REPORT

- For this Machine Learning Project 4, we have downloaded the VQA dataset available at https://visualqa.org/vqa_v1_download.html as per the Assignment instructions.

- Please note that due to computational and run time limitations, we have used 1000 images for training the model and 100 images for validating the model.

- Further, we have used Google Colaboratory environment to develop the Visual Question Answering Model.

We have done data analysis, extracting the data, and converting it into csv prior to building the model - please refer to below screenshots

```
[154] # Here we are merging both the questions and annotations dataframes on the image_id and question_id columns
      training_data = pd.merge(training_questions_df,
                               training_annotations_df,
                               how="inner",
                               left_on=["image_id", "question_id"],
                               right_on=["image_id", "question_id"])
```

```
[155] validation_data = pd.merge(validation_questions_df,
                                 validation_annotations_df,
                                 how="inner",
                                 left_on=["image_id", "question_id"],
                                 right_on=["image_id", "question_id"])
```

```
[156] # A peek at the merged Dataset
      training_data.head()
```

	question	image_id	question_id	question_type	multiple_choice_answer	answers	answer_type
0	What shape is the bench seat?	487025	4870250	what	curved	[{'answer': 'oval', 'answer_confidence': 'yes'...	other
1	Is there a shadow?	487025	4870251	is there a	yes	[{'answer': 'yes', 'answer_confidence': 'yes',...	yes/no
2	Is this one bench or multiple benches?	487025	4870252	is this	1	[{'answer': '1', 'answer_confidence': 'yes', '...	other
3	Is this a modern train?	78077	780770	is this a	no	[{'answer': 'no', 'answer_confidence': 'yes', ...	yes/no
4	What color is the stripe on the train?	78077	780771	what color is the	white	[{'answer': 'white', 'answer_confidence': 'yes...	other

```
[157] # A peek at the merged Dataset
      validation_data.head()
```

	question	image_id	question_id	question_type	multiple_choice_answer	answers	answer_type
0	What is the table made of?	350623	3506232	what is the	wood	[{'answer': 'wood', 'answer_confidence': 'yes'...	other
1	Is the food napping on the table?	350623	3506230	is the	no	[{'answer': 'no', 'answer_confidence': 'yes', ...	yes/no
2	What has been upcycled to make lights?	350623	3506231	what	kettles	[{'answer': 'kettles', 'answer_confidence': 'y...	other
3	Is this an Spanish town?	8647	86472	is this an	no	[{'answer': 'yes', 'answer_confidence': 'maybe...	yes/no
4	Are there shadows on the sidewalk?	8647	86470	are there	yes	[{'answer': 'yes', 'answer_confidence': 'yes',...	yes/no

Finally after building training data and validation data by merging questions and annotations for both validation and training data. We also did Subset Questions and Annotations for the 1000 training Images and 100 validation images

```
[165] train_df.to_csv("/content/drive/MyDrive/VisualQuestionAnswering/data/train_df.csv", index=False)
      val_df.to_csv("/content/drive/MyDrive/VisualQuestionAnswering/data/val_df.csv", index=False)
```

Below are the answers for the questions asked in the Assignment PDF:

1) *Indicate Where you Implemented each of the Component:*

a) Image Encoder:

Below image contains the code modules that encode the Image Data:

```
[178] def load_image(image_path):
          img = tf.io.read_file(image_path)
          img = tf.image.decode_jpeg(img, channels=3)
          img = tf.image.resize(img, (image_width, image_height))
          img = tf.keras.applications.vgg19.preprocess_input(img)
          img = img * (1./255)
          return img, image_path
```

```
[179] def VGG19_Top():
          model = tf.keras.applications.VGG19(include_top=False, weights="imagenet", input_shape=(image_width, image_height, 3))
          input_layer = model.input
          hidden_layer = model.layers[-1].output
          model = tf.keras.Model(input_layer, hidden_layer)
          return model
```

```
      def generateImageFeatures(images, filename):
          model = VGG19_Top()
          all_image_dict = {}
          img_ds = tf.data.Dataset.from_tensor_slices(images)
          img_ds = img_ds.map(load_image, num_parallel_calls=tf.data.experimental.AUTOTUNE).batch(16)

          for batch_img, batch_path in img_ds:
              batch_img_features = model(batch_img)

              for img_features, path in zip(batch_img_features, batch_path):
                  image_path = path.numpy().decode("utf-8")
                  #image_path = image_path.replace(imageDirectory,imageNumpyDirectory).replace('.jpg',"")
                  #np.save(image_path, img_features.numpy())
                  all_image_dict[image_path] = img_features.numpy()

          with open(data_directory + filename, "wb") as handle:
              pickle.dump(all_image_dict, handle, protocol=pickle.HIGHEST_PROTOCOL)
              print("Pickled Data")
          return
```

This is where the image is getting encoded:

```
[183] generateImageFeatures(train_all_image_path, "train_all_image_dict.pickle")
      generateImageFeatures(val_all_image_path, "val_all_image_dict.pickle")

      Pickled Data
      Pickled Data
```

b) Question Encoder:

For Encoding the question, we have tokenized it first, followed by padding sequences operation from the tensorflow module. Further, we have used Embedding Layer with the LSTM Module:

```
[187] #tokenization
      tokenizer = tf.keras.preprocessing.text.Tokenizer(oov_token = "", filters = '!"#$%&()*+.,-/:;=?@[\]^_`{|}~ ')
      tokenizer.fit_on_texts(X_train['question'].values)
      train_question_seqs = tokenizer.texts_to_sequences(X_train['question'].values)
      val_question_seqs = tokenizer.texts_to_sequences(X_val['question'].values)

      print("Number of words in tokenizer:", len(tokenizer.word_index))
      ques_vocab = tokenizer.word_index

      #Padding
      #tokenizer.word_index[''] = 0
      #tokenizer.index_word[0] = ''
      question_vector_train = tf.keras.preprocessing.sequence.pad_sequences(train_question_seqs, padding='post')
      question_vector_val = tf.keras.preprocessing.sequence.pad_sequences(val_question_seqs,padding='post',maxlen=question_vector_train.shape[1])

      Number of words in tokenizer: 1990
```

```
question_emb = tf.keras.layers.Embedding(input_dim = len(tokenizer.word_index) + 1, output_dim=300, name = "Embedding_Layer",
                                         embeddings_initializer = tf.keras.initializers.RandomNormal(mean=0, stddev=1, seed=23))(question_input)

question_lstm = tf.keras.layers.LSTM(1024,
                                     kernel_initializer = tf.keras.initializers.glorot_uniform(seed=26),
                                     recurrent_initializer = tf.keras.initializers.orthogonal(seed=54),
                                     bias_initializer=tf.keras.initializers.zeros())(question_emb)

question_flatten = tf.keras.layers.Flatten(name="Flatten_lstm")(question_lstm)


image_question = tf.keras.layers.Multiply()([image_dense_2, question_flatten])
```

c) Fusion Encoder:

We have used keras.layers.Multiply to fuse both the question and image data. Below piece of code contains this logic:

```
image_question = tf.keras.layers.Multiply()([image_dense_2, question_flatten])


image_question_dense_1 = tf.keras.layers.Dense(1000, activation = tf.nn.relu,
                                               kernel_initializer = tf.keras.initializers.he_uniform(seed=19))(image_question)

image_question_dense_2 = tf.keras.layers.Dense(1000, activation = tf.nn.relu,
                                              kernel_initializer = tf.keras.initializers.he_uniform(seed=28))(image_question_dense_1)

output = tf.keras.layers.Dense(len(ans_vocab), activation=tf.nn.softmax,
                               kernel_initializer = tf.keras.initializers.glorot_normal(seed=15))(image_question_dense_2)

# Create Model
model = tf.keras.models.Model(inputs = [image_input, question_input], outputs = output)

# Compile
model.compile(optimizer="adam", loss="categorical_crossentropy", metrics=["accuracy"])
return model
```

Please Note that all the above pieces of code can be found in the “Visual Question Answering Model.ipynb” code file.

Further, we have followed the below folder hierarchy for this project:

- /content/drive/MyDrive/VisualQuestionAnswering : is the Main Project Folder
- /content/drive/MyDrive/VisualQuestionAnswering/data : Contains the data files such as Questions and Annotations Data for both training and validation data
- /content/drive/MyDrive/VisualQuestionAnswering/data/train2014 : Contains training Images
- /content/drive/MyDrive/VisualQuestionAnswering/data/val2014 : Contains validation images
- /content/drive/MyDrive/VisualQuestionAnswering/Model : Contains the Model.

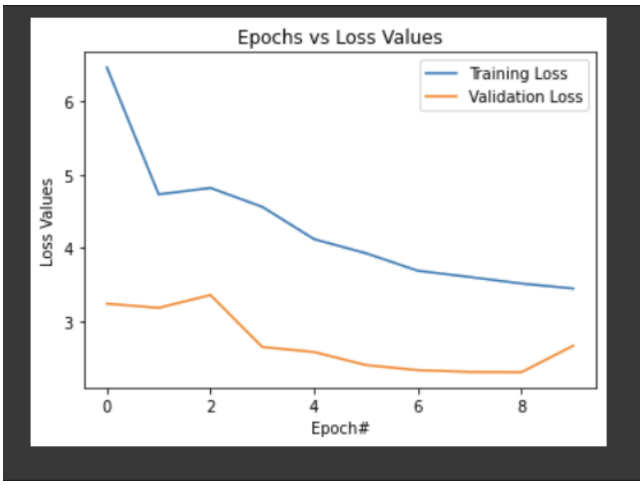
The main project “/content/drive/MyDrive/VisualQuestionAnswering” folder contains the below IPython Notebook: Visual Question Answering Model.ipynb

2) **Validation Accuracy on the Validation Set**: 26.33%

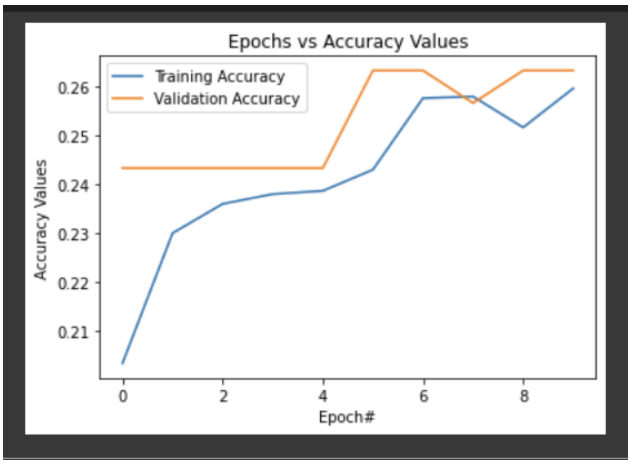
Please note that we have used only 1000 images and about 3000 questions to train the model and further we used 100 images and 300 questions to validate the model across 10 epochs with a learning rate of 0.01. Hence the low accuracy value.

Further we have trained the model with all possible classes available.

Below are the training and validation set loss values across 10 epochs:



Below are the training and validation set accuracy values across 10 epochs:

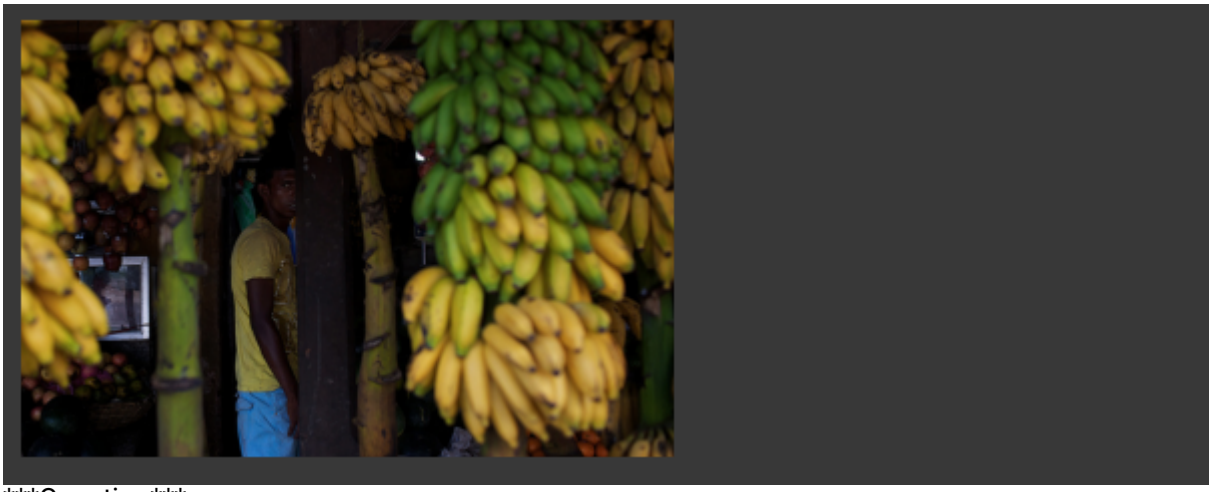


Except for a dip in accuracy at the 7th epoch, most of the time, accuracy was found to be increasing.

3) As requested we have randomly sampled 30 image-question pairs from the validation dataset and plotted the Image, and Question Attention Maps as well. Below are a couple of samples for your reference:

<Start ——>

Image Index: 242
Plotting the Image



Question

What color is the man's shirt?

