Multimodal Learning and Reasoning for Visual Question Answering

MSc Project Demonstration

Vasin Srisupavanich

MSc Artificial Intelligence 19/20

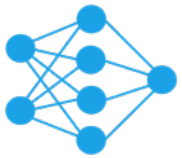
Supervisor: Dr Srinandan Dasmahapatra

2nd Examiner: Professor Eric Rogers

UNIVERSITY OF Southampton

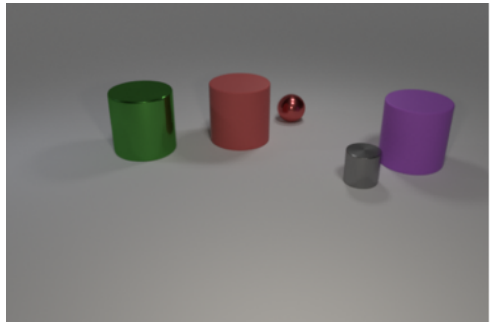
Motivation

Current artificial neural networks are very good at pattern recognition but not so good at reasoning.



cat

Object recognition



Object detection

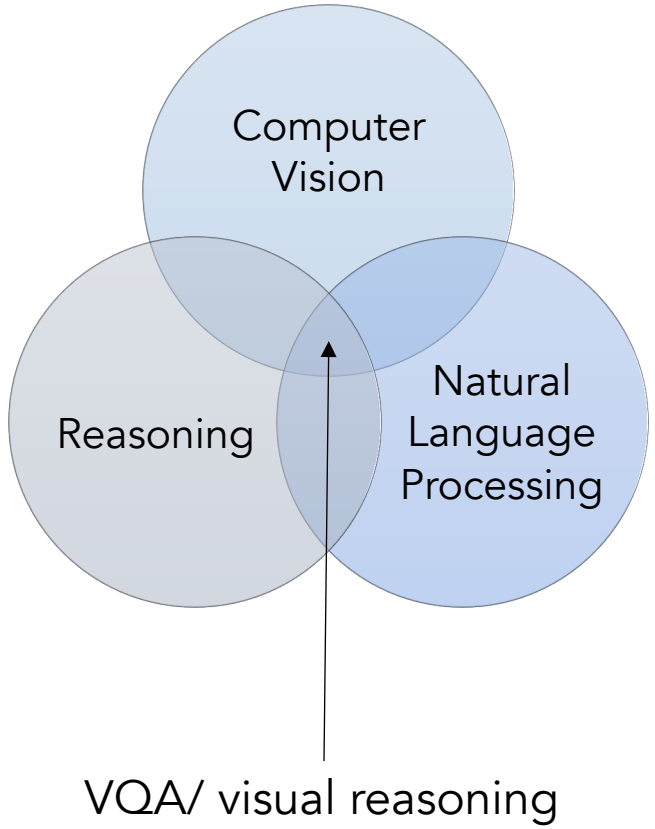
Overview

Visual Question Answering (VQA)

- What colour are her eyes?
- What is the mustache made of?

Visual Reasoning

- Is there any fruit to the left of the tray the cup is on top of?
- Are the napkin and the cup the same colour?



Aims and Objectives

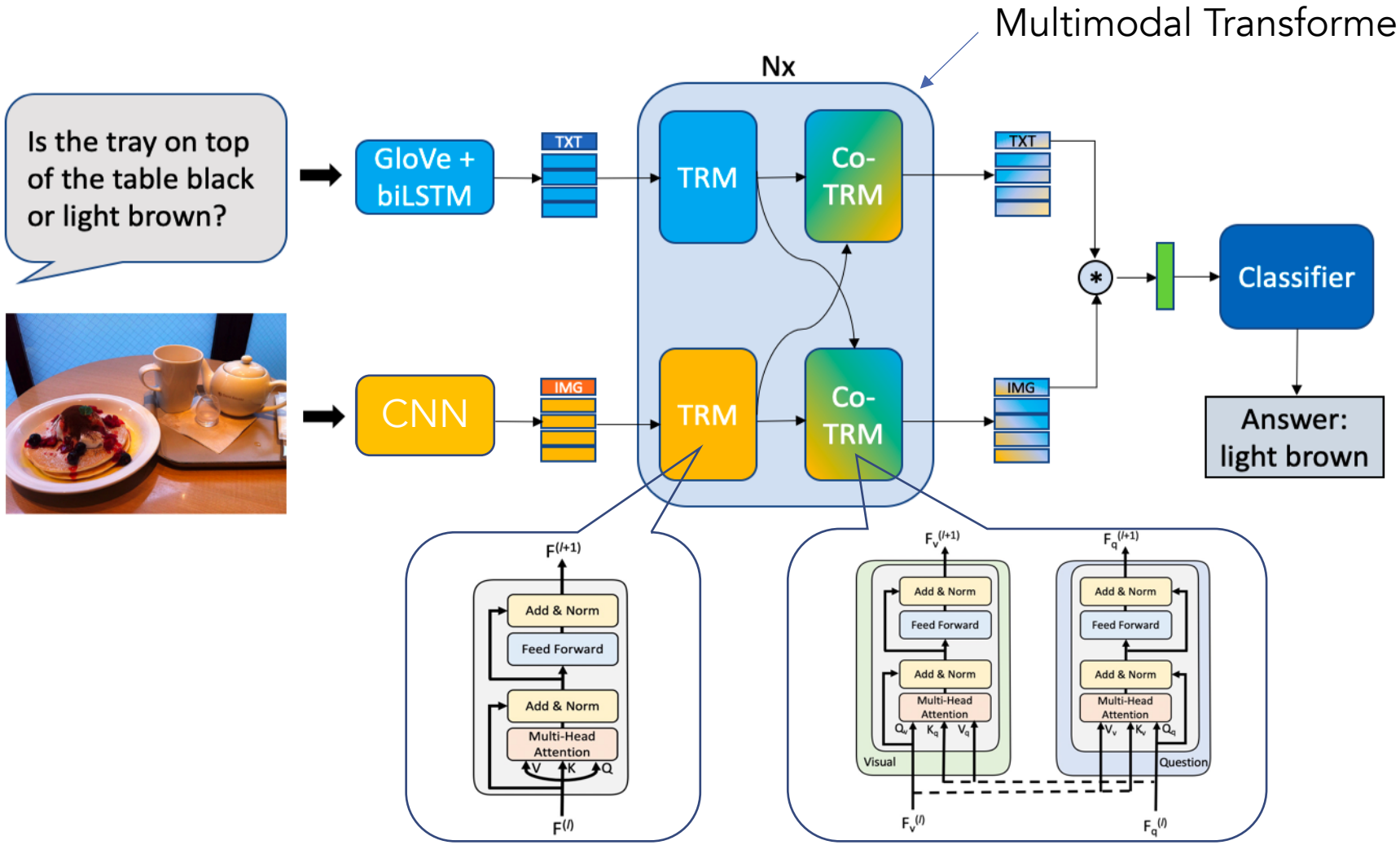
Objectives

- Study existing deep learning models for VQA and visual reasoning
- Explore recent approaches which extend the reasoning capabilities of artificial neural networks

