How clever is the FiLM model, and how clever can it be?

Alexander Kuhnle[0000−0002−2361−115X], Huiyuan Xie[0000−0001−8344−9724], and Ann Copestake[0000−0003−0347−946X]

Department of Computer Science and Technology
University of Cambridge, United Kingdom
{aok25,hx255,aac10}@cam.ac.uk

Abstract. The FiLM model achieves close-to-perfect performance on the diagnostic CLEVR dataset and is distinguished from other such models by having a comparatively simple and easily transferable architecture. In this paper, we investigate in more detail the ability of FiLM to learn various linguistic constructions. Our results indicate that (a) FiLM is not able to learn relational statements straight away except for very simple instances, (b) training on a broader set of instances as well as pretraining on simpler instance types can help alleviate these learning difficulties, (c) mixing is less robust than pretraining and very sensitive to the compositional structure of the dataset. Overall, our results suggest that the approach of big all-encompassing datasets and the paradigm of *"the effectiveness of data"* may have fundamental limitations.

Keywords: VQA · synthetic data · evaluation · deep learning

1 Introduction and related work

The task of Visual Question Answering (VQA) lies at the intersection of Computer Vision and Natural Language Processing. It generalizes the vision tasks of object detection and recognition to arbitrary visual-linguistic inferences, limited only by what can be queried by language.

In reaction to various issues that allowed comparatively naive models – for instance, a text-only system ignoring visual information and solely relying on language statistics – to achieve competitive performance on the popular VQA Dataset [1, 6, 15], abstract and (semi-)automatically generated datasets were introduced [10, 12, 20, 22]. Their motivation is to provide diagnostic tasks, with the aim of analyzing core abilities for visually grounded language understanding, like spatial reasoning or counting. CLEVR [10] is the most widely adopted of these, and several systems have now achieved near-perfect performance on it [8, 11, 18, 16, 19, 9, 14, 22].

One of the advantages of CLEVR is that it annotates questions from a set of instance types, like *"count"* or *"compare attribute"*, which makes a more detailed evaluation and model comparison possible. Building on the *"unit-testing"* proposal of [13] and related work for reading comprehension such as the bAbI

tasks [21], which take this idea of targeted evaluation further, we analyzed the FiLM model [16] on the ShapeWorld evaluation framework [12]. In doing so, we aim to investigate whether its close-to-perfect performance on CLEVR translates to ShapeWorld data as expected, and to shed more light on the strengths and weaknesses of FiLM.

Why FiLM? Arguably, it is one of the simplest models on that performance level for CLEVR. In particular, it does not rely on the semantic program trees underlying its instances, as compared to [8, 14, 11]. The first two strictly require the CLEVR-specific program vocabulary, which is different from the one used by ShapeWorld to generate data. The latter is agnostic to the vocabulary, but still sensitive to the size of the vocabulary, which is bigger for ShapeWorld[1]. Moreover, the code is open-source, and in our experiments we found that the model shows robust learning behavior on ShapeWorld data without any tuning of the CLEVR-based hyperparameters.

While FiLM manages to solve many tasks perfectly, it fails to learn anything on almost all datasets consisting of relational statements. We investigate how two approaches – broader training sets including simpler instances, and a version of curriculum learning [5, 3] – can make the difference between no learning at all and perfectly solving these datasets. However, we find that the first approach is very sensitive to details of the dataset structure. These results put into question the common assumption of *"the effectiveness of data"* [7] underlying datasets such as the VQA Dataset [2] (or SQuAD [17] for reading comprehension, or SNLI [4] for language inference): that all necessary abilities for a task can simply be learned from one big all-encompassing dataset, and that more data should lead to improved performance. Curriculum learning, on the other hand, shows promise as a robust approach to solving more complex instances of a task.

2   Experimental setup

2.1   Task

We look at the task of *image caption agreement*, that is, given a visual scene and a statement, decide whether the latter is true for the former. See figure 1 for some examples. The captions here are formal-semantics-style statements and not necessarily good descriptions, which is a vaguer concept and thus not as useful for evaluation. Instead, this task corresponds more to yes/no questions in visual question answering.