1/1 [==============================] - 0s 25ms/step

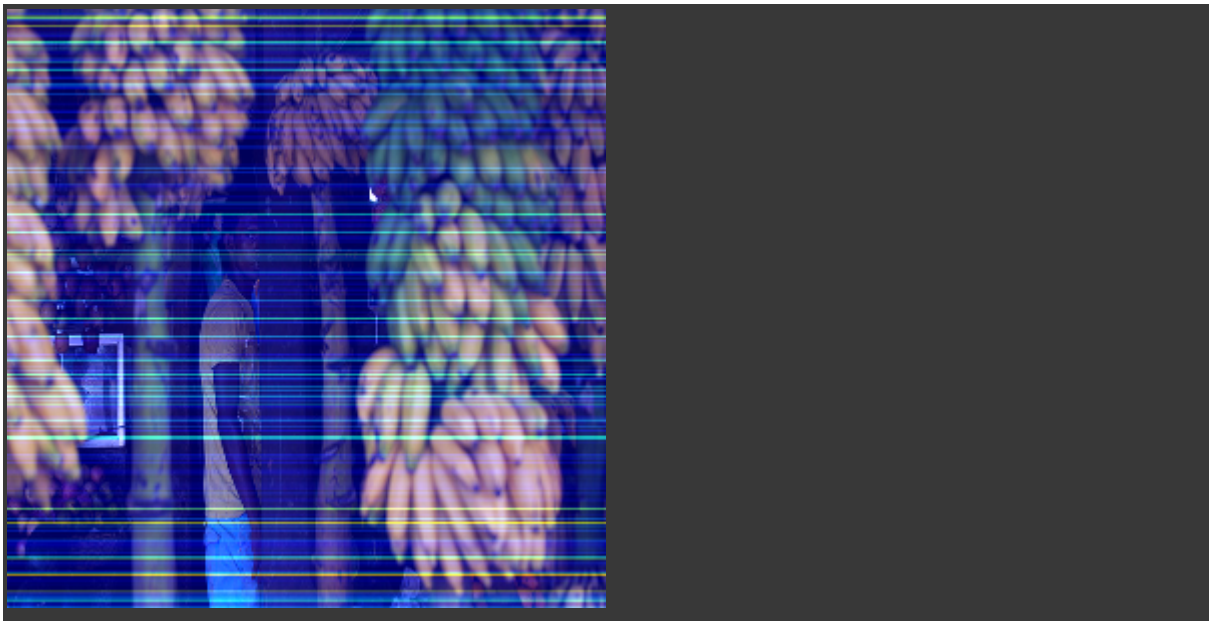
Ground Truth Answer: yellow

Predicted Answer: 2

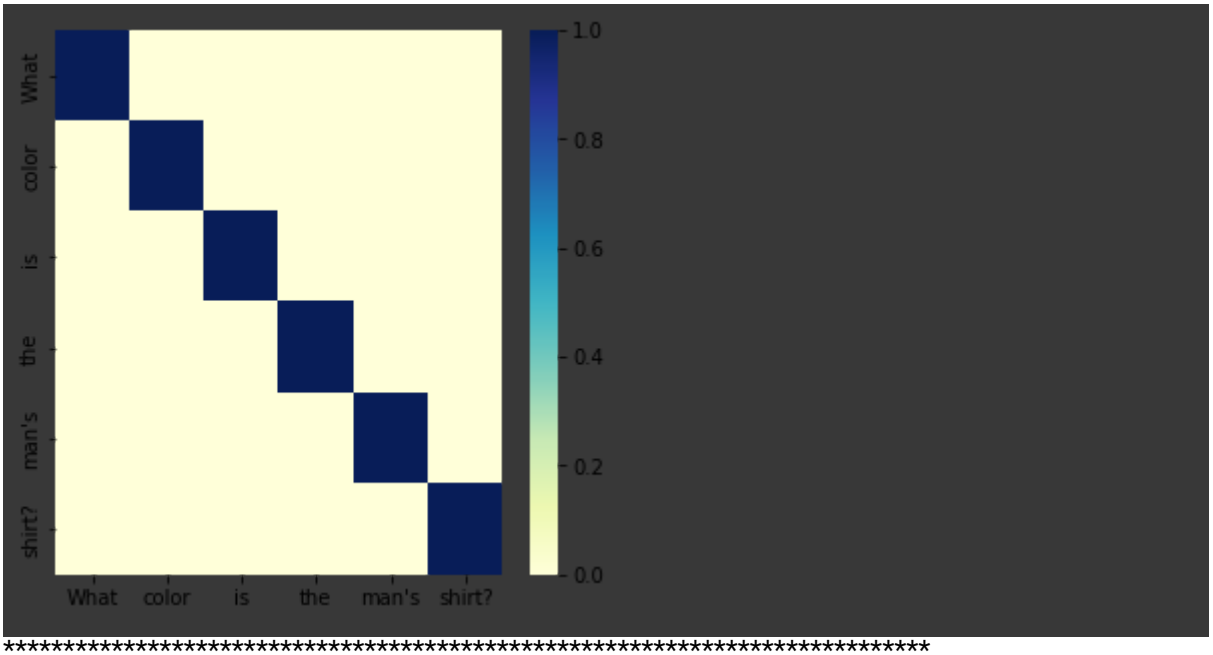
***Confidence with which the Prediction is made**: 19.45518

Image HeatMap

1/1 [==============================] - 0s 31ms/step

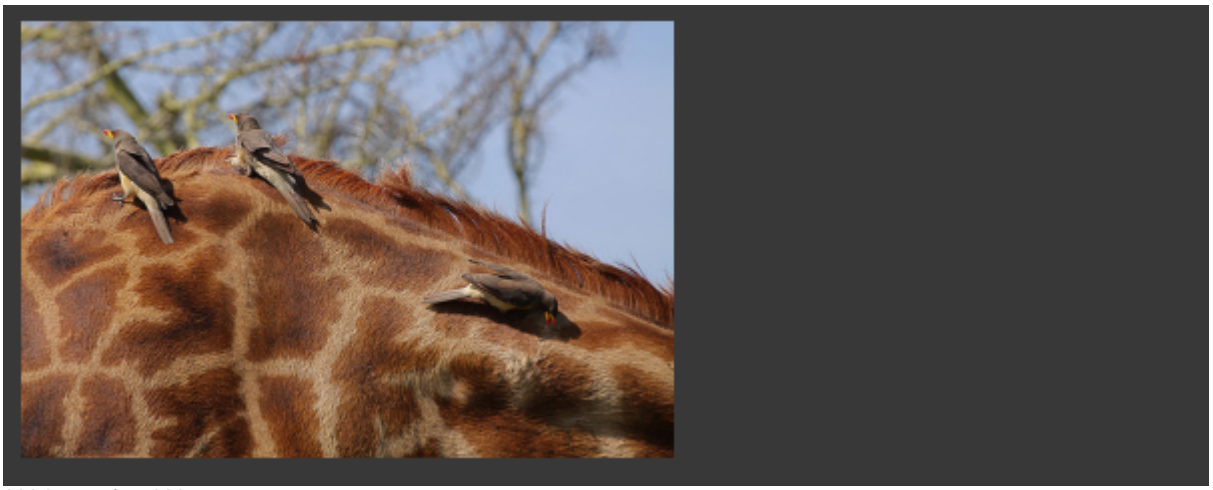


Attention Correlation Visualization

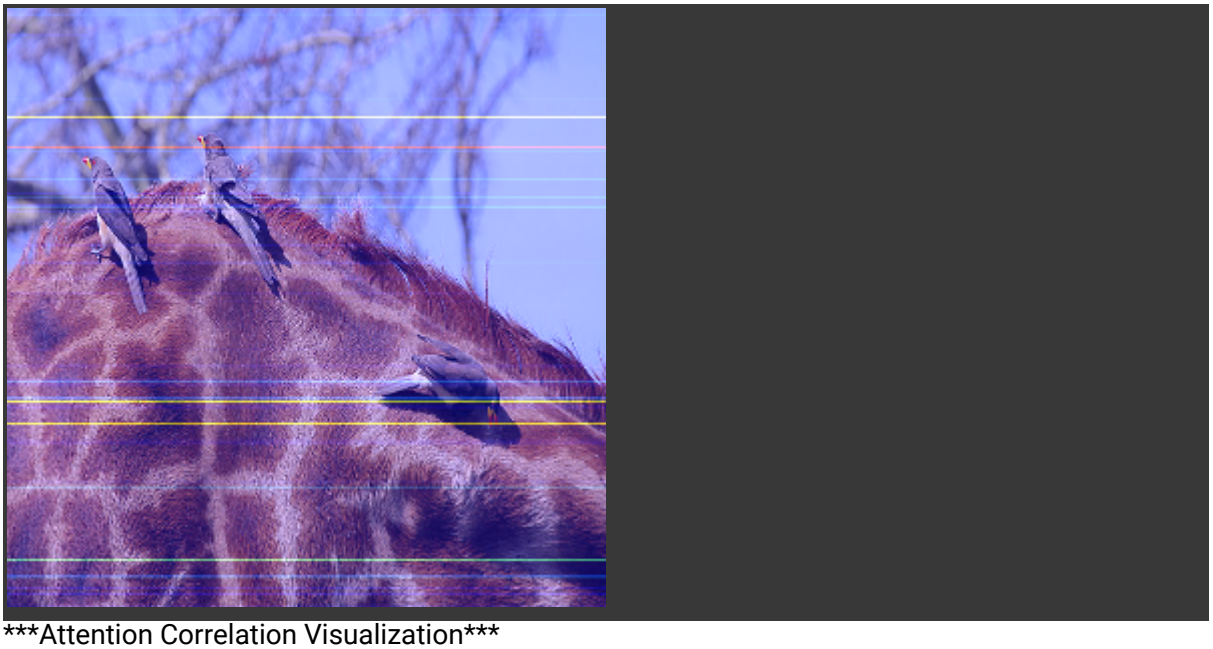


****

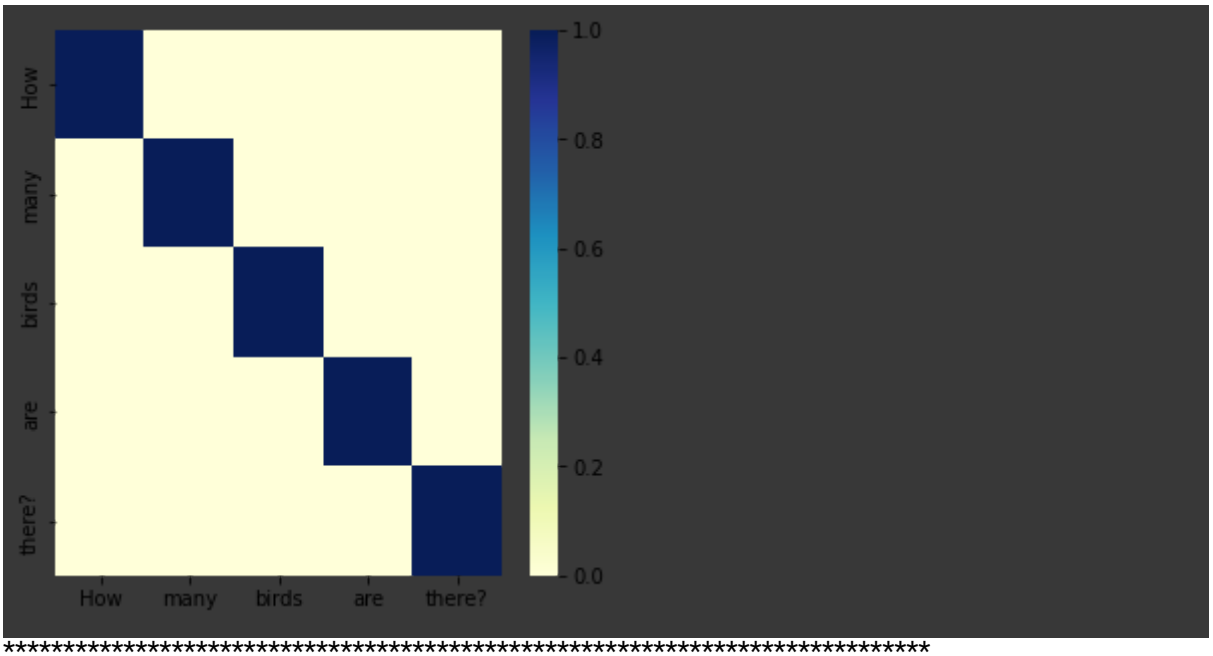
Image Index: 137
Plotting the Image



Question
How many birds are there?
1/1 [==============================] - 0s 21ms/step
Ground Truth Answer: 3
Predicted Answer: 2
***Confidence with which the Prediction is made**: 31.295773
Image HeatMap
1/1 [==============================] - 0s 30ms/step



Attention Correlation Visualization



****

Image Index: 270
Plotting the Image

Question

What is the kid petting?

1/1 [==============================] - 0s 21ms/step

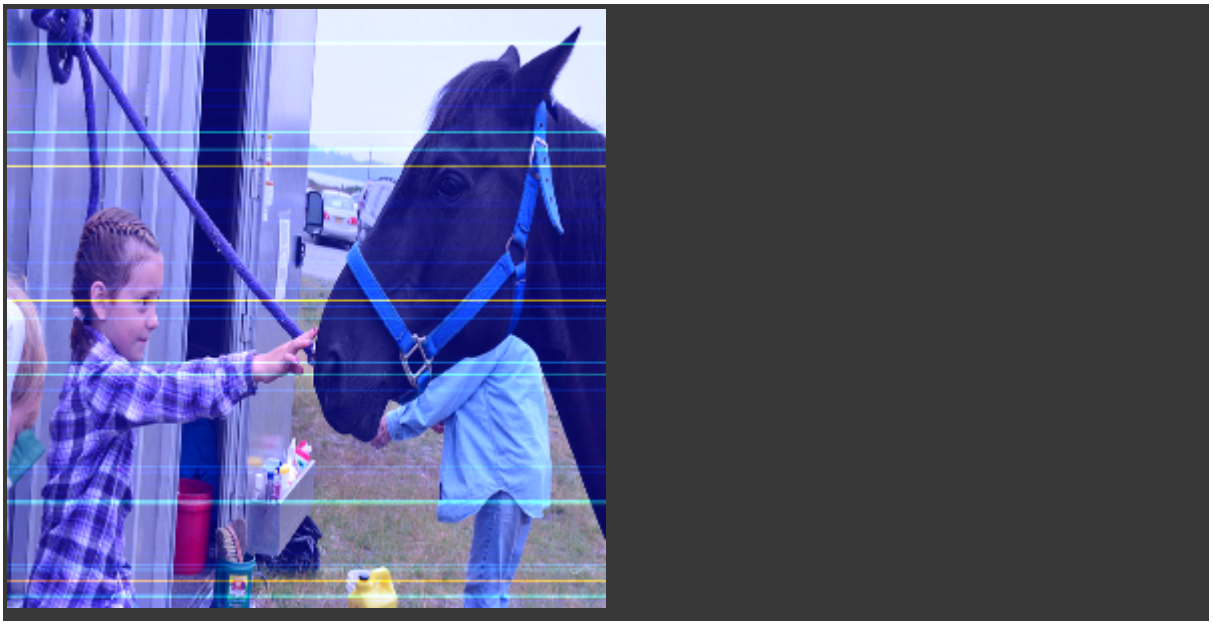
Ground Truth Answer: horse

Predicted Answer: no

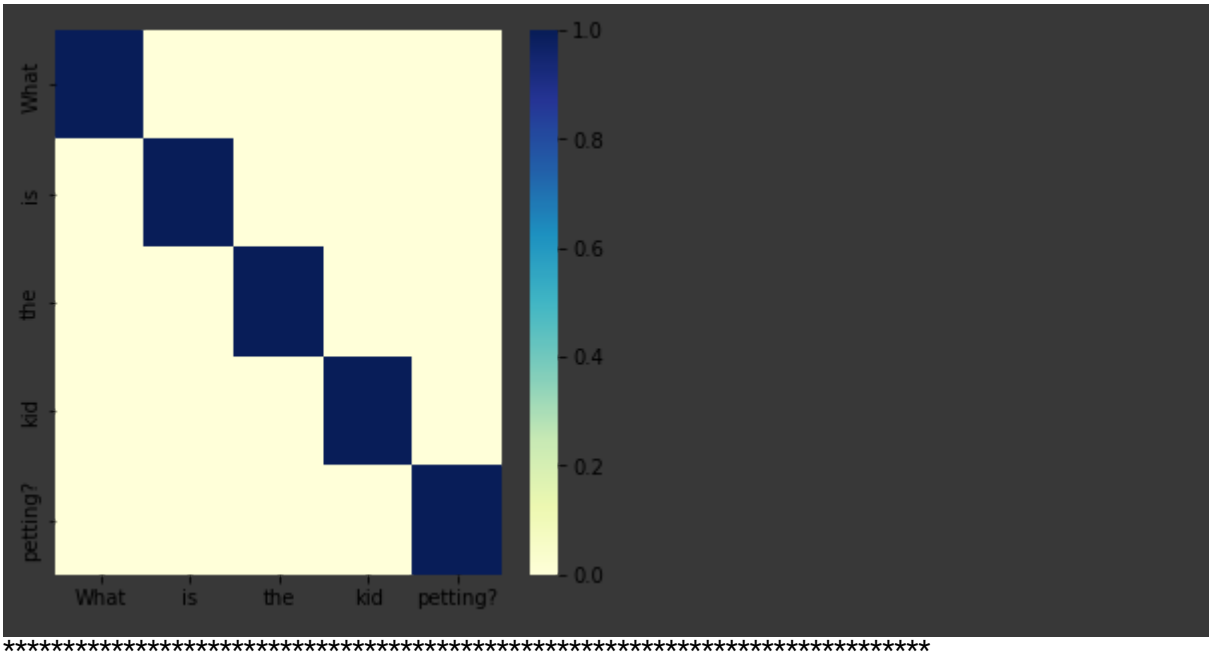
***Confidence with which the Prediction is made**: 4.093178