Aim

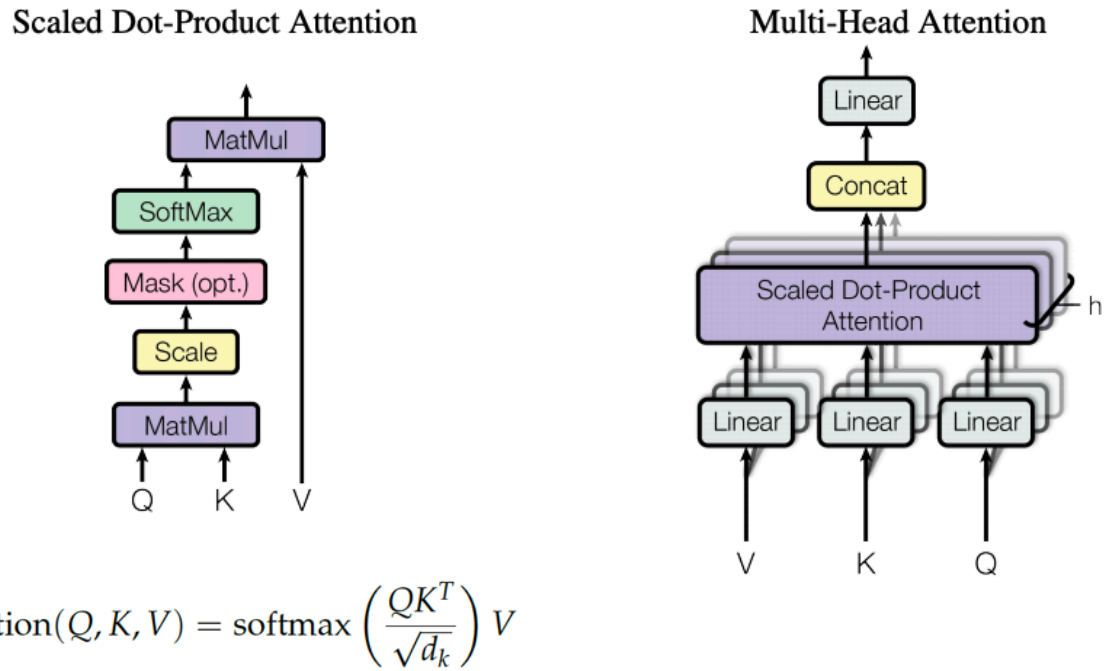
- Develop a new VQA model based on Transformer architecture
  - Utilising self-attention and co-attention mechanism

