Supplementary Material for Linguistically Routing Capsule Network for Out-of-distribution Visual Question Answering

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

1. More Details of the Network Architectures

In this section, we give the detailed architecture of the models evaluated on the CLEVR-CoGenT, the VQA-CP v2 and the VQAv2 datasets.

1.1. CLEVR-CoGenT

As shown in Figure 1, we introduce the linguistically routing into the convolution layers within the 4 residual blocks. The questions and the words are embedded by a GRU. The words embedding are fused with the $14 \times 14 \times 1024$ extracted image feature. The fused feature is fed into 4 residual blocks. Each block contains a linguistically routing convolution with $3 \times 3 \times 144$ kernel, a batch normalization, a multiplication with the transformed question embedding, a ReLU activation, and a residual connection. The classifier convolves the 144-dimensional feature maps to 512 dimensions and feeds the result into two fully connected layers to predict the answer.

1.2. VQA-CP v2 and VQAv2

For the VQA-CP v2 dataset, we modified the Modular Co-Attention Networks (MCAN) [10] and introduce the linguistically routing in the guided-attention blocks. As shown in Figure 2, The question words are embedded with 6 self-attention blocks. Each block has 512 hidden dimension and 8 attention heads. The image is firstly pass through 3 guided-attention blocks. Then we replace the feed-forward layer in the last 3 guided-attention blocks with the linguistically routing feed-forward layer. The guided-attention blocks also have 512 hidden dimension, 8 attention heads, and 36 image objects. The classifier is the same as the MCAN [10]. It performs attention on question words and 36 image objects, then obtains a 1024-dimensional vector. The classifier project the 1024-dimensional vector to 3129-dimension, where the 3129 is the number of the answer candidates.

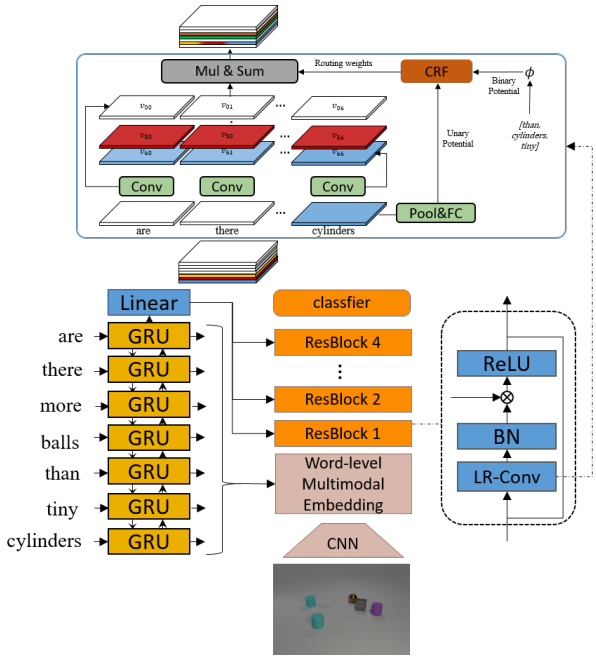


Figure 1: The overall model architecture for the CLEVR-CoGenT dataset.

2. More Experimental Results

We also evaluate our proposed method on the CLEVR dataset to verify its performance on in-domain test data.

CLEVR [4] is a synthesized dataset designed to achieve minimal dataset biases. It consists of $100,000$ images, $853,554$ questions, and the corresponding image scene graphs and questions' functional program layouts. This dataset is similar to the CLEVR-CoGenT but without the swapped color palettes. We use the exactly same model architectures and training hyper-parameters as we used in CLEVR-CoGenT.

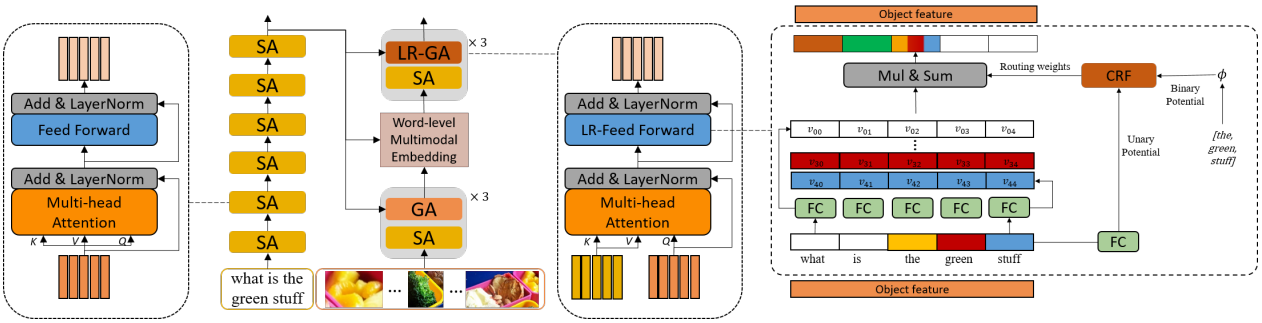


Figure 2: The overall model architecture for the VQA-CP v2 dataset.

Method	Count	Exist	Compare Integer	Compare Attribute	Query	Overall
N2NMN* [2]	68.5	85.7	84.9	88.7	90.0	83.7
IEP* [5]	92.7	97.1	98.7	98.9	98.1	96.9
TbD+reg+hres* [6]	97.6	99.2	99.4	99.6	99.5	99.1
NS-VQA* (270 programs) [9]	99.7	99.9	99.9	99.8	99.8	99.8
CNN+LSTM+SAN [5]	59.7	77.9	75.1	70.8	80.9	73.2
LBP-SIG [11]	61.3	79.6	80.7	76.3	88.6	78.0
Dependency Tree [1]	81.4	94.2	81.6	97.1	90.5	89.3
CNN+LSTM+RN [8]	90.1	97.8	93.6	97.1	97.9	95.5
CNN+GRU+FiLM [7]	94.5	99.2	93.8	99.0	99.2	97.6
MAC [3]	97.1	99.5	99.1	99.5	99.5	98.9
LR-Capsule(ours)	95.6	98.7	97.2	98.8	98.8	97.9

Table 1: Comparison of question answering accuracy on the CLEVR dataset. (*) indicates that the model has been trained with program annotations.

