Reduced Data Scheme on Neuro-Symbolic Concept Learner in Visual Question Answering

Poopa Kaewbuapan

CSC502
Advanced Seminar in Computer Science

December 22nd, 2023

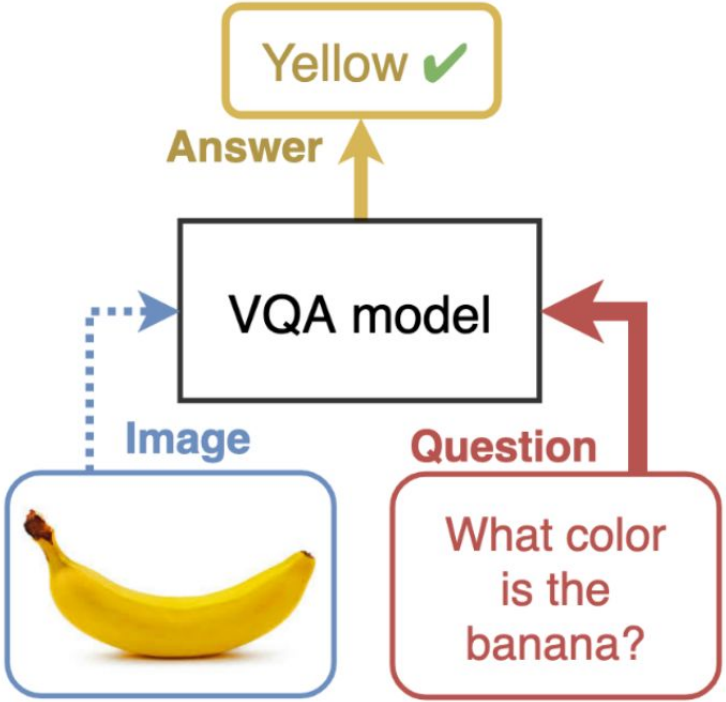
ReducedNSCL in short

Propose a new neuro-symbolic model based on NS-CL architecture that aims to reduce training data to mastered visual question answering task

ReducedNSCL use two attention-based modules to enhanced visual and semantic domain generalization while preserving model interpretability

- Can ReducedNSCL's performance outperformed original NS-CL given equal training data and by what margin?

Visual Question Answering (VQA)



https://visualqa.org/

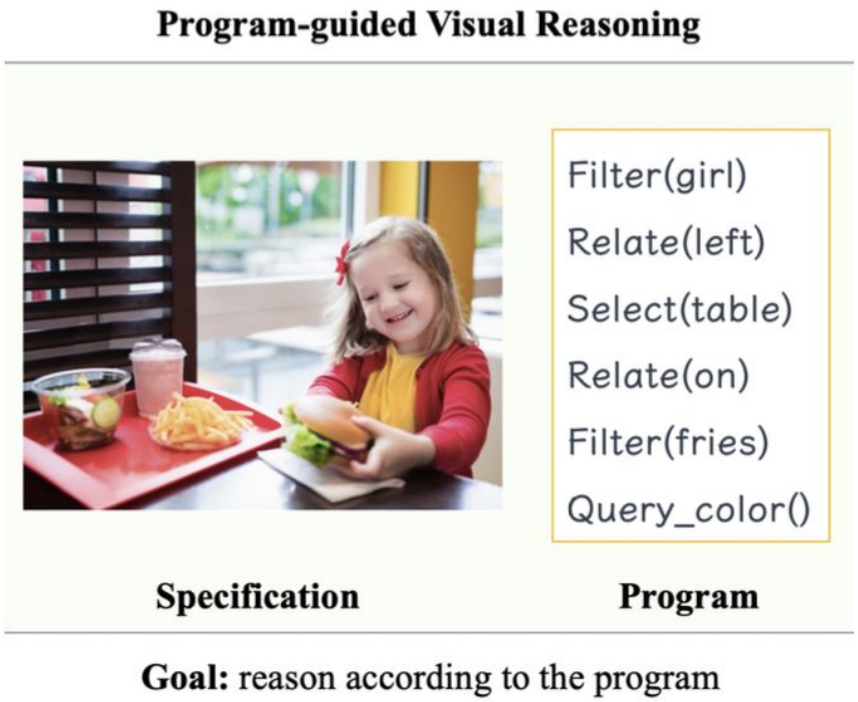


Figure 1: Program VQA (Zhao et al. 2021)

