Vision and Language Integration: Moving beyond Objects

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

Abstract

The last years have seen an explosion of work on the integration of vision and language data. New tasks like Image Captioning and Visual Questions Answering have been proposed and impressive results have been achieved. There is now a shared desire to gain an in-depth understanding of the strengths and weaknesses of those models. To this end, several datasets have been proposed to try and challenge the state-of-the-art. Those datasets, however, mostly focus on the interpretation of *objects* (as denoted by nouns in the corresponding captions). In this paper, we reuse a previously proposed methodology to evaluate the ability of current systems to move beyond objects and deal with attributes (as denoted by adjectives), actions (verbs), manner (adverbs) and spatial relations (prepositions). We show that the coarse representations given by current approaches are not informative enough to interpret attributes or actions, whilst spatial relations somewhat fare better, but only in attention models.

1 Introduction

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.

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]

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

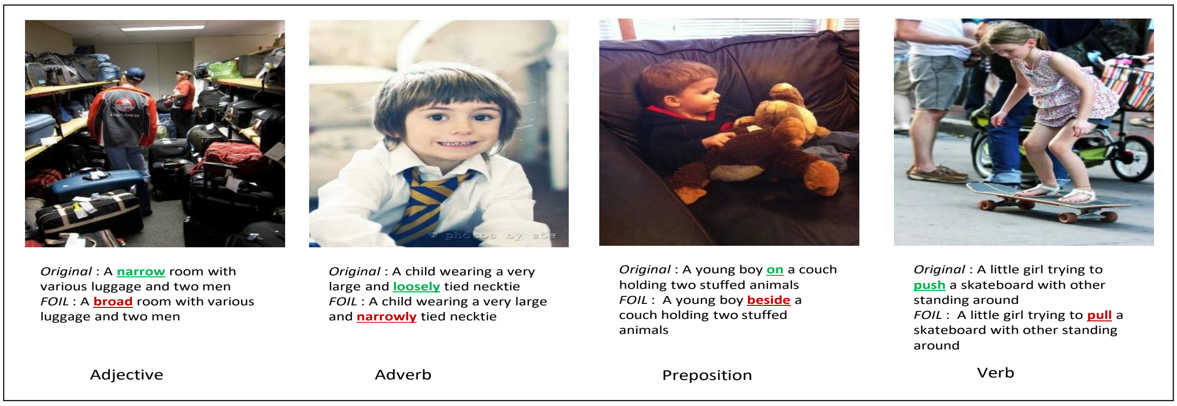
Original : A narrow room with various luggage and two men
FOIL : A broad room with various luggage and two men
Adjective
Original : A child wearing a very large and loosely tied necktie
FOIL : A child wearing a very large and narrowly tied necktie
Adverb
Original : A young boy on a couch holding two stuffed animals
FOIL : A young boy beside a couch holding two stuffed animals
Preposition
Original : A little girl trying to push a skateboard with other standing around
FOIL : A little girl trying to pull a skateboard with other standing around
Verb

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).

2 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). 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.

Blind Model: this model is based on the caption only; in other words, the system does not have access to the visual data. The caption is modelled by an LSTM, fully connected to a hidden layer followed by a softmax to perform the classification. This blind baseline affords an evaluation of the ‘language bias’ of the data (i.e., the phenomenon by which a Language and Vision dataset can be suitably modelled using language only).

Discriminative VQA Models: two VQA models are used, namely the *LSTM + norm I* of Antol et al. (2015) and the Hierarchical Co-Attention model (*HieCoAtt*) of Lu et al. (2016). In *LSTM + norm I*, the text is represented by two stacked LSTMs and the image is represented by a normalisation of the last fully connected layer of VGG network (Simonyan and Zisserman, 2014). Both representations are projected onto a 1024-dimensional feature space. The combination of language and vision features is performed by point-wise multiplication followed by a fully connected and a softmax layer. *HieCoAtt* has a similar architecture, with the addition of an attention layer. Attention is provided to both image and text in alternation, in a hierarchical fashion.

Generative IC Model: we use the IC system of Wang et al. (2016) (henceforth, *IC-Wang*), which generates a word in a caption by considering both past and future contexts, using a bi-directional LSTM. *IC-Wang* consists of three modules: a CNN to encode the image, a text LSTM to encode captions, and a multimodal LSTM for mapping visual and text representations to a common space.

