

CLEVR-Math – Compositional Language, Visual, and Mathematical Reasoning

2nd International Joint Conference on Learning and Reasoning 2022

NeSy workshop - 29 September 2022, London

Adam Dahlgren Lindström, Savitha Sam Abraham

Department of Computing Science



UMEÅ UNIVERSITY

What did we find?

It is **difficult to generalise from one to many reasoning steps** with multimodal data **for both** state of the art **neural and neuro-symbolic methods**

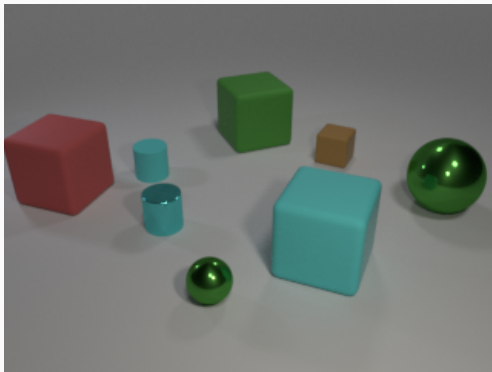
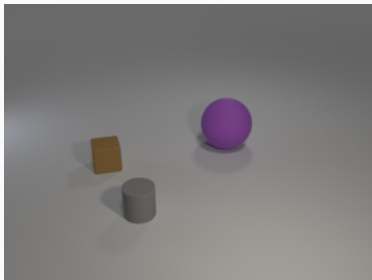
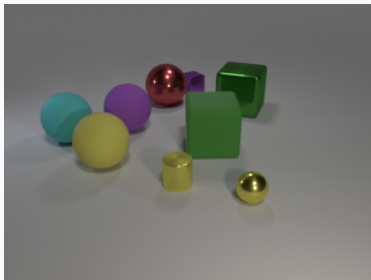


Figure 1: *Take away all large green metallic spheres. Now remove all cyan objects. How many objects are left? (4)*

Test your reasoning

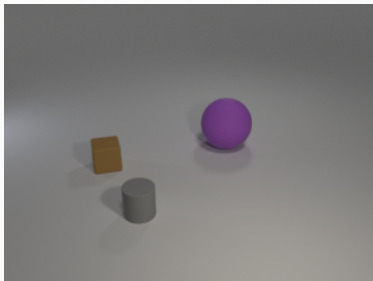


(a) Take away all small purple matte blocks. Subtract all blocks. How many objects are left?



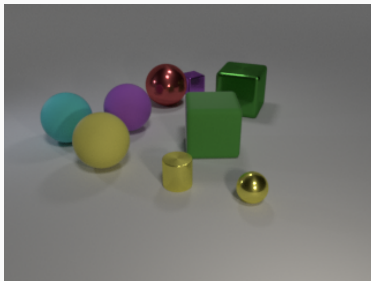
(b) Remove all red metallic objects. Subtract all yellow objects. How many objects are left?

Test your reasoning



(a) Take away all small purple matte blocks. Subtract all blocks. How many objects are left?

Answer: 2



(b) Remove all red metallic objects. Subtract all yellow objects. How many objects are left?

Answer: 5

Motivation

Example of 'reasoning' in the world of neural networks

*You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to **remove the door**. You have a table saw, so you cut the door in half and remove the top half.*



(a) some plants surrounding a lightbulb



(b) a lightbulb surrounding some plants



Figure 5: Teddy bears shopping for groceries in the style of ukiyo-e by DALL-E 2.

Figure 4: Example from the Winoground dataset (Thrush et al., 2022).

Fast and Slow in NeSy - a simple mental model

System 1

Intuitive
Fast
Autopilot

95%



System 2

Reflective
Logical
Takes effort

5%

Problem: Adam has three apples, and Eve has five. Eve gives Adam all her apples. How many apples does Adam have, if he eats one?

Equation: $X = 3 + 5 - 1$



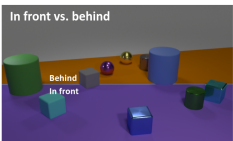
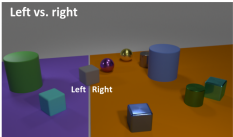
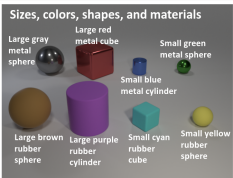
Figure 6: ‘‘Adam holds five apples and Eve gives him three apples, retro style’’ (Ramesh et al., 2022)

Mathematical reasoning tasks have **clear answers** and are **easy to control**. On reasoning tasks **NeSy methods should perform better** than neural methods.

