On transfer learning using a MAC model variant: supplementary material

Vincent Marois **T.S. Jayram** **Vincent Albouy**
Tomasz Kornuta **Younes Bouhadjar** **Ahmet S. Ozcan**
IBM Research AI, Almaden Research Center, San Jose, USA
{vmarois,jayram,tkornut,byounes,asozcan}@us.ibm.com
{vincent.albouy}@ibm.com

Abstract

We introduce a variant of the MAC model (Hudson and Manning, ICLR 2018) with a simplified set of equations that achieves comparable accuracy, while training faster. We evaluate both models on CLEVR and CoGenT, and show that, transfer learning with fine-tuning results in a 15 point increase in accuracy, matching the state of the art. Finally, in contrast, we demonstrate that improper fine-tuning can actually reduce a model's accuracy as well.

A Description of the datasets

Most of the VQA datasets have strong biases. This allow models to learn strategies without reasoning on the visual input [SRB+17]. The CLEVR dataset [JHvdM+17a] was developed to address those issues and come back to the core challenge of visual QA which is testing reasoning abilities. CLEVR contains images of 3D-rendered objects; each image comes with a number of highly compositional questions which fall into different categories. These categories are divided into 5 classes of tasks: Exist, Count, Compare Integer, Query Attribute and Compare Attribute. The CLEVR dataset consists of:

- A training set of 70k images and 700k questions,
- A validation set of 15k images and 150k questions,
- A test set of 15k images and 150k questions about objects,
- Answers, scene graphs and functional programs for all train and val images and questions.

Each object present in the scene is characterized, aside its position, by a set of four attributes:

- 2 sizes: large, small,
- 3 shapes: square, cylinder, sphere,
- 2 material types: rubber, metal,
- 8 color types: gray, blue, brown, yellow, red, green, purple, cyan,

resulting in 96 unique combinations.

Along with CLEVR, the authors [JHvdM+17a] introduced CLEVR-CoGenT (Compositional Generalization Test, CoGenT in short), with a goal of evaluating how well the models can generalize, learn relations and compositional concepts. This dataset is generated in the same way as CLEVR, with two conditions. As shown in Table 1, in Condition A all cubes are gray, blue, brown, or yellow, whereas all cylinders are red, green, purple, or cyan; in Condition B cubes and cylinders swap color palettes. For both conditions, spheres can be of any colors.

The CoGenT dataset contains:

Visually Grounded Interaction and Language (ViGIL) Workshop, NIPS 2018, Montréal, Canada.

- Training set of 70,000 images and 699,960 questions in Condition A,
- Validation set of 15,000 images and 149,991 questions in Condition A,
- Test set of 15,000 images and 149,980 questions in Condition A,
- Validation set of 15,000 images and 150,000 questions in Condition B,
- Test set of 15,000 images and 149,992 questions in Condition B,
- Answers, scene graphs and functional programs for all training and validation images and questions.

Table 1: Colors/shapes combinations present in CLEVR, CoGenT-A and CoGenT-B datasets.

Dataset	Cubes	Cylinders	Spheres
CLEVR	any color	any color	any color
CLEVR CoGenT A	gray / blue / brown / yellow	red / green / purple / cyan	any color
CLEVR CoGenT B	red / green / purple / cyan	gray / blue / brown / yellow	any color

B Full MAC and S-MAC comparison

In Table 2 we present the full comparison between MAC and S-MAC models achieved with our implementations of both models. In the [Row] column we indicate the results that we have analyzed and discussed in the experiments section of the main paper.

Table 2: CLEVR & CoGenT accuracies for the MAC & S-MAC models.

Model	Training			Fine-tuning		Test		Row
	Dataset	Time [h:m]	Acc [%]	Dataset	Acc [%]	Dataset	Acc [%]	
MAC	CLEVR	30:52	96.70	–	–	CLEVR	96.17	(a)
						CoGenT-A	96.22	
						CoGenT-B	96.27	
				CoGenT-A	98.06	CoGenT-A	94.60	
						CoGenT-B	93.28	
				CoGenT-B	98.16	CoGenT-A	93.02	
						CoGenT-B	94.44	
	CoGenT-A	30:52	97.02	–	–	CoGenT-A	96.88	
						CoGenT-B	79.54	
				CoGenT-B	97.91	CoGenT-A	92.06	
						CoGenT-B	95.62	
S-MAC	CLEVR	28:30	95.82	–	–	CLEVR	95.29	(b)
						CoGenT-A	95.47	(d)
						CoGenT-B	95.58	(e)
				CoGenT-A	97.48	CoGenT-A	93.44	
						CoGenT-B	92.31	
				CoGenT-B	97.67	CoGenT-A	92.11	(i)
						CoGenT-B	92.95	(j)
	CoGenT-A	28:33	96.09	–	–	CoGenT-A	95.91	(c)
						CogenT-B	78.71	(f)
				CoGenT-B	96.85	CoGenT-A	91.24	(g)
						CoGenT-B	94.55	(h)