2.2   Datasets

We generated various datasets based on existing configurations in the Shape-World repository. The different datasets are defined by the types of captions they contain, see figure 1 for more details. Each dataset consists of 500k training instances, plus 10k validation and test instances. Training and validation scenes

---

[1] We ran into memory issues when trying to run this model on ShapeWorld data.

generally contain 1-4, 6-9 or 11-14 non-overlapping (unless mentioned otherwise) objects, further constrained depending on the dataset. Test scenes may in addition exhibit the withheld object numbers 5, 10 and 15, and may also contain withheld object types: *"red square"*, *"green triangle"*, *"blue circle"*, *"yellow rectangle"*, *"magenta cross"*, *"cyan ellipse"*. Consequently, the test data follows a slightly different distribution where models is required to generalize to unseen object numbers and new attribute combinations to achieve a comparable score, similar to the CoGenT version of the CLEVR dataset[2] [10].

Existential: "There is a red square.", "A red shape is a square."
Single-shape: same as above, with only one object present
Logical: two existential statements connected by: and, or, if, if and only if
Numbers: zero to five; with modifiers: less/more than, at most/least, exactly, not
Quantifiers: with modifiers as above: no, half, all, a/two third(s), a/three quarter(s)
Relational: left, right, above, below, closer, farther, darker, lighter, smaller, bigger, same/different shape/color
Simple-spatial: the first four spatial relations, with only two objects per scene
Relational-negation: relational plus negated relations
Implicit-relational: left, right, upper, lower, smaller, bigger, darker, lighter, closer, farther (of two potential objects)
Superlatives: superlative forms of the above, of an arbitrary number of objects

Examples for visual scenes   **Examples for true or false statements**

- ○ *"There is a cyan square or a circle is green."*
- ○ *"At least two shapes are green."*
- ○ *"More than half the pentagons are red."*
- ○ *"A red cross is to the left of a yellow shape."*
- ○ *"The left circle is blue."*
- ○ *"The lowermost yellow shape is a circle."*

Fig. 1. *Top*: All basic datasets we experimented with, together with their central words/constructions. *Bottom left*: Two example images. *Bottom right*: Some example captions of different datasets (LOGICAL, NUMBERS, QUANTIFIERS, RELATIONAL, IMPLICIT-RELATIONAL, SUPERLATIVES)

2.3   Models

We focus on the FiLM model [16] here. The image is processed using a six-layer CNN (stride of two after the third and sixth layer) trained from scratch on the task. We found that the common approach of using a pretrained ResNet module did not perform well on our data. The caption as 'question' is processed by a GRU. In four residual blocks, the processed image tensor is linearly modulated

---

[2] Note, however, that CLEVR CoGenT requires stronger generalization skills, as more shape-color combinations per shape/color are withheld.

conditioned on the caption embedding. Following global max-pooling, the classifier module produces the 'answer', i.e. *"true"* or *"false"* in our case. We train the model for 100k iterations in all experiments, using the default hyperparameters. Training performance is measured on the validation set every 1k iterations for the first 10k iterations and every 5k afterwards. We also compare performance on selected datasets to two common baselines [10]: CNN-LSTM and CNN-LSTM-SA. We will release the ShapeWorld-adapted FiLM repository and the generator configurations to create the datasets on acceptance of the paper.

3   Results

Detailed results are presented in figure 2 and 3, unless referred to the appendix.

Initial findings, and what did not work.

(a) In a first experiment, we did not explicitly configure data generation to prevent overlapping objects. This turned out to be a major obstacle for learning in most cases. While FiLM solved EXISTENTIAL (99.7%), performance on NUMBERS stayed at chance level (55.2%) (see appendix A.1).

(b) We experimented with using a fixed/trainable pretrained ResNet module instead of a custom CNN. Both versions of the model reached an accuracy of 65-70% after 100k iterations on EXISTENTIAL, which is substantially lower than our final result of 100% (see appendix A.1).

(c) The FiLM model successfully solves many of our datasets. EXISTENTIAL is mastered after only 10k iterations and at the same speed as the trivial SINGLE-SHAPE variant. LOGICAL, NUMBERS and QUANTIFIERS reach close-to-perfect accuracy after around 60k iterations. The learning curves for these three tasks look remarkably alike and thus suggest a similar learning complexity for the model.

(d) The FiLM model successfully generalizes to the test set in most cases. Only for the simplified variants SINGLE-SHAPE and SIMPLE-SPATIAL, test performance is markedly lower, suggesting that there is not enough incentive to learn a compositional representation, presumably because their simplicity makes overfitting a feasible option.