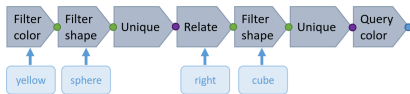
Multimodality is important.

What did we do?

Compositional learning with CLEVR

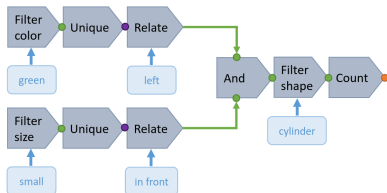


Sample chain-structured question:



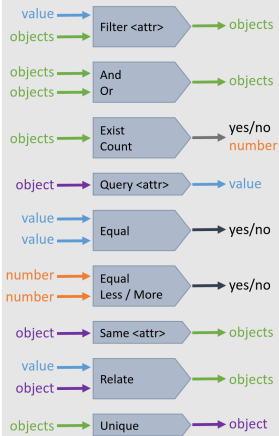
What color is the cube to the right of the yellow sphere?

Sample tree-structured question:



How many cylinders are in front of the small thing and on the left side of the green object?

CLEVR function catalog



Templates

Type	Templates
Remove group	"Remove all How many ... are there?" "Take away X How many objects are there?"
Remove subset	"Remove X How many ... are there?"
Insertion	"Add X How many objects are there?"
Count backwards	"How many ... must be removed to get X ...?" "Take away How many were removed if are X ... left?"
Multi-hop	"Take away all A. Remove all B. How many objects are left?"
Adversarial	"Remove all A . How many B are left?"

More samples from CLEVR-Math

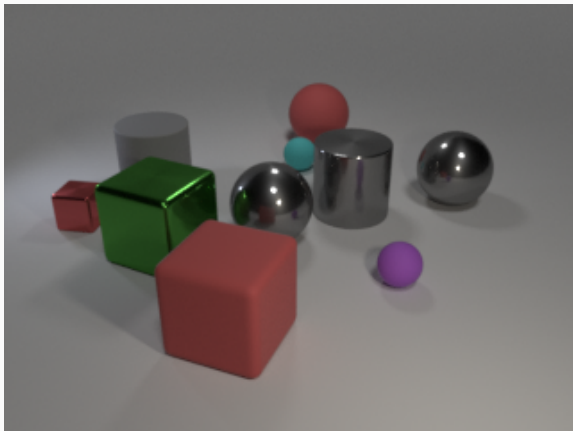
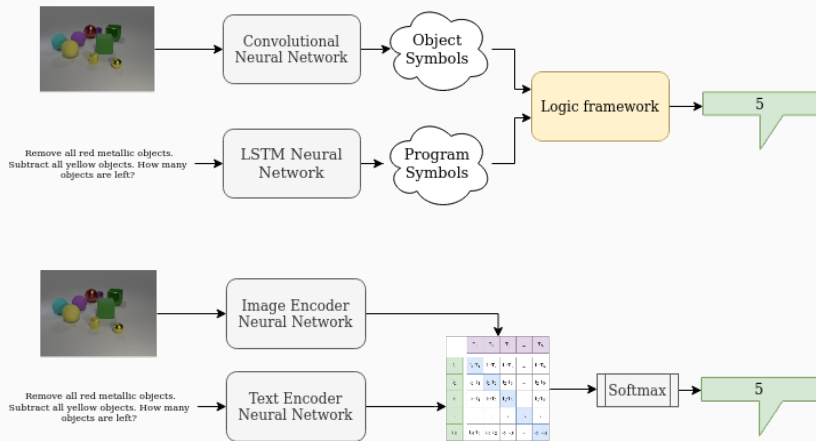


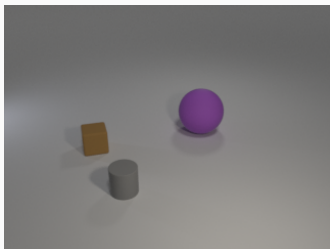
Figure 7: (i) *Remove all gray spheres. How many spheres are there? (3),* (ii) *Take away 3 cubes. How many objects are there? (7),* (iii) *How many blocks must be removed to get 1 block? (2)*

Our experiment with NS-VQA (Yi et al., 2018) and CLIP (Radford et al., 2021)

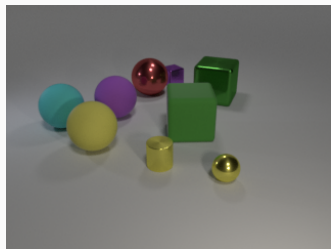


Results

Model	All	Addition	Subtraction	Adversarial	Multihop	Multihop (0-shot)
NS-VQA	0.8840	0.9781	0.9948	0.9957	0.286	0.267
CLIP	0.3464	0.5699	0.3019	0.2848	0.272	0.238



(a) *Subtract all small purple matte blocks. Subtract all blocks. How many objects are left?* was answered by CLIP with 3 instead of 2.