State of the Art on VQA: Pre-trained Vision Transformer

For comprehensive real world dataset VQAv2, ViT-derived (Wang et al. 2022) model are the current SoTA with accuracy of 84.03%

- Pre-trained Transformer model (so-called foundation model) with billions in parameters yield superb performance, at the cost of upsetting interpretability (Bommasani et al. 2022) and training data required
- Single network of deep foundation model might not be so desirable for reduced data scheme

Model	#Layers	Hidden Size	MLP Size	#Parameters				
				V-FFN	L-FFN	VL-FFN	Shared Attention	Total
BEiT-3	40	1408	6144	692M	692M	52M	317M	1.9B

Table 2: Model configuration of BEiT-3. The architecture layout follows ViT-giant [ZKHB21].

Data	Source	Size
Image-Text Pair	CC12M, CC3M, SBU, COCO, VG	21M pairs
Image	ImageNet-21K	14M images
Text	English Wikipedia, BookCorpus, OpenWebText, CC-News, Stories	160GB documents

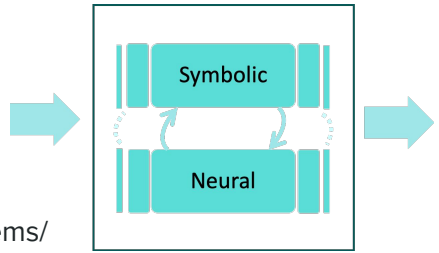
Table 3: Pretraining data of BEiT-3. All the data are academically accessible.

Figure 2: BEiT-3 config and pretraining data (Wang et al. 2022)

Kautz’s Taxonomy

Kautz’s taxonomy classified NeSy (Neuro-Symbolic) system into 5-6 categories, higher number indicate close integration between two paradigms in which symbolic reasoning is more explicit and provide more interpretability

- Type 3 NeSy (Neuro; Symbolic) employs neural network as co-routine to converts non-symbolic input (e.g., pixels) into symbolic input that is manipulated by a symbolic reasoning system
- Clear distinction of neural and symbolic subsystem is why we choose to study Type 3 NeSy



https://harshakokel.com/posts/neurosymbolic-systems/

Reduced Data Goal in NeSy

A review of NeSy research (Hamilton et al. 2022) presented 5 goals that NeSy systems promised over current deep learning paradigm

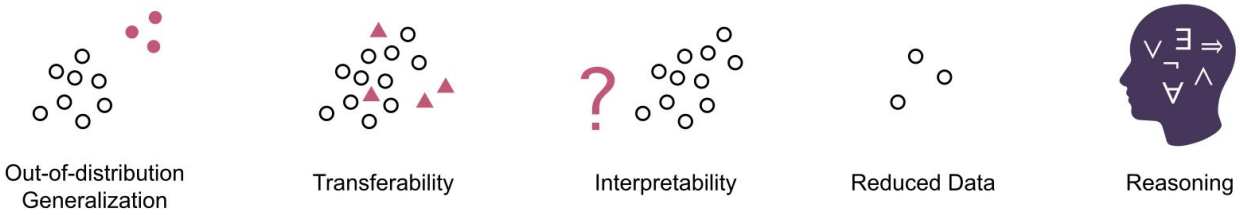
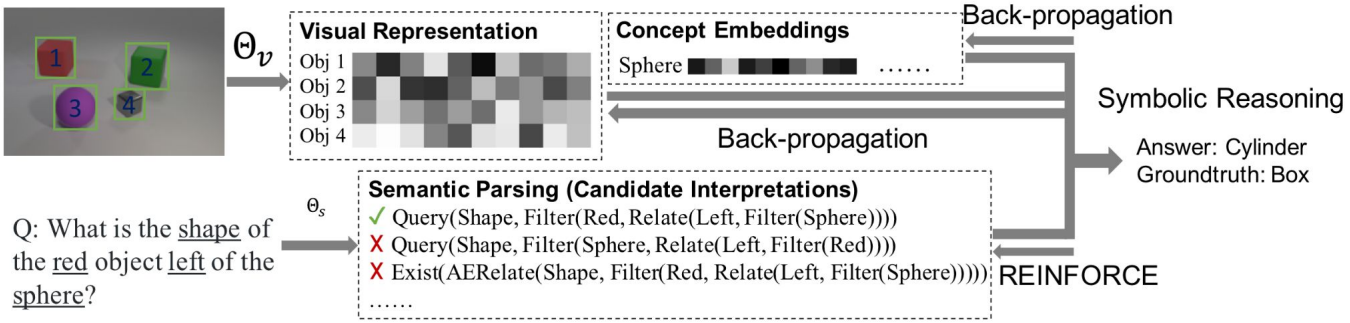