(e) We investigated the performance of two common baselines, CNN-LSTM and CNN-LSTM-SA. While FiLM consistently outperforms both baselines as expected, the supposedly superior CNN-LSTM-SA [23, 10] does not always improve upon the results of CNN-LSTM. However, CNN-LSTM-SA in some cases shows stronger generalization to the test distribution, whereas performance always drops for CNN-LSTM (also see appendix A.2).

Failure to learn relational statements. Surprisingly, we found that, with the exception of SIMPLE-SPATIAL, FiLM struggles to learn anything when trained on the various datasets requiring some form of relational reasoning: RELATIONAL, IMPLICIT-RELATIONAL and SUPERLATIVES (RELATIONAL-LIKE below). We also tried subsets of relations in RELATIONAL (e.g., only spatial relations), with the

Dataset	CNN-LSTM		CNN-LSTM-SA		FiLM	
single-shape	—		—		100.0	87.2
existential	100.0	81.1	100.0	99.7	100.0	99.9
logical	79.7	62.2	76.5	58.4	99.9	98.9
numbers	75.0	66.4	99.1	98.2	99.6	99.3
quantifiers	72.1	69.1	84.8	80.8	97.7	97.0
simple-spatial	81.4	64.8	81.9	57.7	85.1	61.3
relational	—		—		50.6	51.0
implicit-rel	—		—		52.9	53.2
superlatives	—		—		50.8	50.2
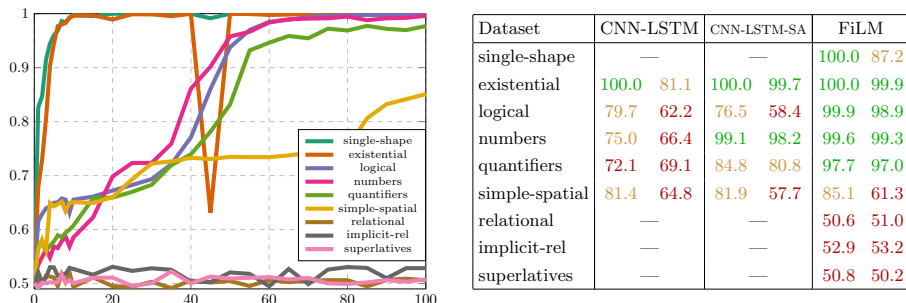
Fig. 2. *Left diagram*: validation performance of the FiLM model trained on various ShapeWorld datasets separately (*x-axis*: iterations in 1000, *y-axis*: accuracy). *Top right table*: final validation (*left*) and test (*right*) accuracy of the trained FiLM models, plus performance of the two baselines on selected datasets (in percent, *green*: $\geq 95\%$, *orange*: $\geq 75\%$, *red*: $< 75\%$)

same result. The only exception is the simplistic two-object SIMPLE-SPATIAL. But even here, learning is comparatively slow and only reaches ~85% after 100k iterations (although the curve indicates that performance is still improving), which further emphasizes the complexity of learning relations for FiLM.

Training on a broader set of instances. Datasets like CLEVR consist of a mix of instances requiring different abilities. Our assumption is that the simpler instances help to stabilize and guide the overall learning process, so that the more complex instances are also learned eventually[3], hence models are able to achieve close-to-perfect performance. We tested this assumption in three setups: First, the FiLM model reaches ~95% accuracy on a dataset augmenting the complex RELATIONAL with the *'pedagogical'* SIMPLE-SPATIAL dataset. Second, when trained on a broader mix of EXISTENTIAL, LOGICAL, NUMBERS, QUANTIFIERS and either of the RELATIONAL-LIKE datasets, instances of the latter dataset are also successfully learned (also see appendix A.3). Finally, in the failure case of NUMBERS for images with overlapping objects, training on a combination with the EXISTENTIAL dataset helps the model to also solve instances of the former.

Improvements via augmenting/mixing are unstable. Further investigation reveals that this *'synergy effect'* of combining different instances is very sensitive to the composition of the training set. For instance, an unbalanced distribution of 45% or 60% SIMPLE-SPATIAL and 55% or 40% RELATIONAL shows no improvement above chance level (see appendix A.5). Similarly, performance stagnates when training on a combination of SIMPLE-SPATIAL and RELATIONAL-NEGATION instead. In the second example above, FiLM sometimes fails to learn mixed datasets with two or more RELATIONAL-LIKE components (see appendix A.4). Of these, RELATIONAL seems to be the most complex for FiLM.