C Comparison of the generalization capabilities on CoGenT

In this section, we present a comparison of our results on generalization capabilities with selected state-of-the-art models. In particular, we focused on three papers reporting state-of-the-art accuracies on CoGenT. These papers introduce the following models: PG+EE [JHvdM$^+$17b], FiLM [PSDV$^+$17] and TbD [MTSM18]. Deeper analysis of these papers revealed that it is likely that different authors used different sets for reporting the scores, which questions the correctness of the comparison. We find that the problem results from the fact that ground truth answers for the test sets are not provided along with these sets; thus subsets of the validation sets were sometimes used for testing. The results of our research are presented in Table 3, where we shortened the names of the datasets. For instance, **A Test Full** means the use of the whole **CoGenT Condition A Test set**, whereas **B Valid 30k** indicates the use of 30.000 samples from **CoGenT Condition B Validation set**. Question marks indicate that the paper does not provide enough information; thus the indicated sets are the ones we assumed were used.

Table 3: Generalization capabilities of selected state-of-the-art models.

Model	Training		Fine-tuning		Test	
(source)	CoGenT set	Acc [%]	CoGenT set	Acc [%]	CoGenT set	Acc [%]
PG+EE ([JHvdM$^+$17b])	A Train Full?	N/A	–	–	A Test Full	96.6
					B Test Full?	73.7
			B Train 30k?	N/A	A Test Full	76.1
					B Test Full	92.7
CNN+GRU+FiLM 0-Shot ([PSDV$^+$17])	A Train Full?	N/A	–	–	A Valid Full?	98.3
					B Valid 120k	78.8
			B Valid 30k	N/A	A Valid Full?	81.1
					B Valid 120k	96.9
TbD + reg ([MTSM18])	A Train Full?	N/A	–	–	A ?	98.8
					B ?	75.4
			B Valid 30k	N/A	A ?	96.9
					B ?	96.3
MAC (our results)	A Train 630k	97.02	–	–	A Valid Full	96.88
					B Valid Full	79.54
			B Valid 30k	97.91	A Valid Full	92.06
					B Valid 120k	95.62
S-MAC (our results)	A Train 630k	96.09	–	–	A Valid Full	95.91
					B Valid Full	78.71
			B Valid 30k	96.85	A Valid Full	91.24
					B Valid 120k	94.55

C.1 The PG+EE model and training methodology

The PG+EE (Program Generator and Execution Engine) [JHvdM$^+$17b] model is composed of two main modules: a Program Generator constructing an explicit, graph-like representation of the reasoning process, and an Execution Engine executing that program and producing an answer. Both modules are implemented by neural networks, and were trained using a combination of backpropagation and REINFORCE [Wil92].

The authors inform that in the first step they trained their models on Condition A, and tested them on both conditions. Next, they fine-tuned these models on Condition B using 3K images and 30K questions, and again tested on both conditions. +We could not ascertain which sets they used for fine-tuning (as Condition B lacks a training set). Being the authors of the CLEVR and CoGenT datasets, it is possible that they generated specific sets. Additionally, as they own the ground truth answers for the test sets, it can be assumed that they reported the accuracies on both Condition A and Condition B test sets, although it is not indicated in the paper.

C.2 The FiLM model and training methodology

Feature-wise Linear Modulation (in short, FiLM) [PSDV$^+$17] is an optional enhancement of a neural network model. The idea is to influence the behavior of existing layer(s) by introducing feature-wise affine transformations which are conditioned on the input. A model composed of CNNs and a GRU with FiLM-enhanced layers achieved state-of-the-art results on both CLEVR and CoGenT, showing improvement over the PG+EE model.

Nonetheless, the authors indicate in the paper that the accuracy reported for Condition B after fine-tuning was obtained on the CoGenT Condition B Validation set, excluding the 30k samples which were used for fine-tuning. This suggests that they probably reported scores for Condition A on the validation set, not the test set.

C.3 The TbD model and training methodology

The TbD (Transparency by Design) network was introduced in [MTSM18]. TbD is composed of a set of visual reasoning primitives relying on attention transformations, allowing the model to perform reasoning by composing attention masks. The authors compare the accuracy of their model tested on CLEVR, alongside with several existing models. In particular, they show improvement over the previously mentioned PG + EE, CNN + GRU + FiLM and MAC models.

They also compare their performance on CoGenT against PG+EE. Nonetheless, the description lacks information about the used sets for training, fine-tuning and testing. They indicate that they used 3k images and 30k questions from the CoGenT Condition B for fine-tuning of their model, but do not explicitly point the set they used the samples from.

C.4 Our MAC and S-MAC models and methodology

Due to the fact that the test sets ground truth labels for both CLEVR and CoGenT aren't publicly available, we decided to follow the approach proposed in [PSDV$^+$17] and split the CoGenT B Validation set. As a result we used 30k samples for fine-tuning and the remaining ones for testing. When testing the model on CoGenT A, we used the whole validation set.

Moreover, as we were using the validation sets for testing, we also splitted the training set and used 90% for training and the remaining 10% for validation during training. +We used the same methodology for all the results presented in this paper, i.e. Table 2 in the main paper, Table 2 and Table 3 here.