Image HeatMap

1/1 [==============================] - 0s 31ms/step



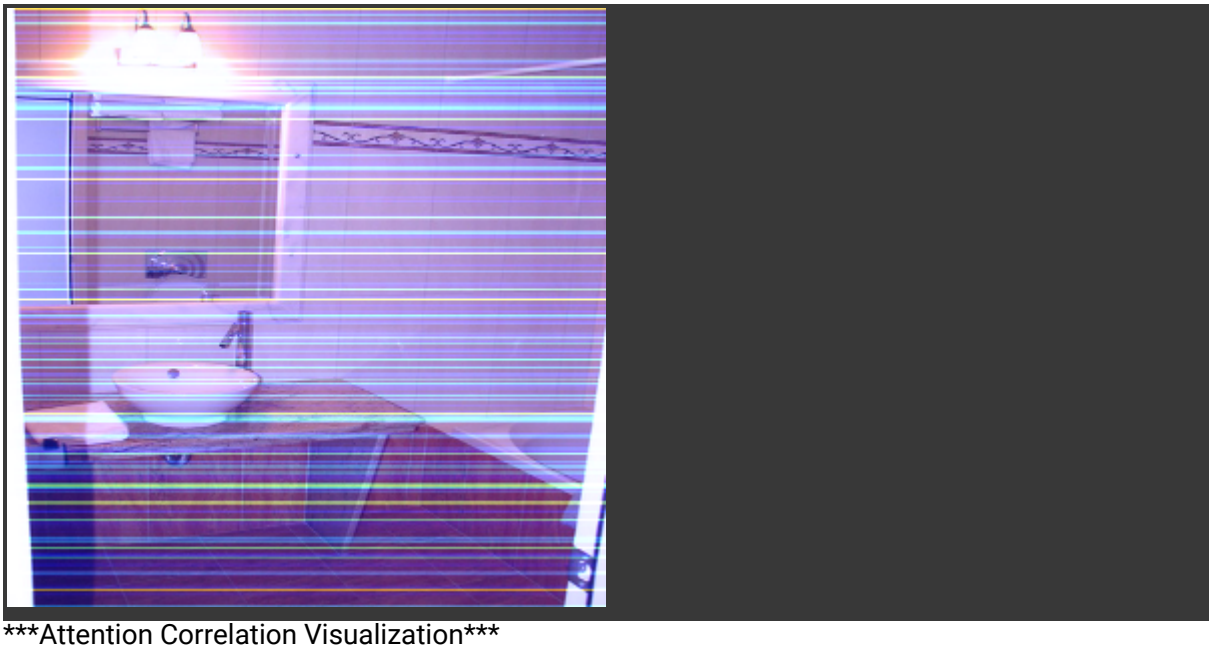
Attention Correlation Visualization



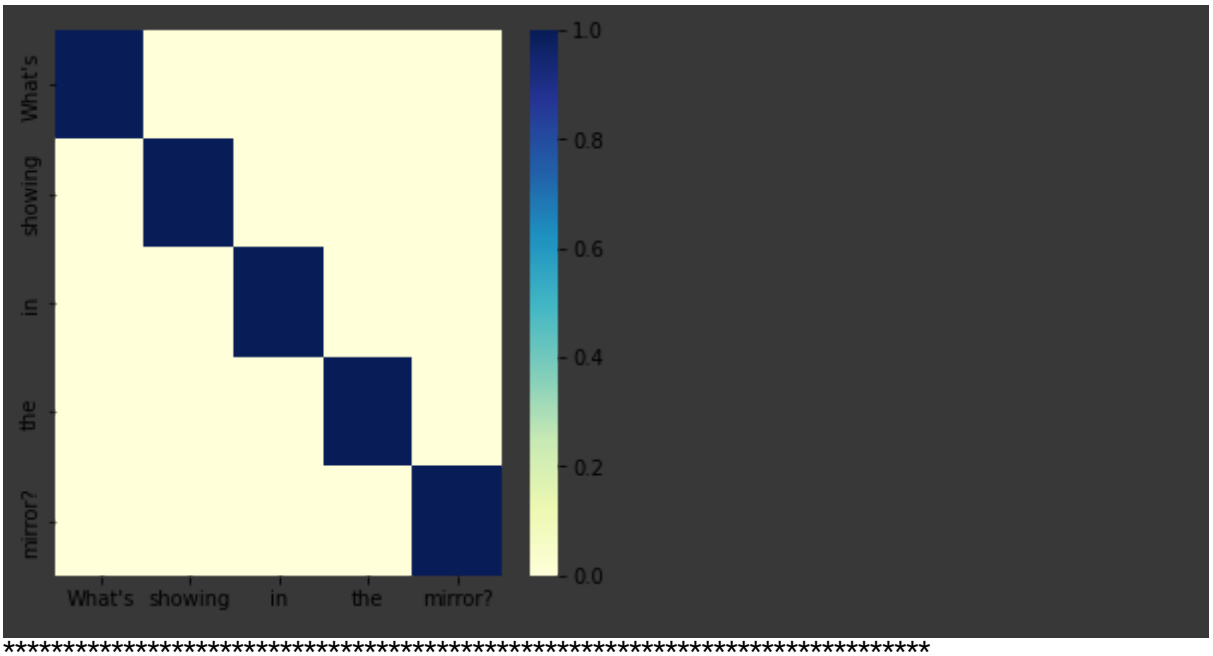
****

Image Index: 179
Plotting the Image

Question
What's showing in the mirror?
1/1 [==============================] - 0s 21ms/step
Ground Truth Answer: shower
Predicted Answer: no
***Confidence with which the Prediction is made**: 7.6544986
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 73
Plotting the Image

Question

What kind of building is behind the bus?

1/1 [==============================] - 0s 22ms/step

Ground Truth Answer: apartment

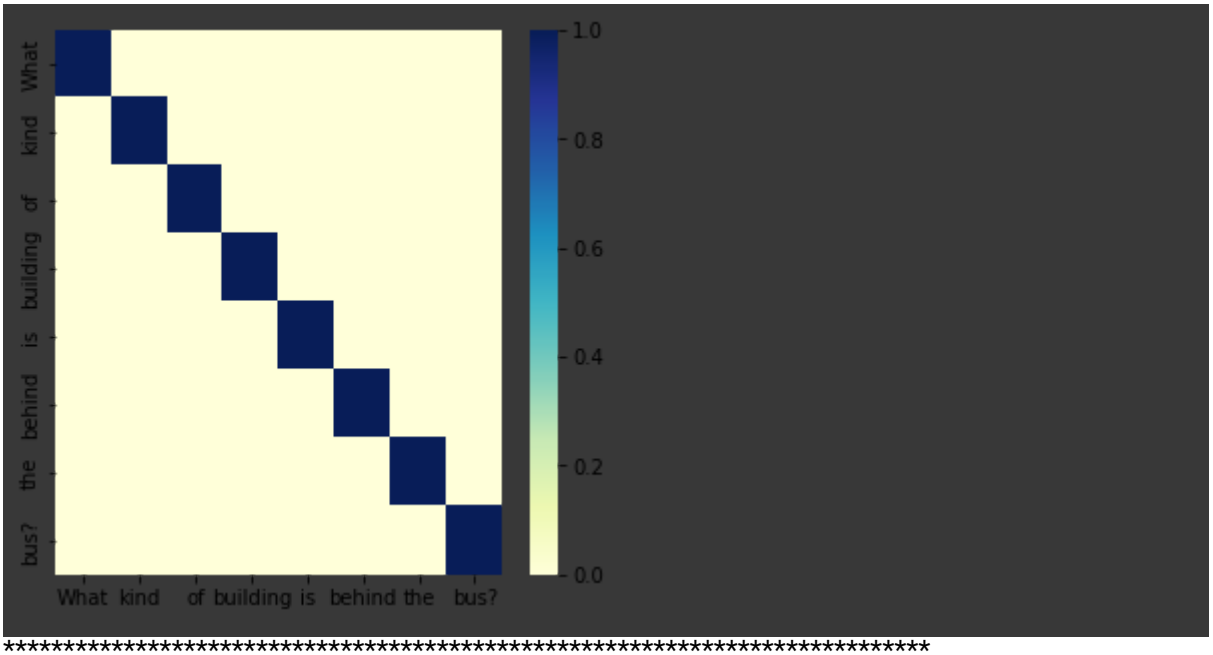
Predicted Answer: no

***Confidence with which the Prediction is made**: 3.6150258

Image HeatMap

1/1 [==============================] - 0s 31ms/step

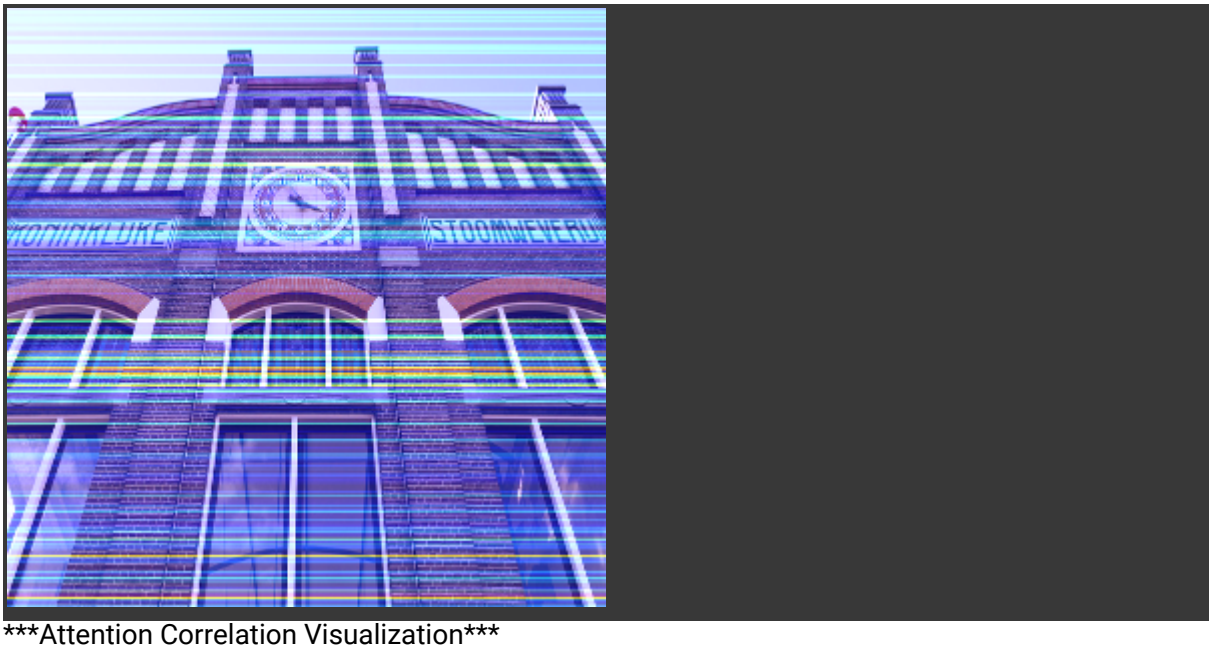
Attention Correlation Visualization



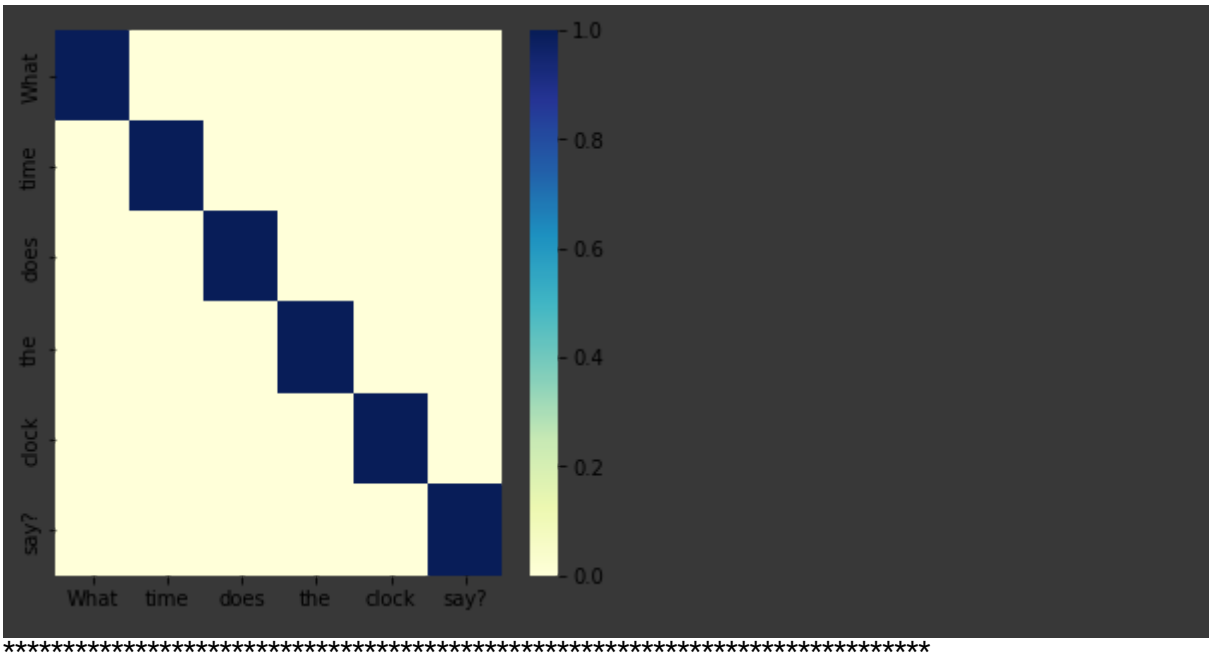
****

Image Index: 195
Plotting the Image

Question
What time does the clock say?
1/1 [==============================] - 0s 24ms/step
Ground Truth Answer: 4:20
Predicted Answer: no
***Confidence with which the Prediction is made**: 5.21289
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 5
Plotting the Image

Question

Is the man in motion?