---

[3] When referring to "simple" and "complex" or "difficult" instances here, we always mean with respect to the ability of the FiLM model to learn these instances.
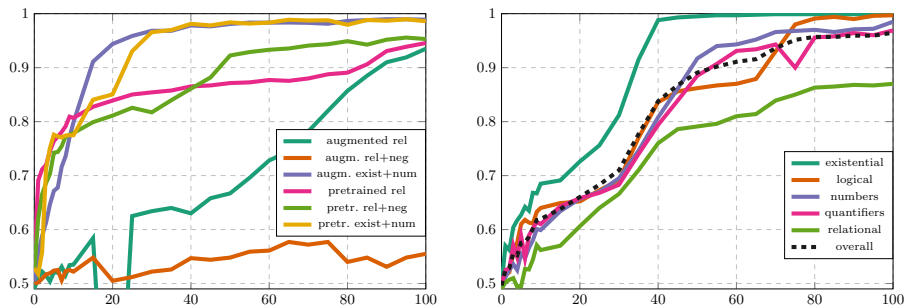
Fig. 3. *Left diagram*: FiLM performance on the RELATIONAL / EXISTENTIAL+NUMBERS (with overlap) dataset, when pretrained on SIMPLE-SPATIAL / EXISTENTIAL instances, or trained on a combination. *Bottom right diagram*: performance per dataset of the FiLM model trained on a broader mix of datasets

The effectiveness of pretraining. In another series of experiments, we investigated whether pretraining on simpler instances can bootstrap a successful learning process on more complex datasets, which is the assumption underlying curriculum learning [5, 3]. For this, we take the model trained for 100k iterations on SIMPLE-SPATIAL and apply it to other RELATIONAL-LIKE datasets. For both RELATIONAL as well as RELATIONAL-NEGATION we observe a sharp increase in performance at the start, reaching ∼95% accuracy after 100k iterations. We particularly want to draw attention to the fact that the pretrained model reaches and eventually surpasses its previous performance level of ∼85% after only 20k/40k iterations, despite the more complex instances. Note also that the model trained on RELATIONAL-NEGATION at some point seems to benefit from this dataset's increased complexity. Finally, we also confirmed that, in the case of overlapping objects, the system pretrained on EXISTENTIAL is subsequently also able to learn added NUMBERS instances.

4   Discussion and conclusion

We have shown how the FiLM model is not able to learn to correctly understand relational statements when trained on a dataset of such statements only. Furthermore, we have investigated two mechanisms which help alleviate these difficulties: augmenting training data with instances that are easier to learn, and pretraining on such simpler instances before moving to more complex ones. The first approach turns out to be very sensitive to the precise composition of the training set, while the second one leads to more reliable improvements.

Augmenting datasets in the limit corresponds to big all-encompassing datasets for general tasks like VQA, where a variety of skills is assumed to be learned implicitly from a lot of input-output pairs. While our results confirm that this is possible (at least for synthetic data), they strongly question the robustness of this process. We saw how otherwise successful learning breaks down when the

combined dataset is too complex or the mixing distribution is chosen wrongly. We emphasize that these findings are based on perfectly clean and controlled abstract data, while the situation is obviously more complex for real-world datasets. Such sensitivity of the learning process to such structural details of the training data is usually not considered, but might be able to explain some of the instability effects that are generally attributed to hyperparameter choice and random seeds. Since it is hard to conceive how real-world data could ever be controlled to the degree possible with synthetic data, we should be far more skeptical of very complex architectures for only a single dataset, and instead encourage the reporting of negative instability/transferability results. As a way forward, our findings suggest the potential of curriculum learning as a more robust alternative to bigger monolithic datasets.