D Illustration of failures of MAC/S-MAC on CLEVR

Following the evaluation of MAC on CoGenT-B, we built a tool which helped us visualizing the attention of the model over the question words and the image, and thus provide insight on some cases of failure.

Fig. 1 presents a question where the model is asked about the shape of the leftmost gray cylinder. The model correctly finds it, as we can see from its visual attention map, and appears to refer to it using its color (*gray*), as we can see from the attention of the question words. Yet, it defaults to predicting the shape as *cube*, because it never saw gray cylinders during training, but instead saw gray cubes.

Fig. 2 presents a similar case, where the model is questioned about the color of the green cube at the back. MAC misses that object, and instead focuses on the nearby gray cylinder. We can hypothesize that MAC missed the green cube as it did not see this combination during training, and thus defaulted to a combination that it knows.

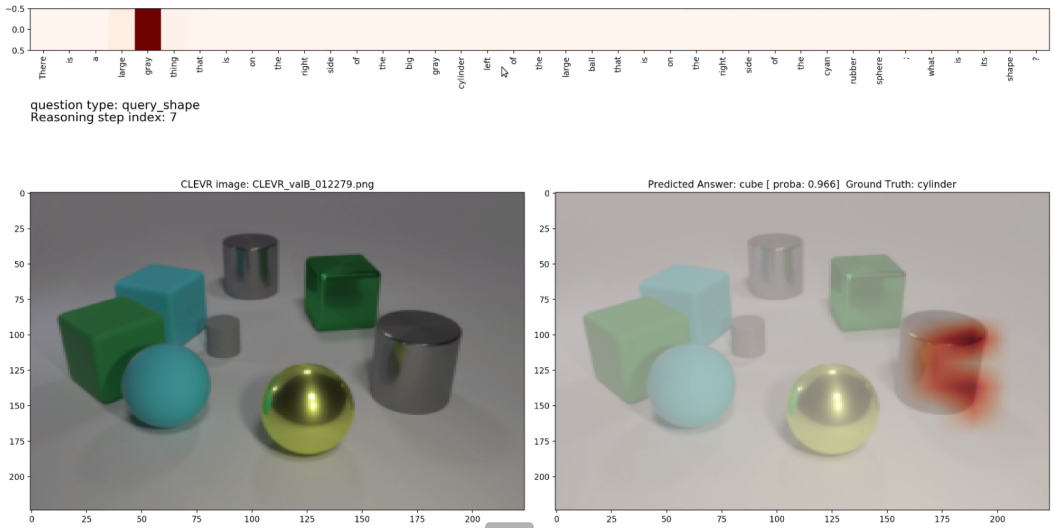

Figure 1: The question reads as: *There is a large gray thing that is on the right side of the big gray cylinder left of the large ball that is on the right side if the cyan rubber sphere; what is its shape?*

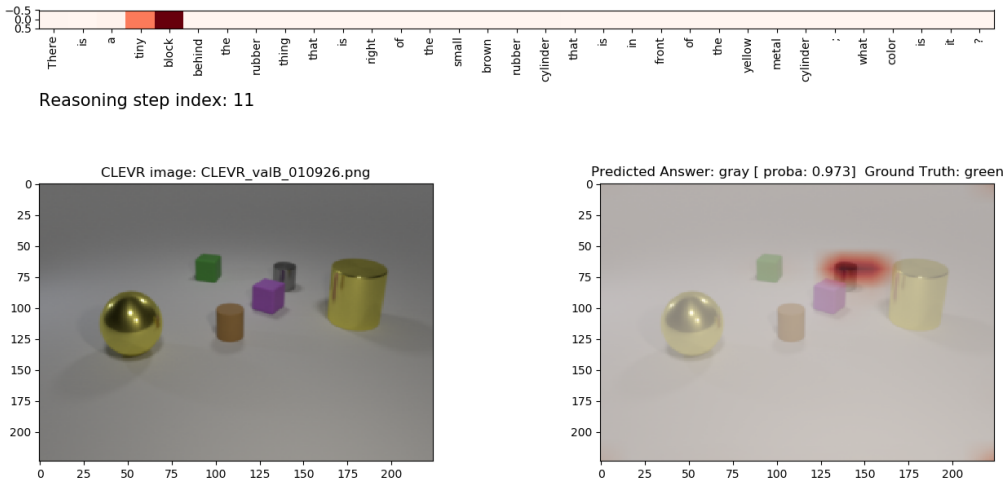

Figure 2: The question reads as: *There is a tiny block behind the rubber thing that is right if the small brown rubber cylinder that is in front of the yellow metal cylinder; what color is it?*

These examples indicate that MAC did not correctly separate the concept of shape from the concept of color, but have a better understanding of the colors (as it found the object of interest in Fig. 1 by its color). This could come from that fact that the shape *sphere* is associated with all possible colors in the dataset.

References

[JHvdM+17a] 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 *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE, 2017.

[JHvdM+17b] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross B Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, pages 3008–3017, 2017.

[MTSM18] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. 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*, pages 4942–4950, 2018.

[PSDV+17] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*, 2017.

[SRB+17] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy P. Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017.

[Wil92] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.