2.1. Results

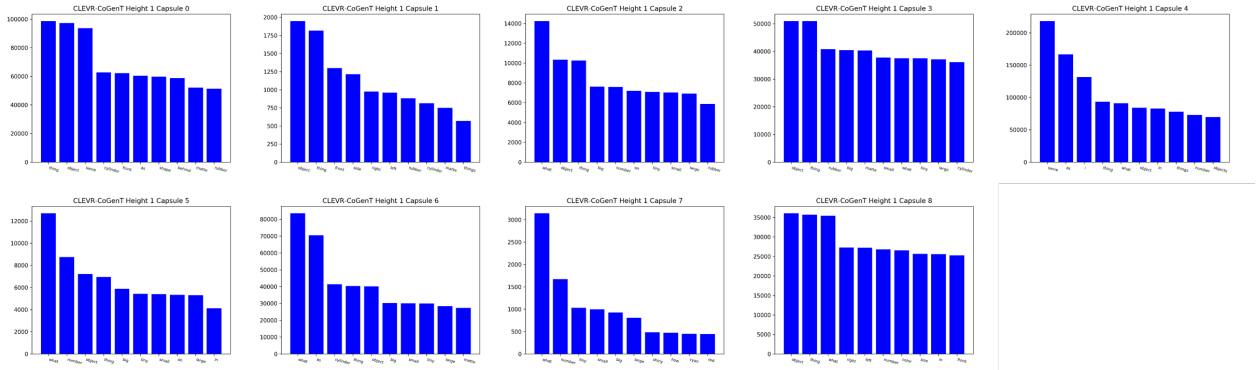
Table 1 shows the performance of all the compared methods on the CLEVR test set. The end-to-end modular network [2], program execution engine [5], transparency by design [6], and neural symbolic visual question answering [9] are referred to as "N2NMN", "IEP", "TbD" and "NS-VQA", respectively. All these methods use the functional programs' layout as the extra training signal. "N2NMN", "IEP" and "TbD" achieve their best results by using all of the program layouts. Although "NS-VQA" uses only 3 samples from each of the 90 question families, it leverages the scene graphs to train a scene parser. "N2NMN', "IEP" and "NS-VQA" also have variants that use different numbers of programs during training, and it has been shown that their performance degrades if fewer program layout examples are used. In contrast, our method can achieve a comparable state-of-the-art performance without using any dataset-specific layout on in-domain test data.

3. Analysis of capsules' Encoded Words

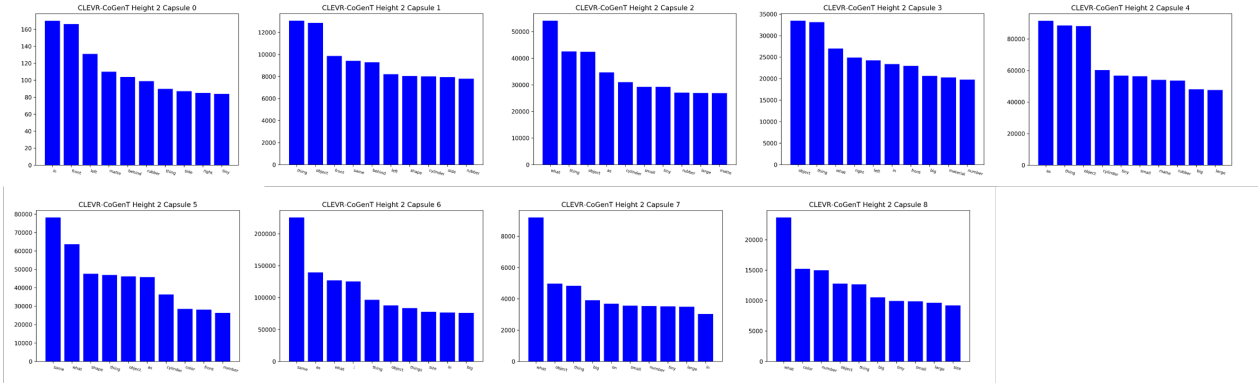
To examine whether the capsules can be activated to represent different samples dynamically, we explore which words are encoded by the capsules in different layers. Specifically, given a capsule, we rank its encoded words by frequency and present the top-10 words in Figure 3 and Figure 4. The x-axis represents the words and the y-axis represents their frequency. We also remove some conjunction, preposition and determiner words, including "the" "a" "an" "of" "is" "are" "it" "there" "that" "do" "does" "to" "have" "has" "or".

As shown in Figure 3, on the CLEVR-CoGenT dataset, the height 1 capsule 0 mostly encodes words that describe shape and material; the height 1 capsule 1 mostly encodes words that indicate the position; the height 1 capsule 3 mostly encodes words that describe the object size. In height 4, the capsule 0 and capsule 5 encode "what" "object" most, the capsule 1 encodes the word "number" and the capsule 2 encodes the "as" "same". It suggests that different capsules may be used to encode different types of questions. However, in height 4, the capsule 8 barely encodes words, and the capsule 4 hasn't been activated at all. At a higher level, it may be better to lower the number of capsules but increase each capsule's capacity to better encode the rich high-level semantic. Similarly in Figure 4, on the VQA-CP v2 dataset, different capsules have encoded

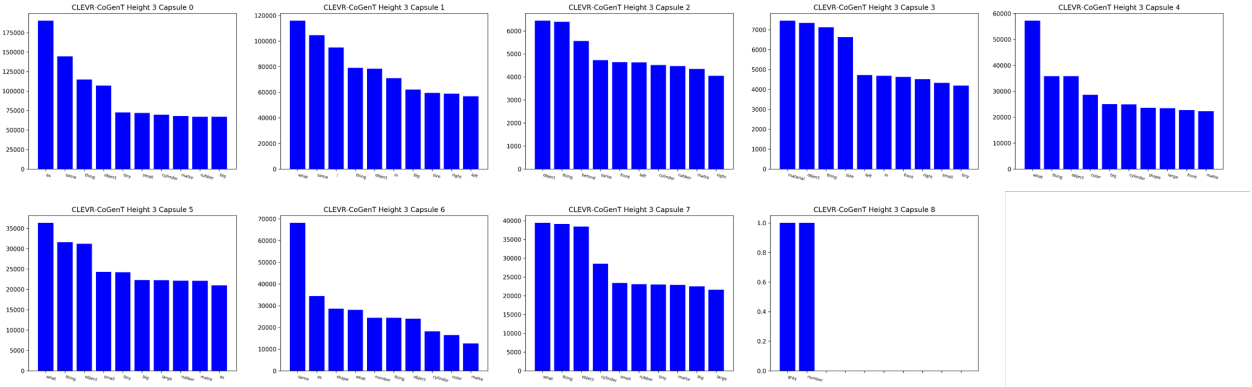
different words. The height 1 capsule 0 mostly encodes the "person" and actions; the height 1 capsule 6 encodes different kinds of animals.