For T1, the models are directly trained to classify a given caption as ‘good’ vs. ‘foil’. For T2 and T3, the model trained on T1 is adopted. For T2, we subsequently occlude one word (Goyal et al. (2016)) at a time and calculate the probability of the new caption to be good vs. foil. The model selects as foil word, the one which has generated the caption with the highest probability. For T3, we regress over all

	no. of unique images		no. of unique datapoints		no. of unique target::foil pairs	
	Train	Test	Train	Test	Train	Test
Noun*	22,101	15,435	73,076	37,381	236	194
Verb	6314	2788	7925	3353	268	219
Adjective	15,640	9009	20,720	11,900	80	62
Adverb	1011	451	1044	475	38	36
Preposition	8733	5551	24,665	15,755	101	89
TOT	22,101	15,435	127,430	68,864	723	600

Table 1: Statistics of the dataset. Here Noun* is a subset of FOIL-COCO used in Shekhar et al. (2017).

the target words on the position of the foil word and select the one which generates the caption with the highest probability to be "good". Due to the generative nature of IC models, adapting *IC-Wang* for the classification purpose is less straightforward. For T1, we generate all possible captions by subsequently predicting one word at a time provided all other words in the caption and the image. We compare the probability of these generated captions with the given caption. When the test caption probability is higher than generated captions probabilities, we classify the given caption as good caption, else as foil caption.

3 Dataset Creation

Following Shekhar et al. (2017), we aim at creating a dataset of images associated with both correct and foil captions, where the latter are obtained by replacing one word in the original text. Expanding on the original paper, our target/foil pairs do not merely consist of nouns. The introduced error can also be an adjective (an object's attribute), a verb (an action), a preposition (a relation between objects) or an adverb (a manner of action). In total, we produce 196,284 datapoints, each corresponding to an <image, original, foil> triple. The starting point for images and correct captions is Microsoft's Common Objects in Context (MS-COCO)(Lin et al. (2014)).

3.1 Creating new target/foil pairs

We describe below our procedure to expand the original dataset with new parts-of-speech.

Verbs: We use three resources: a) VerbOcean, a semi-automatically generated broad-coverage semantic network of verbs extracted from the Web by exploiting a pattern-based approach (Chklovski and Pantel (2004)); b) Computing Lexical Contrast (CLC), a resource of contrasting words selected from direct and indirect WordNet opposites, (Mohammad et al. (2013)); c) SimLex999, a set of related word pairs rated with respect to their similarity (Hill et al. (2016)). From VerbOcean and CLC, we extract all antonyms (e.g., *pull-push*). From SimLex999, we select those pairs with a similarity score lower than the average in the database (e.g., *allowing- preventing*). We end up with 902, 44, and 30 verb pairs from VerbOcean, CLC and SimLex999 respectively.

Adjectives and adverbs: As in the verb case, we use antonyms from CLC and we select pairs from SimLex999 which have a similarity score lower than average. We extract 46 and 127 adjectives pairs from CLC and SimLex999 respectively. All adverbial pairs come from CLC, and amount to 52 datapoints.

Prepositions: We extract prepositions from Berry et al. (1995), divided into three classes: place (e.g., *under*, *below*), direction (e.g., *inside*, *outside*) and device (e.g., *by*, *with*). Using these prepositions, we generate target/foil pairs by coupling prepositions which belong to the same class. We obtain a total of 206 pairs (110, 90 and 6 for place, direction and device respectively).

Nouns: The target/foil noun pairs are built using words that belong to the same category in MS-COCO (e.g., *bird/dog*, from the MS-COCO category ANIMAL). In order to obtain a balanced dataset across the various PoS, we only use a subset of the FOIL-COCO dataset of Shekhar et al. (2017). From

Table 2: 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

the FOIL dataset, we retain the 37,536 images for which foil captions could be generated, using the target/foil pairs extracted from the resources mentioned above. Of the FOIL datapoints generated for the noun pairs, only those containing images used for the other PoS are selected. Hence, the number of unique images of the whole dataset is the same of those used for nouns (see Table 1 for details of the train/test set division.)

We use all word pairs in both directions (e.g. replacing *push* with *pull* and *pull* with *push*). We only use pairs for which target and foil are found in the original captions. This ensures that the model will not learn to recognise a foil caption simply by recording the presence of an unknown word. From each resource, we randomly split the target/foil pairs into training and test sets. The number of unique pairs per PoS is provided in Table 1.

3.2 Foil Caption Generation

From the word pair lists above, foil captions are generated from MS-COCO original captions. The foil captions are generated by replacing nouns are directly extracted from the FOIL dataset by Shekhar et al. (2017). In this case, for each original MS-COCO caption, several foil ones are generated and subsequently filtered using several heuristics. The aim of filtering is to prioritise salient objects in the image, and to minimise the language bias in the data. Ideally, these filters would have to be applied also for the generation of the foil caption for the other PoS, but we found that they reduced the size of our data in an unacceptably small size. As a consequence, the results we report are obviously affected by the language bias, as shown by the reasonable performance of a ‘blind’ model without access to visual data. However, as we will see, our broad claim is not affected by this heightened baseline. Details on the number of the unique images and of of datapoints generated for each PoS are reported in Table 1.