Fig. 14. Neuro-symbolic artificial intelligence goals.

- We choose to evaluate our NeSy model with Reduced Data goal due to direct importance that addressing massive training data usage on foundational models
- NeSy model that reducing training data required to mastered VQA task may implied that the model has genuine understanding of VQA task and able to generalized to diverse datasets in both synthetic and real-world scenes

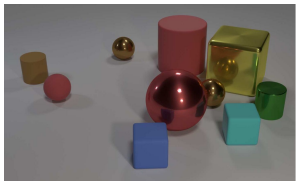
Neuro-Symbolic Concept Learner (NS-CL) (Mao et al. 2019)



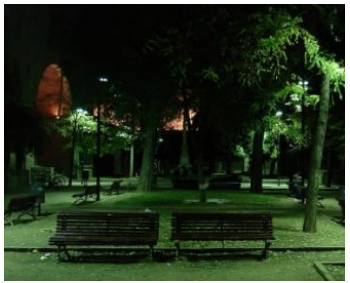
Our framework for Type 3 NeSy that efficiently learned synthetic VQA dataset with reduced data, providing interpretability as a program representation that generated from the question prompt

- Visual Representation module is purely neural (Pre-trained Mask R-CNN) while Semantic Parsing module is responsible for generating symbolic representation by program synthesis
- For every datasets, Semantic Parsing required a human-curated domain-specific language (DSL) that describe all possible operations in that dataset

NS-CL Shortcomings



External semantic knowledge required from human in form of DSL and in addition, a VQA benchmark survey (Zhang et al. 2021) reveals that NS-CL is robust to only synthetic visual domain shifts (significant performance drop in challenging real-world VQA datasets - natural objects are more diverse)



- NS-CL lacks both visual and semantic domain generalization (natural real-world scenes and free-form text)

- Given NS-CL dual modules architecture, by using small attention-based model, we can improve domain generalization that can leads to reduction in training data to mastered VQA task with acceptable pretraining data usage when compared to large single foundation model

Central Hypothesis

"Replacing Semantic Parsing module with attention-based symbolic reasoning model and having Visual Representation module with real-world visual domain generalization allows the new model to achieve competitive accuracy with original NS-CL given equal training data"

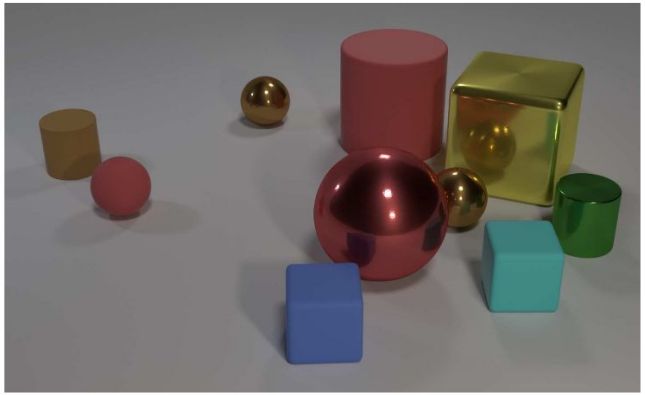
Objectives

1. ReducedNSCL accuracy surpasses over 1% margin when comparing with NS-CL on both VQAv2 (real world) and CLEVR (artificial) dataset on equal training data => Reduced Data & Domain Generalization Goal