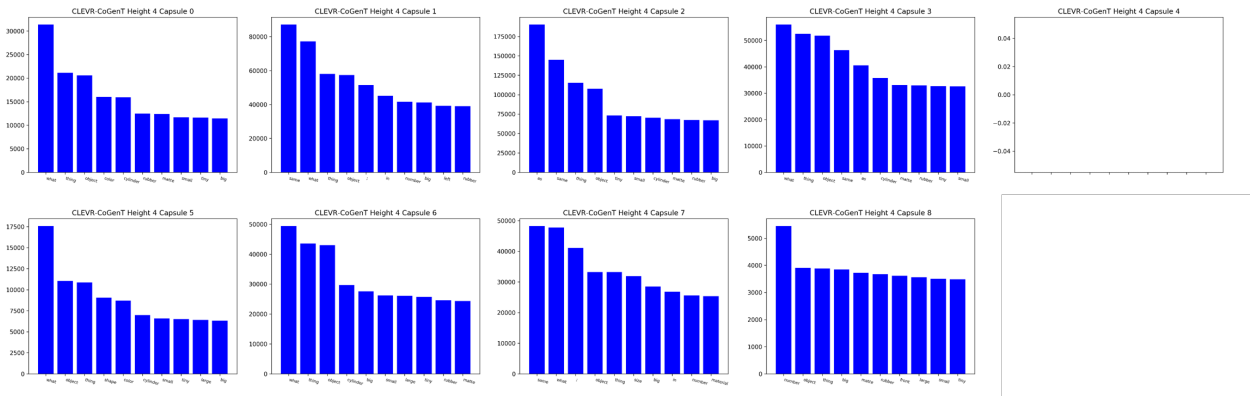
(a) Height 1



(b) Height 2

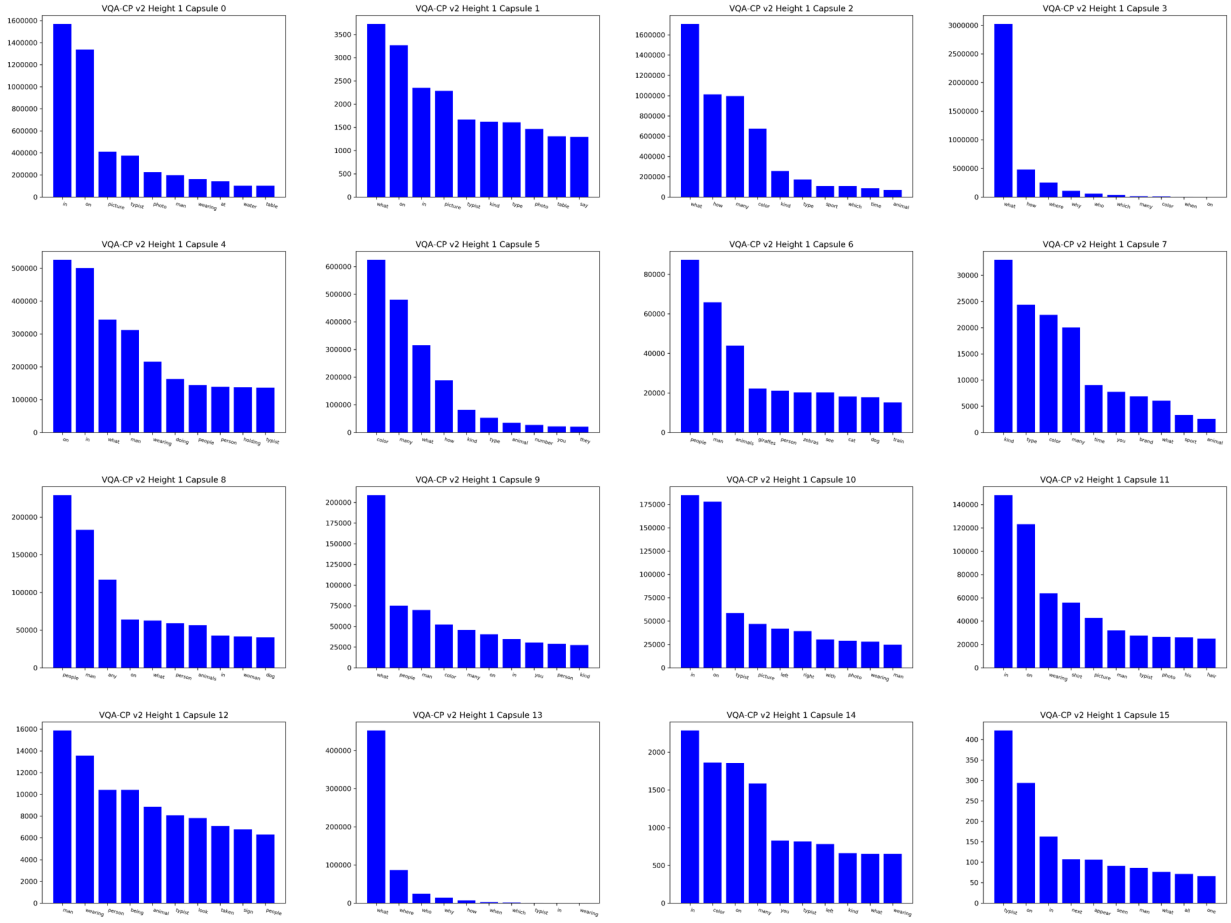


(c) Height 3

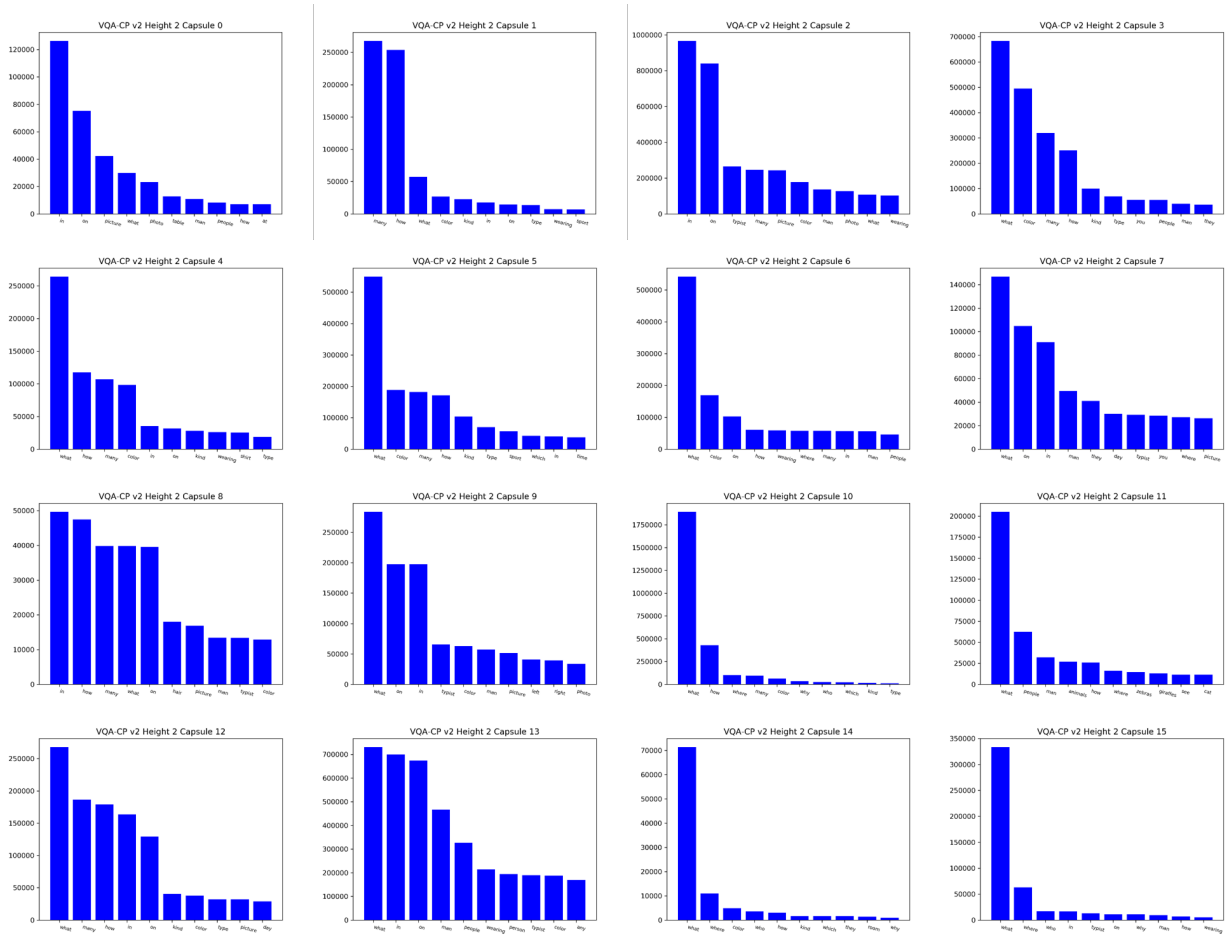