1/1 [==============================] - 0s 24ms/step

Ground Truth Answer: yes

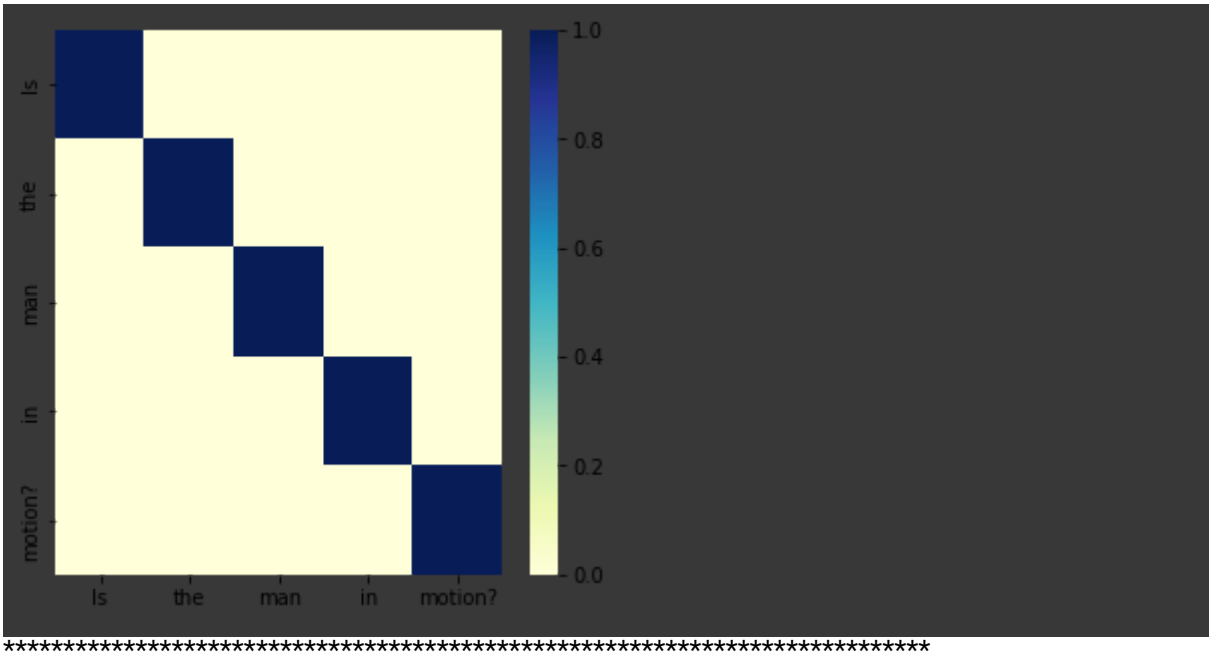
Predicted Answer: yes

***Confidence with which the Prediction is made**: 62.53392

Image HeatMap

1/1 [==============================] - 0s 30ms/step

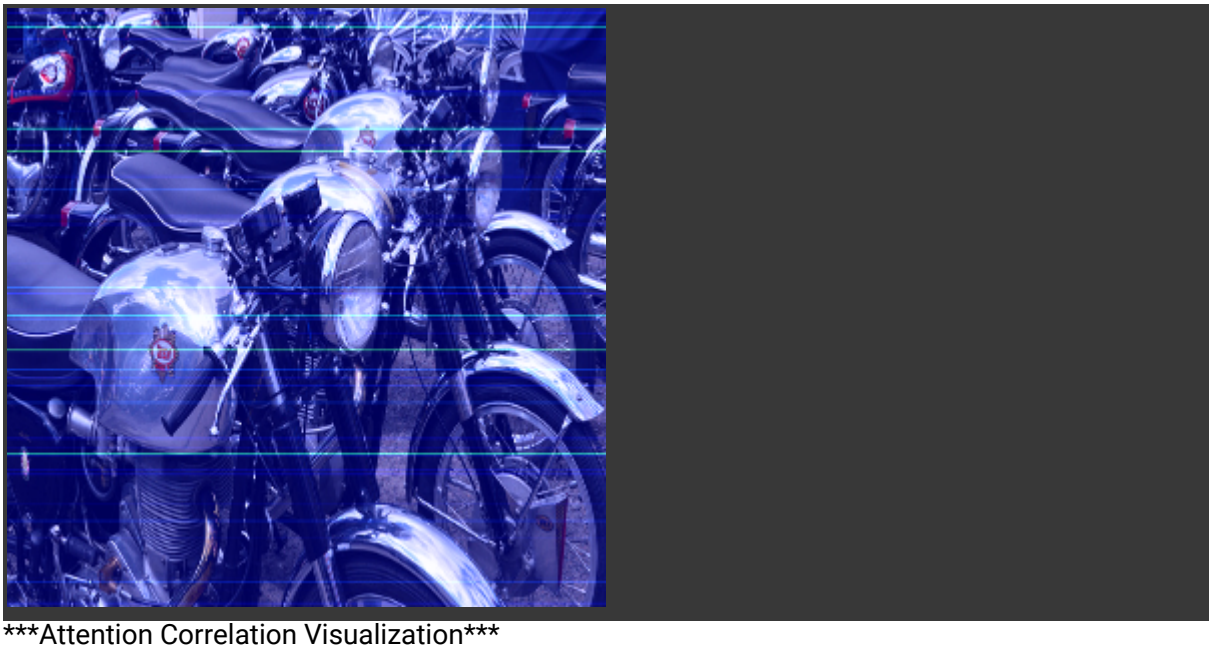
Attention Correlation Visualization



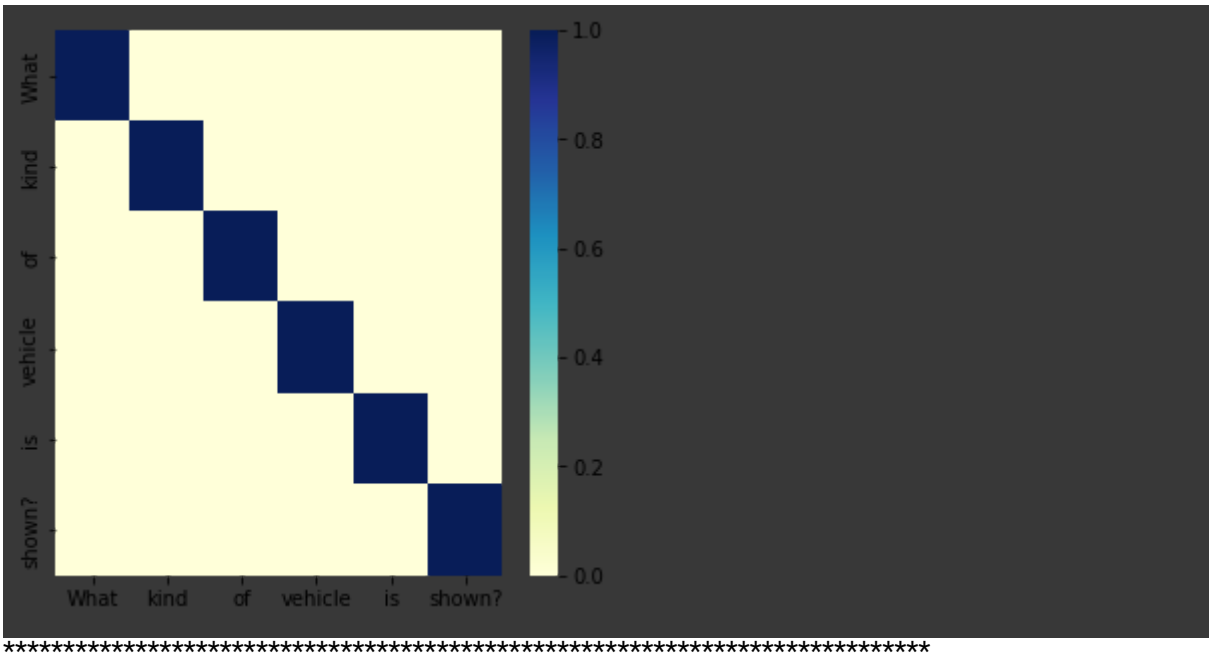
****

Image Index: 191
Plotting the Image

Question
What kind of vehicle is shown?
1/1 [==============================] - 0s 21ms/step
Ground Truth Answer: motorcycle
Predicted Answer: no
***Confidence with which the Prediction is made**: 3.789324
Image HeatMap
1/1 [==============================] - 0s 30ms/step



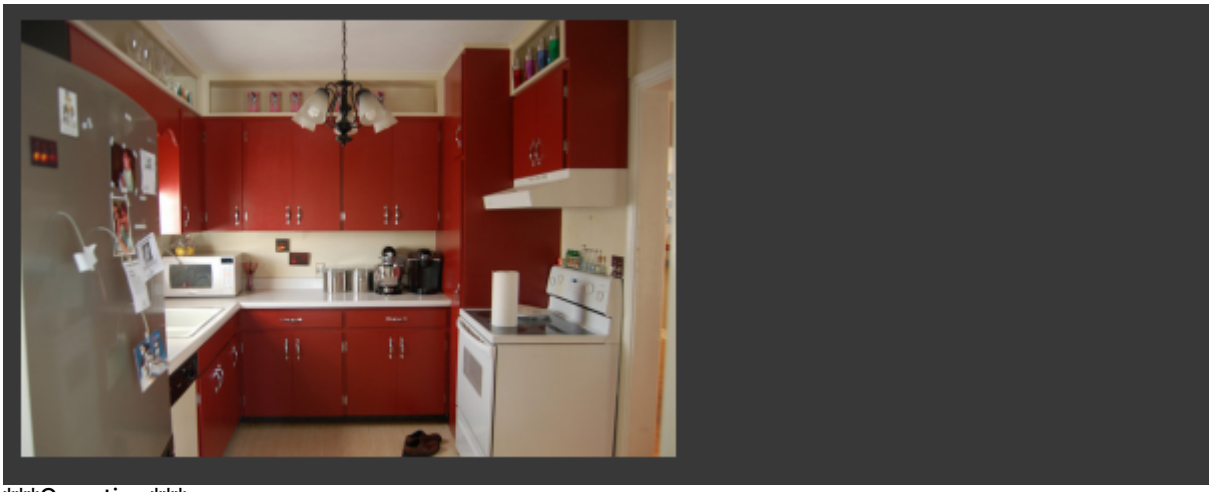
Attention Correlation Visualization



****

Image Index: 247

Plotting the Image



Question

What is hanging from the ceiling?

1/1 [==============================] - 0s 21ms/step

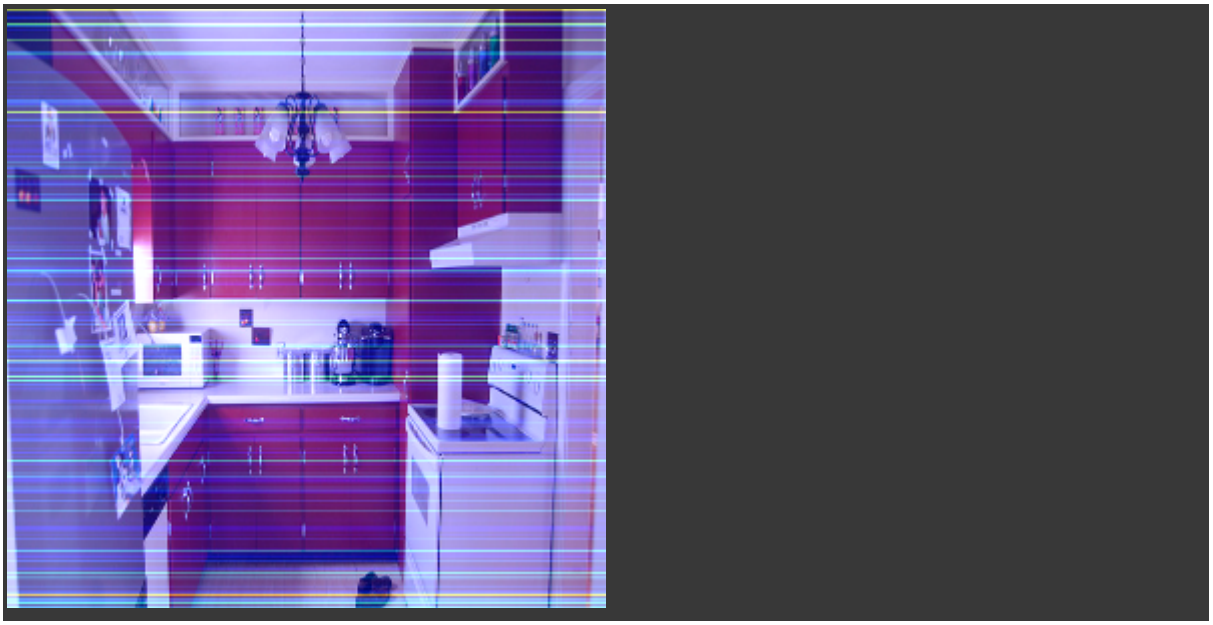
Ground Truth Answer: light fixture

Predicted Answer: no

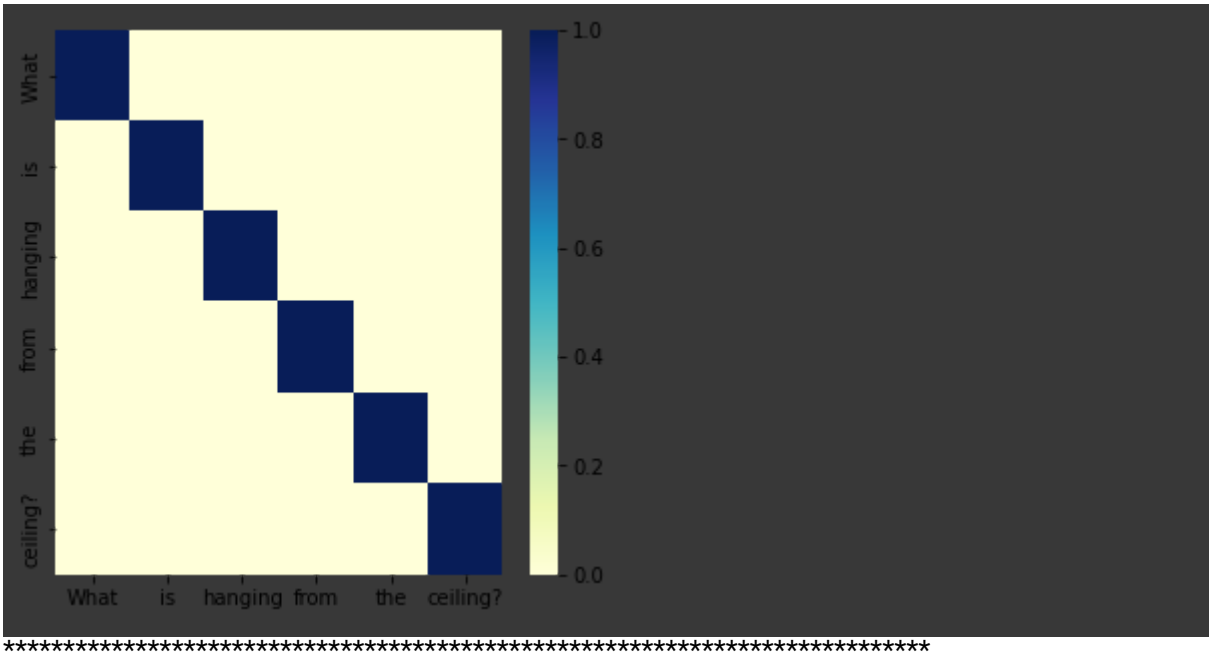
***Confidence with which the Prediction is made**: 4.4942675