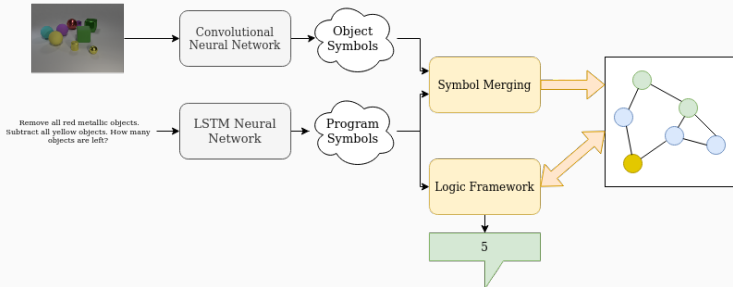


(b) *Subtract all red metallic objects. Subtract all yellow objects. How many objects are left?* was answered with 9 instead of 5 by NS-VQA.

Conclusions

My takeaways

- **Neuro-symbolic methods** (clearly) **better at these reasoning tasks**
- Room for **improvement on how language is used in NeSy-methods**
- **Multimodal multihop reasoning is hard** for state of the art methods
- State representation could be key to better performance for multihop reasoning



- Intermediate scene graphs for each step
- Generate subquestions and apply actions individually
- How does a prompt-like input (like that of CLEVR_Hyp) effect the outcome?



Hugging Face

<https://huggingface.co/datasets/dali-does/clevr-math>



<https://github.com/dali-does/clevr-math>

- Math problem solving (MAWPS (Koncel-Kedziorski et al., 2016), SVAMP (Patel et al., 2021))
- CLEVR-based work (CLEVR_hyp (Sampat et al., 2021), CLEVRER (Yi et al., 2019), CLEVR-Hans (Stammer et al., 2021))
- Visual reasoning (Kandinsky patterns (Holzinger et al., 2019), GQA (Hudson & Manning, 2019))

Confusion matrix

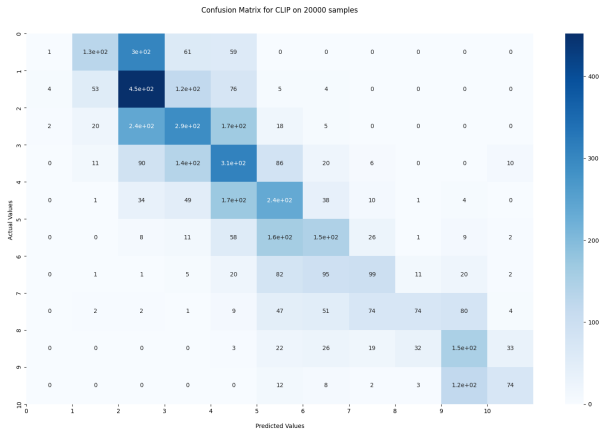













Figure 11: Confusion matrix for CLIP trained on 20 000 samples.

References

-  Holzinger, A., Kickmeier-Rust, M., & Müller, H. (2019). Kandinsky patterns as iq-test for machine learning. *International cross-domain conference for machine learning and knowledge extraction*, 1–14.
-  Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

-  Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., & Hajishirzi, H. (2016). Mawps: A math word problem repository. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1152–1157.
-  Patel, A., Bhattamishra, S., & Goyal, N. (2021). Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.
-  Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

-  Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
-  Sampat, S. K., Kumar, A., Yang, Y., & Baral, C. (2021). CLEVR_HYP: A challenge dataset and baselines for visual question answering with hypothetical actions over images. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3692–3709. <https://doi.org/10.18653/v1/2021.naacl-main.289>
-  Stammer, W., Schramowski, P., & Kersting, K. (2021). Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3619–3629.

-  Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248.
-  Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., & Tenenbaum, J. B. (2019). Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.
-  Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. B. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in Neural Information Processing Systems*, 1039–1050.