(d) Height 4

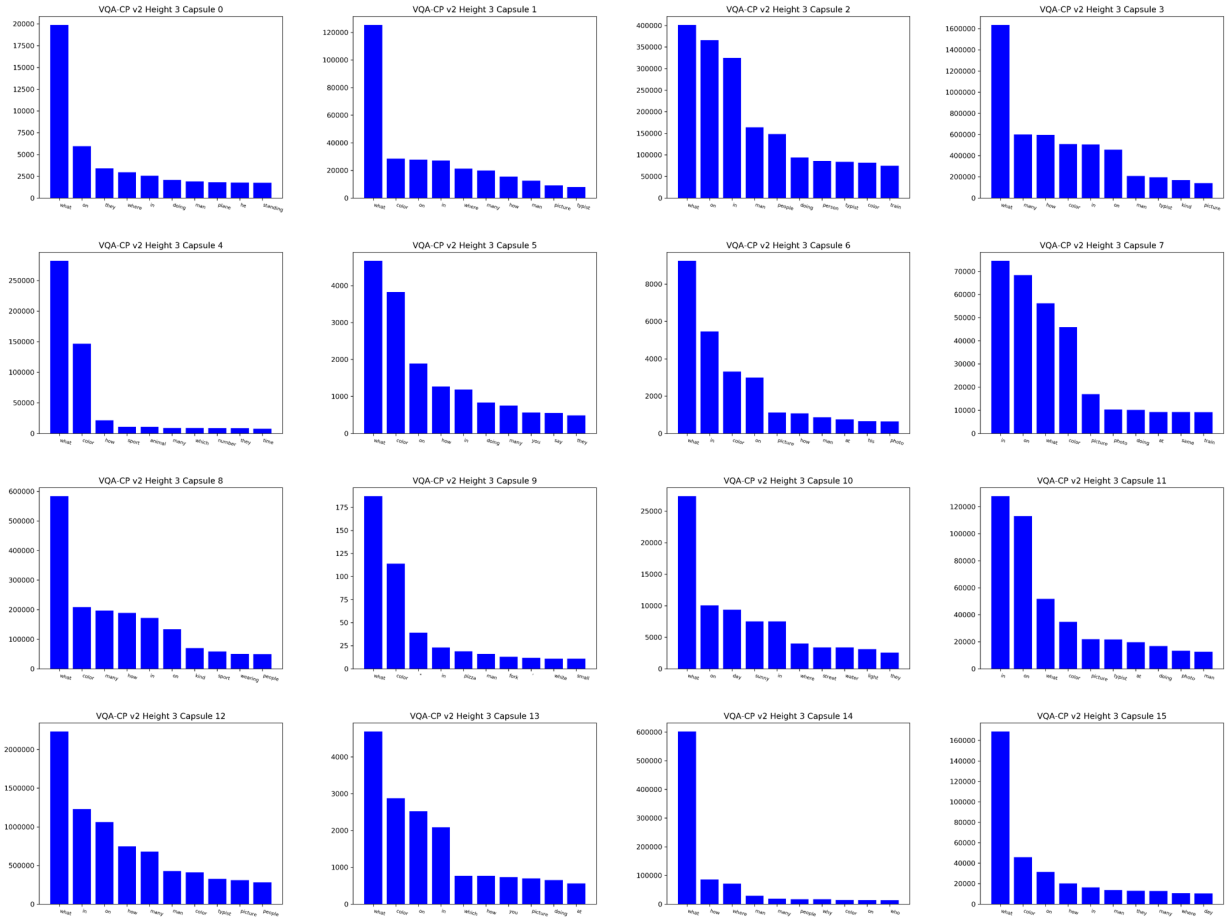
Figure 3: The 10 most frequent words encoded by each capsule on CLEVR-CoGenT.


