Visual Question Answering
Final Project Discussion

Bhavishya, 19D070017
Mukesh Kumar, 190010047
Aditya Kudre, 200070039

30 April, 2023

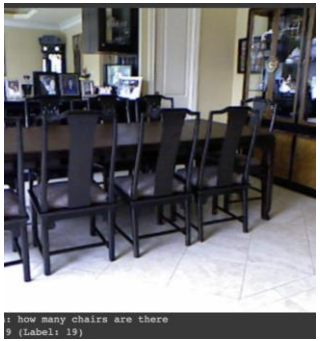
Problem Statement

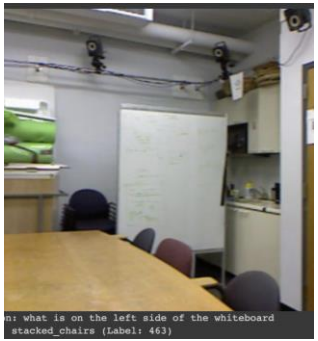
- Given an image and a natural language question related to the image, the objective is to produce a natural language answer correctly.
- Need NLP for two reasons: to understand the question and to generate the answer

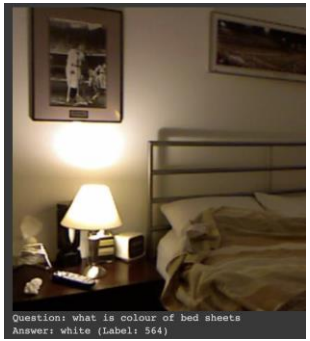
Related work

- Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge
- A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input
- CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning
- VQA: Visual Question Answering
- Stacked Attention Networks for Image Question Answering
- MUTAN: Multimodal Tucker Fusion for Visual Question Answering

Dataset







- DAtaset for QUestion Answering on Real-world images (DAQUAR)
- It contains 9974 training and 2494 test question-answer pairs, based on images from the NYU-Depth V2 Dataset. That means about 9 pairs per image on average.
- Although it is a great initiative, the NYU dataset contains only indoor scenes with, sometimes, lightning conditions that make it difficult to answer the questions. In fact, evaluation on humans shows an accuracy of 50.2%.
- https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/visual-turing-challenge/

Other Datasets (Not used)

1) **The COCO-QA dataset:** *123,287 images coming from the COCO dataset, 78,736 training and 38,948 testing QA pairs*
2) **The VQA dataset:** *In addition to 204,721 images from the COCO dataset, it includes 50,000 abstract cartoon images. There are three questions per image and ten answers per question, that is over 760K questions with around 10M answers.*
3) **CLEVR dataset:** *A training set of 70,000 images and 699,989 question, validation set of 15,000 images and 149,991 questions, and test set of 15,000 images and 14,988 questions*

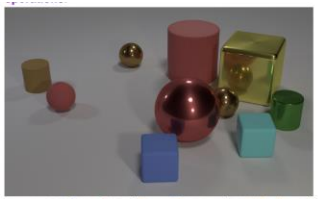
1

COCO-QA: What does an intersection show on one side and two double-decker buses and a third vehicle,?
Ground Truth: Building

