Analyzing Compositionality of Visual Question Answering

Sanjay Subramanian[*] sanjays@allenai.org

Sameer Singh[†] sameer@uci.edu

Matt Gardner[*] mattg@allenai.org

[*] Allen Institute for Artificial Intelligence, Irvine CA
[†] University of California, Irvine CA

Abstract

Since the release of the original Visual Question Answering (VQA) dataset, several newer datasets for visual reasoning have been introduced, often with the express intent of requiring systems to perform compositional reasoning. Recently, transformer models pretrained on large amounts of images and associated text have been shown to perform much better than simple baselines on such compositional reasoning datasets as NLVR2 and GQA. In this paper, we analyze the performance of one of these models, LXMERT, on these two datasets. We show that despite the model's strong quantitative results, it may not be performing compositional reasoning because it does not need many relational cues to achieve this performance and more generally uses relatively little linguistic information. Our analysis utilizes experiments with relational linguistic cues removed, the input reduction technique, and a syntactic probe.

1 Introduction

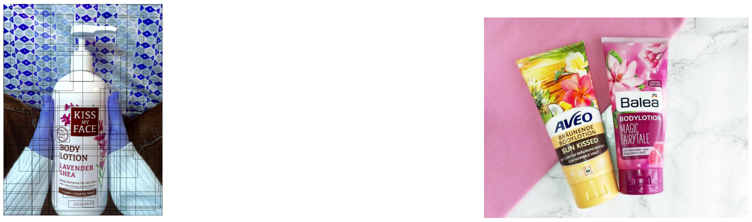
Compositionality is an important aspect of modes of communication employed by humans [Fodor and Lepore, 2002]. Therefore, if machines are to be effective at communicating with humans, machines must be able to do compositional reasoning. Question-answering involving both visual and language inputs offers an effective way to learn and evaluate compositional reasoning [Suhr et al., 2018]. Although early visual question answering datasets (e.g. [Agrawal et al., 2017]) did not directly assess the ability of systems to perform compositional reasoning, more recent datasets such as CLEVR [Johnson et al., 2017] and GQA [Hudson and Manning, 2019a] evaluate compositional reasoning via synthetically generated questions. A separate line of work, comprising primarily of the NLVR (Natural Language for Visual Reasoning) and NLVR2 datasets [Suhr et al., 2017, 2018], also evaluated compositional reasoning but used natural language. The images in the NLVR dataset are synthetically generated, while in NLVR2 each example consists of a sentence and two real photos.

Recently, several transformer models have achieved state-of-the-art (or near state-of-the-art) performance on some of these compositional VQA datasets when fine-tuned after pretraining on large amounts of image and text data [Tan and Bansal, 2019, Li et al., 2019]. Given the strong quantitative performances of the models, a natural question arises – are these models doing compositional reasoning?

In this work, we move toward an answer to this question by providing a preliminary analysis of the results of one transformer model, LXMERT, on the NLVR2 and GQA datasets. We find that without most relational cues, LXMERT can still achieve nearly the same performance on the NLVR2 dataset, and that seemingly difficult sentences can actually be easy for a model due to the images paired with them. In general, we find that LXMERT uses minimal linguistic information. Figure 1 shows an example in which, the model predicts the same result on all four examples for this sentence

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Label: True



Original Instance: “[CLS] the left and right image contains no more than three bottles of lot ##ion. [SEP]”

Reduced Instance: “[CLS] ~~the~~ ~~left~~ and right ~~image~~ ~~contains~~ no more ~~than~~ ~~three~~ bottles ~~of~~ lot ~~##ion~~. [SEP]”

Figure 1: Example instance from NLVR2 (with left and right images from a single image pair) which the model predicts correctly, along with the *reduced input* for which the model makes the same prediction. That is, the model makes the same prediction for all image pairs associated with this sentence when the sentence is reduced. This suggests that the model ignores important content words, like “three” in this case.

without the number “two” as with the number “two”. We use the following techniques to reach these conclusions:[1]

- We modify the NLVR2 and GQA datasets by masking or dropping selected tokens important for object relations and re-evaluate LXMERT.
- We apply input reduction [Feng et al., 2018], a method to maximally remove tokens that retain the model’s prediction, to LXMERT on the NLVR2 dataset.
- We probe the syntactic knowledge in LXMERT and compare it to BERT [Devlin et al., 2018]. This test allows us to assess the amount of linguistic information learned by LXMERT during its pre-training and contrast it with a model that has been shown previously to capture linguistic knowledge.

