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

Summary

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

The MAC Model [HM18]

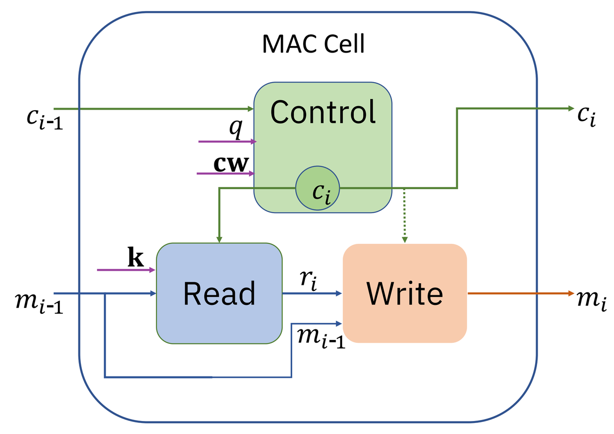

Figure 1: The MAC cell [HM18].

- MAC network: a recurrent model performing sequential reasoning. At each step, it analyzes the question and shifts the attention over the image.
- Recurrent MAC cell: consists of a control unit, a read unit & a write unit. The control unit updates the control state c_i & drives the attention over the question words.
- The read unit, guided by c_i extracts information from the image. The write unit uses this information to update the memory state m_i.

Simplified Mac Model (S-MAC)

Based on two heuristic simplifications:

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

This allows, with a careful reorganization, to apply a single linear layer to the *knowledge base* (feature map extracted from the image) prior to all the reasoning steps and work with this projection throughout the reasoning steps.

MAC | **S-MAC**

Control unit: The question q is first made *position-aware* in each reasoning step using an i-dependent projection: $q_i = U_i^{[d\times 2d]} q + b_i^{[d]}$.

MAC:

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

S-MAC:

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

Read and write units:

MAC:

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

S-MAC:

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

Model	Read Unit	Write Unit	Control Unit
MAC	787,969	524,800	525,313
S-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: Paper on arXiv.

Figure 3: How to reproduce the experiments.

The CLEVR & CoGenT datasets

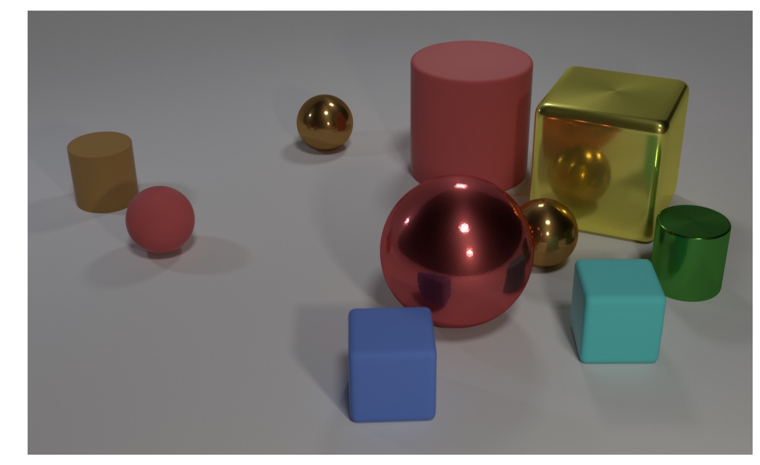

Figure 4: *How many objects are either small cylinders or red things?*

- The authors [JHvdM[+]17] also introduced CLEVR-CoGenT, to evaluate how well a model can learn relations and compositional concepts.
- Similar to CLEVR, but with two conditions, as follows:

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.