Image HeatMap

1/1 [==============================] - 0s 30ms/step

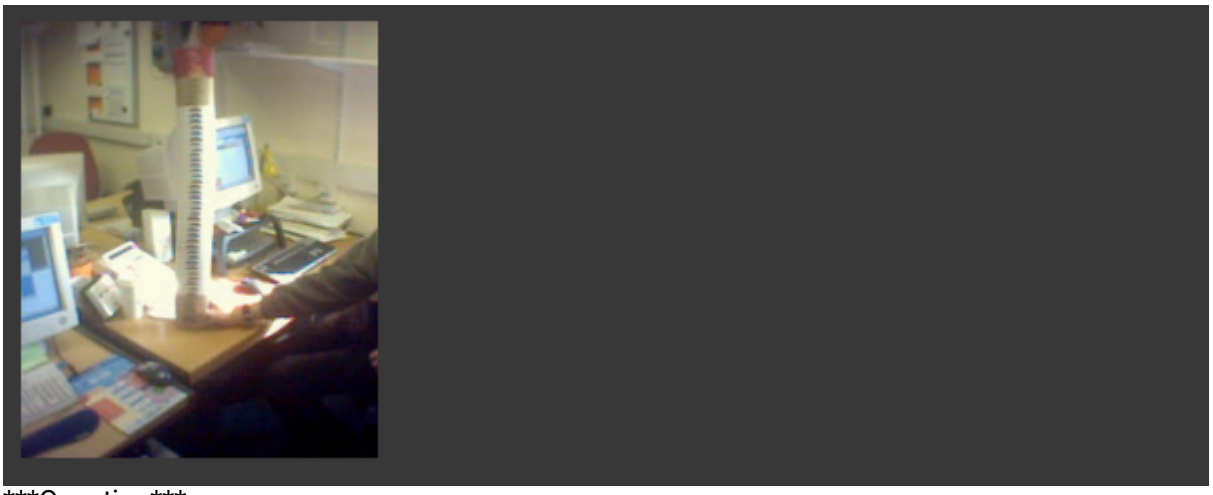


Attention Correlation Visualization

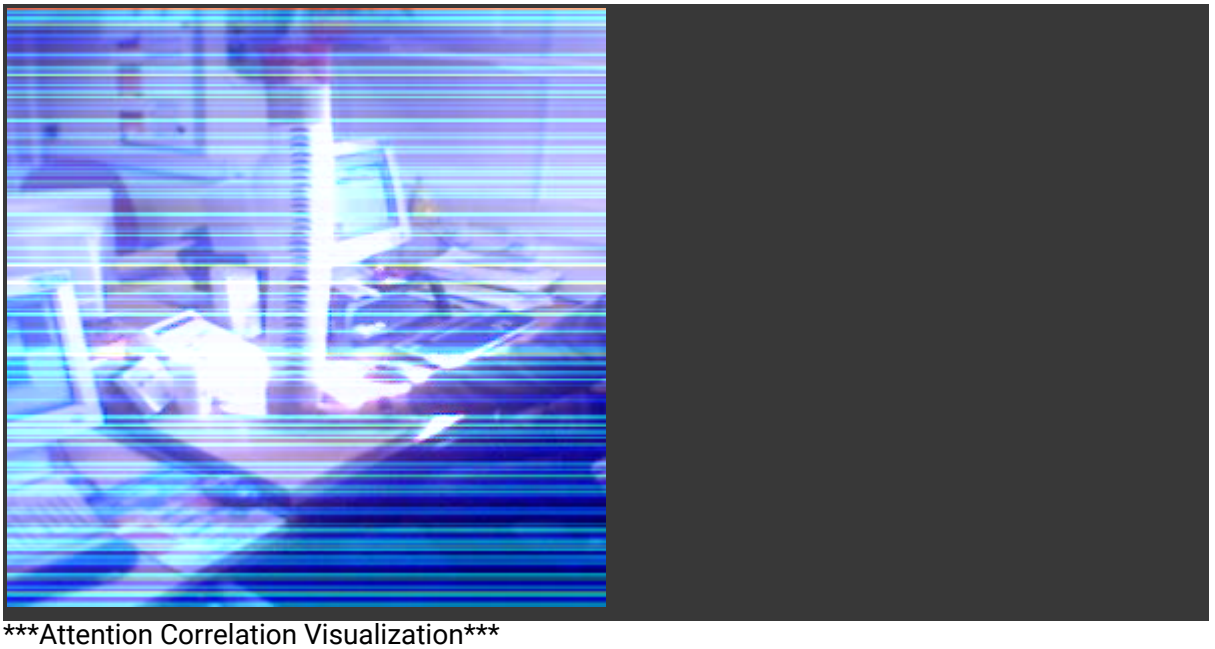


****

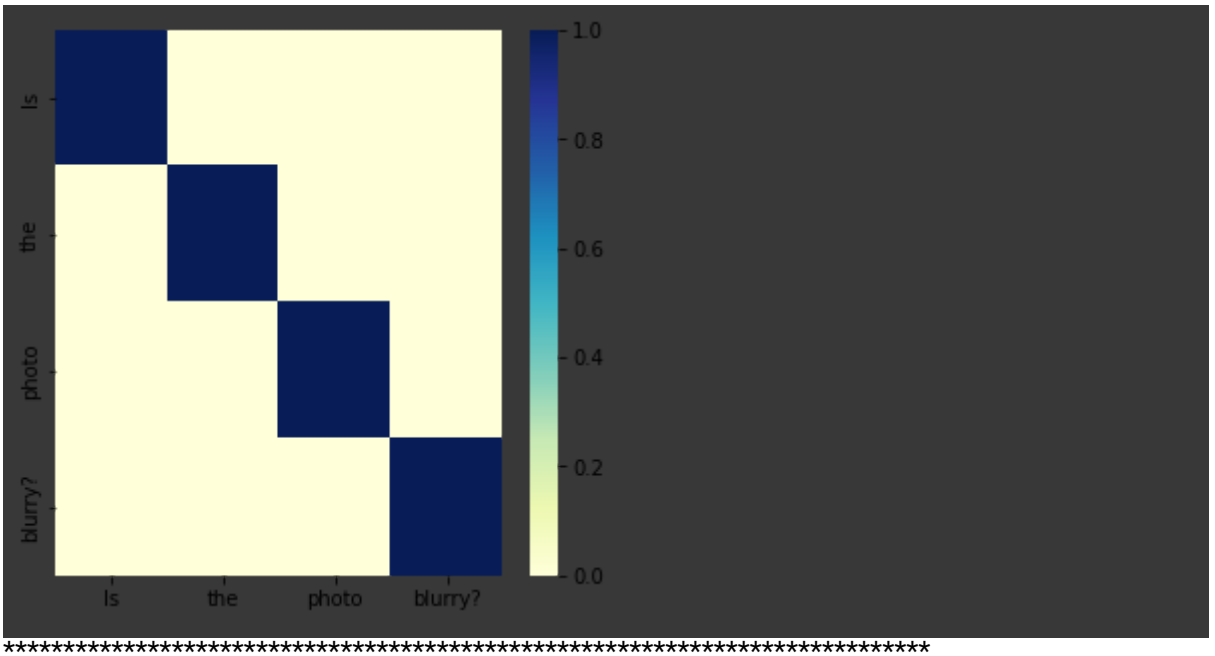
Image Index: 140
Plotting the Image



Question
Is the photo blurry?
1/1 [==============================] - 0s 21ms/step
Ground Truth Answer: yes
Predicted Answer: yes
***Confidence with which the Prediction is made**: 61.971634
Image HeatMap
1/1 [==============================] - 0s 31ms/step

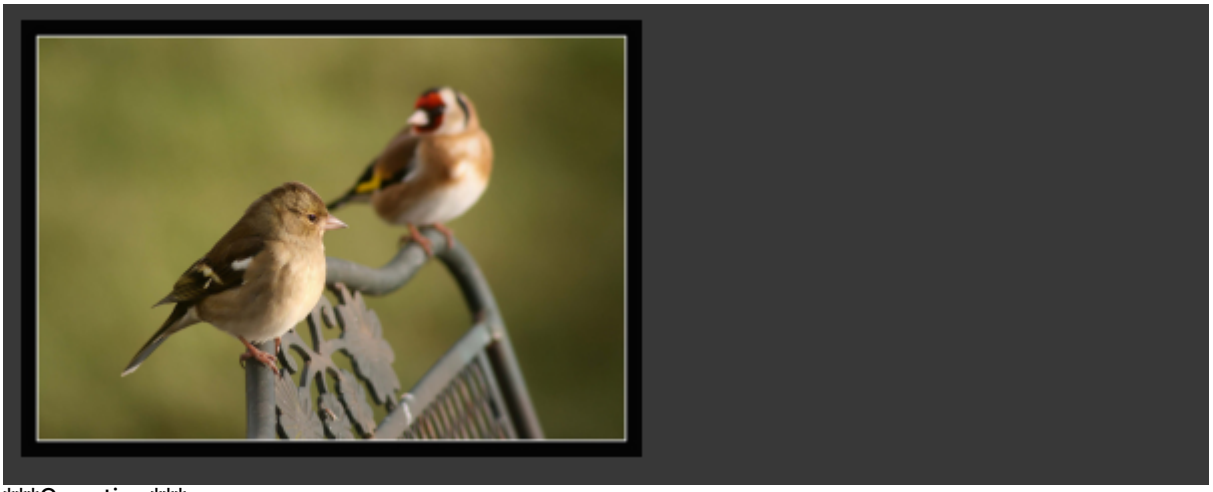


Attention Correlation Visualization



****

Image Index: 235
Plotting the Image



Question

Are these water birds?

1/1 [==============================] - 0s 23ms/step

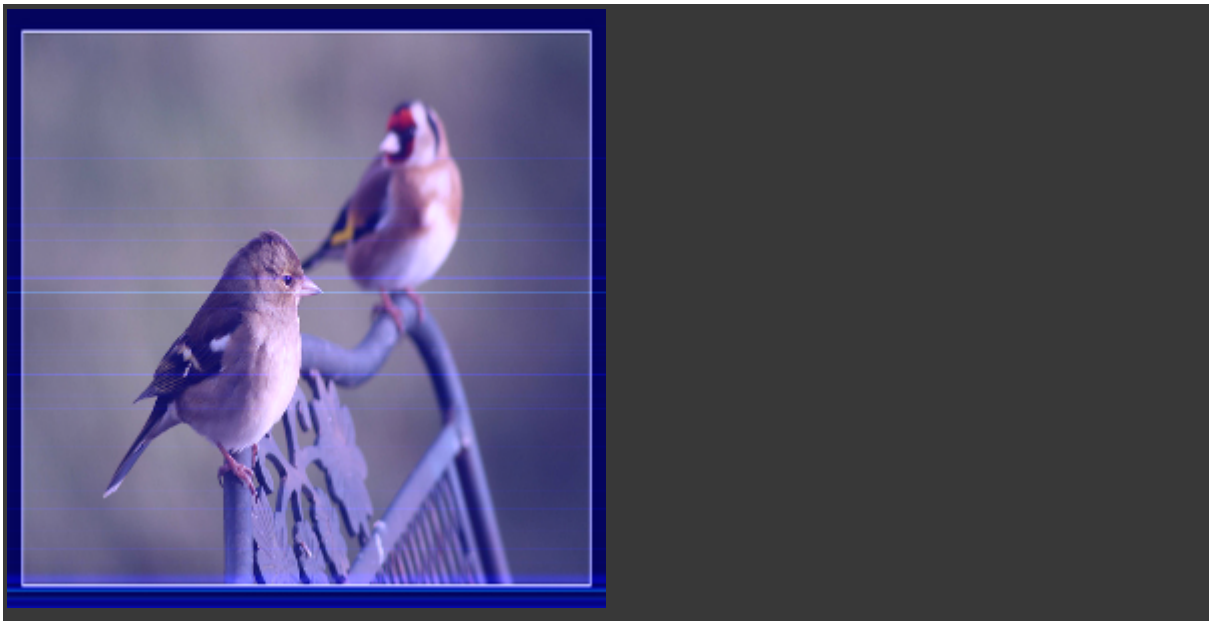
Ground Truth Answer: no

Predicted Answer: yes

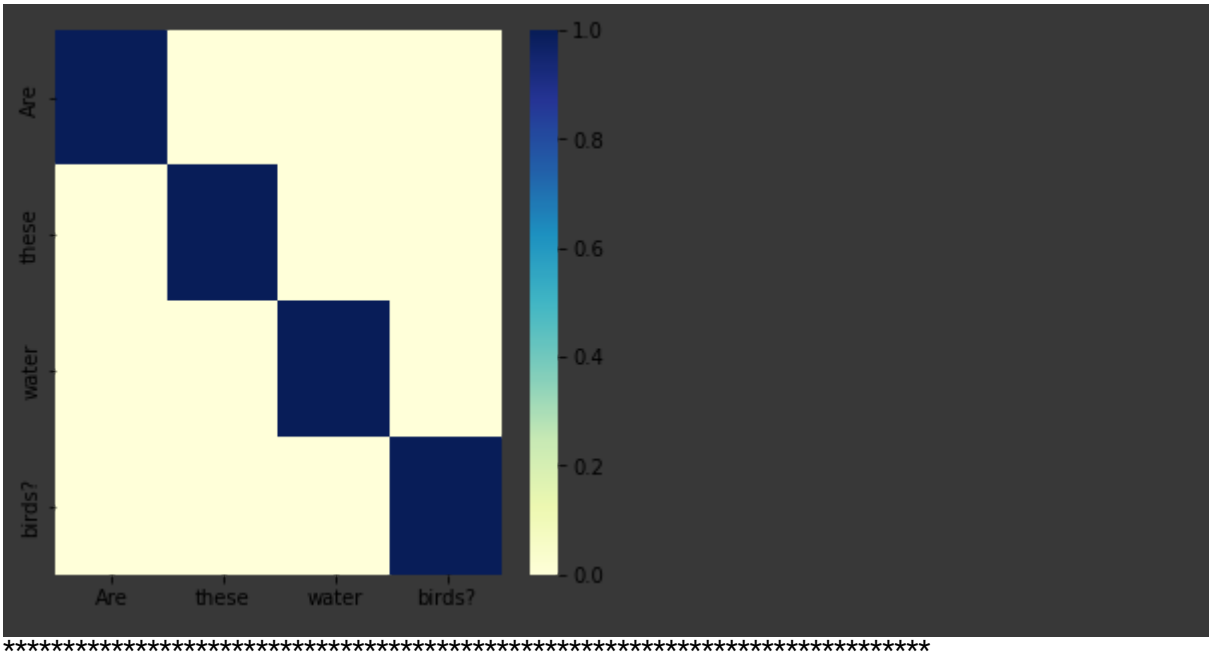
***Confidence with which the Prediction is made**: 67.0864