VQA-CP v2 Height 1 Capsule 0
VQA-CP v2 Height 1 Capsule 1
VQA-CP v2 Height 1 Capsule 2
VQA-CP v2 Height 1 Capsule 3
VQA-CP v2 Height 1 Capsule 4
VQA-CP v2 Height 1 Capsule 5
VQA-CP v2 Height 1 Capsule 6
VQA-CP v2 Height 1 Capsule 7
VQA-CP v2 Height 1 Capsule 8
VQA-CP v2 Height 1 Capsule 9
VQA-CP v2 Height 1 Capsule 10
VQA-CP v2 Height 1 Capsule 11
VQA-CP v2 Height 1 Capsule 12
VQA-CP v2 Height 1 Capsule 13
VQA-CP v2 Height 1 Capsule 14
VQA-CP v2 Height 1 Capsule 15

(a) Height 1


VQA-CP v2 Height 2 Capsule 0
VQA-CP v2 Height 2 Capsule 1
VQA-CP v2 Height 2 Capsule 2
VQA-CP v2 Height 2 Capsule 3
VQA-CP v2 Height 2 Capsule 4
VQA-CP v2 Height 2 Capsule 5
VQA-CP v2 Height 2 Capsule 6
VQA-CP v2 Height 2 Capsule 7
VQA-CP v2 Height 2 Capsule 8
VQA-CP v2 Height 2 Capsule 9
VQA-CP v2 Height 2 Capsule 10
VQA-CP v2 Height 2 Capsule 11
VQA-CP v2 Height 2 Capsule 12
VQA-CP v2 Height 2 Capsule 13
VQA-CP v2 Height 2 Capsule 14
VQA-CP v2 Height 2 Capsule 15