Model Details

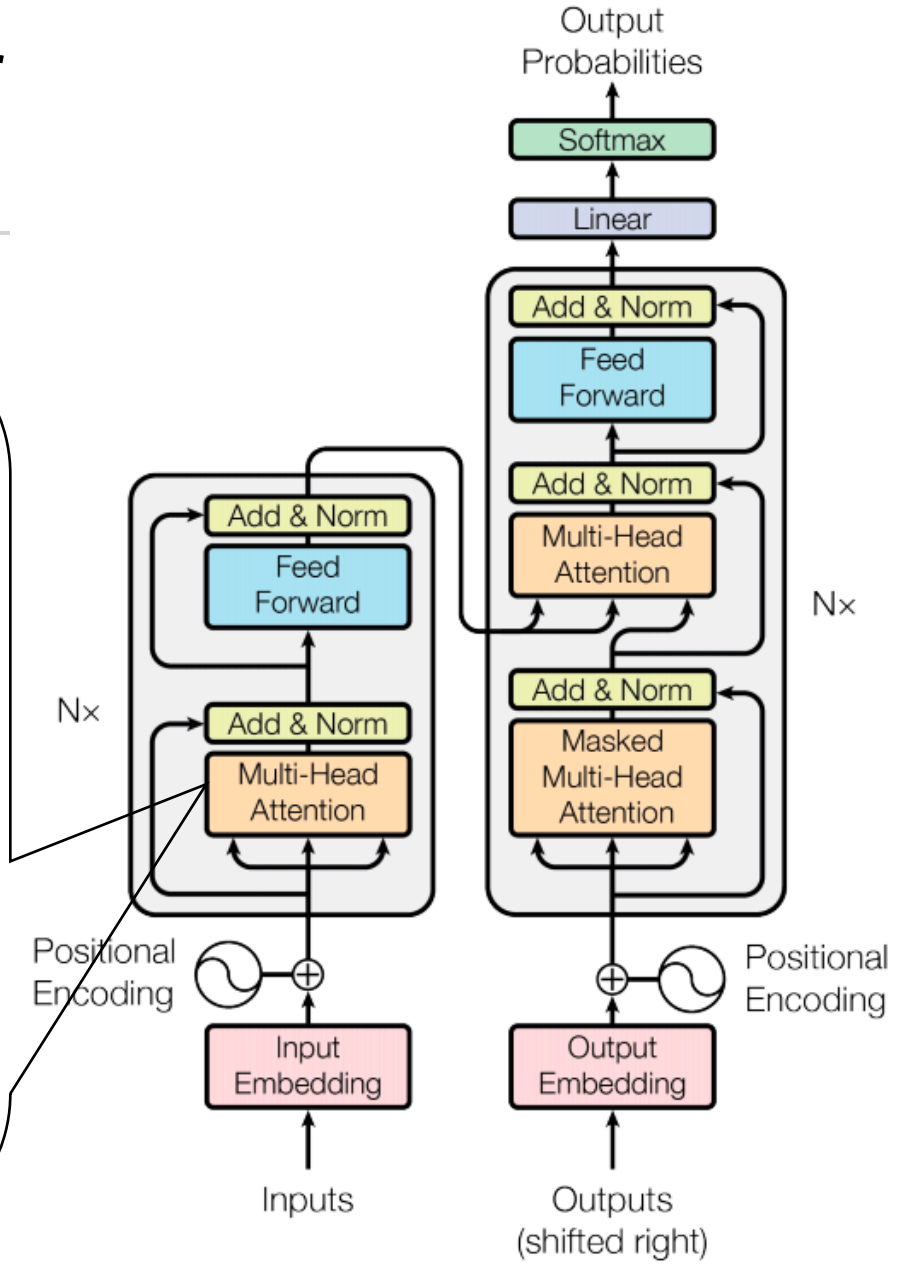


Background - Transformer

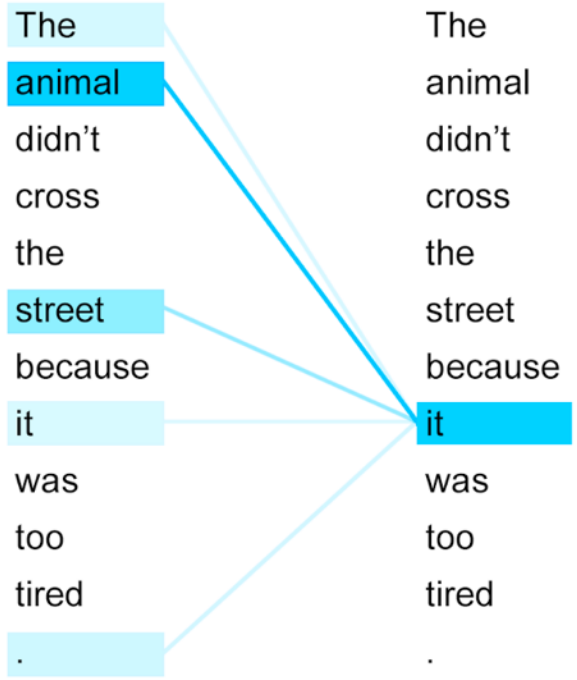
Self-attention



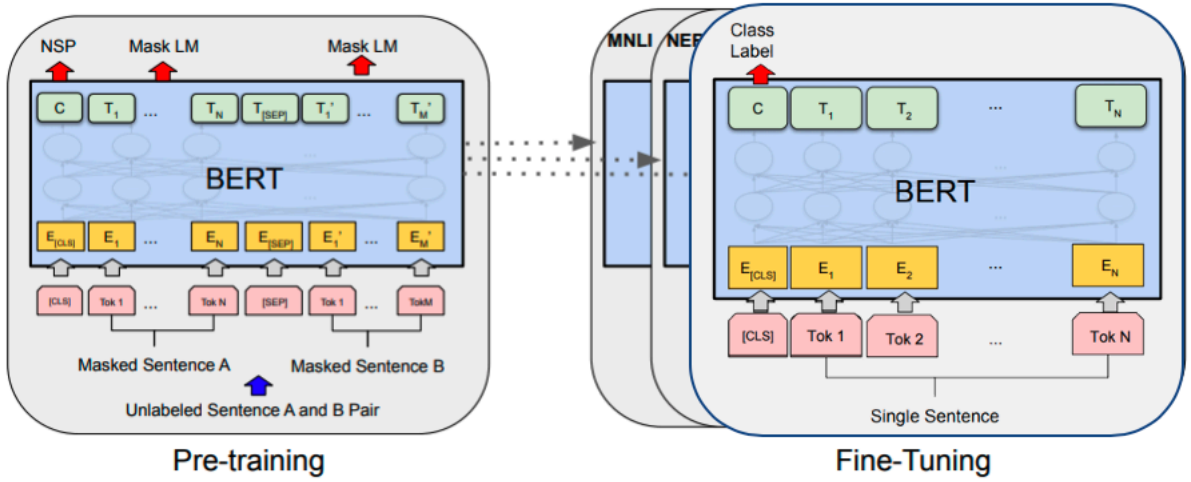
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Background – Transformer (Cont.)



Visualisation of Self-attention

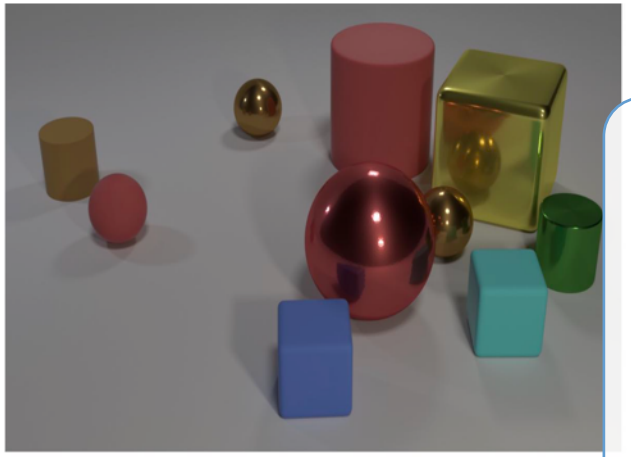


BERT

Experiment

Datasets

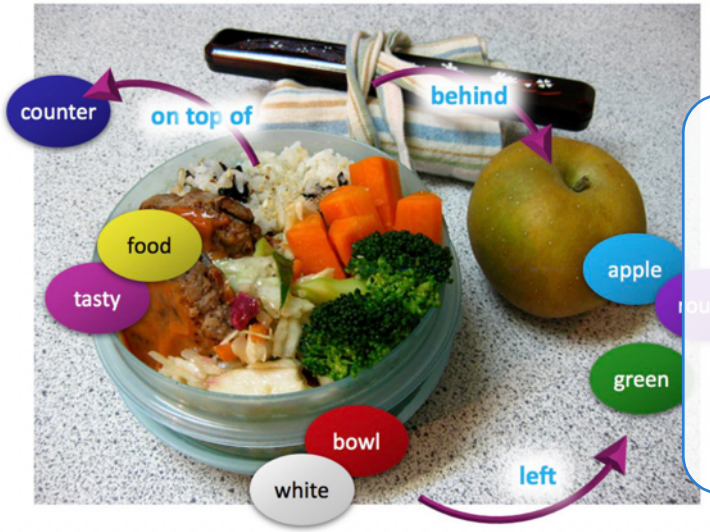
CLEVR



- 850K automatically generated questions
- 100k synthetic images of 3D-rendered objects

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 **red** things?

GQA



- 20M compositional questions
- 113k real-world images

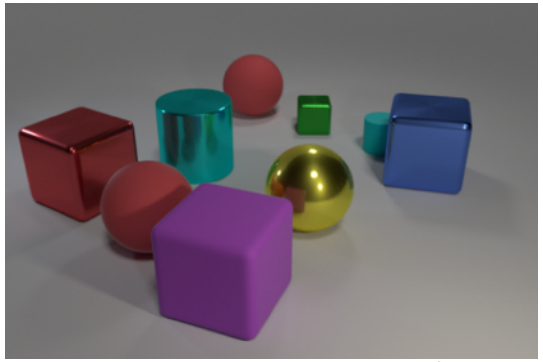
*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

Results - CLEVR dataset

Model	Count	Compare Numbers	Exist	Query Attribute	Compare Attribute	Overall
Human [49]	86.7	96.6	86.5	95.0	96.0	92.6
Q-type Baseline [49]	34.6	50.2	51.0	36.0	51.3	41.8
LSTM [49]	41.7	61.1	69.8	36.8	51.8	46.8
CNN+LSTM [49]	43.7	65.2	67.1	49.3	53.0	52.3
SAN (baseline model) [24]	59.7	77.9	75.1	80.9	70.8	73.2
N2NMN* [50]	68.5	85.7	84.9	90.0	88.7	83.7
IEP* [49]	92.7	98.7	97.1	98.1	98.9	96.9
Relation Networks [57]	90.1	97.8	93.6	97.9	97.1	95.5
FiLM [58]	94.3	99.3	93.4	99.3	99.3	97.6
NS-CL† [30]	98.2	99.0	98.8	99.3	99.1	98.9
MAC [29]	97.1	99.1	99.5	99.5	99.5	98.9
Our Model	95.5	98.3	99.1	99.5	99.0	98.3

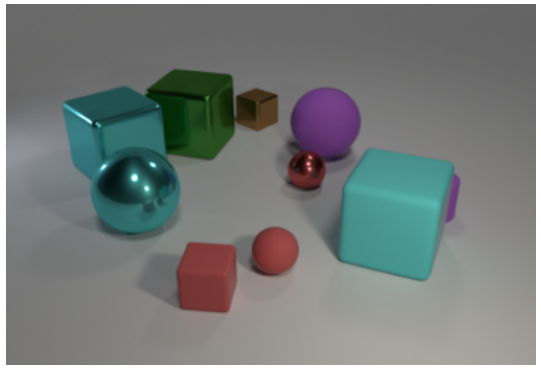
TABLE 5.2: Results of CLEVR dataset (validation set). (∗) denotes use of extra supervisory information through functional program. (†) denotes use of pre-defined symbolic functions and attribute templates.