2. ReducedNSCL still retains symbolic concept grounding (Bounding box, Probabilistic readout etc.) => Interpretability Goal

Benchmarking Datasets: CLEVR (Johnson et al. 2017)

- Synthetic visual scenes with computer-generated questions (NS-CL performed best)

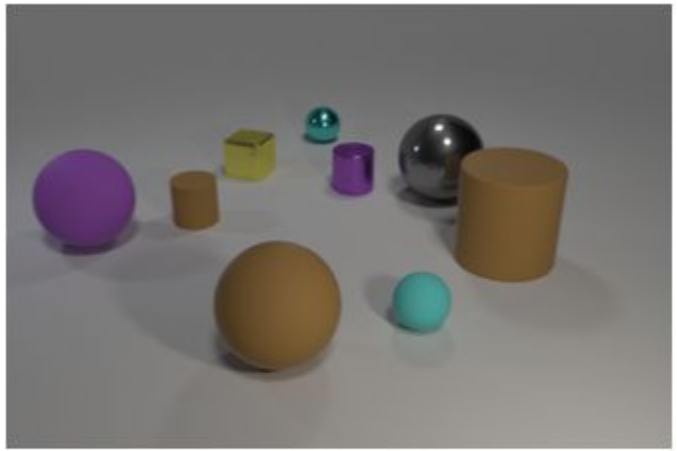


Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

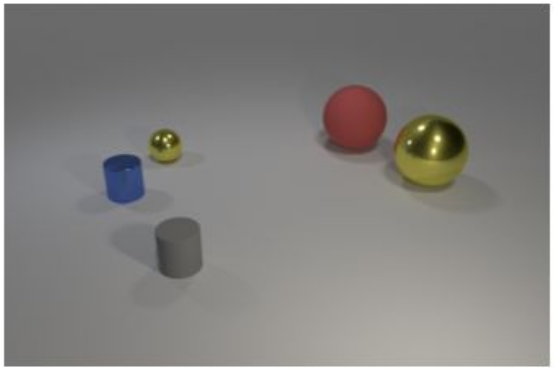
Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.

Benchmarking Datasets: CLEVR-Humans (Johnson et al. 2017)

- Synthetic visual scenes with human-created questions



Q: *Are* all *the balls small?*
A: *no*



Q: Two items share *a color, a material, and a shape; what is the size of the* rightmost *of* those items? **A:** *large*

Benchmarking Datasets: VQAv2 (Goyal et al. 2017)

- Real-world visual scenes with human-created questions
- Most challenging - no bounding-box ground truth

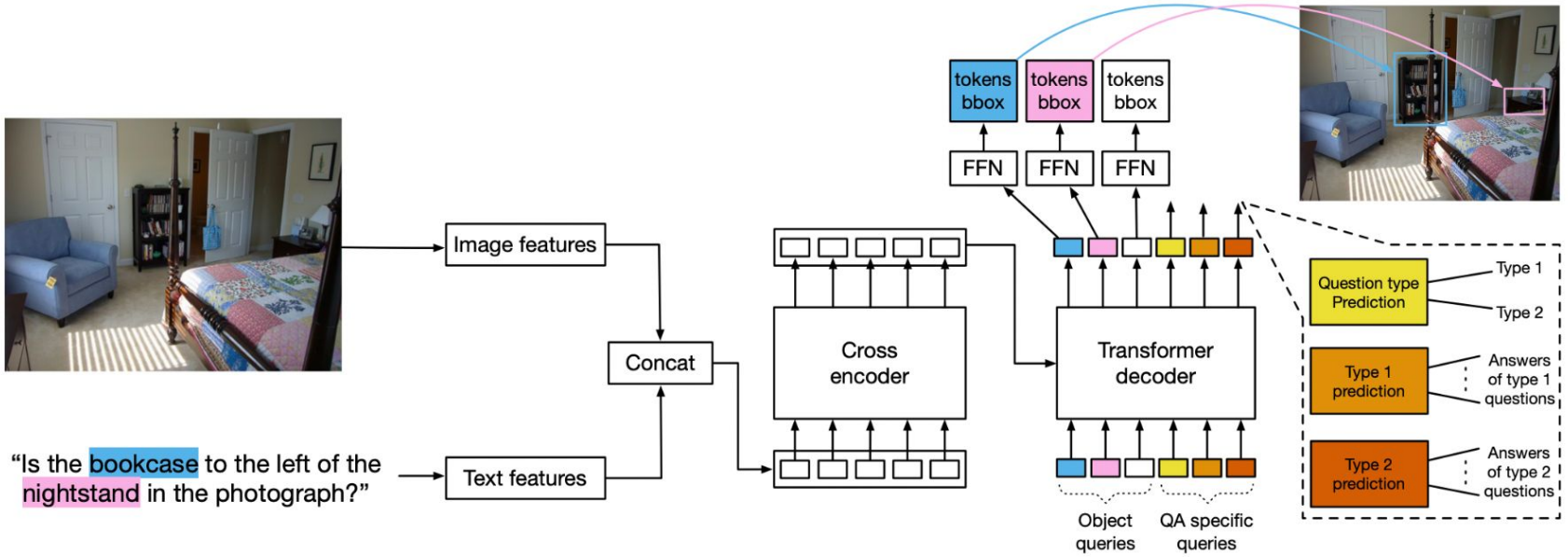
Visual Representation Candidate: MDETR (Kamath et al. 2021)

