On transfer learning using a MAC model variant

Vincent Marois, T.S. Jayram, Vincent Albouy, Tomasz Kornuta, Younes Bouhadjar, Ahmet S. Ozcan
{vmarois,jayram,tkornut,byounes,asozcan}@us.ibm.com,{vincent.albouy}@ibm.com

IBM Research — Almaden Research Center

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.
- We demonstrate that improper fine-tuning can reduce a model's accuracy as well.

The MAC Model [HM18]

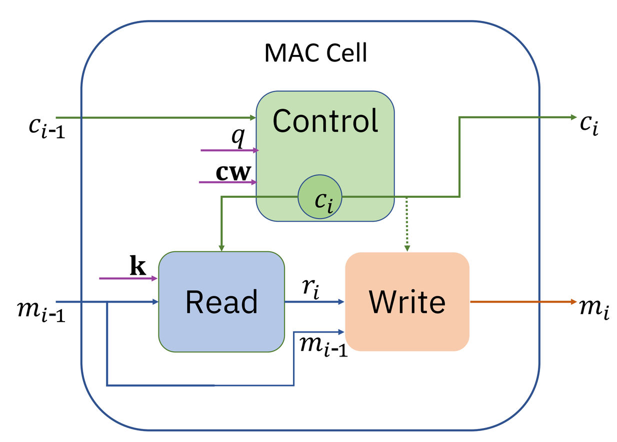


Figure 1: The MAC cell [HM18]

- The MAC network is a recurrent model that performs sequential reasoning; at each step the model analyzes the question and shifts the attention over the image
- The core of the model is the MAC cell, supported with an input unit that processes the question and image pair, and output unit which produces the answer.
- The input unit uses an LSTM to process the question and CNN layers to extract a feature map from the image.

Simplified Mac Model (S-MAC)

Our proposed modification to the MAC network is based on two heuristic simplifications:

- First, we observe that, taking the MAC cell equations as a whole, consecutive linear layers (with no activation in-between) can be combined as one linear layer.
- Secondly, we assume that dimension-preserving linear layers are invertible so as to avoid information loss.

Control unit: For both models, the question q is first transformed in each step of the reasoning using a *position-aware* linear layer depending on i: $q_i = U_i^{[d\times 2d]} q + b_i^{[d]}$.

MAC

$$cq_i = W_{cq}^{[d\times 2d]}[c_{i-1}, q_i] + b_{cq}^{[d]} \quad \text{(c1)}$$
$$ca_{is} = W_{ca}^{[1\times d]}(cq_i \odot \mathbf{cw}_s) + b_{ca}^{[1]} \quad \text{(c2.1)}$$
$$cv_{is} = \text{softmax}(ca_{is}) \quad \text{(c2.2)}$$
$$\mathbf{c}_i = \sum_s cv_{is}\, \mathbf{cw}_s \quad \text{(c2.3)}$$

S-MAC

$$cq_i = W_{cq}^{[d\times d]} c_{i-1} + q_i \quad \text{(c1)}$$
$$ca_{is} = W_{ca}^{[1\times d]}(cq_i \odot \mathbf{cw}_s) \quad \text{(c2.1)}$$
$$cv_{is} = \text{softmax}(ca_{is}) \quad \text{(c2.2)}$$
$$\mathbf{c}_i = \sum_s cv_{is}\, \mathbf{cw}_s \quad \text{(c2.3)}$$

Read and write units:

MAC

$$I_{ihw} = (W_m^{[d\times d]}\mathbf{m}_{i-1} + b_m^{[d]}) \odot (W_k^{[d\times d]}\mathbf{k}_{hw} + b_k^{[d]}) \quad \text{(r1)}$$
$$I'_{ihw} = W_{I'}^{[d\times 2d]}[I_{ihw}, \mathbf{k}_{hw}] + b_{I'}^{[d]} \quad \text{(r2)}$$
$$ra_{ihw} = W_{ra}^{[1\times d]}(\mathbf{c}_i \odot I'_{ihw}) + b_{ra}^{[1]} \quad \text{(r3.1)}$$
$$rv_{ihw} = \text{softmax}(ra_{ihw}) \quad \text{(r3.2)}$$
$$\mathbf{r}_i = \sum_s rv_{ihw}\, \mathbf{k}_{hw} \quad \text{(r3.3)}$$
$$\mathbf{m}_i = W_{rm}^{[d\times 2d]}[\mathbf{r}_i, \mathbf{m}_{i-1}] + b_{rm}^{[d]} \quad \text{(w1)}$$

S-MAC

$$I_{ihw} = m_{i-1} \odot k_{hw} \quad \text{(r1)}$$
$$I'_{ihw} = W_{I'}^{[d\times d]} I_{ihw} + b_{I'}^{[d]} + \mathbf{k}_{hw} \quad \text{(r2)}$$
$$ra_{ihw} = W_{ra}^{[1\times d]}(\mathbf{c}_i \odot I'_{ihw}) \quad \text{(r3.1)}$$
$$rv_{ihw} = \text{softmax}(ra_{ihw}) \quad \text{(r3.2)}$$
$$\mathbf{r}_i = \sum_s rv_{ihw}\, \mathbf{k}_{hw} \quad \text{(r3.3)}$$
$$\mathbf{m}_i = W_{rm}^{[d\times 2d]}\mathbf{r}_i + b_{rm}^{[d]} \quad \text{(w1)}$$

- Simplifications results in a 10% speed up in training time.

