Building a *bagpipe* with a *bag* and a *pipe*: Exploring Conceptual Combination in Vision [*]

Sandro Pezzelle and **Ravi Shekhar**[†] and **Raffaella Bernardi**
CIMeC - Center for Mind/Brain Sciences, University of Trento
[†]DISI, University of Trento
{firstname.lastname}@unitn.it

Abstract

This preliminary study investigates whether, and to what extent, conceptual combination is conveyed by vision. Working with noun-noun compounds we show that, for some cases, the composed visual vector built with a simple additive model is effective in approximating the visual vector representing the complex concept.

1 Introduction

Conceptual combination is the cognitive process by which two or more existing concepts are combined to form new complex concepts (Wisniewski, 1996; Gagné and Shoben, 1997; Costello and Keane, 2000). From a linguistic perspective, this mechanism can be observed in the formation and lexicalization of compound words (eg. *boathouse*, *swordfish*, *headmaster*, etc.), a widespread and very productive linguistic device (Downing, 1977) that is usually defined in literature as the result of the composition of two (or more) existing and free-standing words (Lieber and Štekauer, 2009). Within both perspectives, scholars agree that the composition of concepts/words is something more than a simple addition (Gagné and Spalding, 2006; Libben, 2014). However, additive models turned out to be effective in language, where they have been successfully applied to distributional semantic vectors (Paperno and Baroni, to appear).

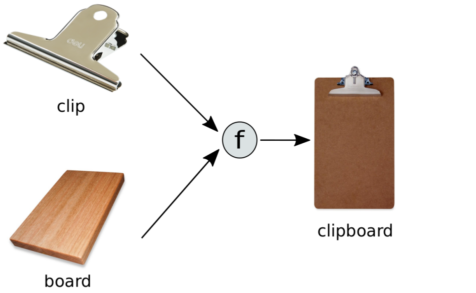

Figure 1: Can we obtain a *clipboard* by combining *clip* and *board* with a compositional function *f*?

Based on these previous findings, the present work addresses the issue of whether, and to what extent, conceptual combination can be described in vision as the result of adding together two single concepts. That is, can the visual representation of *clipboard* be obtained by using the visual representations of a *clip* and a *board* as shown in Figure 1? In order to investigate this issue, we experiment with visual features that are extracted from images representing concrete and imageable concepts. More precisely, we use noun-noun compounds for which ratings of imageability are available. The rationale for choosing NN-compounds is that composition should take advantage from dealing with concepts for which clear, well-defined visual representations are available, as it is the case of nouns (representing objects). In particular, we test whether a simple additive model can be applied to vision in a similar fashion to how it has been done for language (Mitchell and Lapata, 2010). We show that for some NN-compounds the visual representation of the whole can be obtained by simply summing up its parts. We also discuss cases where the model fails and provide conjectures for more suitable approaches. Since, to our knowledge, no datasets of images labeled with NN-compounds are currently available, we manually build and make available a preliminary dataset.

[*]We are grateful to Marco Baroni and Aurélie Herbelot for the valuable advice and feedback. This project has received funding from ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research.

Proceedings of the 5th Workshop on Vision and Language, pages 60–64,
Berlin, Germany, August 12 2016. ©2016 Association for Computational Linguistics

2 Related Works

Recently, there has been a growing interest in combining information from language and vision. The reason lies on the fact that many concepts can be similar in one modality but very different in the other, and thus capitalizing on both information turns out to be very effective in many tasks. Evidence supporting this intuition has been provided by several works (Lazaridou et al., 2015; Johnson et al., 2015; Xiong et al., 2016; Ordonez et al., 2016) that developed multimodal models for representing concepts that outperformed both language-based and vision-based models in different tasks. Multimodal representations have been also used for exploring compositionality in visual objects (Vendrov et al., 2015), but compositionality was intended as combining two or more objects in a visual scene (eg., an apple and a banana) and not as obtaining the representation of a new concept based on two or more existing concepts.

