Project AI: Multimodal Representations Learned via Dialogue Interaction

Ece Takmaz (11823453) **Nikos Kondylidis** (11853913)[1]

Abstract

We explore the effects of dialogue interaction in learning multimodal representations combining visual models and language models. Additionally, we inspect Visual Question Answering and Image Captioning models that do not consist of multiple levels of interaction. We compare the performance of these models at the FOIL classification task, where the aim is to figure out, given an image and a caption, whether that caption is an original caption or a faulty one. We found out that the fine-tuned interactive model outperformed the other models at the FOIL classification task with a test set accuracy of 62.99%.

1. Introduction

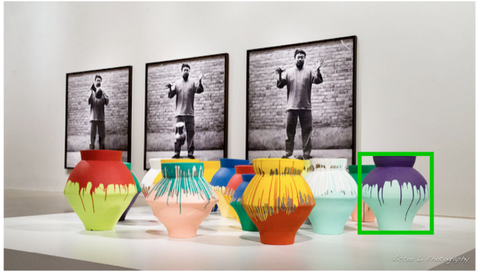
The recent advancements in Deep Learning research has provided the academia and the industry with an increase in the accuracy and performance of the tasks that involve computer vision and image analysis, as well as natural language processing. For instance, Convolutional Neural Networks (CNN) result in very good accuracies at image-related tasks (Krizhevsky et al., 2012; Donahue et al., 2014). In addition, capturing the sequential and intricate nature of natural languages is made better with the help of Recurrent Neural Networks (RNN), and more specifically Long-Short Term Memory (LSTM) networks are well-documented in capturing sequential nature of phenomena such as language (Sutskever et al., 2014; Hochreiter & Schmidhuber, 1997).

In this project, we explore multimodal tasks that utilize both image and language related components in order to achieve high accuracies. These tasks combine language models with visual models. An important question here is whether these models make use of both modalities well enough in learning.

To obtain good results, these systems must visually ground linguistic input. Such a grounding would require image comprehension and computational linguistics at a level that allows for the extraction of representations that account for the crucial information in the image and the linguistic input.

Examples to these multimodal tasks are Visual Question Answering (VQA) and Image Captioning (IC). The fusing of information from different modalities is essential in these tasks that lie in the intersection of image comprehension and computational linguistics. Nevertheless, VQA and IC are non-interactive in that they cannot account for multiple levels of two-way communication.

Our hypothesis in this project is that with the help of interaction, Interactive Visual Dialogue (IVD) models could provide an increase in the capability to capture visual and linguistic representations [1].



Questioner	**Oracle**
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Figure 1. GuessWhat?! IVD example, where, previous answer - question pairs affect the generation of the current question in finding the target object (de Vries et al., 2017)

Building on the previous questions and answers through dialogues, representations may have a better capability of capturing the visual and linguistic information.

In Section 2, we summarize the literature on Visual Question Answering, Image Captioning and Visual Dialogue Models. In Section 3 the goal and the approach of this project is described. Following Section 4 describes the nature of the

[1] Supervisors: Assoc. Prof. Raquel Fernández, Dr. Elia Bruni and Aashish Venkatesh.
Ece Takmaz <ece.takmaz@student.uva.nl>
Nikos Kondylidis <kondilidisn9@gmail.com>.

Project AI 2018, MSc Artificial Intelligence, Univ. of Amsterdam

[1] Project repository: `https://github.com/ecekt/ProjAI_FOIL`

FOIL task that we use to evaluate the multimodal representation abilities of models. Later, in Section 5, the models that are being put to test are described. Section 6, includes the way that the models were evaluated, while Section 7 depicts the results. Furthermore, Section 9 refers to some details on practical issues in Artificial Intelligence (AI) and Section 10 summarizes over our work.

2. Related Work

A dataset that tests the coherence between visual and linguistic representations generated by a model was initially built by Shekhar et al.. This dataset introduced FOIL tasks, which are described in Section 4, where a model is used in order to classify whether a combination of image and caption make sense, or whether an obvious inconsistency exist when these representations are combined. Some models that were put to test were originally trained for tasks that combined visual and linguistic features, such as Image Captioning and Visual Question Answering tasks (Antol et al., 2015).

In addition, there exists another form of tasks (de Vries et al., 2017), that require a model to repeatedly ask questions based on an image and previous question-answer pairs, until the model can identify a target object from the image. Visual and linguistic features of models trained on this task such as GDSE (Venkatesh et al., 2018), are expected to affect one another more than models trained on simpler tasks that do not include dialogue interaction.

