Coloring Objects: Adjective-Noun Visual Semantic Compositionality

Dat Tien Nguyen[(1,2)] **Angeliki Lazaridou**[(2)] **Raffaella Bernardi**[(2)]
[(1)]EM LCT, [(2)]University of Trento/ Italy
name.surname@unitn.it

Abstract

This paper reports preliminary experiments aiming at verifying the conjecture that semantic compositionality is a general process irrespective of the underlying modality. In particular, we model compositionality of an attribute with an object in the visual modality as done in the case of an adjective with a noun in the linguistic modality. Our experiments show that the concept topologies in the two modalities share similarities, results that strengthen our conjecture.

1 Language and Vision

Recently, fields like computational linguistics and computer vision have converged to a common way of capturing and representing the linguistic and visual information of atomic concepts, through vector space models. At the same time, advances in computational semantics have lead to effective and linguistically inspired approaches of extending such methods from single concepts to arbitrary linguistic units (e.g. phrases), through means of vector-based semantic composition (Mitchell and Lapata, 2010).

Compositionality is not to be considered only an important component from a linguistic perspective, but also from a cognitive perspective and there has been efforts to validate it as a general cognitive process. However, in computer vision so far compositionality has received limited attention. Thus, in this work, we study the phenomenon of *visual compositionality* and we complement limited previous literature that has focused on event compositionality (Stöttinger et al., 2012) or general image structure (Socher et al., 2011), by studying models of attribute-object semantic composition.

In a nutshell, our work consists of learning vector representations of attribute-object (e.g., "red car", "cute dog" etc.) and objects (e.g., "car", "dog", "truck", "cat" etc.) and by using those compute the representation of new objects having similar attributes ("red truck", "cute cat" etc.). This question has both theoretical and applied impact. The possibility of developing a visual compositional model of attribute-object, on the one hand, could shed light on the acquisition of such ability in humans; how we learn attribute representation and compose them with different objects is still an open question within the cognitive science community (Mintz and Gleitman, 2002). On the other hand, computer vision systems could become generative and be able to recognize unseen attribute-object combinations, a component especially useful for object recognition and image retrieval.

2 Visual Compositional Model

As our source of inspiration regarding the type of compositionality, we use the *Lexical Functional* model (LF) (Baroni and Zamparelli, 2010), under which adjectives, in linguistic compositionality, are represented as linear functions (i.e., matrix of weights). Concretely, each adjective function f^W_{adj} is induced from corpus-observed vectors of adjective-noun phrases $w_i \in W_{phrase}$ and noun $w_j \in W_{noun}$, e.g., $\langle (w_{red\ car}, w_{car}), (w_{red\ flag}, w_{flag}), \ldots \rangle$, by solving the least-squares regression problem:

$$\underset{f^W_{adj} \in \mathbb{R}^{d \times d}}{\arg\min} ||W_{phrase} - f^W_{adj} W_{noun}||$$

Proceedings of the 25th International Conference on Computational Linguistics, pages 112–114,
Dublin, Ireland, August 23-29 2014.

In this work, we propose to import the LF method in the visual modality, aiming at developing a *Visual Compositional Model*. Similarly to the case of linguistic compositionality, each attribute function f^V_{attr} is induced from image-harvested vector representations of attribute-object $v_i \in V_{phrase}$ and object $v_j \in V_{object}$, e.g. for training the function f^V_{red} the following data can be used $\langle (v_{red\ car}, v_{car}), (v_{red\ flag}, v_{flag}), \ldots \rangle$.

3 Experiments

The visual representations of attribute-objects and objects are created with the PHOW-color features (Bosch et al., 2007) and SIFT color-agnostic features (Lowe, 2004) respectively. The linguistic representations for the adjective-noun W_{phrase} and noun W_{noun} are built with the word2vec toolkit[1] using a corpus of 3 billion tokens.[2] Both visual and linguistic representations consist of 300 dimensions.

In this work, we focus on attributes related to 10 colors (Russakovsky and Fei-Fei, 2012) for a total number of 9699 images depicting 202 unique objects/nouns and 886 unique phrases (attribute-object/adjective-noun). Our experiments are conducted with *aggregated attribute-object* representations obtained by summing the visual vectors extracted from images representing the same attribute-object, The same pipeline is followed for the objects to obtain *aggregated object* vectors.

This work aims at comparing the behavior of the semantically-driven compositionality process across the two modalities. For this reason, we report results on the intersection of V_{phrase} and W_{phrase}, a process that results in 266 attribute-object/adjective-noun items. Furthermore, although the training data for the two modalities are different, the size of the training data is identical, i.e., the f^V_{attr} is trained using the remaining 620 attribute-object items, whereas for the f^W_{adj}, we randomly sample 620 adjective-noun items from the language space.