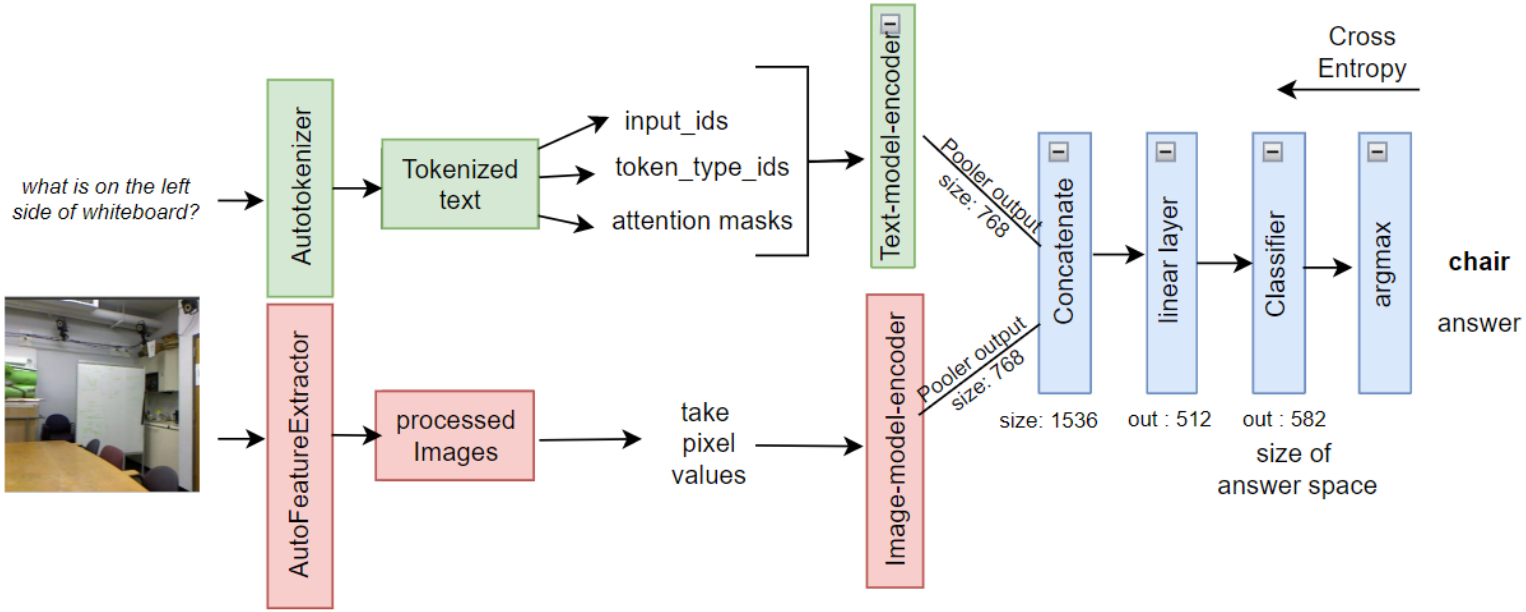
2

Does this man have children?
yes yes yes
yes yes yes

3



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?