Image HeatMap

1/1 [==============================] - 0s 30ms/step



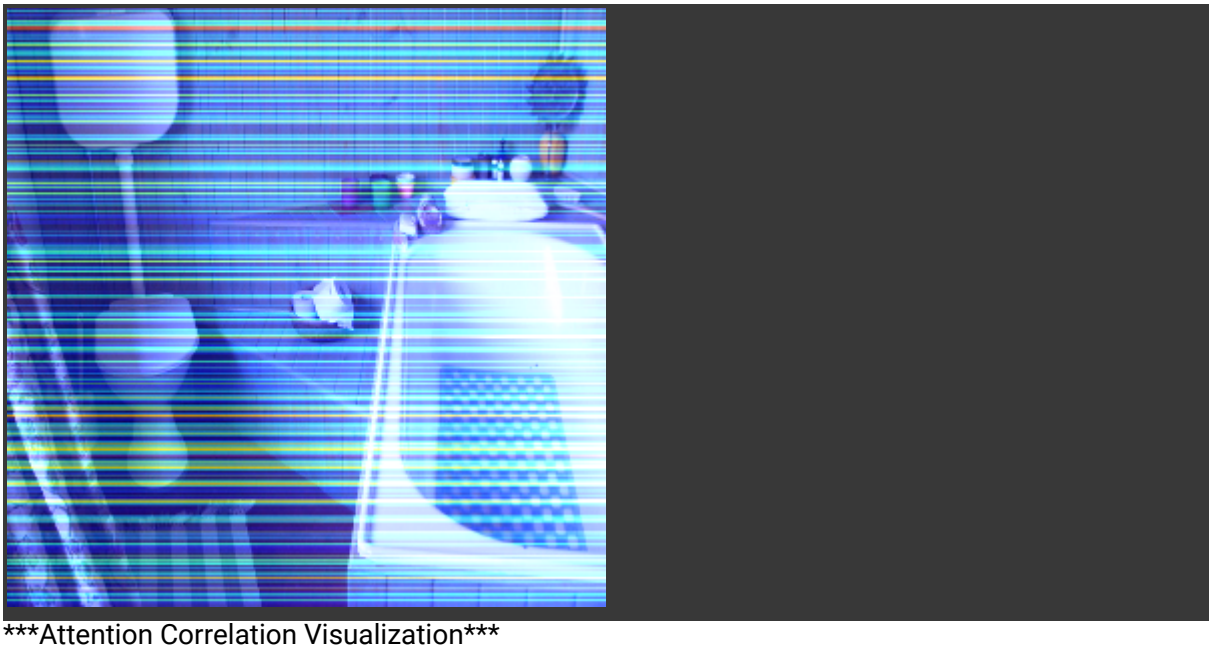
Attention Correlation Visualization



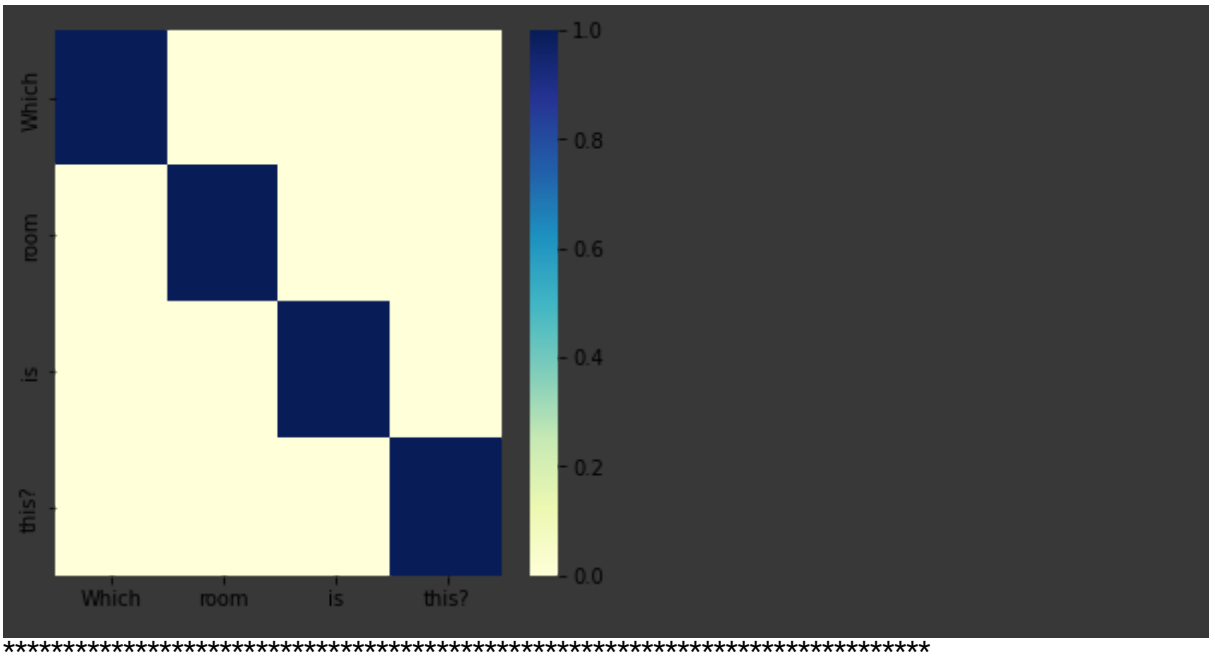
****

Image Index: 116
Plotting the Image

Question
Which room is this?
1/1 [==============================] - 0s 21ms/step
Ground Truth Answer: bathroom
Predicted Answer: no
***Confidence with which the Prediction is made**: 3.0511556
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 285
Plotting the Image

Question

What are the top lights?

1/1 [==============================] - 0s 21ms/step

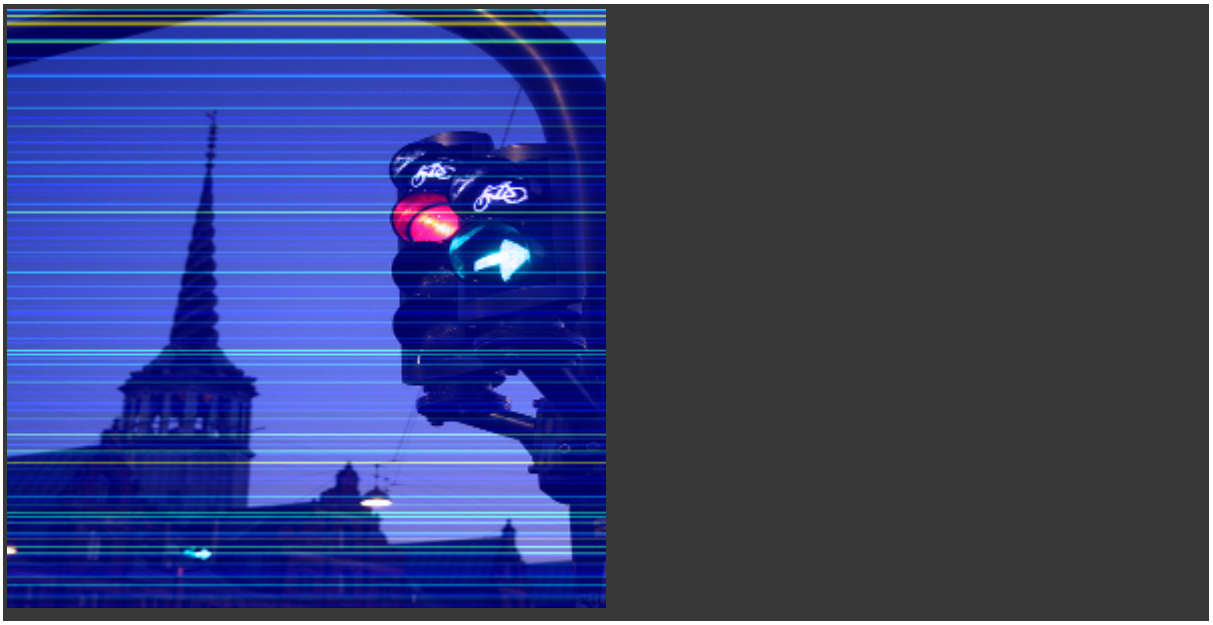
Ground Truth Answer: bike crossing

Predicted Answer: no

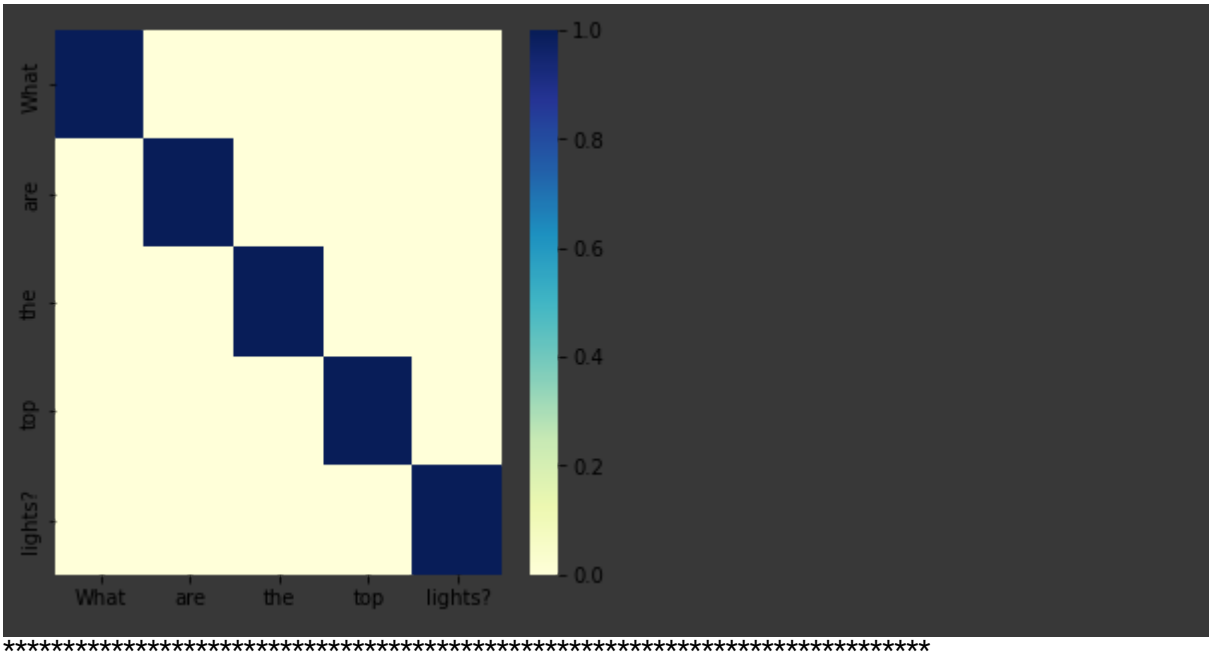
***Confidence with which the Prediction is made**: 4.2673955