2 Removing Linguistic Cues

One way to ascertain whether a cue is important to a model’s predictions is to remove the cue systematically across the dataset and evaluate whether the model’s performance changes significantly. We adopt this approach here to determine whether LXMERT uses relational information to answer questions in NLVR2 and GQA. Specifically, we mask/drop prepositions and verbs, as these contain most, if not all, information about relations among objects in the images. We train on the original dataset and subsequently evaluate on the dataset with masked/dropped prepositions and verbs, and then we train on the dataset with masked/dropped prepositions and verbs and repeat the evaluation. We use the dependency parser and part-of-speech tagger of spaCy [Honnibal and Montani, 2017] to detect the prepositions and verbs, respectively.

We present the results in Tables 1 and 2. For NLVR2, we use accuracy and consistency (same as in the original paper); accuracy denotes the percentage of question-image pairs that are answered correctly, while consistency denotes the percentage of questions for which all question-image pairs are answered correctly. The accuracies for all experiments are clustered close together, though there is some more variation in the consistencies. For GQA, dropping the prepositions and verbs results in a large drop in accuracy for the models trained with the full sentences or masked prepositions/verbs,

[1] Note on Experimental Setup: We perform all of our evaluations on the validation/development sets of NLVR2 and GQA. LXMERT was originally fine-tuned on NLVR2 and GQA with a maximum sequence length of 20. However, the maximum sequence lengths for the sentences occurring in NLVR2 and GQA are 59 and 41, respectively. We set the maximum sequence lengths to 60 and 42 for NLVR2 and GQA, respectively, but do not observe any significant difference in performance compared to the original results.

Table 1: NLVR2: Masking and Dropping Prepositions and Verbs. Columns represent different evaluation modes, and rows represent different training modes.

Training Modification	**Original**		**Masked**		**Dropped**	
	Acc.	Cons.	Acc.	Cons.	Acc.	Cons.
None (Original)	0.745	0.406	0.732	0.387	0.719	0.365
Masked	0.738	0.396	0.732	0.383	0.726	0.378
Dropped	0.734	0.387	0.711	0.351	0.731	0.386

Table 2: GQA: Masking and Dropping Prepositions and Verbs. Columns represent different evaluation modes, and rows represent different training modes.

Training Modification	**Original**	**Masked**	**Dropped**
	Acc.	Acc.	Acc.
None (Original)	0.597	0.555	0.506
Masked	0.595	0.584	0.525
Dropped	0.595	0.564	0.577

suggesting that answering GQA questions containing relational information is more likely to require that relational information than answering NLVR2 questions. Still training the model on sentences with dropped prepositions/verbs recovers most of the drop in performance for GQA.

3 Input Reduction

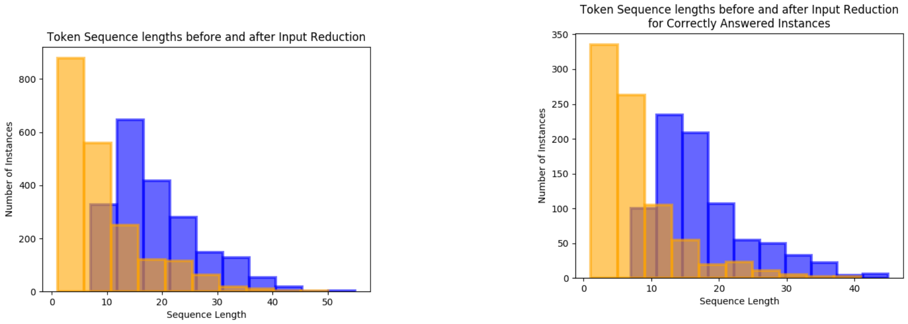
Input reduction [Feng et al., 2018] is a model analysis technique that iteratively removes a token from the input until the model's prediction changes. In open-ended question answering, input reduction often yields modifications that make the input nonsensical to humans yet preserve the model's prediction. We apply the input reduction method to the NLVR2 dataset. Our technique is novel in that we consider all examples with the same sentence to be part of a single input reduction instance – if the output of any of the individual examples changes, input reduction stops. Input reduction has previously been applied to the VQA dataset [Feng et al., 2018], but given that NLVR2 is designed to test compositional reasoning, the effectiveness of input reduction on NLVR2 is still interesting.

