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

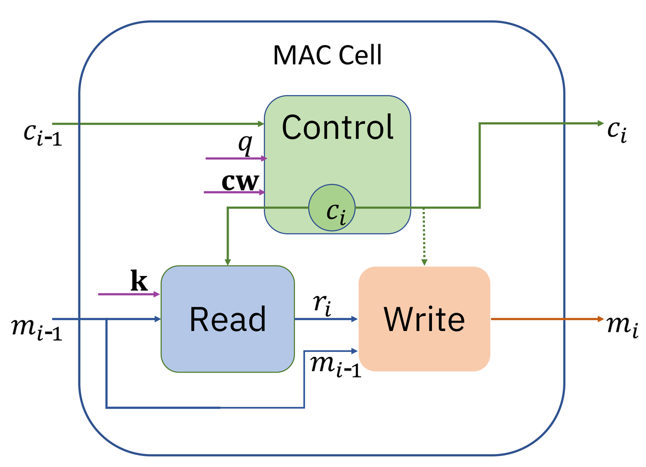
IBM Research

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



(a)

Figure 1: The MAC cell

- The MAC network is a recurrent model that performs sequential reasoning, where each step involves analyzing a part of the question followed by shifting 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 in a word-by-word manner producing a sequence of contextual words and a final question representation.
- The input unit utilizes a pre-trained ResNet followed by two 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 of the MAC cell:

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

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

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

Read and write units:

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

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

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.

Transfer Learning - Experiments

The CoGenT dataset contains:

- 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.

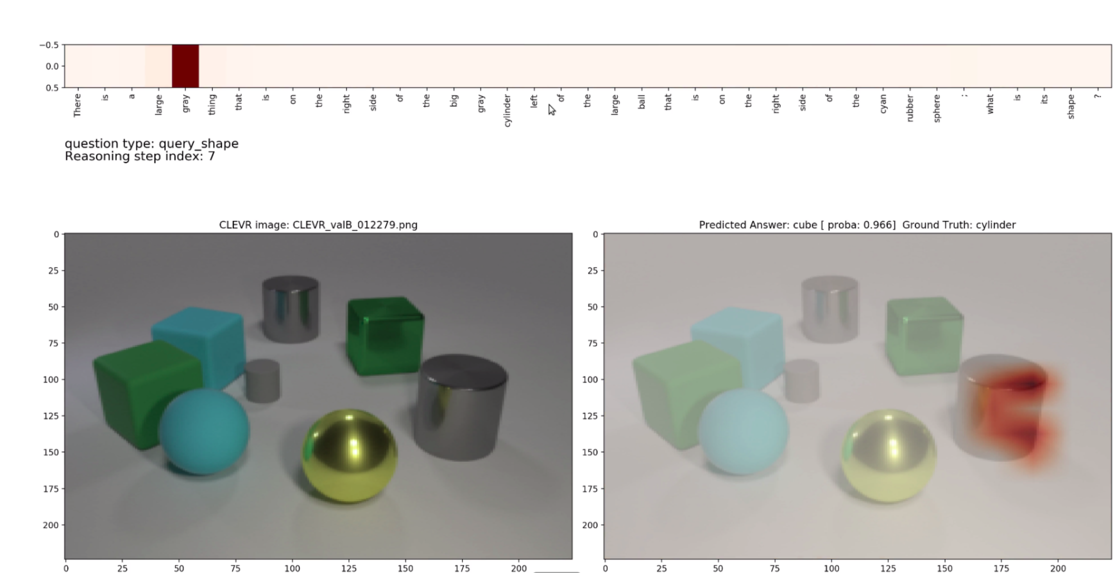
Dataset	Cubes	Cylinders	Spheres
CLEVR	any color	any color	any color
CLEVR CoGenT A	gray / blue / brown / yellow	red / green / purple / cyan	any color
CLEVR CoGenT B	red / green / purple / cyan	gray / blue / brown / yellow	any color

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

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)
	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)
S-MAC	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)

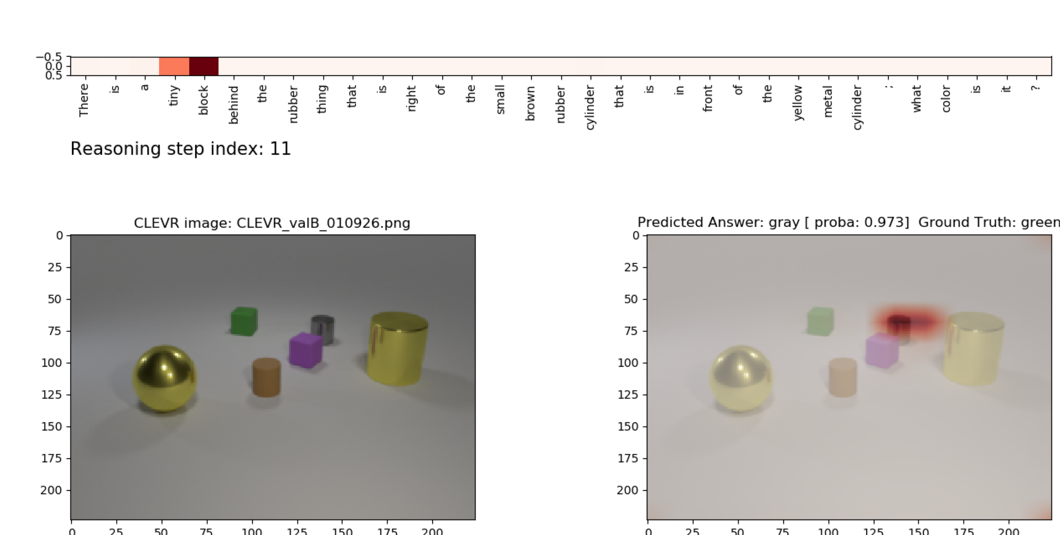
Table 3: Comparing the number of position-independent parameters between MAC & S-MAC cells.

MAC limitations on CLEVR



(a)

Figure 2: 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?*



(a)

Figure 3: 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?*

References