Image HeatMap

1/1 [==============================] - 0s 31ms/step

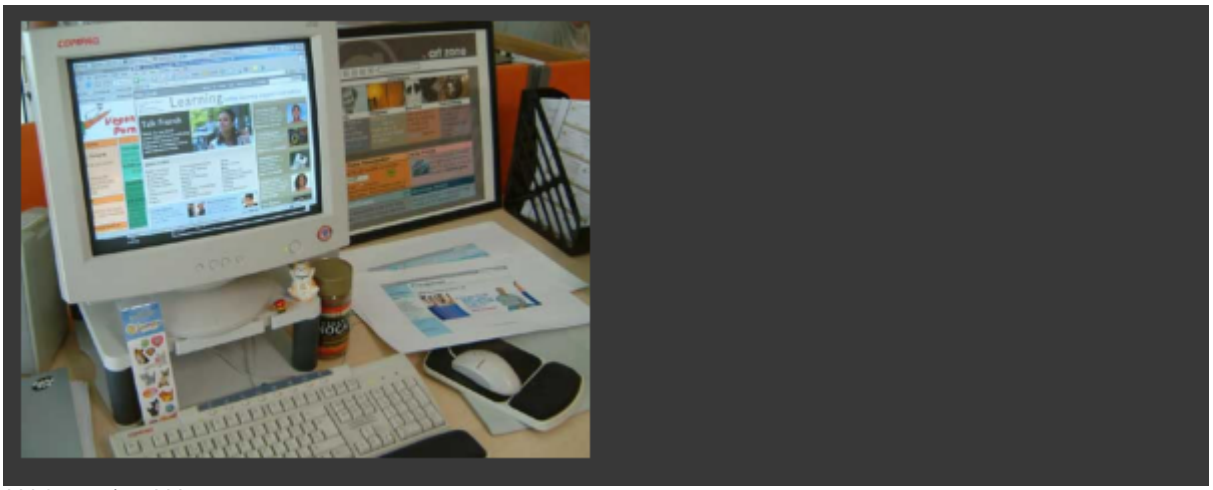


Attention Correlation Visualization

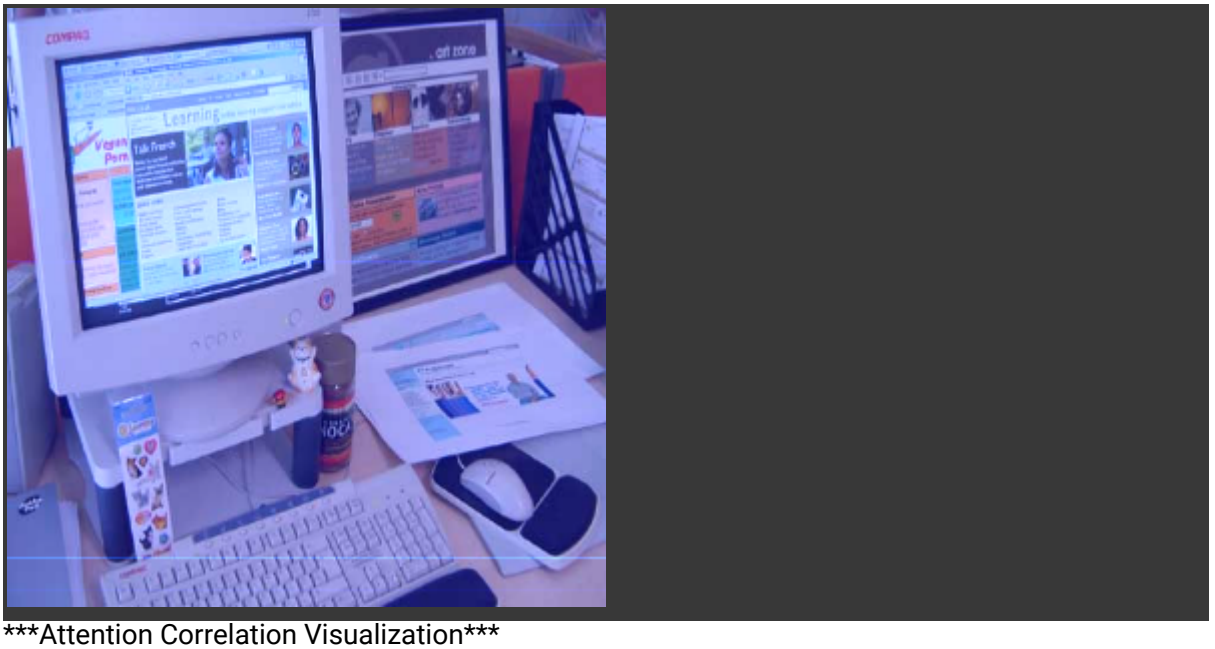


****

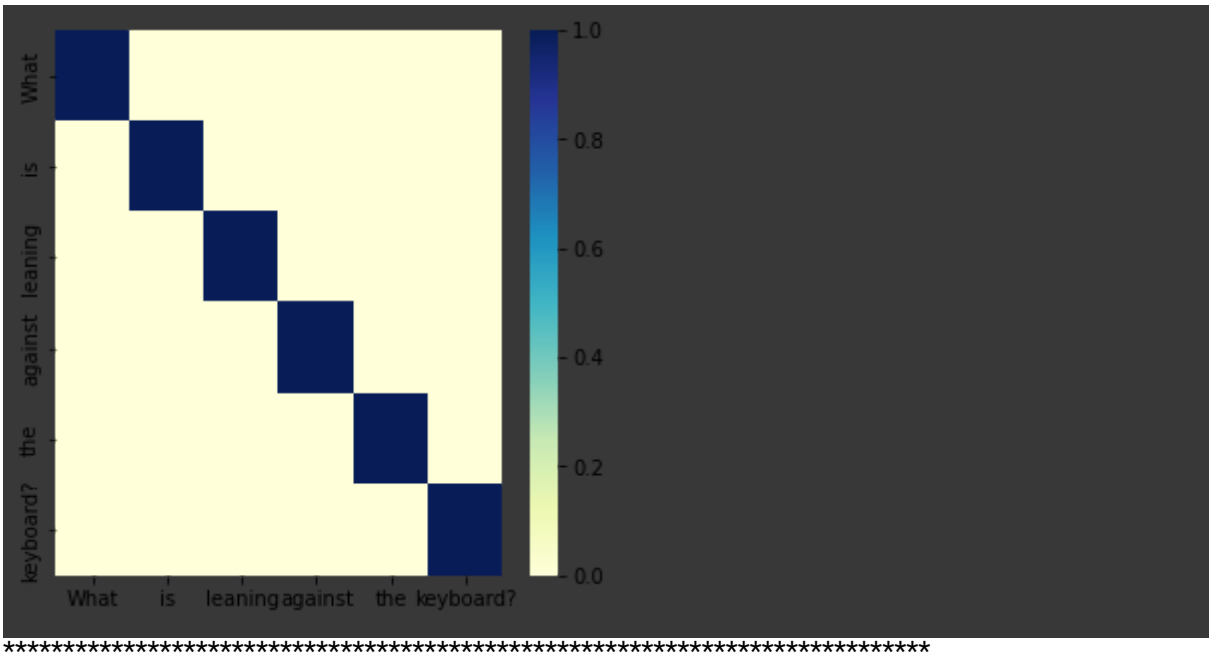
Image Index: 0
Plotting the Image



Question
What is leaning against the keyboard?
1/1 [==============================] - 0s 22ms/step
Ground Truth Answer: stickers
Predicted Answer: no
***Confidence with which the Prediction is made**: 2.914757
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 74
Plotting the Image

Question

Can people go under this street?

1/1 [==============================] - 0s 22ms/step

Ground Truth Answer: yes

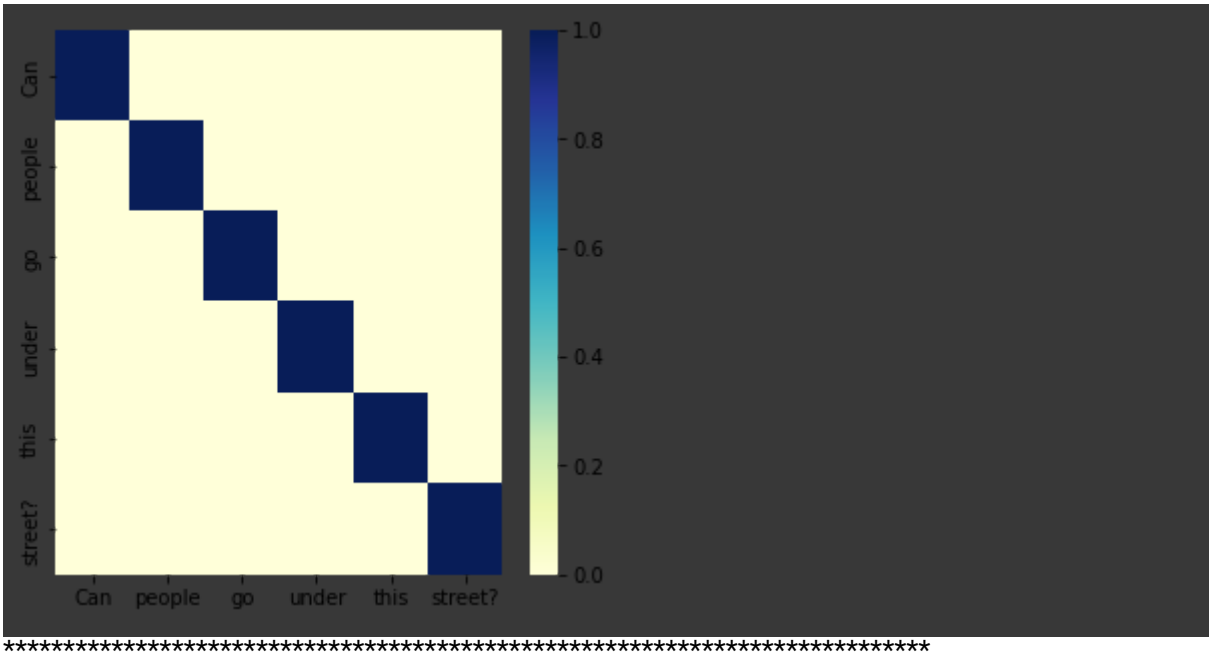
Predicted Answer: yes

***Confidence with which the Prediction is made**: 70.69328

Image HeatMap

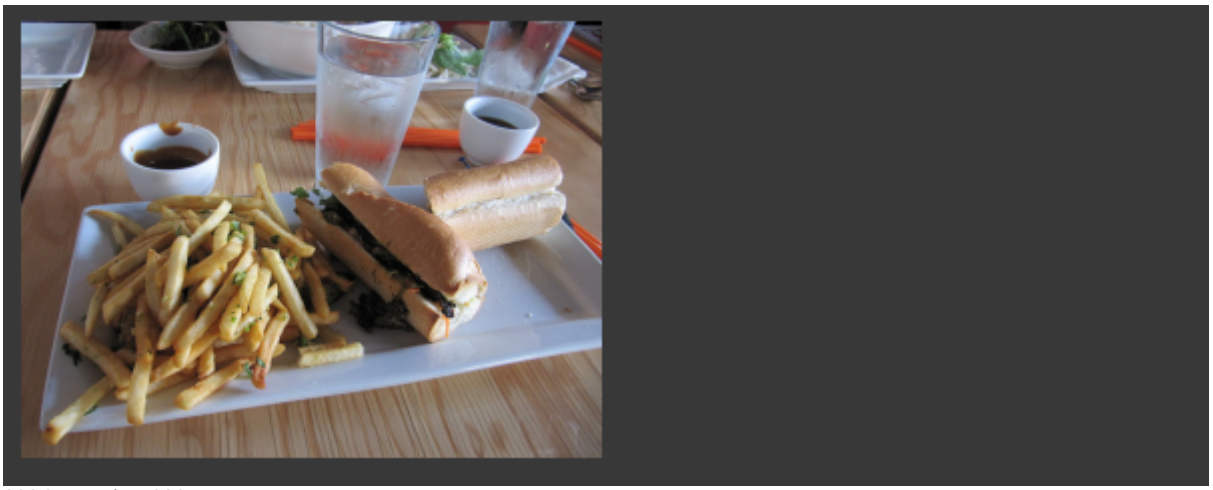
1/1 [==============================] - 0s 30ms/step

Attention Correlation Visualization

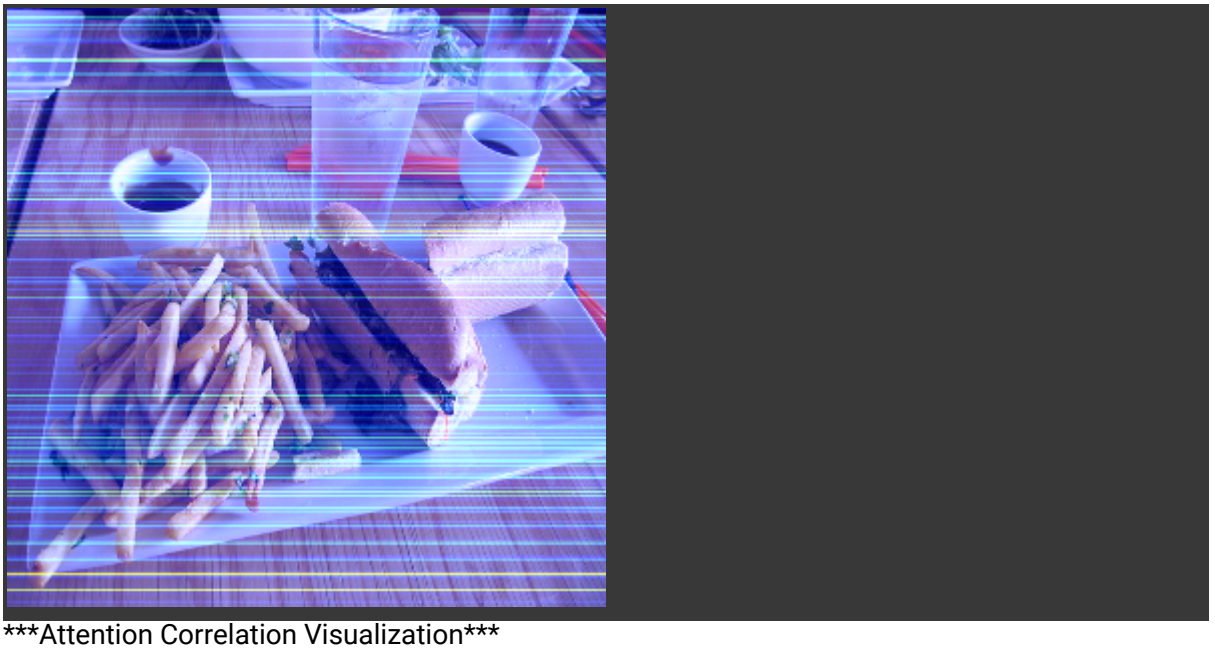


****

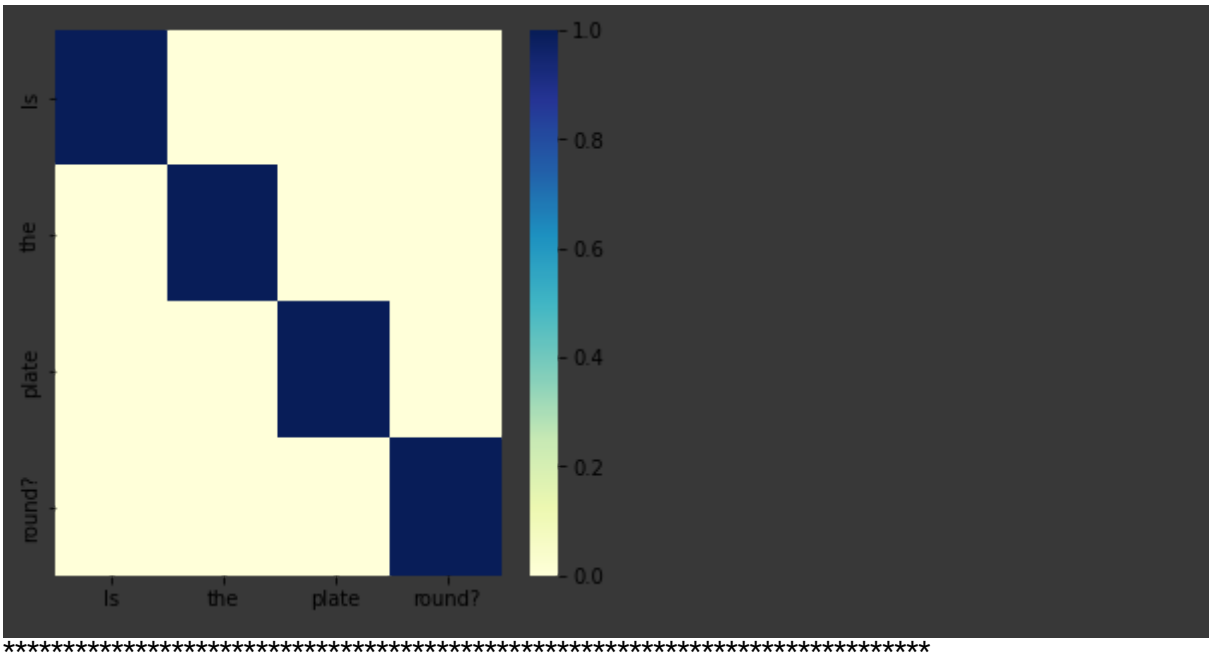
Image Index: 225
Plotting the Image



Question
Is the plate round?
1/1 [==============================] - 0s 22ms/step
Ground Truth Answer: no
Predicted Answer: yes
***Confidence with which the Prediction is made**: 58.418232
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 188
Plotting the Image

Question

Is this a small kitchen?