Pre-trained Transformer object detection model that extends to VQA task by utilizing semantic information to modulate visual space

+ Attention-based model encode denser object informations, enhancing visual generalization power

* For experiment control, Pre-trained ResNet-101 variant will be used and the same being appiled for NS-CL CNN backbone

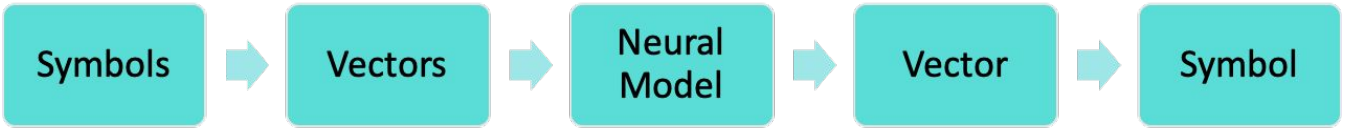
MDETR in Action



Semantic Parsing Candidate: OCCAM (Wang et al. 2021)

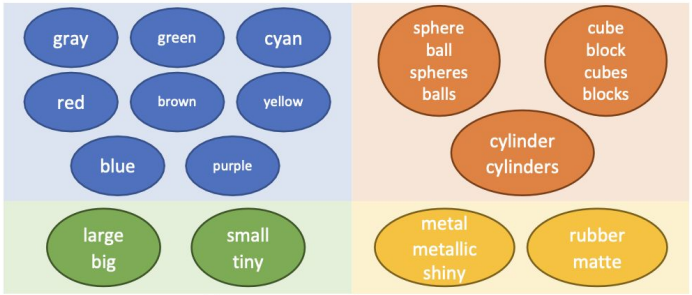
Fully-differentiable approach on symbolic space with concept induction that can discovered hierarchical relationship of concepts

+ Enhancing semantic generalization power (Free form text is supported - automatically learn concepts from induction)
- Consider as Type 1 NeSy (at most only offer symbolic readout) => Suggests to be the worst in interpretability



