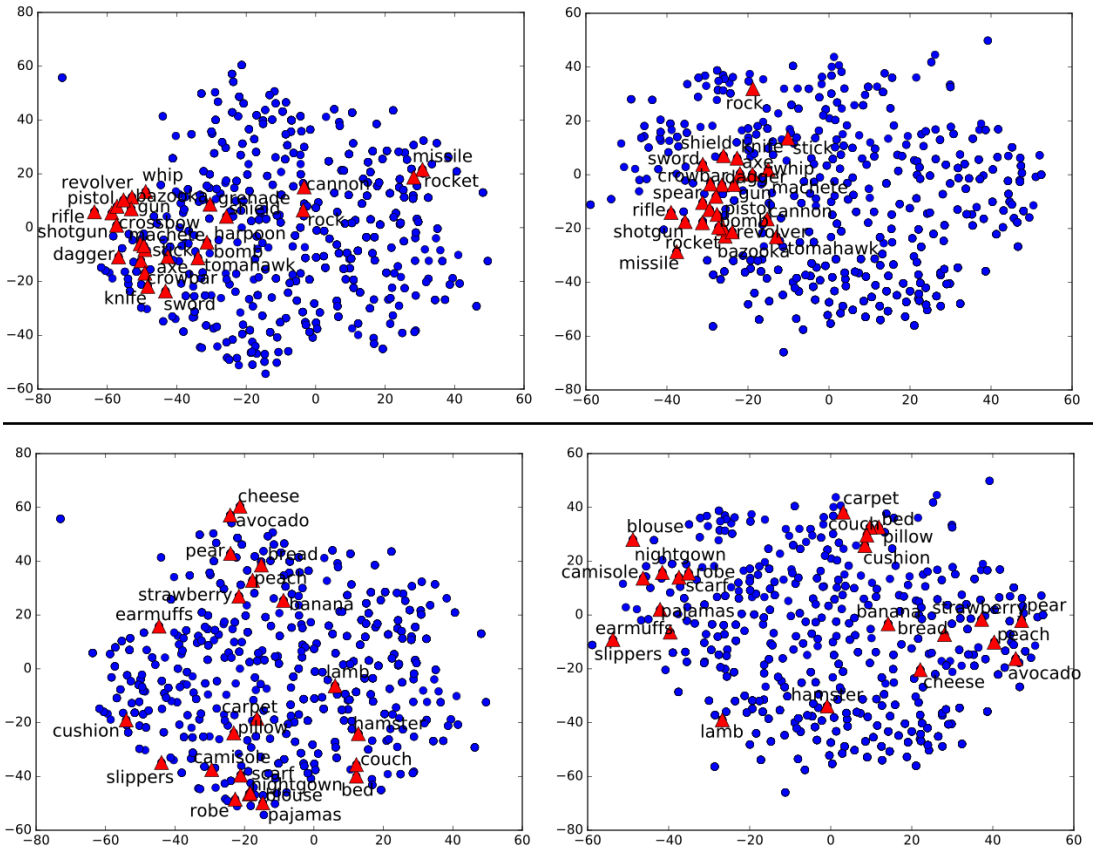

Figure 5: T-SNE visualization (Maaten and Hinton, 2008) of the VIS_{avg} embeddings (left column) and of the GloVe embeddings (right column). Red triangles show the positive class for the attribute *a_weapon* (top row) and *is_soft* (bottom row), while blue circles correspond to the negative class words.

seems to play a role. For example, *a_weapon* and *is_a_bird* are both *taxonomic* attributes although GloVe clearly dominates in the first and vision in the second one (Fig. 3). A closer inspection reveals that

the concepts from *is_a_bird* (e.g., "canary", "chicken", "eagle", "penguin") exhibit an important visual resemblance, whilst those from *a_weapon* (e.g., "rock", "bow", "rifle", "axe") do not. From Fig. 5 (top row) one can observe that the instances of *a_weapon* present a more scattered pattern in the visual embeddings (left) than in the linguistic ones (right). The same applies to *has_a_handle*, which one would perhaps—wrongly—expect that it might be better captured by vision. Contrarily, visual resemblance generally yields vectors that are closer in the visual space, making class boundaries easier to learn.

Even though the performances of classification and regression models (F1 and Spearman, respectively) are markedly correlated (Fig. 2, 3 and 4), there are a few exceptions. Small "contradictions" are plausible considering that classification and regression are two different problems—detection and estimation respectively—in which the learner is exposed to different (output) data. However, just a few bars show an opposite sign, none of them showing extreme opposite values.

5 Conclusions and Future Work

Overall, the present study adds evidence to the fact that visual and textual representations encode different semantic aspects of concepts. Crucially, we find that neither vision nor language are superior to the other in grasping every aspect of meaning, but they rather dominate in different attribute types. More concretely, vision proves generally better at capturing *form_and_surface*, *color* and *motion* attributes while language proves better at *encyclopedic* and *function* attributes. However, even within these general trends, we find that there are important exceptions in which visual resemblance among the concepts to which an attribute applies seems to play an important role. As an additional finding, we find that vision and language present a surprisingly similar pattern of predictive capacity for the different attributes types (Fig. 2), and that neither vision nor language succeed at capturing certain attribute types (e.g., *tactile* or *sound*). This suggests that other perceptual modalities can further improve the representations.

Taken together, we conclude that fine-grain attribute understanding can benefit from the integration of visual and textual representations. Thus, our results align with previous multimodal research that evaluates vision and language in coarse-grain tasks such as concept similarity and conclude that multimodal representations outperform the unimodal baselines (Lazaridou et al., 2015; Silberer and Lapata, 2014; Kiela and Bottou, 2014). Ultimately, our findings provide insights that can help building better multimodal representations by taking into account the types of semantic knowledge that vision and language selectively capture. In turn, better representations can improve automatic language understanding by providing a more human-like and perceptually grounded processing. Furthermore, we believe that the taxonomic knowledge encoded in visual representations can be further exploited towards building methods that automatically identify taxonomic relationships between concrete (perceptible) nouns, offering potential alternatives to WordNet (Fellbaum, 1998). Finally, recent work in computer vision showed that better fine-grain semantic understanding of objects can be achieved if images are segmented into the objects' individual parts (Vedaldi et al., 2014). We believe that this is an interesting direction to extend our work.

Acknowledgements

This work has been supported by the CHIST-ERA EU project MUSTER[3].

References

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

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

[3] http://www.chistera.eu/projects/muster

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *ECCV*, pages 404–417. Springer.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *ACL*, pages 136–145. ACL.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE.

Koen Deschacht and Marie-Francine Moens. 2009. Using the latent words language model for semi-supervised semantic role labeling. In *EMNLP*.

Koen Deschacht, Jan De Belder, and Marie-Francine Moens. 2012. The latent words language model. *Computer Speech & Language*, 26(5):384–409.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. IEEE.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157. IEEE.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. Some like it hot-visual guidance for preference prediction. *arXiv preprint arXiv:1510.07867*.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *ACL*, volume 2, pages 726–730.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL*, pages 721–732.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *JAIR*, 37(1):141–188.

Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, David Weiss, et al. 2014. Understanding objects in detail with fine-grained attributes. In *CVPR*, pages 3622–3629.