Qualitative results - CLEVR



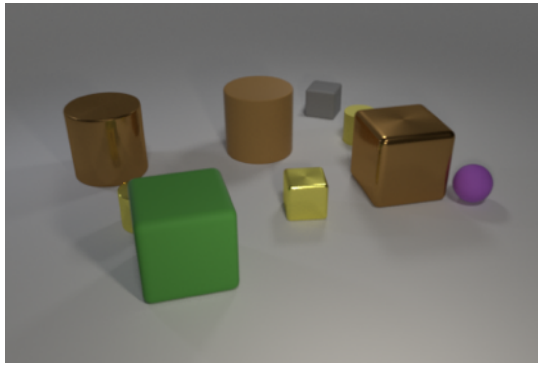
How many other things are there of the same shape as the green thing?

Answer: 3 Prediction: 3



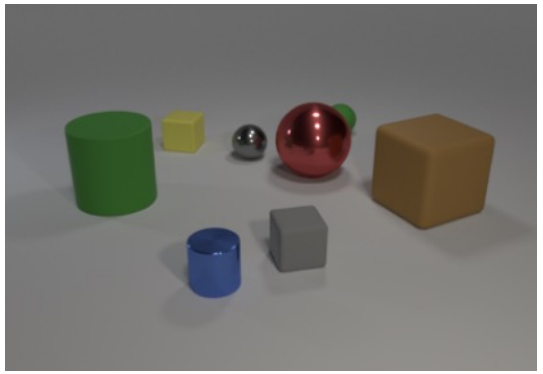
Are there any large balls that have the same material as the tiny red cube?

Answer: yes Prediction: yes



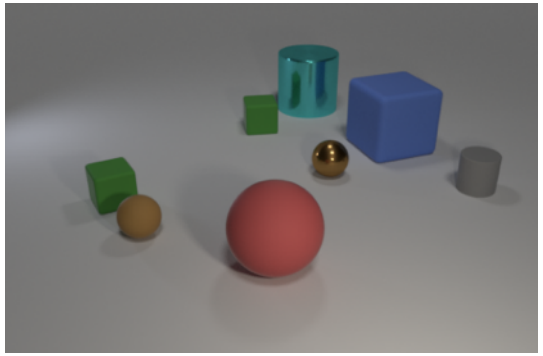
What number of objects are the same colour as the tiny metal cylinder?

Answer: 2 Prediction: 1



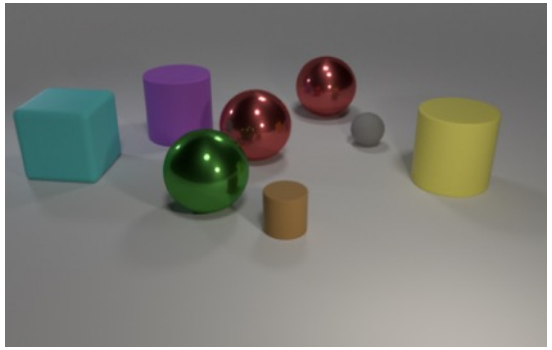
There is a small gray matte thing: what shape is it?

Answer: cube Prediction: cube



Is the number of big matte blocks greater than the number of tiny red metal cubes?

Answer: yes Prediction: yes



What number of large shiny balls are left of the small cylinder and behind the matte cube?

Answer: 1 Prediction: 0

Results - GQA dataset

Model	Binary	Open	Consistency	Validity	Plausibility	Distribution	Accuracy
Human [6]	91.20	87.40	98.40	98.90	97.20	-	89.30
LSTM [6]	61.90	22.69	68.68	96.39	87.30	17.93	41.07
CNN+LSTM [6]	63.26	31.80	74.57	96.02	84.25	7.46	46.55
BottonUp [26]	66.64	34.83	78.71	96.18	84.57	5.98	49.75
MAC [29]	71.23	38.91	81.59	96.16	84.48	5.34	54.06
MCAN [32]	76.49	42.45	87.36	96.98	84.47	1.29	58.38
LXMERT [38]	77.76	44.97	92.84	96.30	85.19	8.31	60.34
NSM [47]	78.98	49.25	93.25	96.41	84.28	3.71	63.17
Our Model	73.73	40.87	86.86	96.01	84.20	5.78	56.28

TABLE 5.4: Results of GQA dataset (test set)

Qualitative results - GQA

Are both the sky and the shirt the same color?
Answer: no Prediction: no

What vegetables are inside the bowl the cookies are to the right of?
Answer: beans Prediction: beans

Where is the man?
Answer: lawn Prediction: grass

Is the white bag to the left or to the right of the chair?
Answer: left Prediction: left

What is the meat on the pizza called?
Answer: pepperoni Prediction: pepperoni

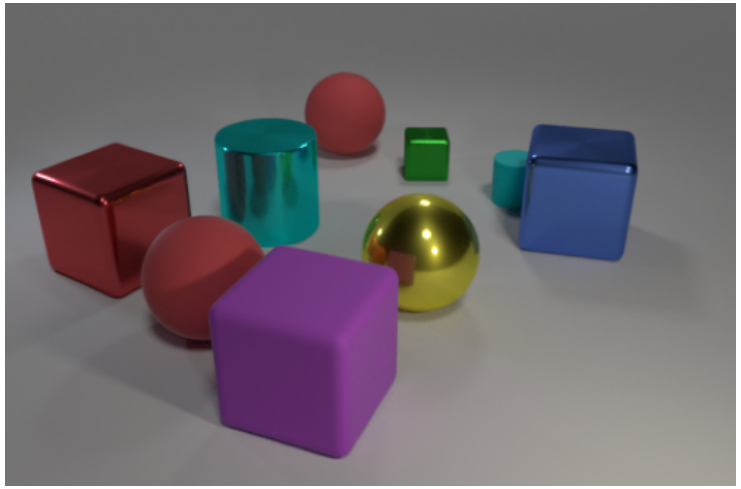
Which color is the table the lamp is above?
Answer: dark Prediction: brown

Analysis

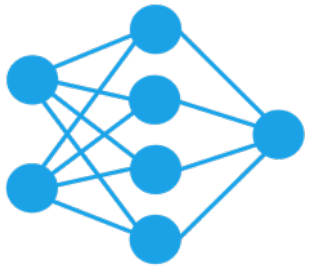
Analysis

- Analysis of the model on CLEVR dataset
- Visualising the attention distributions produced by the model during the inference process
  - To obtain insights into the underlying reasoning process of our model
  - Understand why the model make mistakes (error analysis)
- Attention distribution of the [IMG] and [TXT] tokens across all the layers and attention heads.

Analysis – Human Behaviour



How many other things are there of the same shape as the green thing?