https://harshakokel.com/posts/neurosymbolic-systems/

OCCAM Concept Model and Induced Symbolic Space

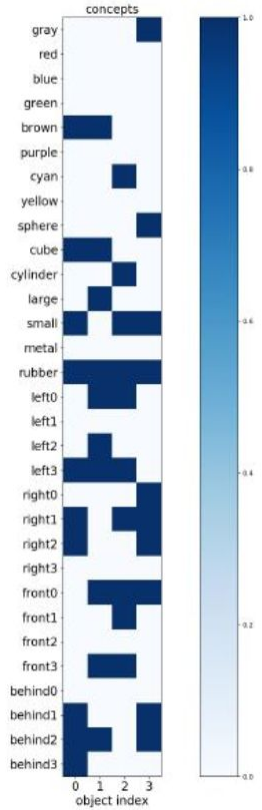


(a) unary concepts/super-concepts

(b) binary concepts/super-concepts

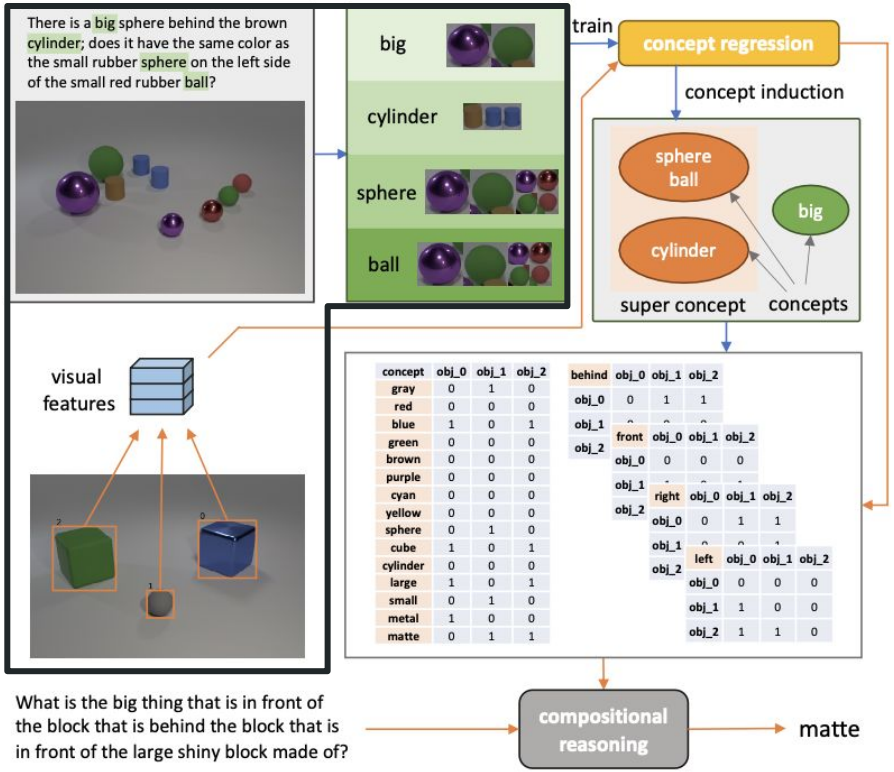
	small	tiny	big	cube	ball	⋯
	1	1	0	1	0	⋯
	0	0	1	1	0	⋯
	1	1	0	0	1	⋯
	0	0	1	0	1	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