2.1. Visual Question Answering

VQA concerns the task of generating an answer given an image and a question related to this image (Antol et al., 2015). In order to achieve reasonable outcomes in this task, it is essential for a system to extract both visual and linguistic information and fuse them in a way that allows for the selection or generation of the correct answer. To this end, several features from images can be extracted implicitly or explicitly. These features range from object presence in the image to spatial relations of object instances (Xu & Saenko, 2015; Jabri et al., 2016). Additionally, VQA requires the comprehension of syntax and semantics of posed questions and related answers. In general, VQA models obtain the representation of the image features from a CNN and the representation of the question from the hidden states of an LSTM or GRU. An important aspect of the model here is the way visual and linguistic representations are combined. An example combination could be the concatenation and further feeding of the representations into another network for the generation of an answer.

(Lu et al., 2016) introduced a model that uses convolutional layers both for extracting linguistic and visual information by the use of convolutional layers. Regarding linguistic information, three linguistic representations are being maintained, at word level, phrase level and finally at question level. On the other hand, only one representation is maintained regarding the visual information. Then, attention can be applied to linguistic and visual representations in parallel or asynchronously. As the authors utilize three linguistic representations, three attention models are being trained in total.

2.2. Image Captioning

IC is the task of generating a relevant caption given an image. In general, similar to IC models obtain the representation of the image through a CNN and couples it with the LSTM that takes human-annotated captions.

In addition, there is also research on retrieval tasks in this domain. They either take a caption and retrieve top-k possible images or they are fed an image and rank and retrieve captions. Specifically, a model that was designed in order to generate a caption, given an image, by the use of deep bidirectional LSTMs has been introduced by Wang et al..

2.3. Interactive Visual Dialogue Models

An interactive visual dialogue problem, assumes a task that requires modeling multi-level interactive behaviour that can take place in a dialogue. The model needs to ask questions until a goal has been reached. For example, the GuessWhat?! (de Vries et al., 2017) game starts by selecting an object from an image and an Interactive Visual Dialogue model has to narrow down the possible objects until it finds the selected one, by forming questions. The forming of a question strongly depends on the previous question-answer pairs, so that the new question aims to increase the amount of total information gathered at that point. In this way, interactive visual dialogue models are expected to perform better on the FOIL task.

The Grounded Dialogue State Encoder model (GDSE) (Venkatesh et al., 2018), as depicted in Figure 2, given an image representation and a conversation representation, generates a question that aims to narrow down the possibly selected objects. The model consists of a question generator model (QGen), an oracle that provides the answers to the questions and a Guesser. The guesser decides when the conversation contains enough information so that the model should try guessing the object, instead of asking for more information regarding the object.

3. Research questions

One of the main concerns is how to make sure that the multimodal models capture the intricate relation between the visual and linguistic modalities. In addition, we would like

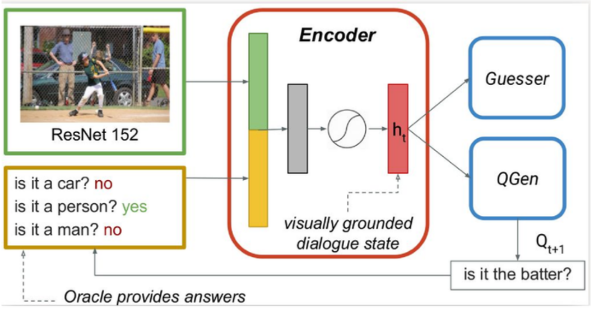


Figure 2. Overall architecture of the GDSE model (Venkatesh et al., 2018)

to observe whether the addition of interaction in the form of dialogues would provide any improvements in capturing multimodal information compared to the non-interactive models of VQA and IC.

Models that were successfully trained on tasks that involve the concept of dialogue, are expected to have better understanding of linguistic information. This is expected, because in order to actively participate in a dialogue, understanding what has been previously stated is necessary. That said, we will compare two models that combine linguistic and visual information, but only one of them will be trained on a task that includes active participation in a dialogue.

In order to address the first question, Shekhar et al. proposed the FOIL dataset and the FOIL tasks as a diagnostic tool, which are explained in the next section.