In our experiment, we ran the input reduction method on all sentences in the development set. Of the 2018 sentences in the development set, there are 819 sentences such that the model gives correct predictions for all of the image pairs associated with these sentences. Figure 2 shows the histograms for token sequence lengths before and after input reduction. Table 3 shows examples of reductions for sentences for which the model gave correct answers for all image pairs associated with the sentence. In these examples, omission of various pieces of relational (e.g. "in it" or "in front of") and even numeric ("three") information does not change the model's predictions on any image pair.

4 Probing Syntax Information

Finally, we present results from training a syntax probe on top of the representations of LXMERT using the data from NLVR2. We extract the dependency parse of each sentence using the spaCy dependency parser [Honnibal and Montani, 2017] and use the word-pair probe introduced by [Hewitt and Manning, 2019]. This probe consists of a linear projection of the transformer's representations into a vector space with smaller dimension. The probe is trained so that the distances between pairs of projected word vectors align with corresponding distances in the parse tree. These representations are frozen (after only pre-training; i.e. without fine-tuning on NLVR2 data). We also include results for BERT-base (using the final layer representations) for a comparison with a model that has been shown to capture linguistic information in its representations [Hewitt and Manning, 2019, Liu et al., 2019]. This comparison indicates the extent to which LXMERT learns linguistic knowledge in pre-training, which is relevant to its ability to do compositional reasoning.

[2] To obtain cross-modal outputs, we feed the left image of the example as input as well.



(a) All Sentences

(b) Sentences for which all sentence-image pairs were answered correctly by the model

Figure 2: Histograms showing token sequence lengths before and after input reduction for all sentences in NLVR2 development set. **Blue** corresponds to lengths before input reduction. **Orange** corresponds to lengths after input reduction.

Table 3: NLVR2 Selected Input Reduction examples of sentences for which the model was correct on all associated image-pairs.

Original	**Reduced**
[CLS] the left and right image contains no more than three bottles of lot ##ion. [SEP]	[CLS] and right no more bottles lot
[CLS] an image shows a man sitting in front of a computer screen. [SEP]	[CLS] man sitting screen
[CLS] at least one human is wearing eye glasses. [SEP]	eye
[CLS] exactly three white ducks are standing in a row on dry ground. [SEP]	[CLS] exactly three white ducks row
[CLS] at least 2 vulture ##s are sitting in a tree in one of the pictures. [SEP]	[CLS] least 2 vulture
[CLS] a black dug beetle is pushing a ball of dung in one image, and is without one in the other. [SEP]	dug beetle pushing ball dung
[CLS] a silver spoon has cookie dough in it. [SEP]	[CLS] silver spoon cookie

The evaluation metric that we use is the average Spearman correlation between distances between words in the dependency parse tree and the distances between the corresponding projected vectors that are produced by the probe.[3] The results are shown in Table 4. The fact that BERT performs better in the probing task can be explained by the difference between the pre-training data used for BERT and that used for LXMERT. Whereas BERT was pre-trained on the 800M-token BooksCorpus [Zhu et al., 2015] and English Wikipedia, LXMERT was pre-trained on visual question answering datasets and image captioning datasets (with 100M tokens). It can be inferred that sentences in LXMERT's pre-training data are not as linguistically rich or complex as sentences in BERT's pre-training data.

[3] As in work of Hewitt and Manning [2019], the average is across sentences with lengths between 5 and 50.

Table 4: Results of syntax probe on NLVR2 dataset.

Probe	**Avg. Spearman Correlation**
BERT Last Layer	0.842
LXMERT Language-only Transformer Last Layer	0.734
LXMERT Cross-Model Transformer Last Layer[2]	0.638

5 Conclusion

Our experiments suggest that LXMERT does not use much relational information to make predictions and relies on relatively few linguistic cues. To improve compositional reasoning on NLVR2, future work can consider introducing more inductive bias in the model or more supervision in fine-tuning as in previous work [Hudson and Manning, 2018, Hu et al., 2017, Hudson and Manning, 2019b]. To improve LXMERT's linguistic knowledge, future work should consider pre-training such a model on a larger, linguistically richer body of text.

References

Jerry A Fodor and Ernest Lepore. *The compositionality papers*. Oxford University Press, 2002.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123 (1):4–31, 2017.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 2019a.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Shi Feng, Eric Wallace, II Grissom, Mohit Iyyer, Pedro Rodriguez, Jordan Boyd-Graber, et al. Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017.

Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*, 2019b.