Answer: 3

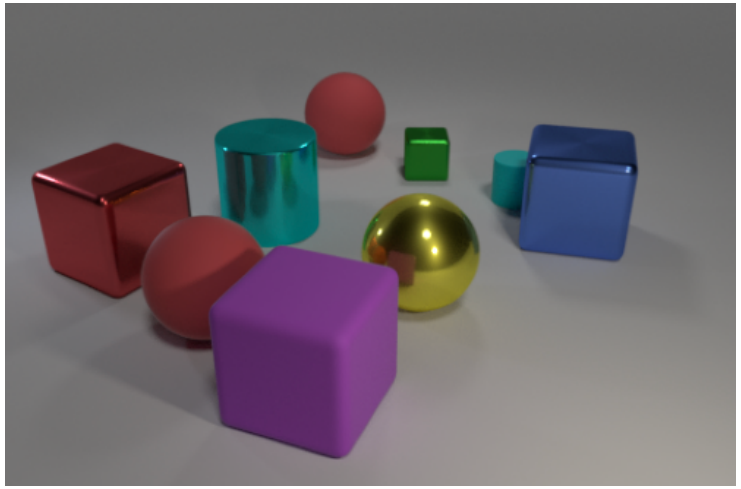
Find green thing

Identify its shape

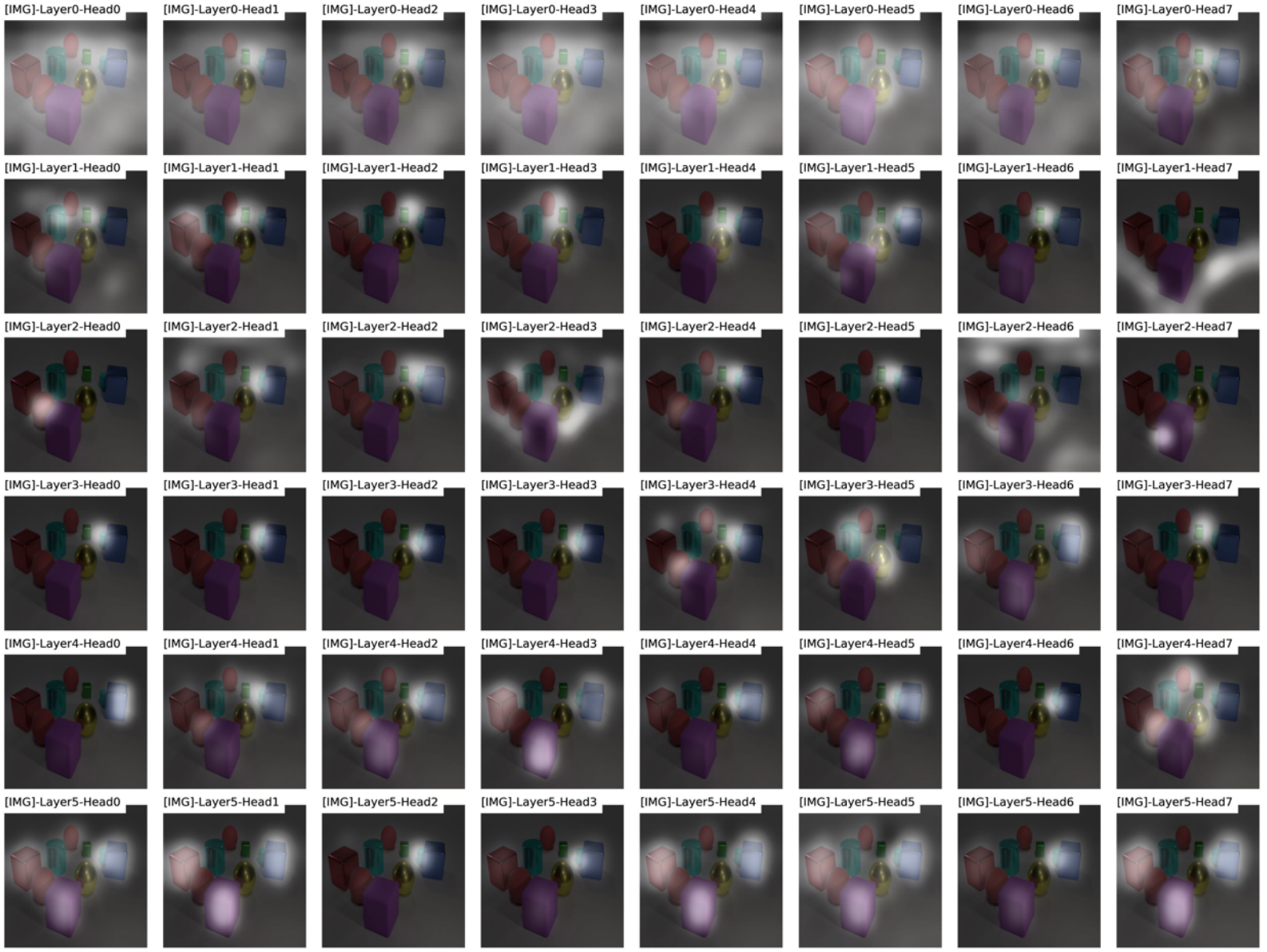
Locate other objects with the same shape

Count all the located objects

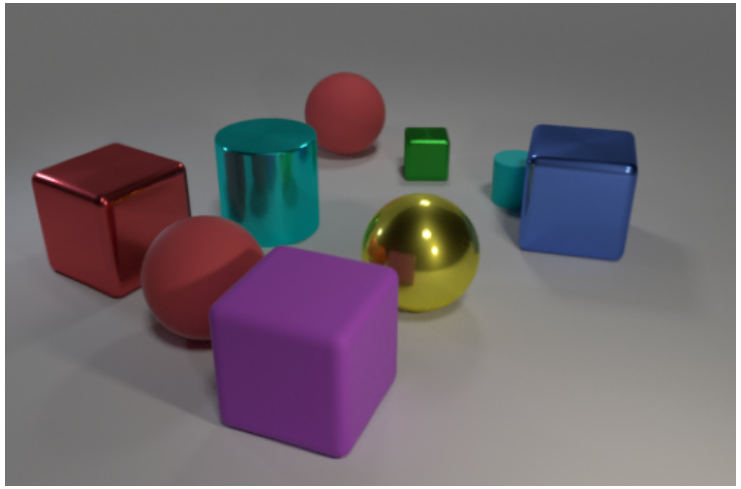
Analysis (1)



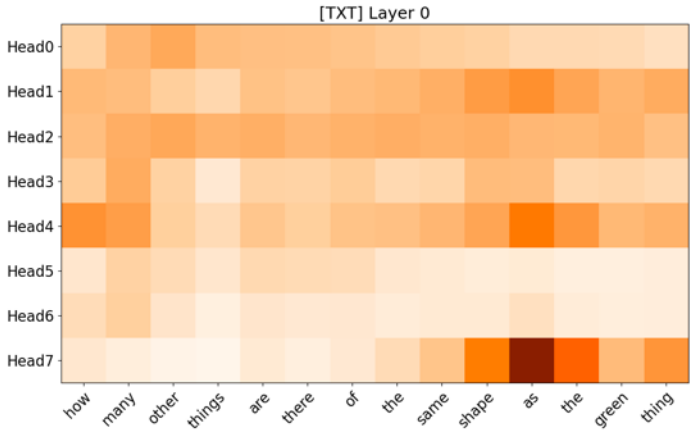
How many other things are there of the same shape as the green thing?
Answer: 3
Prediction: 3