	left	right	front	behind
	0	1	1	0
	1	0	0	1
	1	0	1	0
⋮	⋮	⋮	⋮	⋮



ReducedNSCL: MDETR+OCCAM

MDETR

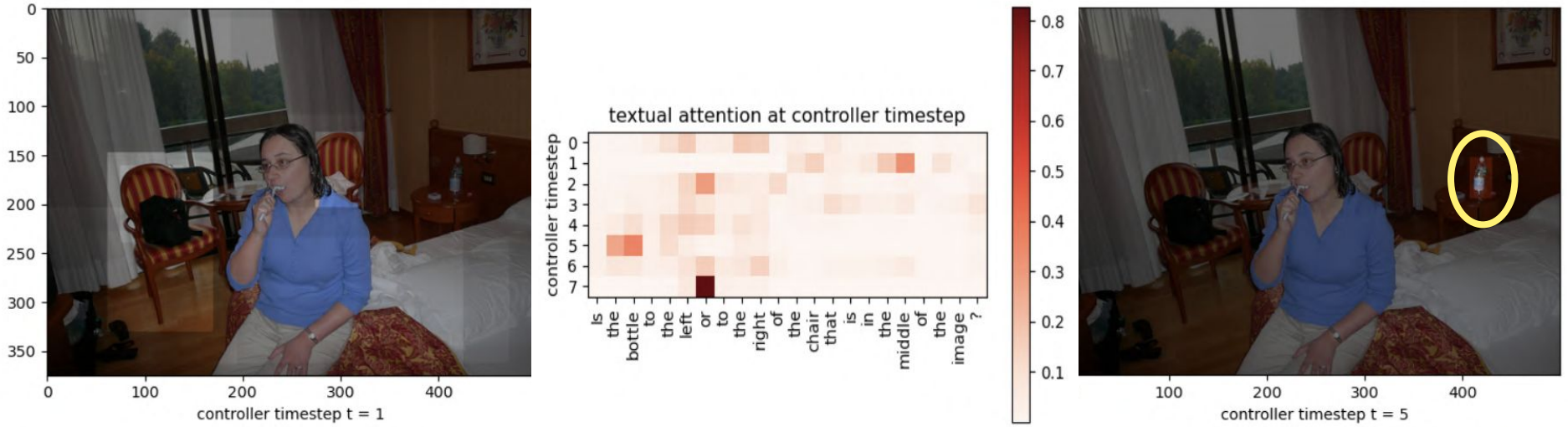


Methodology

1. Implement ReducedNSCL model by integrating OCCAM (Semantic Module) with pre-trained MDETR (Visual Module) and writing CLEVR-Humans and VQAv2 DSL for NS-CL
2. Training ReducedNSCL and NS-CL with equal training data using benchmark datasets
3. Testing ReducedNSCL and NS-CL performance on each benchmark dataset by accuracy metric (% correct on testing data)

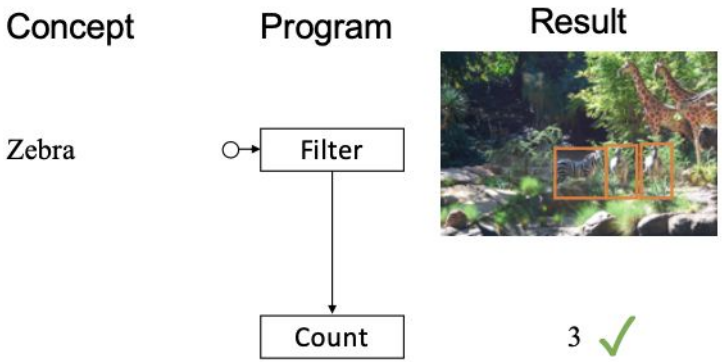
- If accuracy of ReducedNSCL surpassed NS-CL on same training data amount => ReducedNSCL “reduced data” on training phase
- Also do training and testing with 10% training data and 1% training data to demonstrate few-shot sample efficiency => ReducedNSCL accuracy should surpassed NS-CL in 10% and 1% training data case

OCCAM Interpretability (Type 1 NeSy)



NS-CL Interpretability (Type 3 NeSy)

Q: How many zebras are there?



Expected Difficulties

Implementation Overhead

- OCCAM implementation is closed-source
  - Request original authors for source code
  - Need to implement from scratch using original paper
- NS-CL required DSL for CLEVR-Humans and VQAv2 datasets to be able to benchmark

Result Evaluation

- Comparison of Interpretability between different NeSy types is purely qualitative => How to measure model's knowledge gain from symbolic representation?

References

[1] K. Hamilton, A. Nayak, B. Božić, and L. Longo, “Is neuro-symbolic AI meeting its promises in natural language processing? A structured review,” Semantic Web, no. Preprint, pp. 1–42, 2022.

[2] R. Bommasani et al., “On the opportunities and risks of foundation models,” arXiv preprint arXiv:2108.07258, 2021.

[3] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, “The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision,” arXiv preprint arXiv:1904.12584, 2019.

[4] Z. Wang et al., “Interpretable visual reasoning via induced symbolic space,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1878–1887.

[5] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “Mdetr-modulated detection for end-to-end multi-modal understanding,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1780–1790.

[6] M. Zhang, T. Maidment, A. Diab, A. Kovashka, and R. Hwa, “Domain-robust vqa with diverse datasets and methods but no target labels,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7046–7056.

Reduced Data Scheme on Neuro-Symbolic Concept Learner in Visual Question Answering

"Thank you for listening!"