eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Defoiling Foiled Image Captions

Pranava Madhyastha, Josiah Wang and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{p.madhyastha, j.k.wang, l.specia}@sheffield.ac.uk

Abstract

We address the task of detecting foiled image captions, i.e. identifying whether a caption contains a word that has been deliberately replaced by a semantically similar word, thus rendering it inaccurate with respect to the image being described. Solving this problem should in principle require a fine-grained understanding of images to detect linguistically valid perturbations in captions. In such contexts, encoding sufficiently descriptive image information becomes a key challenge. In this paper, we demonstrate that it is possible to solve this task using simple, interpretable yet powerful representations based on explicit object information. Our models achieve state-of-the-art performance on a standard dataset, with scores exceeding those achieved by humans on the task. We also measure the upper-bound performance of our models using gold standard annotations. Our analysis reveals that the simpler model performs well even without image information, suggesting that the dataset contains strong linguistic bias.

1 Introduction

Models tackling vision-to-language (V2L) tasks, for example Image Captioning (IC) and Visual Question Answering (VQA), have demonstrated impressive results in recent years in terms of automatic metric scores. However, whether or not these models are actually learning to address the tasks they are designed for is questionable. For example, Hodosh and Hockenmaier (2016) showed that IC models do not understand images sufficiently, as reflected by the generated captions. As a consequence, in the last few years many diagnostic tasks and datasets have been proposed aiming at investigating the capabilities of such models in more detail to determine whether and how these models are capable of exploiting visual and/or linguistic information (Shekhar et al., 2017b; John-son et al., 2017; Antol et al., 2015; Chen et al., 2015; Gao et al., 2015; Yu et al., 2015; Zhu et al., 2016).

FOIL (Shekhar et al., 2017b) is one such dataset. It was proposed to evaluate the ability of V2L models in understanding the interplay of objects and their attributes in the images and their relations in an image captioning framework. This is done by replacing a word in MSCOCO (Lin et al., 2014) captions with a 'foiled' word that is semantically similar or related to the original word (substituting *dog* with *cat*), thus rendering the image caption unfaithful to the image content, while yet linguistically valid. Shekhar et al. (2017b) report poor performance for V2L models in classifying captions as foiled (or not). They suggested that their models (using image embeddings as input) are very poor at encoding structured visual-linguistic information to spot the mismatch between a foiled caption and the corresponding content depicted in the image.

In this paper, we focus on the foiled captions classification task (Section 2), and propose the use of *explicit object detections* as salient image cues for solving the task. In contrast to methods from previous work that make use of word based information extracted from captions (Heuer et al., 2016; Yao et al., 2016; Wu et al., 2018), we use explicit object category information directly extracted from the images. More specifically, we use an interpretable *bag of objects* as image representation for the classifier. Our hypothesis is that, to truly 'understand' the image, V2L models should exploit information about *objects* and their relations in the image and not just global, low-level image embeddings as used by most V2L models.

Our main contributions are:

1. A model (Section 3) for foiled captions classification using a simple and interpretable

object-based representation, which leads to the best performance in the task (Section 4);

2. Insights on upper-bound performance for foiled captions classification using gold standard object annotations (Section 4);

3. An analysis of the models, providing insights into the reasons for their strong performance (Section 5).

Our results reveal that the FOIL dataset has a very strong linguistic bias, and that the proposed simple object-based models are capable of finding salient patterns to solve the task.

2 Background

In this section we describe the foiled caption classification task and dataset.

We combine the tasks and data from Shekhar et al. (2017b) and Shekhar et al. (2017a). Given an image and a caption, in both cases the task is to learn a model that can distinguish between a **REAL** caption that describes the image, and a **FOIL**ed caption where a word from the original caption is swapped such that it no longer describes the image accurately. There are several sets of 'foiled captions' where words from specific parts of speech are swapped:

- **Foiled Noun:** In this case a noun word in the original caption is replaced with another similar noun, such that the resultant caption is not the correct description for the image. The foiled noun is obtained from list of object annotations from MSCOCO (Lin et al., 2014) and nouns are constrained to the same super-category;

- **Foiled Verb:** Here, verb is foiled with a similar verb. The similar verb is extracted using external resources;

- **Foiled Adjective and Adverb:** Adjectives and adverbs are replaced with similar adjectives and adverbs. Here, the notion of similarity again is obtained from external resources;

- **Foiled Preposition:** Prepositions are directly replaced with functionally similar prepositions.