Even though some research in visual compositionality has been carried out for part segmentation tasks (Wang and Yuille, 2015), we focus on a rather unexplored avenue. To our knowledge, the closest work to ours is represented by Nguyen et al. (2014), who used a compositional model of distributional semantics for generating adjective-noun phrases (eg., a *red car* given the vectors of *red* and *car*) both in language and vision. According to their results, a substantial correlation can be found between observed and composed representations in the visual modality. Moving from these results, the present study addresses the issue of whether, and to which extent, a compositional model can be applied to vision for obtaining noun-noun combinations, without relying on linguistic information.

3 Dataset

To test our hypothesis, we used the publicly available dataset by Juhasz et al. (2014). It contains 629 English compounds for which human ratings on overall imageability (ie., a variable measuring the extent to which a compound word evokes a nonverbal image besides a verbal representation) are available. We relied on this measure for carrying out a first filtering of the data, based on the assumption that the more imageable a compound, the clearer and better-defined its visual representation. As a first step, we selected the most imageable items in the list by retaining only the ones with an average score of at least 5 points in a scale ranging from 1 (e.g., *whatnot*: 1.04) to 7 (e.g., *watermelon*: 6.95). From this subset, including 240 items, one of the authors further selected only genuine noun-noun combinations, so that items like *outfit* or *handout* were discarded. We then queried each compound and its constituent nouns in Google images and we selected only those items for which every object in the tuple (eg. *airplane*, *air*, and *plane*) had a relatively good visual representation by looking at the top 25 images. This step, in particular, was aimed at discarding the surprisingly numerous cases for which only noisy images (ie., representing brands, products, or containing signs) were available.

From the resulting dataset, containing 115 items, we manually selected those that we considered as compositional in vision. As a criterion, only NN-combinations that can be seen as resulting from either combining an object with a background (e.g., *airplane*: a *plane* is somehow superimposed in the *air* background) or concatenating two objects (e.g., *clipboard*) were selected. Such a criterion is consistent with our purpose, that is finding those cases where visual composition works. The rationale is that there should be composition when both the constituent concepts are present in the visual representation of the composed one. Two authors separately carried out the selection procedure, and the few cases for which there was disagreement were resolved by discussion. In total, 38 items were selected and included in what we will heceforth refer to as **compositional group**. Interestingly, the two visual criteria followed by the annotators turned out to partly reflect the kind of semantic relation implicitly tying the two nouns. In particular, most of the selected items hold either a noun2 HAS noun1 (eg., *clipboard*) or a noun2 LOCATED noun1 (eg., *cupcake*) relation according to Levi (1978).

In addition, 12 other compounds (eg., *sunflower*, *footstool*, *rattlesnake*, etc.) were randomly selected from the 115-item subset. We will heceforth refer to this set as the **control group**, whereas we will refer to the concatenation of the two sets (38+12=50 items) as the **full group**. For each compound in the full group, we manually searched images representing it and each of its constituents nouns in Google images. One good image, possibly showing the most prototypical representation of that concept according to the authors' experience, was selected. In total, 79 images for N-constituents plus 50 images for NN-compounds (129 in total) images were included in our dataset.[1]

4 Model

In order to have a clear and interpretable picture of what we obtain when composing visual features of nouns, in this preliminary study we experimented with a simple additive compositional model. Simple additive models can be seen as weighting models applying the same weight to both elements involved. That is, when composing *waste* and *basket*, both nouns are considered as playing the same (visual) role with respect to the overall representation, ie. *wastebasket*. Intuitively enough, we expect this function being effective in approximating visual representations of complex concepts where the parts are still visible (eg., *clipboard*). In contrast, we don't expect good results when the composition requires more abstract, subtle interactions between the nouns (eg., *cannonball*).

To directly compare vision against language, we applied the same compositional function to the linguistic vectors (extracted from large corpora of texts) representing the same dataset. What we expected from such a comparison is a different and possibly complementary behavior: since linguistic vectors encode contexts in which the target word is very likely to occur, language could be more effective in modulating abstract interactions (ie., *cannonball*), whereas vision might be possibly better in composing grounded concepts (ie., *clipboard*). As a consequence, we expect language performing better in the control group, but differently from vision in the compositional group.

