Vision and Language Integration: Objects and beyond

Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot,
Moin Nabi, Enver Sangineto, Raffaella Bernardi
University of Trento, Trento, Italy
{firstname.lastname}@unitn.it

Motivation

Nouns are a crucial component of natural language sentences. It is not a coincidence that children first learn to use nouns and only afterwords expand their vocabulary with verbs, adjectives and other parts of speech Waxman et al. (2013). Interestingly, the same development has taken place with Language and Vision models. Object classification has long been the main concern of the computer vision field, only then followed by action classification shared tasks. Recently, more ambitious competitions have been proposed, aiming to evaluate models' ability to connect whole sentences to images, through both Image Captioning (IC) or Visual Question Answering (VQA) tasks. Progress in this area has seemed swift and impressive, but the community is now scrutinising the results to understand whether enthusiasm is warranted. Several diagnostic datasets have been proposed with this goal in mind, highlighting various flaws in existing tasks Johnson et al. (2017); Zhang et al. (2015). Our paper is a contribution to these efforts, showing that the field may have moved too fast from noun to sentence interpretation, overlooking difficulties in understanding other parts-of-speech.

Proposal

Our paper expands the existing FOIL dataset Shekhar et al. (2017). FOIL consists of a set of images matched with captions containing one single mistake. The mistakes are always nouns referring to objects not actually present in the image. The work demonstrates that the language and vision modalities are not truly integrated in current computational models, as they fail to spot the mistake in the caption and to correct it appropriately (humans, on the other hand, obtain almost 100% accuracy on those tasks). In the present paper, we exploit the FOIL strategy to evaluate Language and Vision models on a larger set of possible mismatches between language and vision. Beside considering nouns as possible 'foil' words, we also consider verbs, adjectives, adverbs and prepositions, as illustrated in Figure 1. The results obtained by state-of-the-art systems on this data demonstrate that current models are indeed little able to move beyond object understanding.[1]

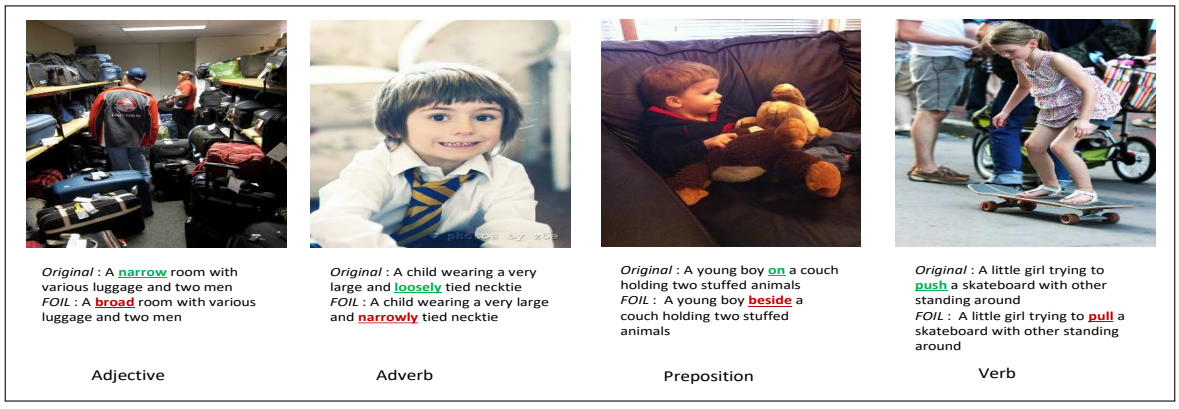

Figure 1: Sample image, corresponding original caption and the generated foil caption for the different parts of speech. The model has to be able to classify the caption as 'correct' or 'foil' (Task 1); detect the foil word in the foil caption (see words highlighted in red) (Task 2); and correct the foil word with an appropriate replacement (see words highlighted in green) (Task 3).

[1]The data will be made available at: https://foilunitn.github.io/.