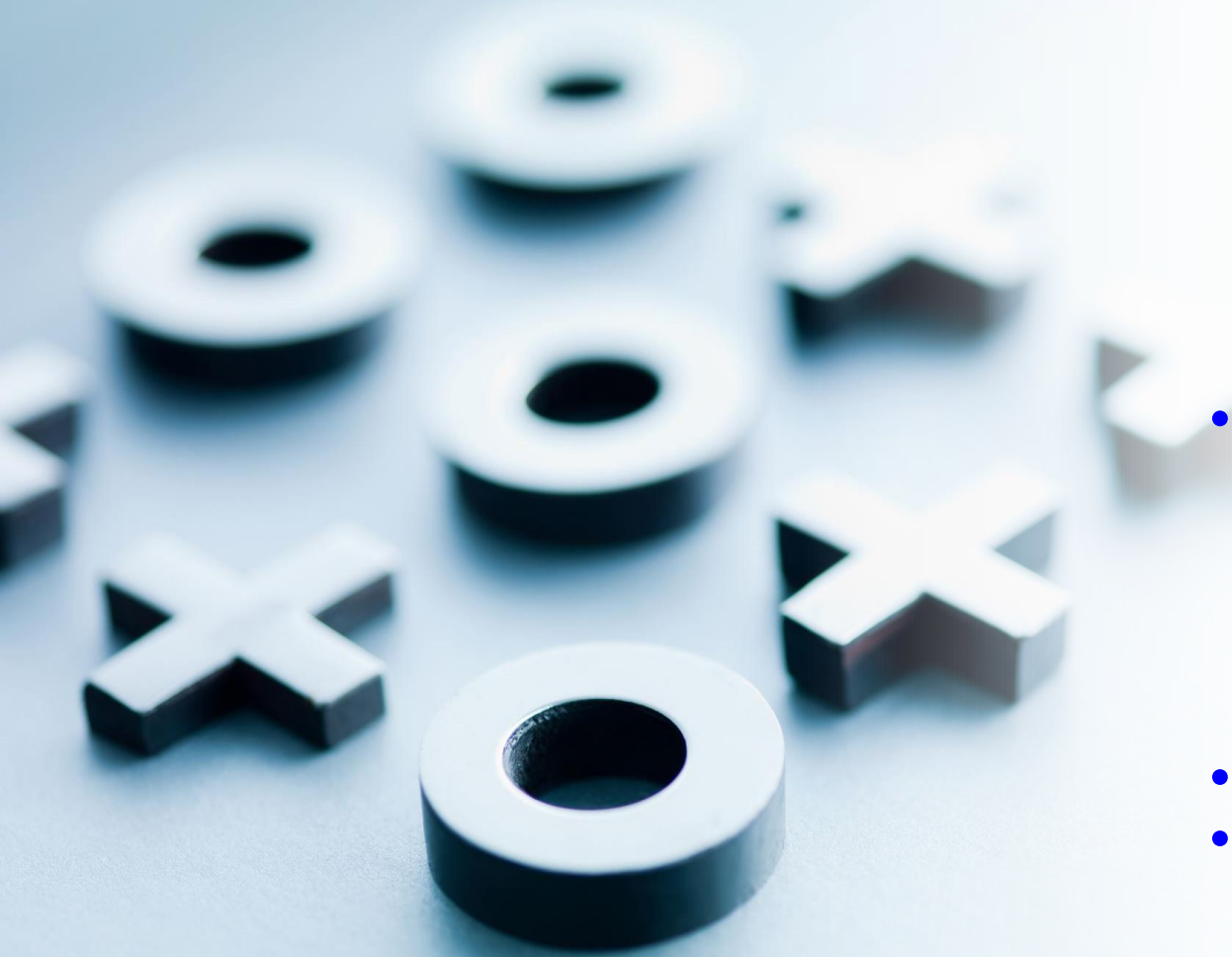
Workflow, Architecture, Technique

Pre-Processing

```
what is on the wall,spot_light,image1070
what is above the cupboards,books,image110
what is the object left of the shelf,soft_toy,image1001
what is above the drawer,mirror,image1031
what is on the shelf,bag,image612
what is in front of the bed,night_stand,image1033
what is to the left of the sofa,lamp,image488
what is around the table,chair,image467
```

```
what is on the right side of the notebook on the desk in the image4 ?
plastic_cup_of_coffee
what is on the right and left and in front of the papers on the desk in the image4 ?
notebook
what is on the desk and behind the black cup in the image4 ?
bottle
how many bottles are on the desk in the image4 ?
11
what is in front of the papers and notebook and bottles in the image4 ?
chair
what is on the left side of the cabinet and on the right side of the chair in the image5 ?
whiteboard
what is in front of the door and on the right of the table in the image5 ?
chair
how many chairs are on the right side of the table in the image5 ?
3
```

- We made sure: size -> (560,425)
- Question -> lowercase, no special character
- Prepared an all answer's space (size-> 582)
- Bifurcated into train: test=80:20



Evaluation Metric

- WUPS Score (Wu-Palmer similarity score)
  - measure the similarity between two words based on their distance in a semantic tree
- Accuracy
- F1_score

Results and Analysis

Text Transformer	Image Transformer	Accuracy	F1	WUPS	No of Trainable parameter	Total Training Time
Bert	Vit (google)	0.230	0.0189	0.281	~197 million	53 minutes
Bert	DeiT (Facebook)	0.244	0.0248	0.295	~197 million	55 minutes
Bert	Beit (microsoft)	0.206	0.0154	**0.297**	~196.3 million	52 minutes
RoBERTa (Robust)	Vit (google)	0.246	**0.0259**	0.289	~212 million	54 minutes
RoBERTa	DeiT (Facebook)	**0.246**	0.0281	0.295	~212 million	65 minutes
RoBERTa	Beit (microsoft)	0.227	0.0190	0.276	~211.4 million	47 minutes
ALBERT (Lite)	Vit (google)	0.191	0.0136	0.246	~99 million	47 minutes
ALBERT	DeiT (Facebook)	0.166	0.0162	0.221	~99 million	51 minutes
ALBERT	Beit (microsoft)	0.113	0.0059	0.168	~98.4 million	51 minutes

Hyperparameters: #Epochs=5,Batch=32,Learning Rate=5e-5,Seed=12345

Demo