3.1 Analysis of Language and Visual Semantic Spaces

This experiment aims at assessing the degree to which language and vision share commonalities. To this end, we compute the cosine similarities between all possible combination of objects (resp., nouns) and perform a correlation analysis of the similarity of the corresponding pairs in the two lists resulting in **0.45** Spearman correlation – e.g., we correlate the similarity between v_{cat} and v_{dog} with that between w_{cat} and w_{dog}. For instance, "goat" and "sheep" are highly similar in both spaces, whereas "whale" and "bird" are similar only linguistically, whereas "blackboard" and "chair" are similar only visually. The same experiment is performed between all possible combinations of attribute-object/adjective-noun items, e.g. we correlate the similarity between $v_{white\ cat}$ and $v_{black\ dog}$ with that between $w_{white\ cat}$ and $w_{black\ dog}$, resulting in **0.33** Spearman correlation (see Table 1).

Overall, our results suggest that the topologies of the semantic spaces are similar in the two modalities. Furthermore, since this phenomenon is also apparent in the cases of attribute-object and adjective-noun pairs, this alludes to the possibility of transferring approaches of semantic compositionality from the linguistic to the visual modality.

	High Visual	**Low Visual**
High Linguistic	goat-sheep, jaguar- lion black bag - brown bag, brown bear - yellow dog	baboon-transporter, bird-whale blue grass - blue van, gray whale - white deer
Low Linguistic	ball-horse, blackboard-chair red strawberry - white ball, white bear - yellow dog	baboon-sofa, backboard-panda black bag - green bridge, green table - yellow stick

Table 1: Similar and dissimilar concepts in the language and vision space.

3.2 Semantically-driven composition for attribute-object representations

The findings of the previous experiment suggest a high correlation between the visual attribute-attribute representations and the corpus-harvested adjective-noun representations. An interesting question that arises is whether we could approximate such visual representations of complex visual units, similarly to

[1] https://code.google.com/p/word2vec/

[2] http://wacky.sslmit.unibo.it, http://www.natcorp.ox.ac.uk

how is done in Computational Linguistics for approximating the text-based representations of adjective-noun phrases. Thus, this experiment is designed in order to assess the validity of the semantically-driven compositionality approach in the visual domain. Results are reported in Table 2. Since we expect that the quality of the aggregated vectors depends on the numbers of available images, we report results for subsets of the original data set that differ on the number of images per phrase.

By means of the LF composition method sketched in Section 2, we obtain the compositional representations of attribute-object (V^{comp}_{phrase}) and adjective-noun (W^{comp}_{phrase}) items. We then perform the correlation analyses between the similarities obtained in the composed visual space V^{comp}_{phrase} with: 1) the equivalent image-harvested representations V_{phrase}, 2) the equivalent corpus-derived linguistic representations W_{phrase}, 3) the equivalent compositionally-derived linguistic representations W^{comp}_{phrase}.

Overall, the correlation between V^{comp}_{space} and V_{space} suggests that the visual compositionality of attribute-object can account, to some extend, for the visual semantics of the respective image, and it further improves with the number of images we consider for obtaining the *aggregated vectors* of the visual phrases. Finally, as expected, the correlations between V^{comp}_{space} although lower than the ones reported in Section 3.1, i.e., 0.22 vs 0.32, are still non negligible.

	all phrases	$>$ 10 images	$>$ 20 images	$>$ 30 images
V^{comp}_{phrase} - V_{phrase}	0.24	0.40	0.53	**0.58**
V^{comp}_{phrase} - W_{phrase}	0.10	0.22	0.19	**0.23**
V^{comp}_{phrase} - W^{comp}_{phrase}	0.04	0.05	**0.18**	0.10

Table 2: Spearman correlations between the similarities in the V^{comp}_{phrase} and other semantic spaces.

4 Conclusions

In this work, we have experimented with semantically-driven compositionality of attributes with objects in the visual modality, by adopting an out-of-the-box composition method from the computational semantics literature. Our preliminary results have shown that the visual representations of attribute-objects when obtained compositionally reflect properties similar not only to the ones found in representations harvested automatically from images, but also from those extracted from text corpora. These results show that semantic compositionality might be a general process irrespective of the underlying modality. We have just scratched the surface on this topic and in the future we plan to experiment with a larger variety of attributes and use and design alternative visual compositional models.

Acknowledgements

The second and third author acknowledge ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES). We thank the 3 anonymous reviewers for their comments, Marco Baroni and Elia Bruni for their constant and useful feedback.

References

[Baroni and Zamparelli2010] Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, 1183–1193.

[Bosch et al.2007] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Image classification using random forests and ferns. In *Proceedings of ICCV*, 1–8.

[Lowe2004] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110.

[Mintz and Gleitman2002] Toben H. Mintz and Lila R. Gleitman. 2002. Adjectives really do modify nouns: the incremental and restricted nature of early adjective acquisition. *Cognition*, 84:267–293.

[Mitchell and Lapata2010] Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

[Russakovsky and Fei-Fei2012] Olga Russakovsky and Li Fei-Fei. 2012. Attribute learning in large-scale datasets. In *Trends and Topics in Computer Vision*, 1–14. Springer.

[Socher et al.2011] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of ICML*, 129–136.

[Stöttinger et al.2012] J. Stöttinger, J.R.R. Uijlings, A.K. Pandey, N. Sebe, and F. Giunchiglia. 2012. (unseen) event recognition via semantic compositionality. In *CVPR*.