The FOIL methodology

We follow the methodology highlighted in Shekhar et al. (2017), which consists of replacing a single word in a human-generated caption with a 'foil' item, making the caption unsuitable to describe the original image. Given such replacements, the system should be able to perform three tasks: a) a *classification task* (T1): given an image and a caption, the model has to predict whether the caption is correct or inappropriate for the image (evaluating whether the model has a coarse understanding of the linguistic and visual inputs and their relations); b) a *foil word detection task* (T2): given an image and a foil caption, detect the foil word in the caption (evaluating whether the model reaches a fine-grained representation of the linguistic input); a *foil word correction task* (T3): given an image, a foil caption and the foil word, the model has to correct the mistake (verifying whether the model reaches a fine-grained representation of the image).

Models and Results

Four models are tested on tasks 1-3: one baseline (a 'blind' model), and three state-of-the-art models from the Visual Question Answering (VQA) and Image Captioning (IC) literature, namely the *LSTM + norm I* of Antol et al. (2015) and the Hierarchical Co-Attention model (*HieCoAtt*) of Lu et al. (2016) and the IC system of Wang et al. (2016) (henceforth, *IC-Wang*). Results are reported in the Tables below.

Table 1: Classification Task (T1). Overall (both original and foil captions) accuracy. Chance level 50%.

	Noun	Verb	Adjective	Adverb	Preposition	Total
Blind	57.39	77.90	83.10	54.62	70.88	75.48
LSTM + norm I	63.17	78.37	83.81	**55.84**	73.70	77.11
HieCoAtt	**64.46**	**81.79**	**86.00**	53.40	**74.91**	**79.09**
IC-Wang	47.59	34.93	28.67	44.92	32.68	31.58

Table 2: Classification Task (T1). Accuracy results of the foil captions only. Chance level 50%.

	Noun	Verb	Adjective	Adverb	Preposition
Blind	23.18	57.11	76.99	18.73	54.32
LSTM + norm I	36.17 (**+12.99**)	59.49 (+2.3)	77.48 (+0.49)	20.42 (+1.69)	57.53 (+3.21)
HieCoAtt	38.22 (**+15.04**)	57.94 (+0.83)	80.05 (+3.06)	14.73	61.92 (+7.6)
IC-Wang	43.32 (**+20.16**)	13.98	4.3	23.87 (+5.14)	21.43

Table 3: Foil Detection Task (T2) and Foil Correction Task (T3).

	Foil Detection Task (T2)					Foil Correction Task (T3)				
	Noun	Verb	Adj.	Adv.	Prep.	Noun	Verb	Adj.	Adv.	Prep.
Chance	23.25	**21.72**	**21.72**	**21.72**	21.72	1.38	0.22	2.04	2.04	4.34
LSTM + norm I	26.32	7.96	4.06	9.68	6.46	4.7	1.14	1.33	0.36	1.54
HieCoAttn	**38.79**	3.57	2.34	9.26	6.09	4.21	0.98	**2.48**	0.24	1.47
IC-Wang	27.59	8.67	9.23	12.56	**26.56**	**22.16**	**9.1**	1.61	**3.44**	**7.78**

Our results show that none of the current SoA models achieve a fine-grained representation for all components of a sentence, including attributes and relations: attention models may have the right components to detect location (e.g., see locative prepositions), but some image captioning systems probably provide a better language model, in particular for closed-class words.

References

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh (2015). VQA: Visual Question Answering. In *Proceedings of ICCV*. `https://github.com/VT-vision-lab/VQA_LSTM_CNN`.

Johnson, J., B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick (2017). CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of CVPR*.

Lu, J., J. Yang, D. Batra, and D. Parikh (2016). Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Proceedings of NIPS 2016*. `https://github.com/jiasenlu/HieCoAttenVQA`.

Shekhar, R., S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, and R. Bernardi (2017). FOIL it! Find One mismatch between Image and Language caption. In *Proceedings of ACL*. https://arxiv.org/abs/1705.01359.

Wang, C., H. Yang, C. Bartz, and C. Meinel (2016). Image captioning with deep bidirectional LSTMs. In *Proceedings of ACM Multimedia* , pp. 988–997.

Waxman, S., X. Fu, S. Arunachalam, E. Leddon, K. Geraghty, and H. joo Song (2013). Are nouns learned before verbs? infants provide insight into a longstanding debate. *Child Dev Perspect 7*(3).

Zhang, P., Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh (2015). Yin and yang: Balancing and answering binary visual questions. *arXiv preprint arXiv:1511.05099*.