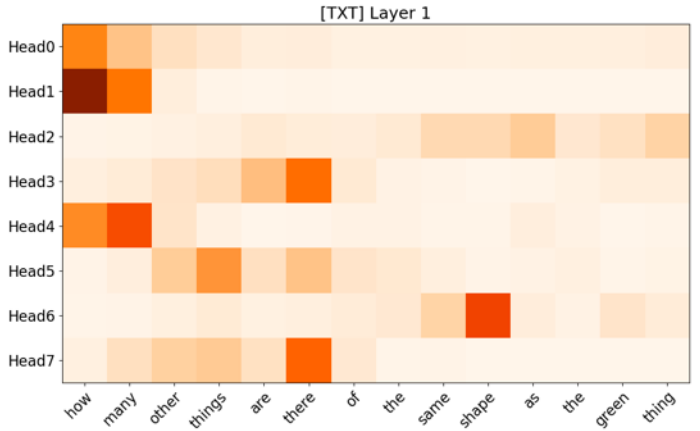


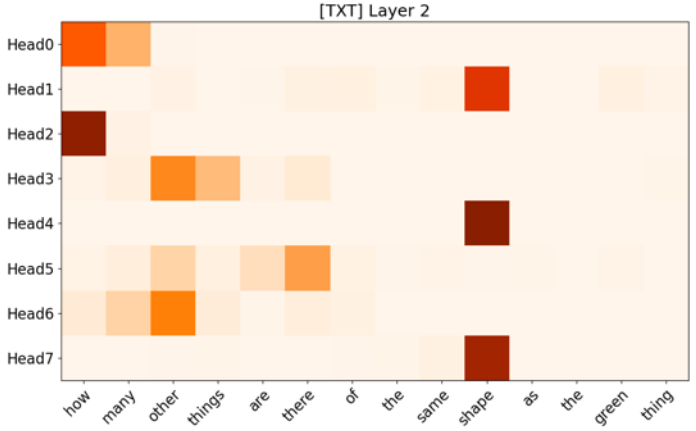
Analysis (1)

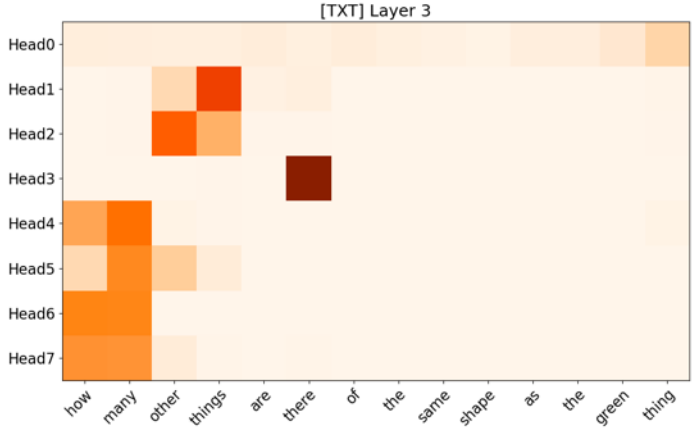


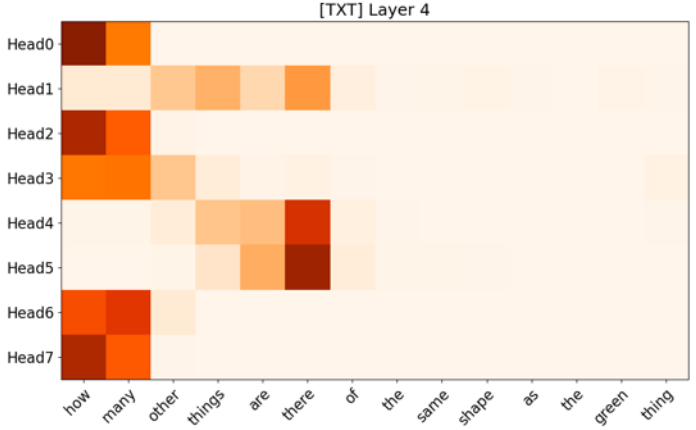
How many other things are there of the same shape as the green thing?
Answer: 3
Prediction: 3