4. FOIL Dataset and FOIL tasks

4.1. FOIL.v1

The first version of the FOIL dataset FOIL.v1 has been created with the aim of diagnosing multimodal systems and assessing their capability to capture image-related information and the linguistic input (Shekhar et al., 2017). Shekhar et al. had created FOIL.v1 using the image and the human-annotated captions from the 2014 split of the MS-COCO dataset (Lin et al., 2014).

MS-COCO 2014 dataset contains 83K/41K train/validation images and for each image there are 5 human-annotated captions. With the dataset, other information is also made available such as image IDs, caption IDs, object categories and supercategories, object instances in the image and their spatial properties.

To generate the FOIL dataset for these tasks, the authors employ the MS-COCO and process it through 4 main steps. First, they generate word pairs (original target - foil target) from the same supercategory for replacement. Then, randomly these pairs are separated for replacements in the training sentences and the test sentences.

Third step is where they actually alter the captions after a minimal application of preprocessing. They obtain the visually salient target objects, which are category words that were mentioned by more than 1 annotator. Using the replacement pairs that were generated previously, they replace the salient object word with a potential foil word. Here, the authors also take into account the possibility that the selected word might actually occur in the image. In order to eliminate such a possibility, they discard any foil target, if it has been mentioned by at least one of the annotators. After this phase of substitution, they input the images and captions into NeuralTalk (Karpathy & Fei-Fei, 2017). In this way, they obtain the probabilities for the sentences to select the hardest foil captions and prune the easier examples.

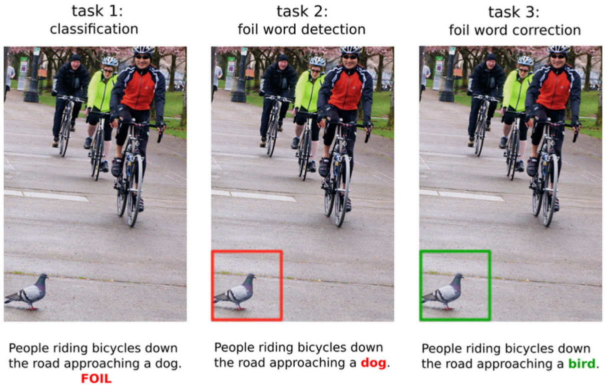

task 1:
classification
task 2:
foil word detection
task 3:
foil word correction
People riding bicycles down the road approaching a dog.
FOIL
People riding bicycles down the road approaching a dog.
People riding bicycles down the road approaching a bird.

Figure 3. FOIL tasks. Image credit: (Shekhar et al., 2017)

There are 3 main tasks that underlie this idea of diagnosing such systems as depicted in Figure 3:

- **Task 1**: Classify whether a caption is correct or foil
- **Task 2**: Find the foil word if a caption is a foil one
- **Task 3**: Change the foil word to correct the caption

To achieve high accuracies in these tasks, models should combine visual information and the semantics behind a given caption, to assess whether it is a correct caption or a foil one.

In this project, to see if the models indeed combine visual and linguistic information, we have focused on **the task of classifying whether a sentence is an original caption or a foil caption**, i.e. Task 1. Rationale behind this is that if a model is not able to incorporate information from two modalities, then it is expected that the model would fail in this task. On the other hand, a model that can obtain and blend sufficient information from two modalities is expected to outperform the former model.

4.2. Issues with FOIL.v1

FOIL.v1 has only been tested with IC and VQA tasks (i.e. non-interactive), whereas in this study we are also testing it with an interactive model as a contribution.

The main drawback of the FOIL.v1 dataset has been indicated by Madhyastha et al.. The authors claimed that blind models that take only the captions as input can perform much better than chance level in the FOIL tasks (above 85%). This indicates that text-only models are enough to detect whether a sentence is a foil or an original caption.

We have inspected the word frequencies and noticed that there seems to be a frequency bias for the target words, where some target words occurred mostly in foil sentences and some vice versa. Even with this information, it is possible to obtain better than chance level performance at FOIL Task 1.

4.2.1. FOIL.v2

To address the bias in the FOIL.v1 set, we started with replicating the steps in the creation of the FOIL set. Instead of using the 2014 split of MS-COCO, we utilized the 2017 split, which contains 118K/5K for train/validation examples. Additionally, instead of pruning for the hardest foil captions, we opted for keeping multiple foil captions as a means not to cause any further constraints that might elicit bias in the set. We also tried balancing the frequencies of target words in the foil and original sentences. Even though this subset was performing a little above chance level with a blind LSTM trained for Task 1, it was a rather small subset with only 16000 sentences in the training set, so that we decided not to use it.