Acknowledgments

A   Learning curves for other experiments

A.1   Overlapping objects & pretrained ResNet module
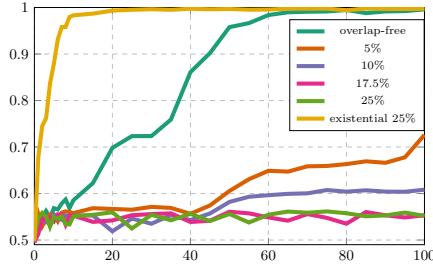
Performance of FiLM on NUMBERS,
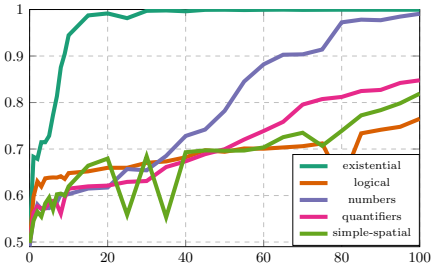controlled for max area overlap
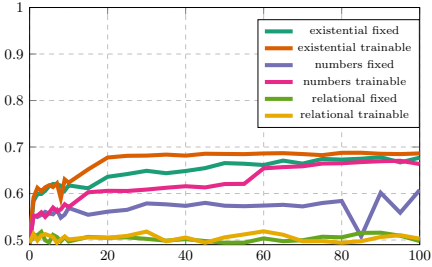
Performance of FiLM using a
pretrained ResNet module

A.2   Baselines: CNN-LSTM and CNN-LSTM-SA
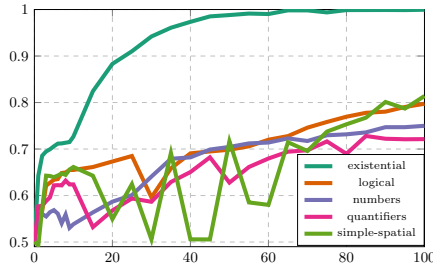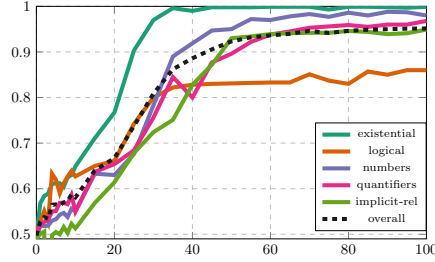

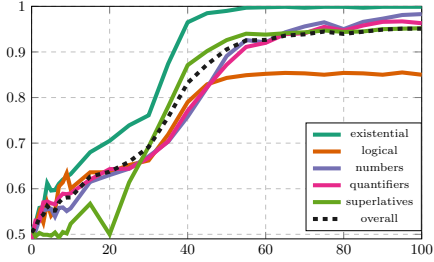Performance of CNN-LSTM on
various datasets

Performance of CNN-LSTM-SA on
various datasets

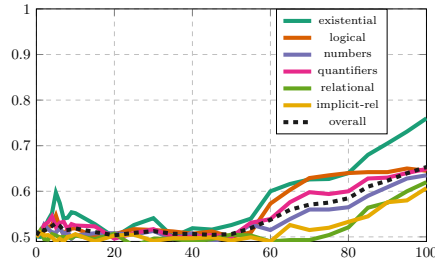### A.3   Combinations with implicit-relational and superlatives

Per-dataset performance on a mix
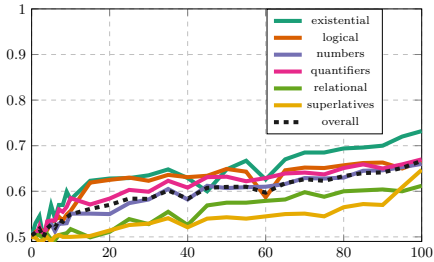including IMPLICIT-RELATIONAL

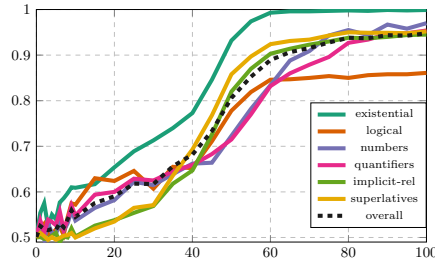Per-dataset performance on a mix
including SUPERLATIVES

A.4   Combinations with multiple relational-like datasets
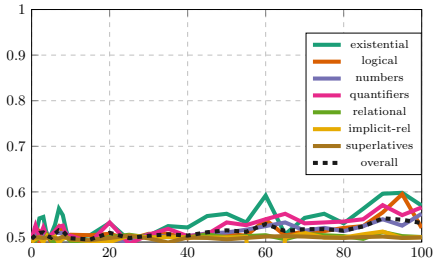

Performance on a mix including
RELATIONAL and IMPLICIT-REL.