Model	Read Unit	Write Unit	Control Unit
MAC	787,969	524,800	525,313
simplified MAC	263,168	262,656	263,168
Reduction by [%]	67%	50%	50%

Table 1: Comparing the number of position-independent parameters between MAC & S-MAC cells.

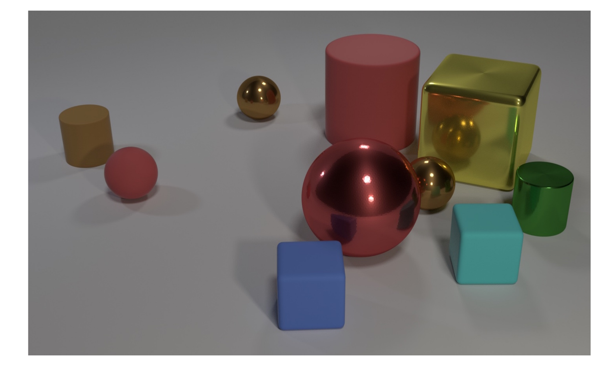
Links

Figure 2: Documentation.

Figure 3: GitHub repo.

Datasets - CLEVR and CoGenT

The CLEVR task:



· *How many objects are either small cylinders or red things?*

- Along with CLEVR, the authors [JHvdM+17] introduced CLEVR-CoGenT
- The goal is to evaluate how well the models can generalize, learn relations and compositional concepts.
- This dataset is generated in the same way as CLEVR, with two conditions, A and B. as shown in Table 2.

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

Table 2: Colors/shapes combinations present in CLEVR, CoGenT-A and CoGenT-B datasets.

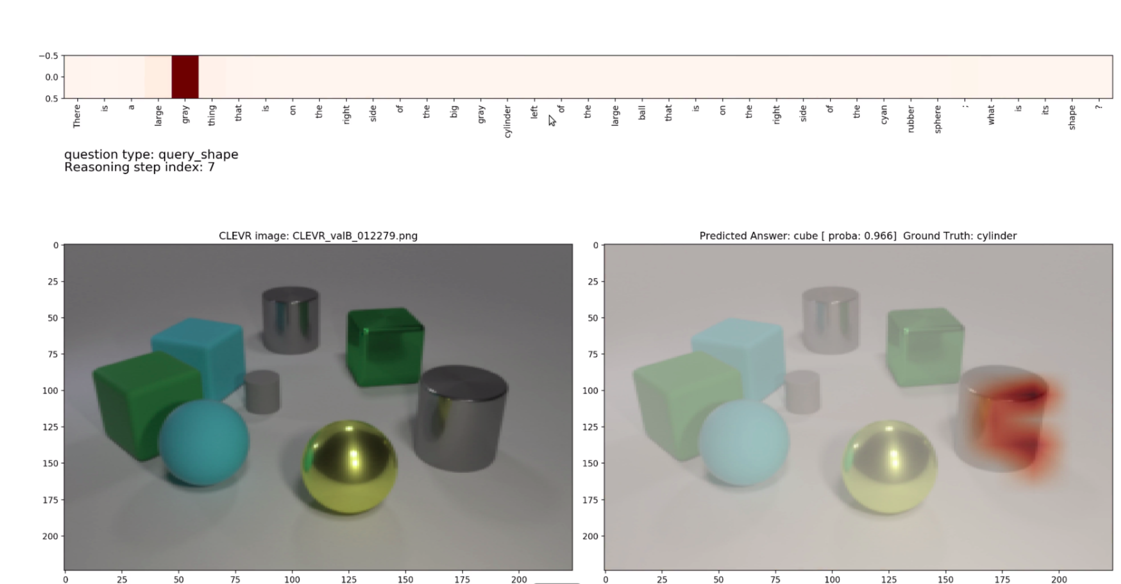
Transfer Learning - Experiments

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)
S-MAC	CLEVR	28:30	95.82	–	–	CLEVR	95.29	(b)
	CoGenT-A	28:33	96.09	–	–	CoGenT-A	95.91	(c)
	CLEVR	28:30	95.82	–	–	CoGenT-A	95.47	(d)
						CoGenT-B	95.58	(e)
	CoGenT-A	28:33	96.09	–	–	CogenT-B	78.71	(f)
				CoGenT-B	96.85	CoGenT-A	91.24	(g)
						CoGenT-B	94.55	(h)
	CLEVR	28:30	95.82	CoGenT-B	97.67	CoGenT-A	92.11	(i)
						CoGenT-B	92.95	(j)

- Our experiments on zero-short learning show that the MAC model has poor performance in line with the other models in the literature.
- With fine-tuning, the MAC model matches state of the art accuracy
- Remains an interesting problem to investigate how we can train it to disentangle the concepts of shape and color.
- Experiments can be reproduced by following the **mi-prometheus** documentation

MAC drawbacks on CLEVR



- 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? Predicted answer: Cylinder - Truth: Cube*

References

[HM18] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. *International Conference on Learning Representations*, 2018.

[JHvdM+17] 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.