4.2.2. CommonFOIL.v1

While FOIL.v2 generation still continues, Shekhar et al. has provided us with a subset of the FOIL.v1, which is obtained by taking the common images and captions that contain common vocabulary items with the GuessWhat?! dataset, VQA dataset and FOIL.v1. This dataset is called CommonFOIL.v1 and it helped reduce the language bias to obtain chance level accuracy (50%-55%) in a blind LSTM model that we implemented for this reason. It includes 10390 training images, 46044 captions, half of them being original captions and the other half of them being foil captions.

5. Models

5.1. Hierarchical Co-Attention Model for VQA

The Hierarchical Co-Attention model, extracts 3 levels of linguistic representations and then applies attention techniques between them and the visual feature representation (Lu et al., 2016). The words of the question are being propagated through convolutional layers so that linguistic information is being extracted, on word, phrase and question level. At the same time representation of visual information is also being extracted by the use of convolution layers. Finally, attention techniques are being applied on all combinations of linguistic and visual representations. The Hierarchical Co-Attention model achieved state-of-the-art results on the VQA dataset; hence, it was selected as the VQA model to test with the FOIL.v1 dataset (Shekhar et al., 2017).

5.2. GDSE - Interactive Visual Dialogue

The GDSE model uses convolutional layers to extract visual features and an LSTM for encoding the linguistic features obtained from the dialogue history consisting of previous questions and answers. (Venkatesh et al., 2018). Then, the concatenation of those features is being decoded, in a way so that it generates a question. The GDSE model achieved state-of-the-art results on the GuessWhat!? task, 50.8% surpassing the previous best model's accuracy by 10%.

6. Experiments

A models behaviour depends on its architecture and on the nature of the task that it was trained to perform. We utilize the aforementioned models in 3 different configurations: Frozen, Fine-tuned and Fully-trained. In the **Frozen** setting, we keep all the layers intact and only replace the last fully-connected (FC) layer and train only that part on the binary FOIL sentence classification task. In the **Fine-tuned** setting, we keep the weights that were obtained through the pre-training on the task in the respective paper and fine-tune all the layers on the FOIL classification task. By doing so, we assume that the values of the trained weight, are a good weight initiation, and then the model is trained with a smaller learning rate than the learning rate it was originally trained, so that the weight values will not be completely altered. **Fully-trained** configuration, on the other hand, trains all the layers from scratch on FOIL sentence classification.

7. Results

We ran the models explained in Section 5. Overall performance of the trained models are provided in Table 1. At the time of writing this report, we did not have the results for the fine-tuned HiCoAttn model, which is left empty in the table. In addition, the outcome of fully-training the GDSE model is NA (not applicable). We would like to keep its interactive nature, since if we fully train it on the FOIL set, this element is eliminated from the model.

In general, we notice that GDSE has higher accuracy than the non-interactive model. In addition, more in-depth training could provide higher accuracies in the FOIL caption classification task.

Table 1. Test accuracies for the models

Model	Frozen	Fine-Tuned	Fully-trained
HiCoAttn	50.23%	-	62.2%
GDSE	57.86%	62.99%	NA

Below, we list possible metrics that can be used in further analyses. Instead of simply employing the overall task performance, for a more thorough comparison of the interactive and non-interactive models, metrics below can be utilized.

- Overall accuracy
- Accuracy for the FOIL sentences
- Accuracy for the Original sentences
- Accuracy per caption length
- Accuracy per supercategory
- Correlation between accuracy and absolute and relative positions of the target word in FOIL sentences
- Correlation between target word frequencies in the training set and accuracies per object
- Correlation between the number of object instances in the image and the accuracy
- Correlation between the target object area in FOIL sentences and the accuracies

In Table 2, we provide the separate accuracies for the FOIL and original captions. We observe that overall and for the original captions, fine-tuned interactive model gives the best outcomes, whereas for the FOIL sentences, frozen non-interactive model outperforms the other models.

Table 2. Breakdown of the overall accuracy

Model	Overall	Original	Foil
GDSE-Frozen	57.86	73.15	**42.57**
GDSE-Fine-Tuned	**62.99**	**84.20**	41.77
VQA-Frozen	50.23	68.76	31.71

In Table 3, instead of accuracy, we provide the results of correlation between accuracy and a subset of the metrics for further analysis.