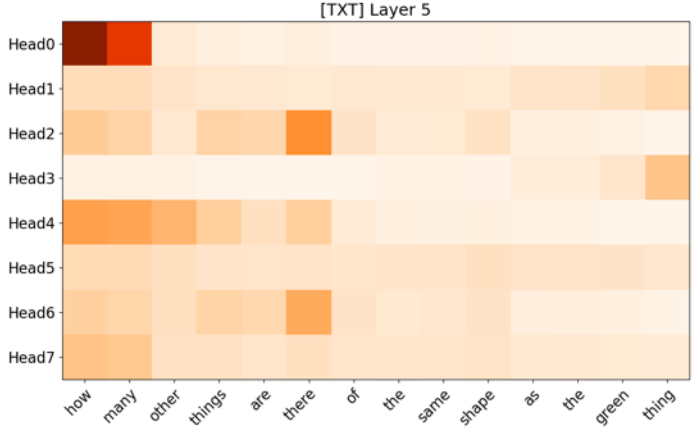




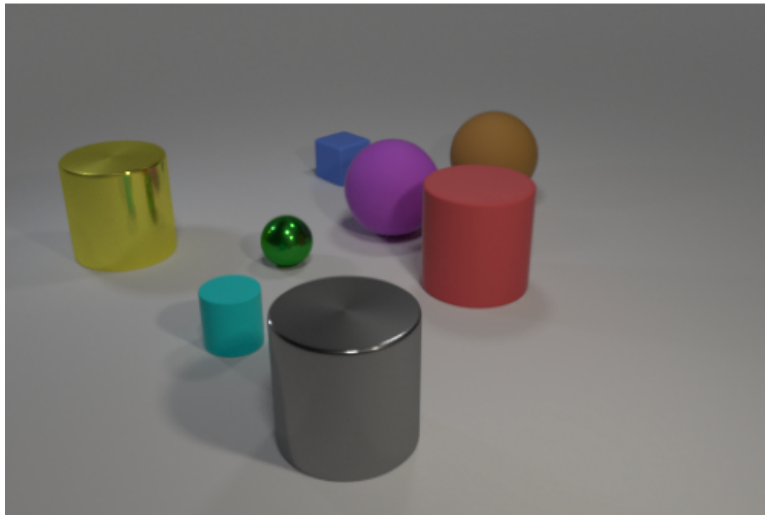




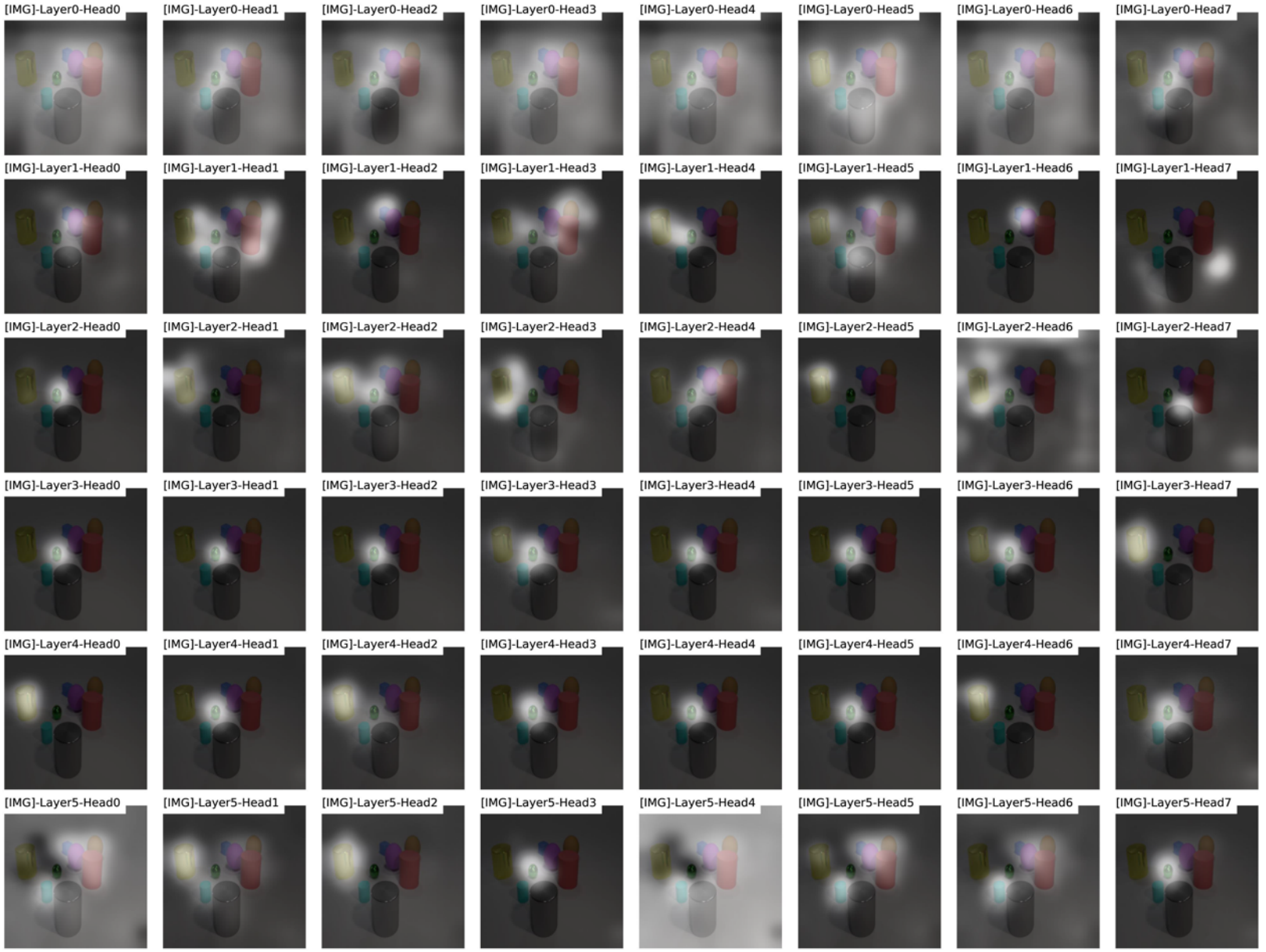




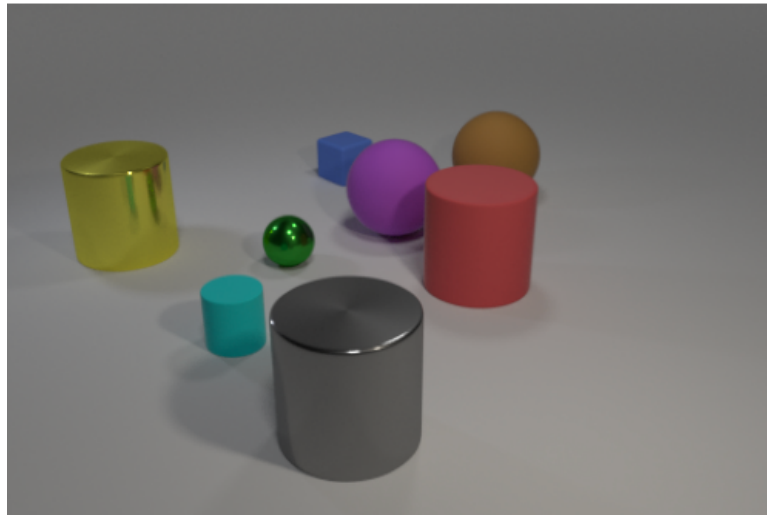
Analysis (2)



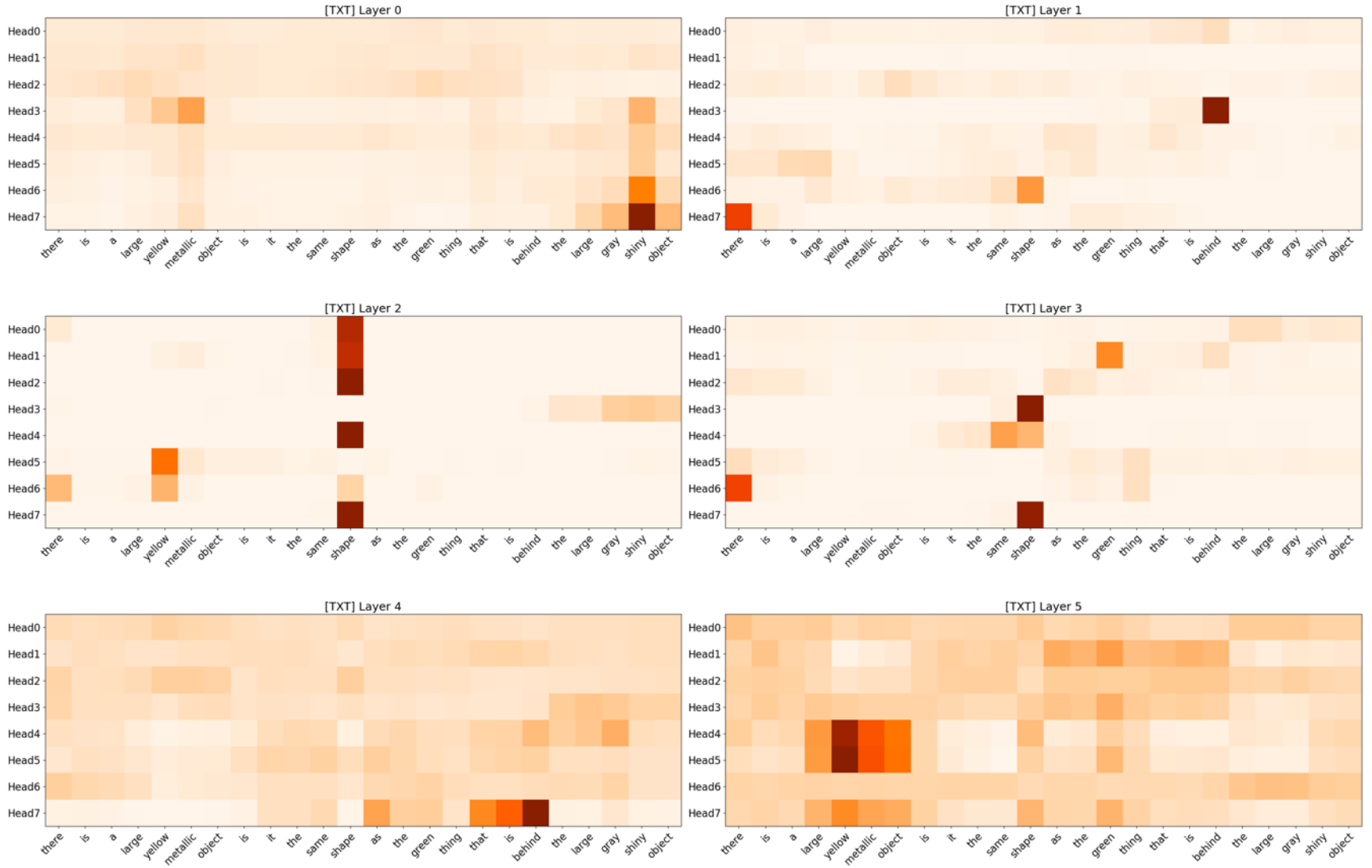
There is a large yellow metallic object; is it the same shape as the green thing that is behind the large gray shiny object?
Answer: no
Prediction: no