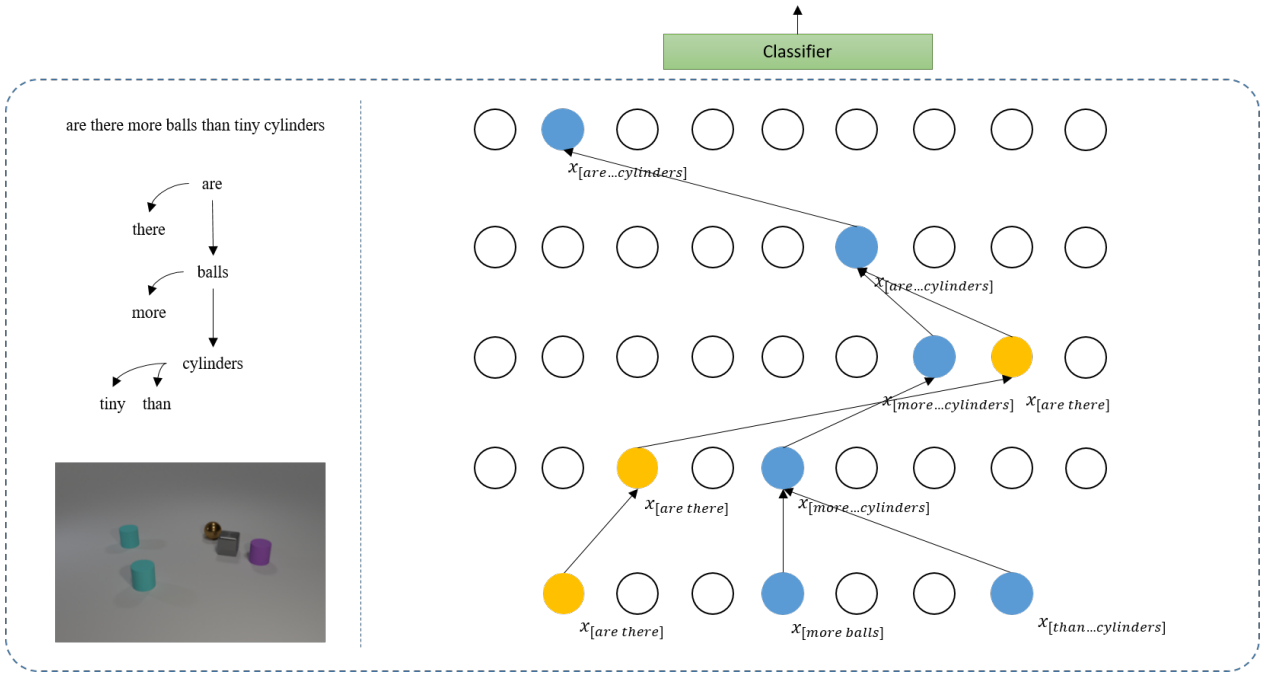
(b) Height 2

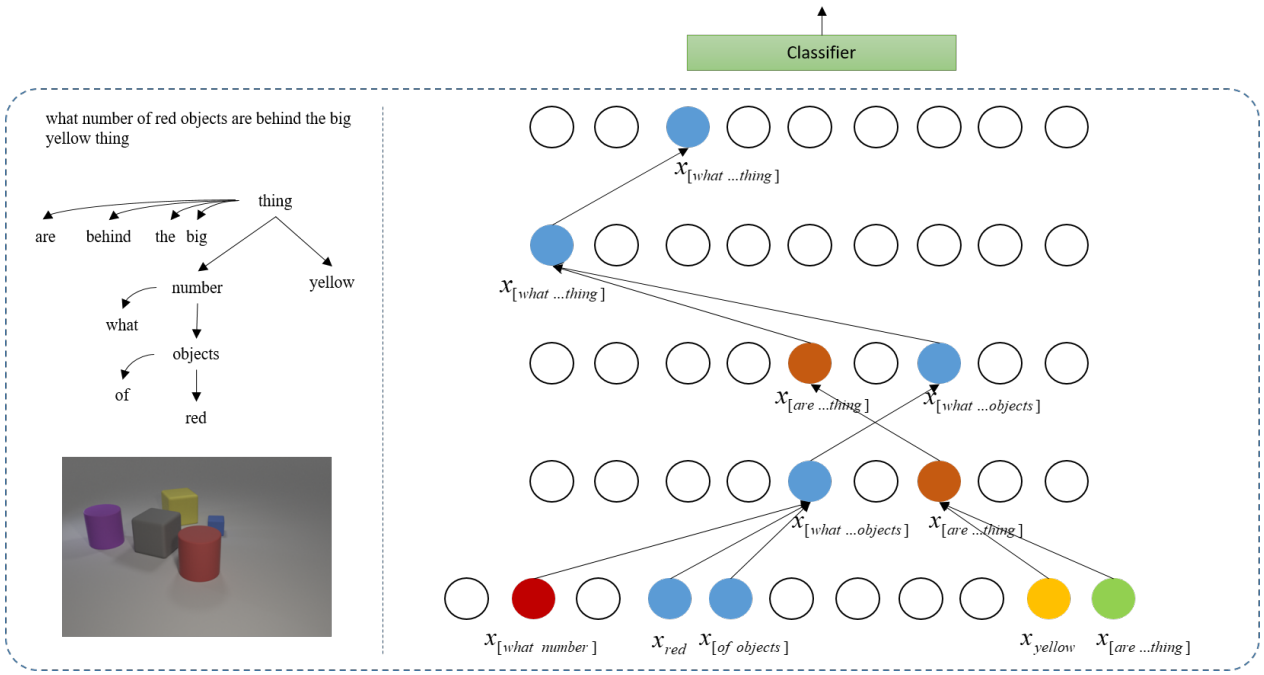


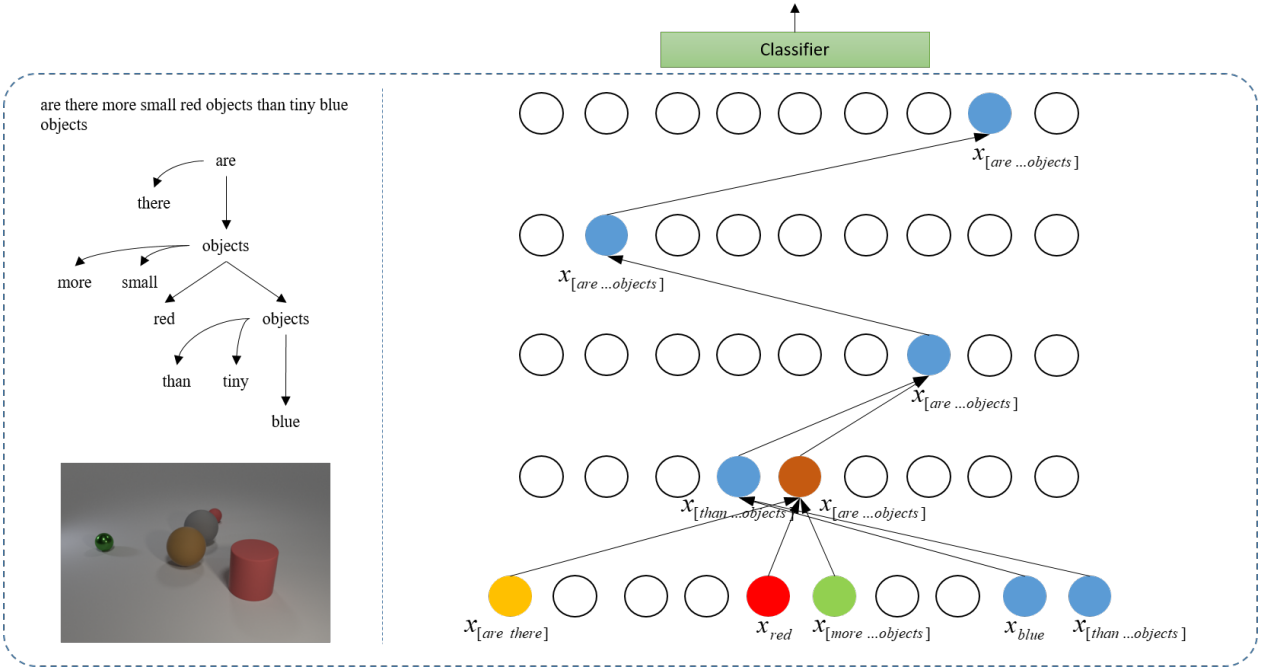
(c) Height 3

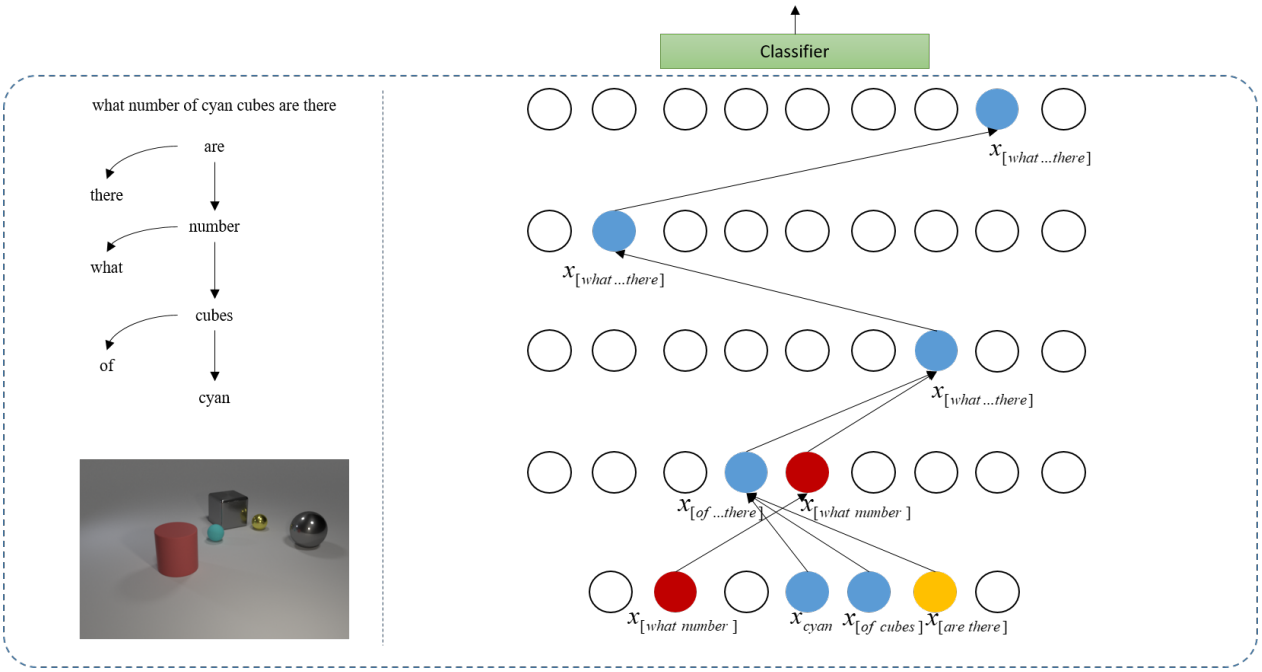
Figure 4: The 10 most frequent words encoded by each capsule on VQA-CP v2.