Only the correlation between accuracy and caption length for the frozen interactive model and the correlation between accuracy and absolute position for the fine-tuned interactive model seem significant.

Table 3. Correlations with accuracy per model. CL: Caption Length, AP: Absolute Position. TF: Target word frequency in the training set, NO: Number of objects in the image. * indicates significance.

	CL	AP	TF	NO
GDSE-frozen	-0.475 *	-0.176	0.04	0.1
GDSE-fine-tuned	-0.358	0.468 *	0.2	-0.11
VQA-frozen	0.211	0.346	0.04	-0.121

Since the values for the metrics were distributed non-normally, we take Spearman's rank correlation coefficients into account, Pearson's coefficients are also listed in tables 5, 6, 7 and 8 in the Appendix. For instance, the distribution of the caption lengths in the test set can be observed in Figure 4 in the Appendix. The histogram indicates a skewed distribution where there are very few sentences longer than around 20 words. Similarly, for the distribution of absolute positions (index of the target word in a foil caption), we observe a skewed histogram as in Figure 5. We also obtained the relative positions, which are positions normalized with respect to caption length. It also was not a normal distribution as can be seen in Figure 6.

In addition to accuracy, we also inspect Precision and Recall separately for FOIL and Original sentences. The outcomes for this analysis can be seen in Table 4.

Table 4. Precision and Recall for Foil and Original Sentences. PF: Precision for Foil, PO: Precision for Original, RF: Recall for Foil, RO: Recall for Original

	PF	PO	RF	RO
GDSE-frozen	0.613	0.56	**0.426**	0.732
GDSE-fine-tuned	**0.726**	**0.591**	0.418	**0.842**
VQA-frozen	0.504	0.502	0.317	0.688

It can be observed that, GDSE models, in particular the fine-tuned one, outperforms the other models in terms of Precision (Foil:72.6%, Original:59.1%) and Recall (Original: 84.2%).

8. Discussion

It can be claimed that, in general, the fine-tuned GDSE model seems to outperform other models, with few exceptions. The fine-tuned GDSE model is followed by the frozen GDSE model, which then outperforms the VQA model. This trend supports the hypothesis that adding interaction to the models could result in an increase in the performance of these models in the FOIL tasks. Additionally, as in the fine-tuned model, we apply a deeper training procedure, the network is better at classifying captions given images.

We notice that the performance of all the models is much better at identifying original sentences, when compared to their performance at identifying foil sentences. This might be due to the fact that there are more supporting facts when the image and the caption are congruent, as a result, the network is more confident when the caption is original. Whereas, in the case of foil sentences, the classification might not be backed confidently, which results in bad performance, even worse than chance level.

When we inspect the precision and recall values, we notice that again the interactive models outperform the non-interactive VQA model. The main finding here is that the precision for original sentences is lower than the precision for foil sentences. On the other hand, recall values demonstrate that the recall for original sentences is much higher compared to the foil ones. Even though the model is able to pinpoint original sentences precisely, a lot of foil sentences are also classified as original sentences, which negatively contribute to the precision for the foil sentences and the recall of the original sentences. The reason behind this could be similar to the results of accuracy, where the original sentences are backed up by the actual image more confidently in the network.

Correlation tests revealed that for the frozen GDSE model, accuracy and caption length are inversely proportional to each other significantly. This indicates that as the captions get longer, it becomes harder for a model to identify whether they are original or foil as depicted in Table 10. This is reasonable, since longer captions include more words that might be contributing to the confusion of the model.

Additionally, for the fine-tuned GDSE model, accuracy is positively correlated with the absolute target position significantly. Interestingly, for the frozen model we observe a non-significant inverse correlation. The first point could be explained with the fact that there are very few sentences longer than 20 words and they are usually classified correctly in the fine-tuned model as can be seen in Table 11. It might be argued that more words help the fine-tuned model to better discriminate between original and foil captions, whereas a frozen model might only account for sentences that have a target word close to the beginning. Fine-tuned model might have obtained the ability to better account for longer distances.

We did not obtain other significant correlations regarding the number of objects in the image, the frequency of the target words or object supercategories, indicating that they might not be deciding factors in the performance at this task.

9. Practical Issues in Artificial Intelligence

Algorithm Reliability: The codes for the models are provided by the authors of the cited articles and algorithm reliability is assumed. Additionally, further analysis is performed under very specific settings, making Algorithm Reliability checks implicitly present in the code.