The Verb, Adjective, Adverb and Preposition subsets were obtained using a slightly different methodology (see Shekhar et al. (2017a)) than that used for Nouns (Shekhar et al., 2017b). Therefore, we evaluate these two groups separately.

3 Proposed Model

For the foiled caption classification task (Section 3.1), our proposed model uses information from explicit object detections as an object-based image representation along with textual representations (Section 3.2) as input to several different classifiers (Section 3.3).

3.1 Model definition

Let $y \in \{\text{REAL, FOIL}\}$ denote binary class labels. The objective is to learn a model that computes $P(y|I;C)$, where I and C correspond to the image and caption respectively. Our model seeks to maximize a scoring function θ:

$$y = \arg\max \theta(I;C) \qquad (1)$$

3.2 Representations

Our scoring function θ takes in image features and text features (from captions) and concatenates them. We experiment with various types of features.

For the image side, we propose a *bag of objects* representation for 80 pre-defined MSCOCO categories. We consider two variants: (a) **Object Mention**: A binary vector where we encode the presence/absence of instances of each object category for a given image; (b) **Object Frequency**: A histogram vector where we encode the *number* of instances of each object category in a given image.

For both features, we use **Gold** MSCOCO object annotations as well as **Predict**ed object detections using YOLO (Redmon and Farhadi, 2017) pre-trained on MSCOCO to detect instances of the 80 categories.

As comparison, we also compute a standard **CNN**-based image representation, using the POOL5 layer of a ResNet-152 (He et al., 2016) CNN pre-trained on ImageNet. We posit that our object-based representation will better capture semantic information corresponding to the text compared to the CNN embeddings used directly as a feature by most V2L models.

For the language side, we explore two features: (a) a simple **bag of words (BOW)** representation for each caption; (b) an **LSTM** classifier based model trained on the training part of the dataset.

PoS Type	Train^{foil}	Test^{foil}	Total Train	Total Test
Noun	153,229	75,278	306,458	150,556
Verb	6,314	2,788	60,262	33,616
Adjective	15,640	9,009	73,057	42,163
Adverb	1,011	451	53,381	30,738
Preposition	8,733	5,551	77,002	46,018

Table 1: Dataset statistics for different foiled parts of speech. The superscript $foil$ indicates the number of foiled captions.

Our intuition is that an image description/caption is essentially a result of the interaction between important objects in the image (this includes spatial relations, co-occurrences, etc.). Thus, representations explicitly encoding object-level information are better suited for the foiled caption classification task.

3.3 Classifiers

Three types of classifiers are explored: (a) **Multilayer Perceptron (MLP)**: For BOW-based text representations, a two 100-dimensional hidden layer MLP with ReLU activation function is used with cross-entropy loss, and is optimized with Adam (learning rate 0.001); (b) **LSTM Classifier**: For LSTM-based text representations, a uni-directional LSTM classifier is used with 100-dimensional word embeddings and 200-dimensional hidden representations. We train it using cross-entropy loss and optimize it using Adam (learning rate 0.001). Image representations are appended to the final hidden state of the LSTM; (c) **Multimodal LSTM (MM-LSTM) Classifier**: As above, except that we initialize the LSTM with the image representation instead of appending it to its output. This can also be seen as am image grounded LSTM based classifier.

4 Experiments

Data: We use the dataset for nouns from Shekhar et al. (2017b)[1] and the datasets for other parts of speech from Shekhar et al. (2017a) [2]. Statistics about the dataset are given in Table 1. The evaluation metric is accuracy per class and the average (overall) accuracy over the two classes.

Performance on nouns: The results of our experiments with foiled nouns are summarized in Table 2. First, we note that the models that use **Gold**

Feats	Overall	Real	Foil
Blind (LSTM only)†	55.62	86.20	25.04
HieCoAtt†	64.14	91.89	36.38
CNN + BOW MLP	88.42	86.89	89.97
Predict Mention + BOW MLP	94.94	95.68	94.23
Predict Freq + BOW MLP	95.14	95.82	94.48
Gold Mention + BOW MLP	95.83	96.30	95.36
Gold Freq + BOW MLP	96.45	96.04	96.85
CNN + LSTM	87.45	86.78	88.14
Predict Freq + LSTM	85.99	85.17	86.81
Gold Freq + LSTM	87.38	86.62	88.18
Predict Freq + MM-LSTM	87.90	86.73	88.95
Gold Freq + MM-LSTM	89.02	88.35	89.72
Human (majority)†	92.89	91.24	94.52