Analysis (2)



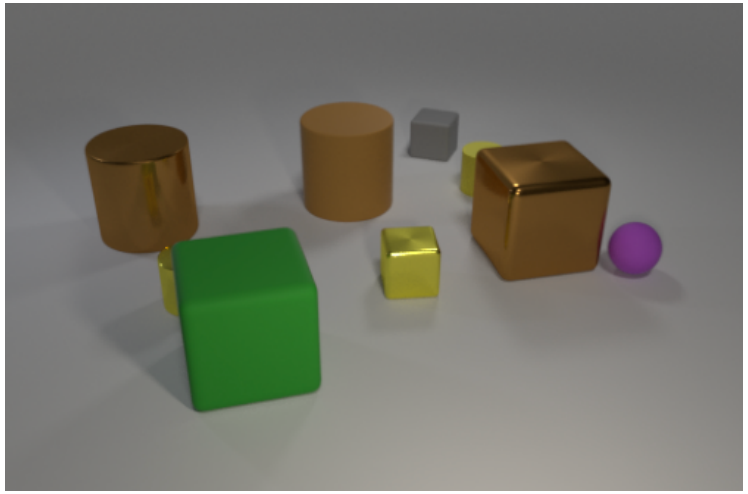
There is a large yellow metallic object; is it the same shape as the green thing that is behind the large gray shiny object?
Answer: no
Prediction: no



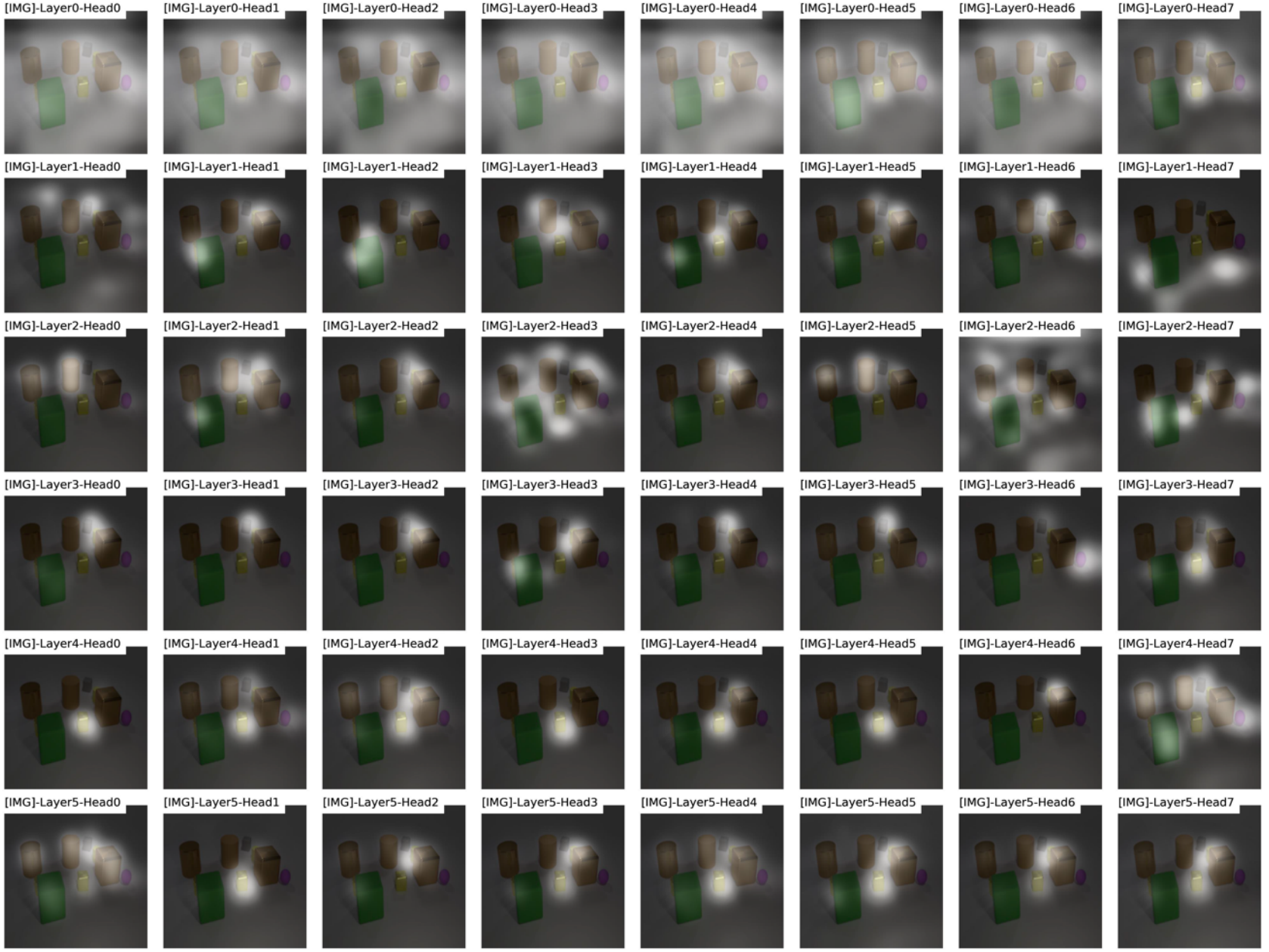
Analysis Summary

- In the first layer of visual attention, the model tends to look at the overall picture.
- It is able to learn the concept of spatial relation (left, right, in front, behind)
- It can also focus on multiple objects with the same attribute at the same time.
- It is adaptable to different difficulty level of question and image.

Error Analysis



What number of objects are the same color as the tiny metal cylinder?
Answer: 2
Prediction: 1



Conclusion

- Visualisations show that the model learns to look around the image, and iteratively focus on relevant objects.
- Shown one effective way to adapt the Transformer architecture for multimodal supervised learning.