Privacy: All datasets have been created for purposes of research and are under public use license.

Data Governance: The produced datasets Common-FOIL.v1 and FOIL.v2 will later be publicly available, and maintained, accompanied by several statistical information. Annotations are owned by the COCO Consortium, licensed under Creative Commons Attribution 4.0 license, whereas the images are subject to the Flickr Terms of Use (currently under Oath Terms of Service). We used the datasets for academic research purposes with no intention of other benefits. Other points might not be relevant in the context of our research.

10. Conclusions and Future Work

In this project, we have explored multimodal tasks that involve visual and linguistic inputs and tried to compare the performance of the non-interactive models to that of the interactive models in the FOIL task.

Overall, fine-tuned interactive model yielded the best accuracy result (62.00%). In the frozen configuration, dialogue model (57.86%) performed better than VQA model (50.23%). As a result, particularly for the frozen models, it can be claimed that interactive dialogue models can capture multimodal representations better than non-interactive models such as VQA models. This trend is expected in the other configurations, as well.

The generation of FOIL.v2 is still in progress, once it is finalized, we are planning to test the existence of the bias again. It might be possible to reduce the bias to some extent; however, unobserved reasons for linguistic bias might still be present in the dataset. It would also be beneficial to perform the other 2 FOIL tasks apart from the classification task, to observe whether the performances of the models in those tasks are in line with their performance in the classification task. For the further analysis of the models, an additional analysis could be conducted on the difficulty of identifying the FOIL object in the image through object detection and classification. We are also going to compare non-interactive and interactive models in terms of these metrics. Since we only had outcomes for dialogue models, we were only able compare within that group of models. In addition, we plan to apply the same evaluation on an image captioning model.

Acknowledgments

We would like to thank Assoc. Prof. Raquel Fernández, Dr. Elia Bruni and Aashish Venkatesh from the University of Amsterdam, The Netherlands, and Assist. Prof. Raffaella Bernardi, Ravi Shekhar and Claudio Greco from the University of Trento, Italy, for their valuable contributions and suggestions.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. *CoRR*, abs/1505.00468, 2015.

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., and Courville, A. C. GuessWhat?! Visual Object Discovery Through Multi-modal Dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667.

Jabri, A., Joulin, A., and van der Maaten, L. Revisiting visual question answering baselines. *CoRR*, abs/1606.08390, 2016.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, April 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2598339.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, USA, 2012. Curran Associates Inc.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

Lu, J., Yang, J., Batra, D., and Parikh, D. Hierarchical Question-image Co-attention for Visual Question Answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 289–297, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9.

Madhyastha, P. S., Wang, J., and Specia, L. Defoiling Foiled Image Captions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 433–438. Association for Computational Linguistics, 2018.

Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., and Bernardi, R. Foil it! Find One Mismatch between Image and Language Caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 255–265. Association for Computational Linguistics, 2017.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Proc. NIPS*, Montreal, CA, 2014.

Venkatesh, A., Baumgrtner, T., Bruni, E., Bernardi, R., and Fernández, R. Jointly Learning to See, Ask and GuessWhat?! Under Review. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Wang, C., Yang, H., Bartz, C., and Meinel, C. Image Captioning with Deep Bidirectional LSTMs. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pp. 988–997, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3603-1.

Xu, H. and Saenko, K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *CoRR*, abs/1511.05234, 2015.

A. Additional figures and tables

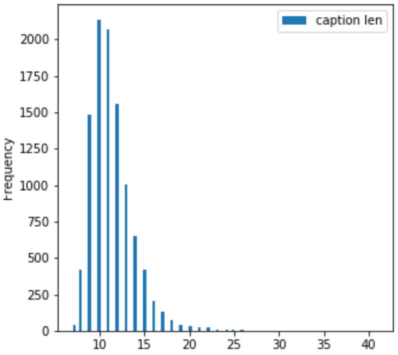


Figure 4. Distribution of caption lengths, normality tests significant

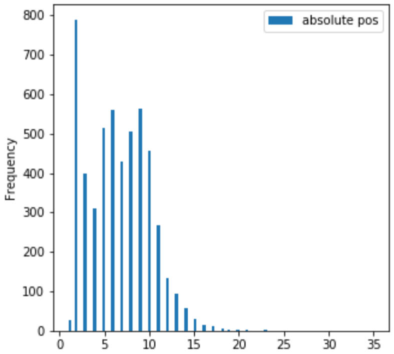


Figure 5. Distribution of absolute positions, normality tests significant

Table 5. Correlation between Accuracy vs. absolute position of the target word in the Foil sentence