Table 2: Accuracy on Nouns dataset. † are taken directly from Shekhar et al. (2017b). HieCoAtt is the state of the art reported in the paper.

bag of objects information are the best performing models across classifiers. We also note that the performance is better than human performance. We hypothesize the following reasons for this: (a) human responses were crowd-sourced, which could have resulted in some noisy annotations; (b) our gold object-based features closely resembles the information used for data-generation as described in Shekhar et al. (2017b) for the foil noun dataset. The models using **Predict**ed bag of objects from a detector are very close to the performance of **Gold**. The performance of models using simple bag of words (**BOW**) sentence representations and an **MLP** is better than that of models that use **LSTM**s. Also, the accuracy of the bag of objects model with **Frequency** counts is higher than with the binary **Mention** vector, which only encodes the presence of objects. The Multimodal LSTM (**MM-LSTM**) has a slightly better performance than **LSTM** classifiers. In all cases, we observe that the performance is on par with human-level accuracy. Our overall accuracy is substantially higher than that reported in Shekhar et al. (2017b). Interestingly, our implementation of **CNN+LSTM** produced better results than their equivalent model (they reported 61.07% vs. our 87.45%). We investigate this further in Section 5.

Performance on other parts of speech: For other parts of speech, we fix the image representation to **Gold Frequency**, and compare results using the **BOW**-based **MLP** and **MM-LSTM**. We also compare the scores to the state of the art reported in Shekhar et al. (2017a). Note that this

	Classifier	Overall	Real	Foil
VB	Gold Freq + BOW MLP	84.03	97.38	70.68
	Gold Freq + MM-LSTM	87.90	99.48	76.32
	HieCoAtt[†]	81.79	-	57.94
ADJ	Gold Freq + BOW MLP	87.74	96.96	78.52
	Gold Freq + MM-LSTM	92.29	85.82	98.77
	HieCoAtt[†]	86.00	-	80.05
ADV	Gold Freq + BOW MLP	54.99	98.49	11.48
	Gold Freq + MM-LSTM	56.55	99.45	13.65
	HieCoAtt[†]	53.40	-	14.73
PREP	Gold Freq + BOW MLP	75.53	92.61	58.45
	Gold Freq + MM-LSTM	89.74	95.59	83.89
	HieCoAtt[†]	74.91	-	61.92

Table 3: Accuracy on Verb, Adjective, Adverb and Preposition datasets, using **Gold Frequency** as the image representation. [†] is the best performing model as reported in Shekhar et al. (2017a).

model does not use gold object information and may thus not be directly comparable – we however recall that only a slight drop in accuracy was found for our models when using predicted object detections rather than gold ones. Our findings are summarized in Table 3. The classification performance is not as high as it was for the nouns dataset. Noteworthy is the performance on adverbs, which is significantly lower than the performance across other parts of speech. We hypothesize that this is because of the imbalanced distribution of foiled and real captions in the dataset. We also found that the performance of **LSTM**-based models on other parts of speech datasets are almost always better than **BOW**-based models, indicating the necessity of more sophisticated features.

5 Analysis

In this section, we attempt to better understand why our models achieve such a high accuracy.

5.1 Ablation Analysis

We first perform ablation experiments with our proposed models over the Nouns dataset (FOIL). We compute image-only models (**CNN** or **Gold Frequency**) and text-only models (**BOW** or **LSTM**), and investigate which components of our model (text or image/objects) contribute to the strong classification performance (Table 4). As expected, we cannot classify foiled captions given only image information (global or object-level), resulting in chance-level performance.

On the other hand, text-only models achieve a

very high accuracy. This is a central finding, suggesting that foiled captions are easy to detect even without image information. We also observe that the performance of **BOW** improves by adding object **Frequency** image information, but not **CNN** image embeddings. We posit that this is because there is a tighter correspondence between the bag of objects and bag of word models. In the case of LSTMs, adding either image information helps slightly. The accuracy of our models is substantially higher than that reported in Shekhar et al. (2017b), even for equivalent models.