Performance on a mix including
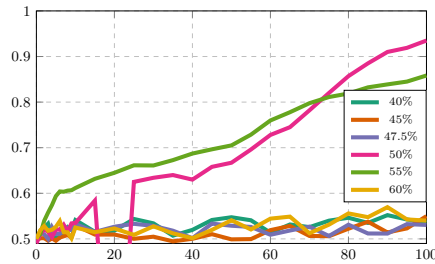RELATIONAL and SUPERLATIVES

Performance on a mix including
IMPLICIT-REL. and SUPERLATIVES

Performance on a mix including all
RELATIONAL-LIKE datasets

A.5   Mixing distribution

Performance on a combination of
SIMPLE-SPATIAL and RELATIONAL,
controlled for the probability of
SIMPLE-SPATIAL

References

1. Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1955–1960. EMNLP 2016, Association for Computational Linguistics, Stroudsburg, PA, USA (2016)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV 2015, IEEE Computer Society, Washington, DC, USA (2015)
3. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 41–48. ICML 2009, ACM, New York, NY, USA (2009)
4. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Màrquez, L., Callison-Burch, C., Su, J. (eds.) Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 632–642. EMNLP 2015, Association for Computational Linguistics, Stroudsburg, PA, USA (2015)
5. Elman, J.L.: Learning and development in neural networks: the importance of starting small. Cognition $\mathbf{48}$(1), 71–99 (1993)
6. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6325–6334. CVPR 2017, IEEE Computer Society, Washington, DC, USA (2017)
7. Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE Intelligent Systems $\mathbf{24}$(2), 8–12 (2009)
8. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: End-to-end module networks for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV 2017, IEEE Computer Society, Washington, DC, USA (2017)
9. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. In: Proceedings of the International Conference on Learning Representations. ICLR 2018 (2018)
10. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2017, IEEE Computer Society, Washington, DC, USA (2017)
11. Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Inferring and executing programs for visual reasoning. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV 2017, IEEE Computer Society, Washington, DC, USA (2017)
12. Kuhnle, A., Copestake, A.: ShapeWorld - A new test methodology for multimodal language understanding. ArXiv e-prints 1704.04517 (2017)
13. Kuhnle, A., Copestake, A.: Deep learning evaluation using deep linguistic processing. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the Workshop on Generalization in the Age of Deep Learning. pp. 17–23. NAACL 2018, Association for Computational Linguistics, Stroudsburg, PA, USA (2018)

14. Mascharka, D., Tran, P., Soklaski, R., Majumdar, A.: Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2018, IEEE Computer Society, Washington, DC, USA (2018)
15. Mudrakarta, P.K., Taly, A., Sundararajan, M., Dhamdhere, K.: Did the model understand the question? ArXiv e-prints 1805.05492 (2018)
16. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: FiLM: Visual reasoning with a general conditioning layer. In: AAAI. AAAI Press, Palo Alto, CA, USA (2018)
17. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. EMNLP 2016, Association for Computational Linguistics, Stroudsburg, PA, USA (2016)
18. Santoro, A., Raposo, D., Barrett, D.G.T., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.P.: A simple neural network module for relational reasoning. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. pp. 4974–4983. Curran Associates, Inc., Red Hook, NY, USA (2017)
19. Suarez, J., Johnson, J., Li, F.: DDRprog: A CLEVR differentiable dynamic reasoning programmer. ArXiv e-prints 1803.11361 (2018)
20. Suhr, A., Lewis, M., Yeh, J., Artzi, Y.: A corpus of natural language for visual reasoning. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. ACL 2017, Association for Computational Linguistics, Stroudsburg, PA, USA (2017)
21. Weston, J., Bordes, A., Chopra, S., Mikolov, T.: Towards AI-complete question answering: A set of prerequisite toy tasks. ArXiv e-prints 1502.05698 (2015)
22. Yang, G.R., Ganichev, I., Wang, X.J., Shlens, J., Sussillo, D.: A dataset and architecture for visual reasoning with a working memory. ArXiv e-prints 1803.06092 (2018)
23. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2016, IEEE Computer Society, Washington, DC, USA (2016)