4.1 Visual Features

Each image in the dataset is represented by visual features extracted by using state-of-the-art technique based on Convolutional Neural Networks (Simonyan and Zisserman, 2014). We used the VGG-19 model pretrained on the ImageNet ILSVRC data (Russakovsky et al., 2015). The model includes multiple convolutional layers followed by max pooling and the top of these are fully connected layers ($fc6$, $fc7$, $fc8$). We used 4096-dimensional visual vectors extracted from the $fc6$ layer, which has shown better performance in image retrieval/matching task (Babenko et al., 2014) compared to other layers. For experimental purpose, we used MatConvNet (Vedaldi and Lenc, 2015) toolbox for features extraction.

4.2 Linguistic Features

Each word in the dataset is represented by a 400-dimension vector extracted from a semantic space[2] built with the CBOW architecture implemented in the word2vec toolkit (Mikolov et al., 2013) and the best-performing parameters in Baroni et al. (2014).

5 Evaluation Measures

To evaluate the compositionality of each NN-compound, we measure the extent to which the composed vector is similar to the corresponding observed one, ie. the vector directly extracted from either texts or the selected image. Hence, first of all we use the standard $Cosine$ similarity measure. The higher the similarity, the better the composition. It could be the case that the composed vector is however less similar to the observed one than it is the closest N-constituent. Thus, similarity by its own is not informative of whether the composition function has provided additional information compared to that conveyed by the closest single noun. In order to take into account this issue, we also compute the similarity between the composed vector and both its N-constituents ($N1$,$N2$). We lower the similarity between the composed and the observed vector by subtracting the similarity between the observed vector and the noun that is closest to it (we call this measure $CompInfo$, since it is informative of the effectiveness of the composition). When the composition operation maps the composed vector closer to the observed vector compared to its constituents in the semantic space, the composition provides more information. In particular, when $CompInfo$ is positive (ie., greater than 0), the composition is considered to be effective.

To further evaluate the compositionality of the nominal compound, we test the effectiveness of the composed vector in the retrieval task. The reason is to double-check the distictiveness of the composed vector with respect to all the objects (ie., 79 N-constituents plus 50 NN-compounds) in

[1] The dataset is publicly available and can be downloaded at `https://github.com/shekharRavi/`

[2] The corpus used for building the semantic space is a 2.8 billion tokens concatenation of the web-derived ukWac, a mid-2009 dump of the English Wikipedia, and the British National Corpus.

Table 1: Compositionality evaluation in Vision and Language.

Dataset	$Avg.Similarity$		$\%(CompInfo > 0)$		$Rec@1$		$Rec@5$	
	Vision	Lang	Vision	Lang	Vision	Lang	Vision	Lang
Full	0.6283	0.407	62%	72%	0.34	0.52	0.76	0.88
Compositional	0.6476	0.429	76.31%	76.31%	0.3947	0.57889	0.8158	0.9211
Control	0.5671	0.3377	16.66%	58.33%	0.1667	0.3333	0.5833	0.75

the semantic space. Using the composed vector as query, we are interested in knowing the rank of the corresponding observed vector. Since for each query there is only one correct item in the whole semantic space, the most informative retrieval measure is Recall. Hence, we evaluate compositionality by $Rec@k$. Since we have already scrutinized the role of the N-constituents with the previous measure, in the retrieval of a NN-compound both its N-constituents are removed from the semantic space. The same evaluation is conducted for both vision and language, thus providing a way to directly compare the two modalities.

6 Results

In Table 1, we report average similarity, percentages of cases where $CompInfo$ is positive (ie., composition is informative), and both $Rec@1$ and $Rec@5$. As can be seen, all measures are significantly higher for the compositional group than for the control group both in visual and linguistic modality. Focusing on vision, the cases in which composition provides additional information compared to the closest N-constituent drops from 76.3% of the compositional group to 16.6% of the control group. Interestingly, the same trend is confirmed by Similarity and Recall measures. This confirms the intuition that for combinations involving either superimposition of an object over a background or object concatenation the composition can be obtained with a simple additive model. It also confirms that a large number of conceptual combinations cannot be composed with a simple additive model, as shown by the randomly choosen items of the control group. Evidence for a real effectiveness of the composition is also provided by the analysis of the neighbors (ie., the closest vectors) of the working cases and their constituent nouns. For example, the observed *wastebasket* is the closest neighbor of the composed *wastebasket*, but it is not even in the top 2 positions in both *waste* (*hail*, *sunshine*) and *basket* (*cup*, *clipboard*).