4 Experiments and Results

4.1 Results and Analysis

Table 2 reports the accuracy of the various models described in §2 for Task T1. The blind model’s accuracy is well above chance level for all PoS, with lower results observed on the captions generated by noun and adverb replacement. Recall that for nouns, the language prior has been minimised, whereas the datapoints generated with verb, adjective and preposition replacements have some language prior that the models can exploit. The comparatively low performance on adverbs may be explained by the fact that all generated target/foil pairs are antonyms which behave very similarly from a distributional point of view (e.g. *upwards/downwards*, *partially/completely*, etc). *HieCoAtt* is the overall best performing model, but we note that it only outperforms the blind model by a few points. These numbers, however, do not show to which extent the models are able to avoid the trap of the dataset: Shekhar et al. (2017) showed that on the FOIL data, models tend to detect correct captions with reasonable accuracy but fail to identify the incorrect ones, leading to a large bias in classification. Taking this insight into account, for the rest of this paper, we focus on the accuracy of the systems in dealing with foil captions, across all three tasks.

Table 3: 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 4: 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**

As shown in Table 3, the blind model's accuracy is still reasonable on T1, but lower than chance for nouns and adverbs. In the case of nouns, the visual input helps obtaining a higher accuracy, whereas this is not the case for the other PoS. This could be due to the ability of vision models to 'see' objects but not their properties (adjectives) or relations (verbs, prepositions). It is a known shortcoming of such systems that they have difficulties in recognising anything that is not straightforwardly defined by a bounding box (Johnson et al. (2017)). *IC-Wang* performs very poorly on verbs, adjectives and prepositions, even though it is the best system for nouns. Other models improve minimally on the baseline, with prepositions getting the best improvement: +7% for *HieCoAtt*. When looking more in detail into this result, we observe that most instances in the preposition data indicate location: it is not surprising that an attention model would perform well on those, since it is trained to focus on particular areas of the image.

For task T2 (see Table 4), all models perform well under baseline on verbs, adjectives and adverbs. *IC-Wang* does however provide some improvement on prepositions. The reason for this may be that the system, being trained to generate sequences, has a better internal language model than other approaches. Whilst a good language model is unlikely to help in the case of content words, we can expect some benefits for function words. This trend has been observed in work on L2 error detection, where mistakes in words from closed classes are easier to spot and correct (Herbelot and Kochmar (2016)).

For task T3 (see Table 4), improvements over the baseline are minimal. *IC-Wang* performs best overall, but at a level well below its achievement on nouns. We do not only confirm that foil correction is hard, but that it is particularly challenging on parts-of-speech that represent attributes or relations. We note that *IC-Wang* improves more on prepositions than on adjectives and adverbs, confirming what was observed in T2 (i.e. closed classes are easier to deal with). But it also provides a good improvement on the verb baseline, which is puzzling given its inability to *spot* verb foils in T2.

5 Conclusion

Language and Vision integration has been studied in a fine-grained, but single-minded way, when focusing on objects (nouns). The level of events (sentences) has also received attention, but through coarse representations. Our work aims to highlight the importance of a fine-grained representation for all components of a sentence, including attributes and relations. Our results show that none of the current SoA models achieve this overall goal: 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.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research.

References

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

Berry, C., A. Brizee, E. Angeli, and M. Ghafoor (1995). Prepositions for Time, Place, and Introducing Objects. `https://owl.english.purdue.edu/owl/owlprint/594/`.

Chklovski, T. and P. Pantel (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *EMNLP*, Volume 4, pp. 33–40.

Goyal, Y., A. Mohapatra, D. Parikh, and D. Batra (2016). Towards Transparent AI Systems: Interpreting Visual Question Answering Models . In *In Proceedings of ICML Visualization Workshop*.

Herbelot, A. and E. Kochmar (2016). Calling on the classical phone: a distributional model of adjective-noun errors in learners English. In *International Conference on Computational Linguistics (COLING)*.

Hill, F., R. Reichart, and A. Korhonen (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

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 *CVPR*.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pp. 740–755. Springer.

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`.

Mohammad, S. M., B. J. Dorr, G. Hirst, and P. D. Turney (2013). Computing lexical contrast. *Computational Linguistics 39*(3), 555–590.

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 *ACL (to appear)*. https://arxiv.org/abs/1705.01359.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

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*.