Experiments & Results

Model	Training			Fine-tuning		Test	
	Dataset	Time [h:m]	Acc [%]	Dataset	Acc [%]	Dataset	Acc [%]
MAC	CLEVR	30:52	96.70	–	–	CLEVR	96.17
S-MAC	CLEVR	28:30	95.82	–	–	CLEVR	95.29
	CoGenT-A	28:33	96.09	–	–	CoGenT-A	95.91
	CLEVR	28:30	95.82	–	–	CoGenT-A	95.47
						CoGenT-B	95.58
	CoGenT-A	28:33	96.09	–	–	CogenT-B	78.71
				CoGenT-B	96.85	CoGenT-A	91.24
						CoGenT-B	94.55
	CLEVR	28:30	95.82	CoGenT-B	97.67	CoGenT-A	92.11
						CoGenT-B	92.95

Table 3: CLEVR & CoGenT accuracies for the MAC & S-MAC models.

- Our experiments on *zero-shot learning* show that both models have poor performance, in line with the other models in the literature.
- With fine-tuning, both MAC models match state-of-the-art accuracy (a 15pts increase).
- **S-MAC presents a 10% speed-up in training time and comparable accuracy.**
- Finetuning CLEVR-trained models on CoGenT-A or -B hurts their generalization capabilities.

→ *Zero-shot learning* (CoGenT-A → CoGenT-B) remains an interesting problem to solve.

Compositional generalization of the MAC model

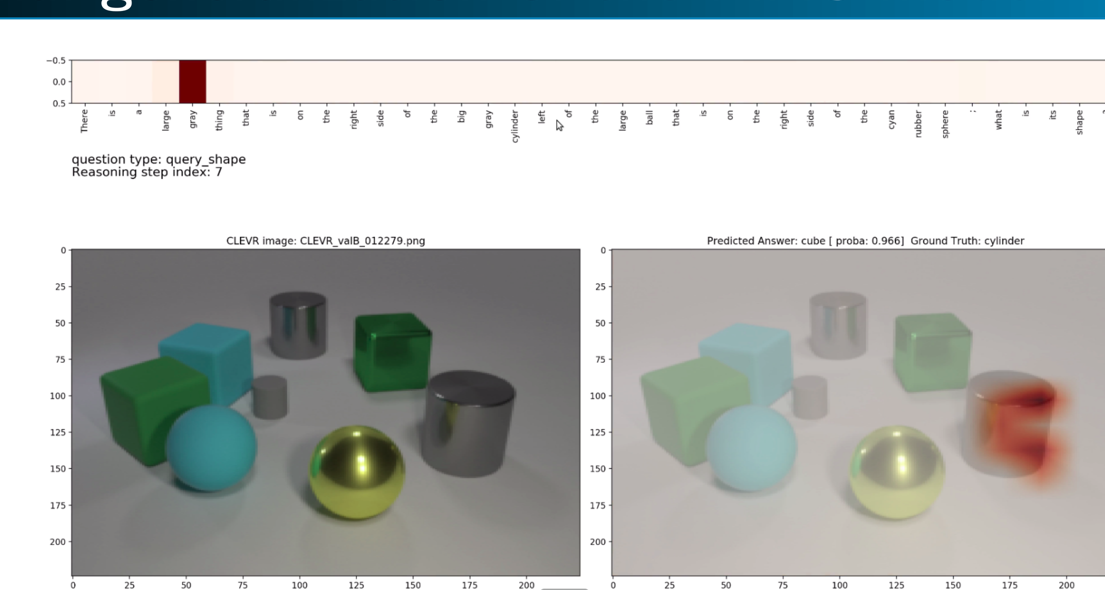

Figure 5: There is a large gray thing that is on the right side of the big gray cylinder left of the large ball that is on the right side if the cyan rubber sphere; what is its shape?

- Asked about the shape of the leftmost **gray cylinder**, the model correctly finds it, (cf. *visual attention map*), and refers to it using its color (*attention over the question words*).
- Yet, predicts the shape as **cube**, as it never saw **gray cylinders** during training, but saw **gray cubes**.

→ Indicate that MAC does not separate shape from color, but has a better understanding of colors (as found the object by its color).

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.