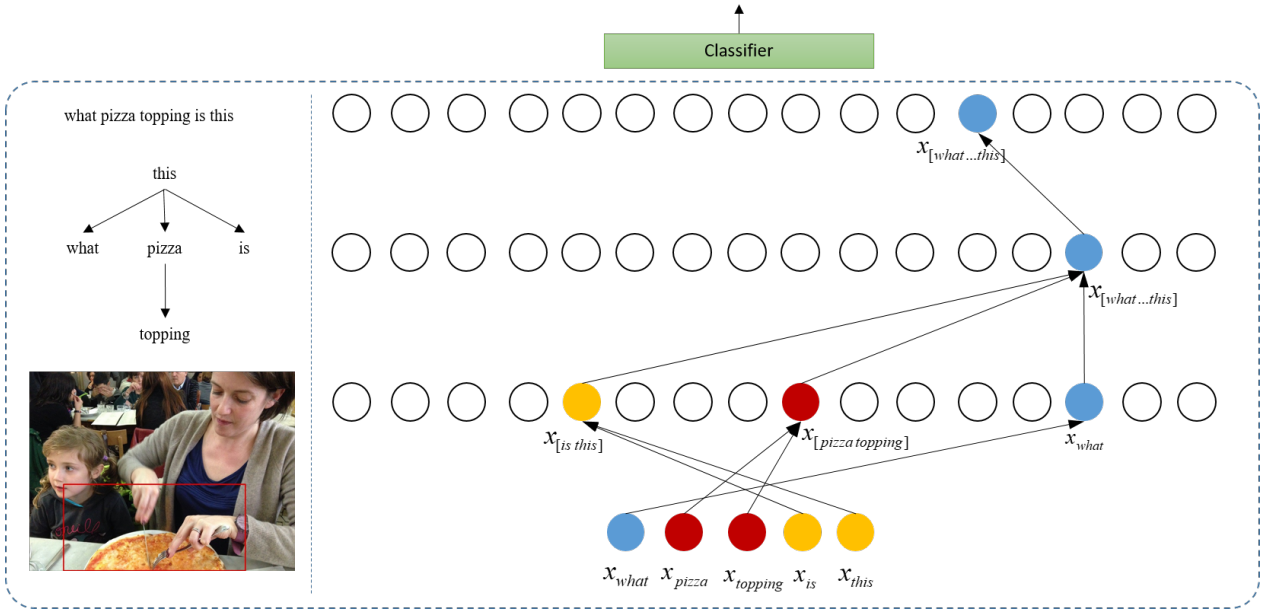
4. More Visualisation Results

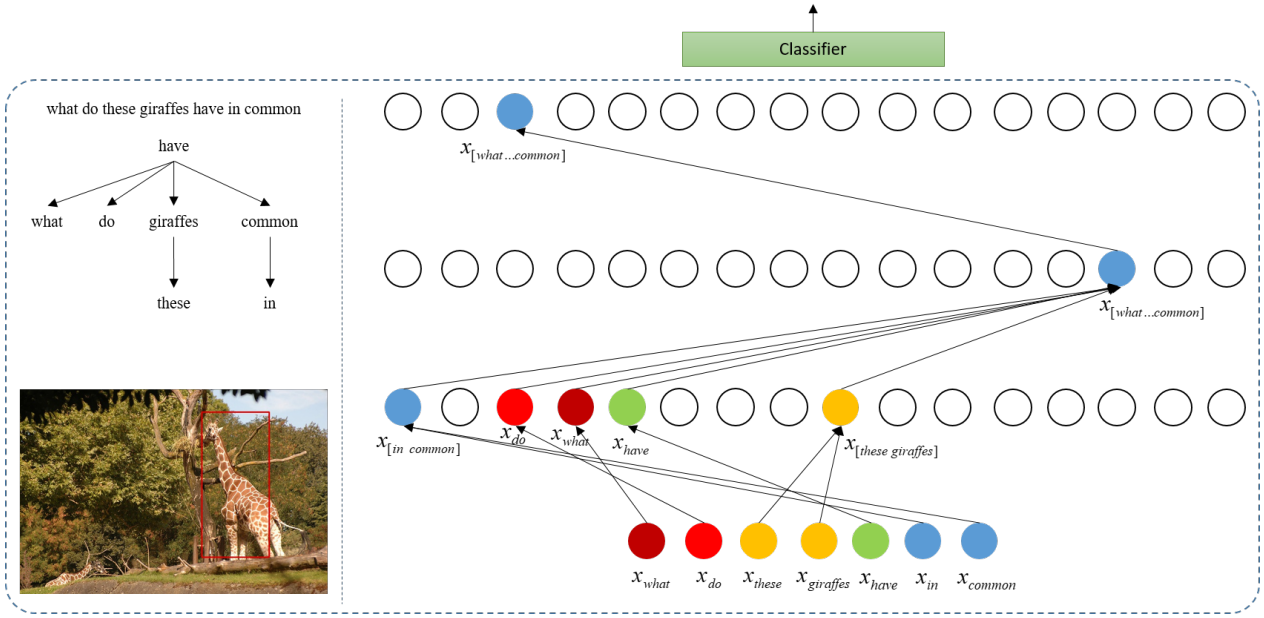


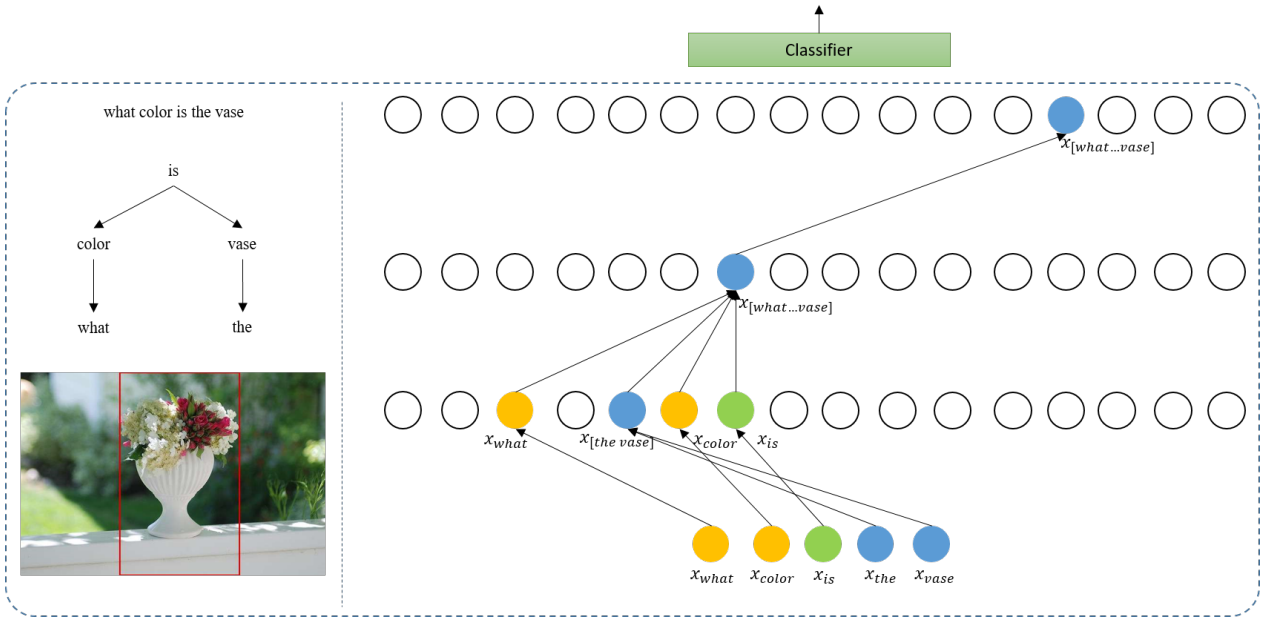


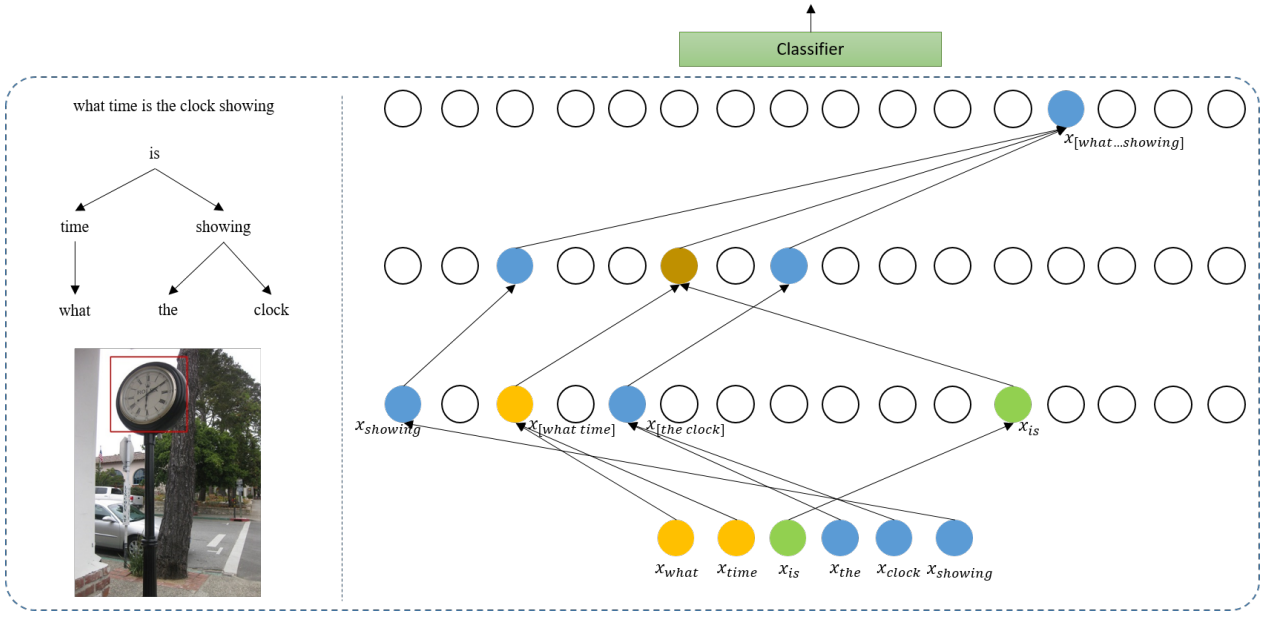

Classifier
are there more small red objects than tiny blue objects
are
there
objects
more
small
red
objects
than
tiny
blue
$x_{[are\ ...objects]}$
$x_{[are\ ...objects]}$
$x_{[are\ ...objects]}$
$x_{[than\ ...objects]}$
$x_{[are\ ...objects]}$
$x_{[are\ there]}$
x_{red}
$x_{[more\ ...objects]}$
x_{blue}
$x_{[than\ ...objects]}$


Classifier
what number of cyan cubes are there
are
there
number
what
cubes
of
cyan
$x_{[what\ ...there]}$
$x_{[what\ ...there]}$
$x_{[what\ ...there]}$
$x_{[of\ ...there]}$
$x_{[what\ number]}$
$x_{[what\ number]}$
x_{cyan}
$x_{[of\ cubes]}$
$x_{[are\ there]}$


Classifier
what pizza topping is this
this
what
pizza
is
topping
$x_{[what\ ...this]}$
$x_{[what\ ...this]}$
$x_{[is\ this]}$
$x_{[pizza\ topping]}$
x_{what}
x_{what} x_{pizza} $x_{topping}$ x_{is} x_{this}


Classifier
what do these giraffes have in common
have
what
do
giraffes
common
these
in
$x_{[what\ ...common]}$
$x_{[what\ ...common]}$
$x_{[in\ common]}$
x_{do}
x_{what}
x_{have}
$x_{[these\ giraffes]}$
x_{what} x_{do} x_{these} $x_{giraffes}$ x_{have} x_{in} x_{common}


Classifier
what color is the vase
is
color
vase
what
the
$x_{[what\ldots vase]}$
x_{what}
$x_{[the\ vase]}$
x_{color}
x_{is}
x_{what}
x_{color}
x_{is}
x_{the}
x_{vase}


Classifier
what time is the clock showing
is
time
showing
what
the
clock
$x_{[what\ldots showing]}$
$x_{showing}$
$x_{[what\ time]}$
$x_{[the\ clock]}$
x_{is}
x_{what}
x_{time}
x_{is}
x_{the}
x_{clock}
$x_{showing}$

References

[1] Qingxing Cao, Xiaodan Liang, Bailing Li, Guanbin Li, and Liang Lin. Visual question reasoning on general dependency tree. In *CVPR*, June 2018. 2

[2] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, 2017. 2

[3] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. 2018. 2

[4] 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 *CVPR*, 2017. 1

[5] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Fei-Fei Li, C. Lawrence Zitnick, and Ross B. Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, 2017. 2

[6] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *CVPR*, 2018. 2

[7] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 2

[8] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy P. Lillicrap. A simple neural network module for relational reasoning. *CoRR*, abs/1706.01427, 2017. 2

[9] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *NIPS*, 2018. 2

[10] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 1

[11] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured attentions for visual question answering. In *ICCV*, 2017. 2