We note, however, that while the trends of image information is similar for other parts of speech datasets, the performance of **BOW** based models are lower than the performance of **LSTM** based models. The anomaly of improved performance of **BOW** based models seems heavily pronounced in the nouns dataset. Thus, we further analyze our model in the next section to shed light on whether the high performance is due to the models or the dataset itself.

	Image	Text	Overall	Real	Foil
	CNN	-	50.01	64.71	35.31
	Gold Freq	-	50.04	53.10	47.00
MLP	-	BOW	89.33	88.32	90.34
	CNN	BOW	88.42	86.89	89.97
	Gold Freq	BOW	96.45	96.04	96.85
LSTM	-	LSTM	85.07	85.52	84.66
	CNN	LSTM	87.38	86.62	88.18
	Gold Freq	LSTM	87.45	86.78	88.14

Table 4: Ablation study on FOIL (Nouns).

5.2 Feature Importance Analysis

We apply Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016) to further understand the strong performance of our simple classifier on the Nouns dataset (FOIL) without any image information. We present an example in Figure 1. We use **MLP** with **BOW** only (no image information) as our classifier. As the caption is correctly predicted to be foiled, we observe that the most important feature for classification is the information on the word *ball*, which also happens to be the foiled word. We further analyzed the chances of this happening on the entire test set. We found that 96.56% of the time the most important classification feature happens to be the foiled word. This firmly indicates that there is a very strong linguistic bias in the training data, despite
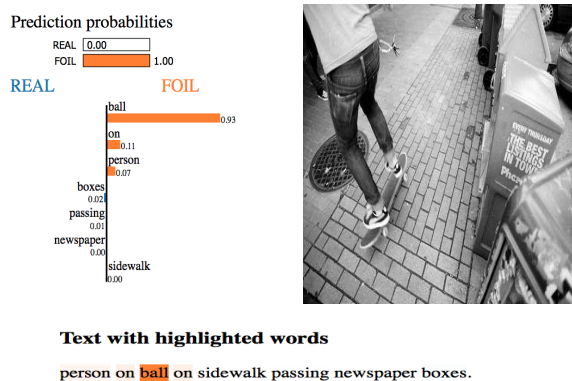
Figure 1: Classifier's prediction for the foiled caption: The classifier is able to correctly classify the foiled caption and uses the foiled word as the trigger for classification.

the claim in Shekhar et al. (2017b) that special attention was paid to avoid linguistic biases in the dataset.[3] We note that we were not able to detect the linguistic bias in the other parts of speech datasets.

6 Conclusions

We presented an object-based image representation derived from explicit object detectors/gold annotations to tackle the task of classifying foiled captions. The hypothesis was that such models provide the necessary semantic information for the task, while this informaiton is not explicitly present in CNN image embeddings commonly used in V2L tasks. We achieved state-of-the-art performance on the task, and also provided a strong upper-bound using gold annotations. A significant finding is that our simple models, especially for the *foiled noun* dataset, perform well even without image information. This could be partly due to the strong linguistic bias in the foiled noun dataset, which was revealed by our analysis on our interpretable object-based models. We release our analysis and source code at https://github.com/sheffieldnlp/foildataset.git.

Acknowledgments

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433. IEEE.

Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. Déjà image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 504–514. Association for Computational Linguistics.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? Dataset and methods for multilingual image question. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2296–2304. Curran Associates, Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 770–778. IEEE.

Hendrik Heuer, Christof Monz, and Arnold W. M. Smeulders. 2016. Generating captions without looking beyond objects. In *ECCV Workshop on Storytelling with Images and Videos*.

Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1988–1997. IEEE.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer International Publishing.

Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 6517–6525. IEEE.

---

[3] Shekhar et al. (2017b) have acknowledged about the bias in our personal communications and are currently working on a fix

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. ACM.

Ravi Shekhar, Sandro Pezzelle, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. Vision and language integration: Moving beyond objects. In *Proceedings of International Conference on Computational Semantics (IWCS)*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. FOIL it! Find One mismatch between Image and Language caption. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 255–265. Association for Computational Linguistics.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. *Transactions on Pattern Analysis and Machine Intelligence*.

Li Yao, Nicolas Ballas, Kyunghyun Cho, John R. Smith, and Bengio Yoshua. 2016. Oracle performance for visual captioning. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 141.1–141.13. BMVA Press.

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual Madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2461–2469. IEEE.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question aswering in images. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 4995–5004. IEEE.