	Pearson	Spearman
GDSE-Frozen	-0.456 *	-0.176
GDSE-Fine-tuned	0.533 *	0.468 *
VQA-Frozen	0.115	0.346

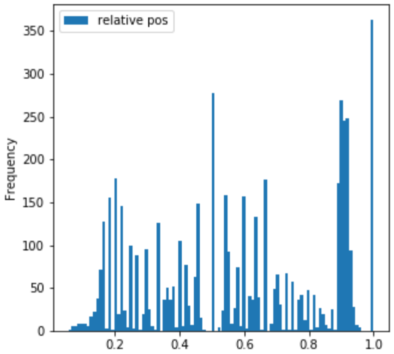


Figure 6. Distribution of relative positions, normality tests significant

Table 6. Correlation between Accuracy vs. caption length

	Pearson	Spearman
GDSE-Frozen	-0.383 *	-0.475 *
GDSE-Fine-tuned	-0.456 *	-0.358
VQA-frozen	0.305	0.211

Table 7. Correlation between Accuracy vs. word frequency in the training captions

	Pearson	Spearman
GDSE-Frozen	-0.01	0.04
GDSE-Fine-tuned	0.11	0.2
VQA-frozen	0.02	0.04

Table 8. Correlation between Accuracy vs. number of objects in the images

	Pearson	Spearman
GDSE-Frozen	0.084	0.1
GDSE-Fine-tuned	-0.199	-0.11
VQA-frozen	0.038	-0.121

Table 9. Correlation between Accuracy per supercategory

Supercategory	GDSE-frozen	GDSE-finetuned	VQA-frozen
Accessory	0.705	0.66	0.314
Animal	0.421	0.419	0.383
Appliance	0.42	0.348	0.326
Electronic	0.428	0.363	0.409
Food	0.275	0.291	0.355
Furniture	0.5	0.518	0.355
Indoor	0.482	0.461	0.234
Kitchen	0.521	0.514	0.296
Outdoor	0.6	0.1	0.3
Sports	0.39	0.402	0.261
Vehicle	0.419	0.393	0.23

Table 10. Caption lengths and average accuracies per model

Caption length	Count	GDSE-Frozen	GDSE-Fine-tuned	VQA-Fine-tuned
7	40	0.57	0.66	0.5
8	424	0.59	0.62	0.5
9	1484	0.59	0.62	0.5
10	2134	0.58	0.6	0.51
11	2066	0.59	0.63	0.5
12	1554	0.63	0.71	0.54
13	1006	0.58	0.65	0.52
14	650	0.57	0.73	0.5
15	420	0.55	0.56	0.53
16	202	0.49	0.62	0.46
17	128	0.56	0.67	0.49
18	74	0.56	0.6	0.5
19	44	0.58	0.66	0.47
20	36	0.6	0.55	0.4
21	28	0.5	0.67	0.5
22	24	0.47	0.61	0.47
23	10	0.7	0.7	0.4
24	12	0.61	0.61	0.5
25	8	0.63	0.63	0.5
26	10	0.5	0.5	0.5
27	2	0.5	0.5	0.5
28	4	0.5	0.7	0.8
29	2	0.25	0.5	0.5
30	4	0.5	0.5	0.5
31	2	0	0.5	0.5
41	2	0.5	0.5	0.75

Table 11. Absolute positions and average accuracies per model

Absolute position	GDSE-frozen	GDSE-fine tuned	VQA-frozen
1	0.296	0.296	0.22
2	0.401	0.36	0.31
3	0.436	0.411	0.31
4	0.452	0.333	0.33
5	0.463	0.482	0.33
6	0.405	0.437	0.33
7	0.438	0.427	0.3
8	0.41	0.461	0.32
9	0.402	0.402	0.32
10	0.447	0.414	0.31
11	0.44	0.437	0.3
12	0.396	0.44	0.34
13	0.474	0.411	0.28
14	0.414	0.379	0.31
15	0.567	0.567	0.4
16	0.688	0.688	0.31
17	0.308	0.385	0.38
18	0.4	0.4	0.4
19	0.5	1	0.25
20	0	0.5	0.5
21	0.5	1	0.5
22	0	0	1
23	0.333	0.667	0.33
35	0	1	0