By comparing vision and language, two main differences emerge. First, the average similarity in each group is significantly lower in language compared to the visual modality. That is, the composed and the observed vectors are on average closer in vision than in language[3]. Second, a different drop in the percentage of working cases can be observed between the compositional and the control group in language and vision. Whereas the percentage of working cases in the compositional group is exactly the same between the two modalities (76.3%), the performance in the linguistic control group is significantly higher than in its visual counterpart (ie., 58.3% vs 16.6%). That is, randomly choosen items are not compositional in vision, but compositional to some extent in language. Interestingly, the same percentage of working cases (76.3%) between the two modalities in the compositional group does not result from the same items. To illustrate, *bagpipe* turns out to be compositional in vision but not in language, whereas *corkscrew* is compositional in language but not in vision. Consistently with our hypothesis, *corkscrew* would require more than the grounded information provided by the visual representations of *cork* and *screw*. In contrast, summing together *bag* and *pipe* gives something similar to a *bagpipe* in vision, but not in language.

7 Conclusions

A simple additive model is effective in generating composed representations that approximate the observed representations for NN-combinations made up by either superimposed or concatenated objects. On the other hand, the same method cannot be applied to the full range of NN-compounds,

[3] One could think that this difference is due to the different setting used for the two modalities: the visual vectors encode one image vs. the linguistic vectors encode all the contexts in which the word is used. However, this is in not the case, since we have observed the same behavior (for the cases where compositionality works in vision) on a previous study carried out on large image datasets.

as the results on the control group reveal. This suggests that new compositional methods (perhaps capitalizing on both language and vision) are required to solve this task for all cases. In this light, we believe our dataset is a good starting point for any future investigation.

References

Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural codes for image retrieval. In *Computer Vision–ECCV 2014*, pages 584–599. Springer.

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 (1)*, pages 238–247.

Fintan J Costello and Mark T Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349.

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.

Christina L Gagné and Edward J Shoben. 1997. Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):71.

Christina L Gagné and Thomas L Spalding. 2006. Conceptual combination: Implications for the mental lexicon. In *The representation and processing of compound words*, chapter 7, pages 145–168. Oxford University Press Oxford, New York.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2015. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*.

Barbara J Juhasz, Yun-Hsuan Lai, and Michelle L Woodcock. 2014. A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior research methods*, pages 1–16.

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

Judith N Levi. 1978. *The syntax and semantics of complex nominals*. Academic Press, New York.

Gary Libben. 2014. The nature of compounds: A psychocentric perspective. *Cognitive neuropsychology*, 31(1-2):8–25.

Rochelle Lieber and Pavol Štekauer, editors. 2009. *The Oxford Handbook of Compounding*. Oxford University Press, New York.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Dat Tien Nguyen, Angeliki Lazaridou, and Raffaella Bernardi. 2014. Coloring objects: adjective-noun visual semantic compositionality. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 112–114.

Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2016. Learning to name objects. *Communications of the ACM*, 59(3):108–115.

Denis Paperno and Marco Baroni. to appear. When the whole is less than the sum of its parts: How composition affects PMI values in distributional semantic vectors. Accepted for publication in Computational Linguistics.

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. *International Journal of Computer Vision*, 115(3):211–252.

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

Andrea Vedaldi and Karel Lenc. 2015. *MatConvNet – Convolutional Neural Networks for MATLAB*. Proceeding of the ACM Int. Conf. on Multimedia.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

Jianyu Wang and Alan L Yuille. 2015. Semantic part segmentation using compositional model combining shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797.

Edward J Wisniewski. 1996. Construal and similarity in conceptual combination. *Journal of Memory and Language*, 35(3):434–453.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. *arXiv preprint arXiv:1603.01417*.