1/1 [==============================] - 0s 24ms/step

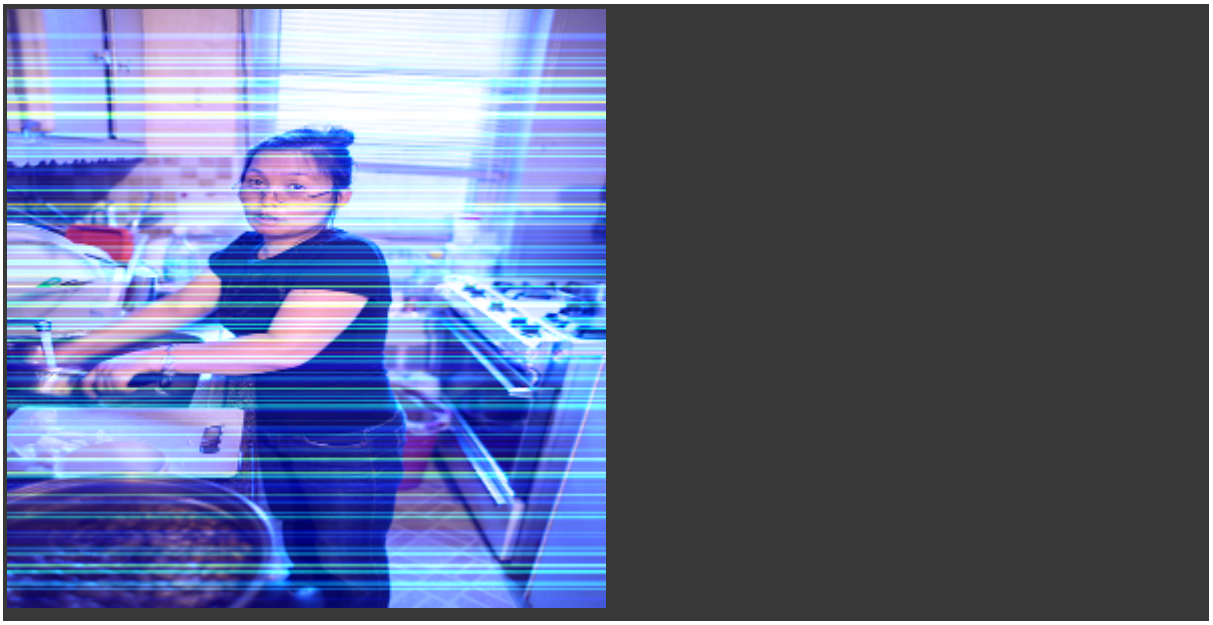
Ground Truth Answer: yes

Predicted Answer: yes

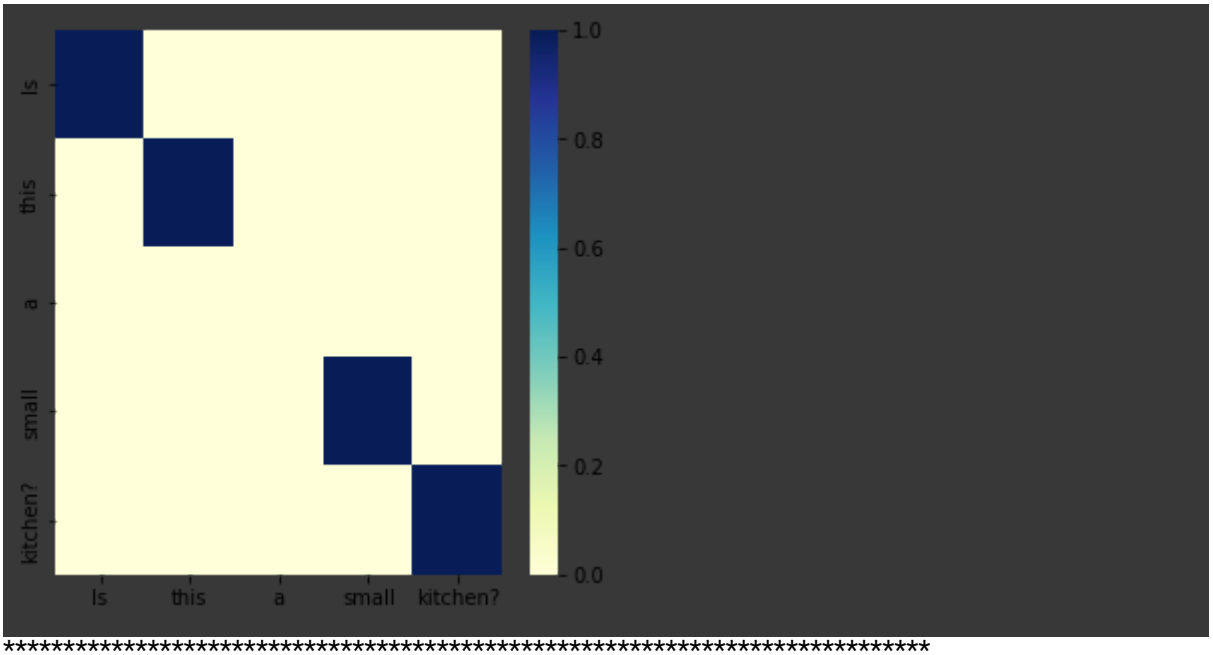
***Confidence with which the Prediction is made**: 68.876144

Image HeatMap

1/1 [==============================] - 0s 32ms/step

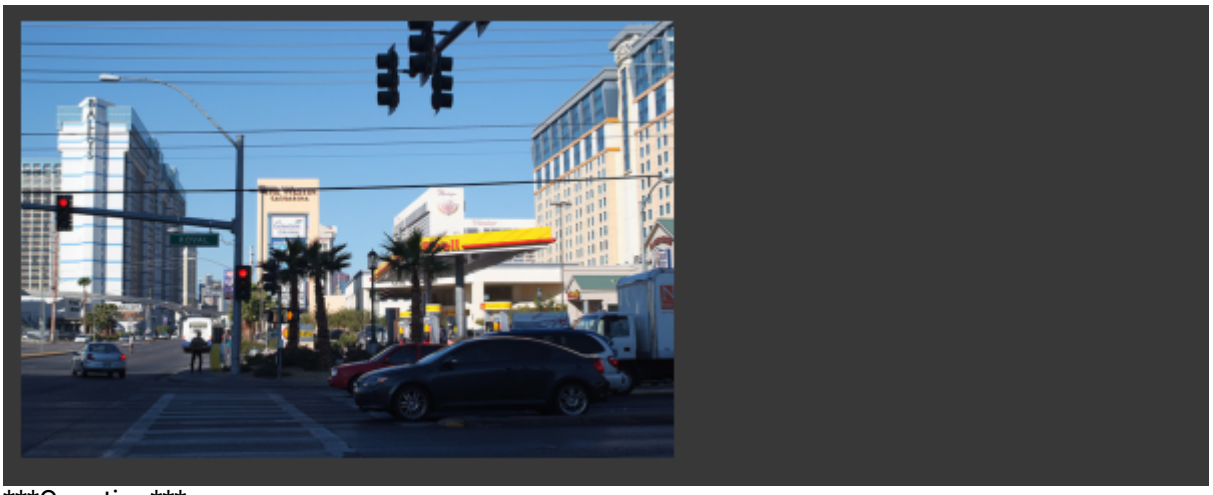


Attention Correlation Visualization



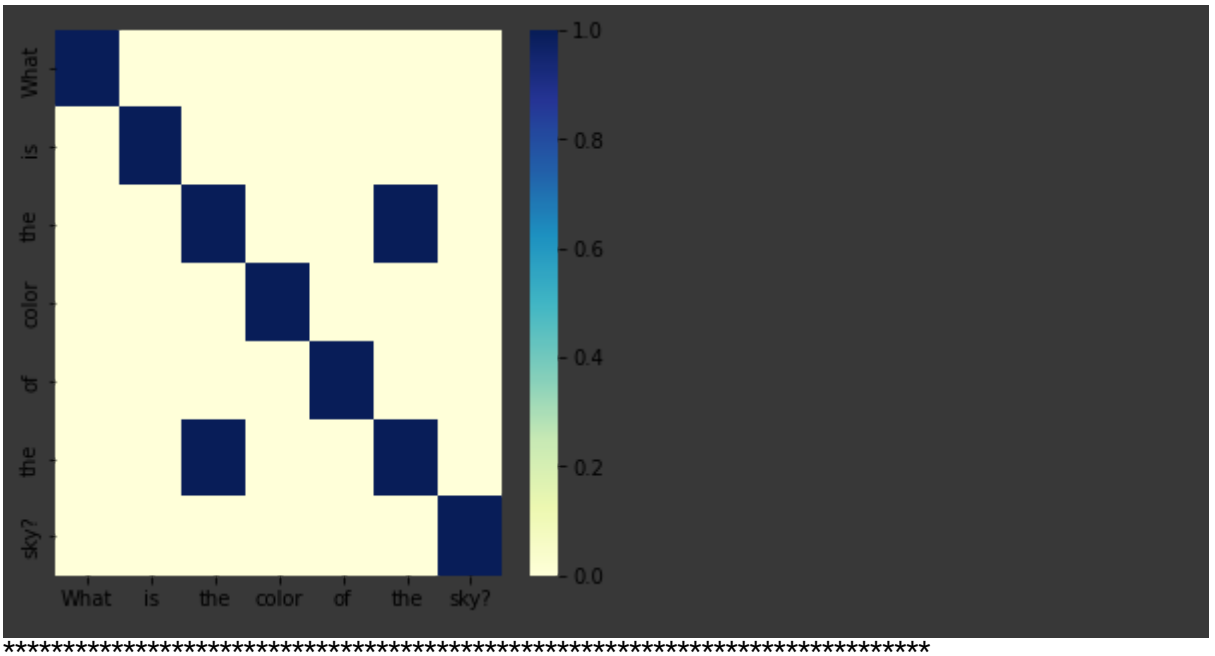
****

Image Index: 83
Plotting the Image



Question
What is the color of the sky?
1/1 [==============================] - 0s 21ms/step
Ground Truth Answer: blue
Predicted Answer: white
***Confidence with which the Prediction is made**: 18.04677
Image HeatMap
1/1 [==============================] - 0s 32ms/step

Attention Correlation Visualization



****

Image Index: 173
Plotting the Image

Question

What game is she playing?

1/1 [==============================] - 0s 22ms/step

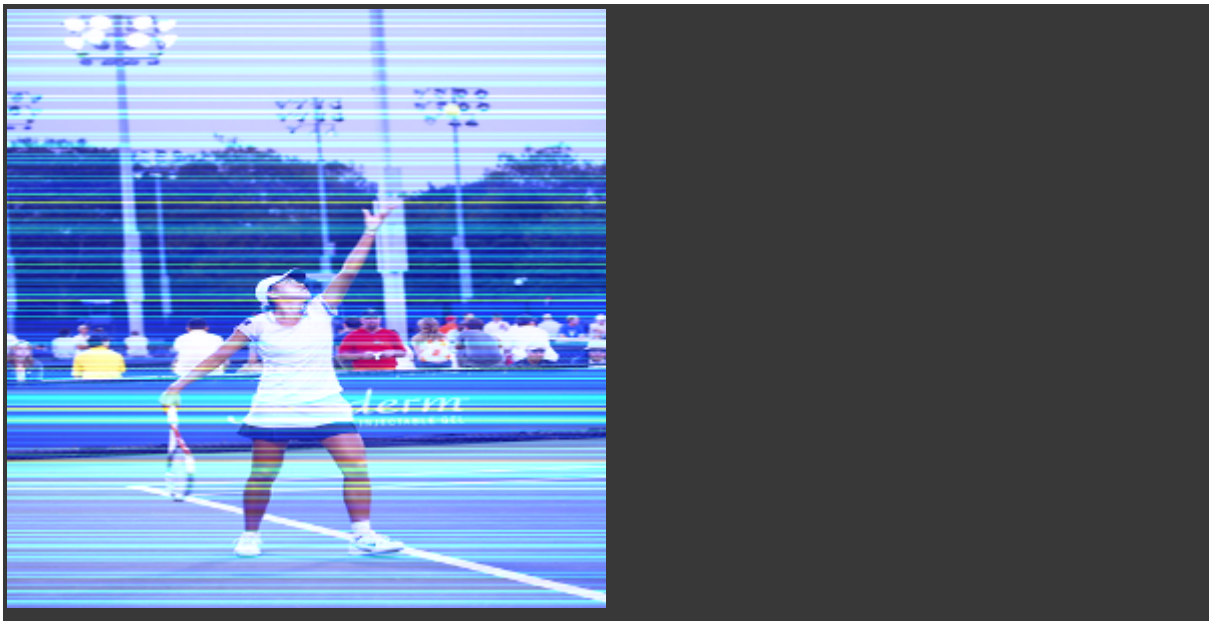
Ground Truth Answer: tennis

Predicted Answer: no

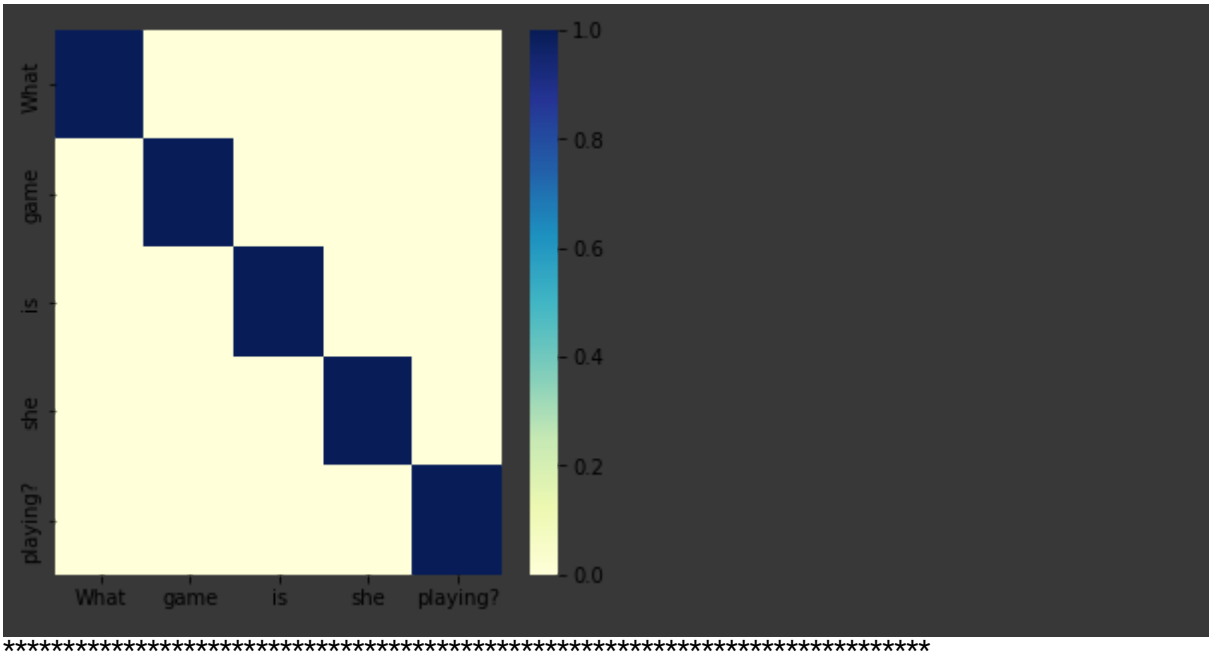
***Confidence with which the Prediction is made**: 4.5689025

Image HeatMap

1/1 [==============================] - 0s 32ms/step

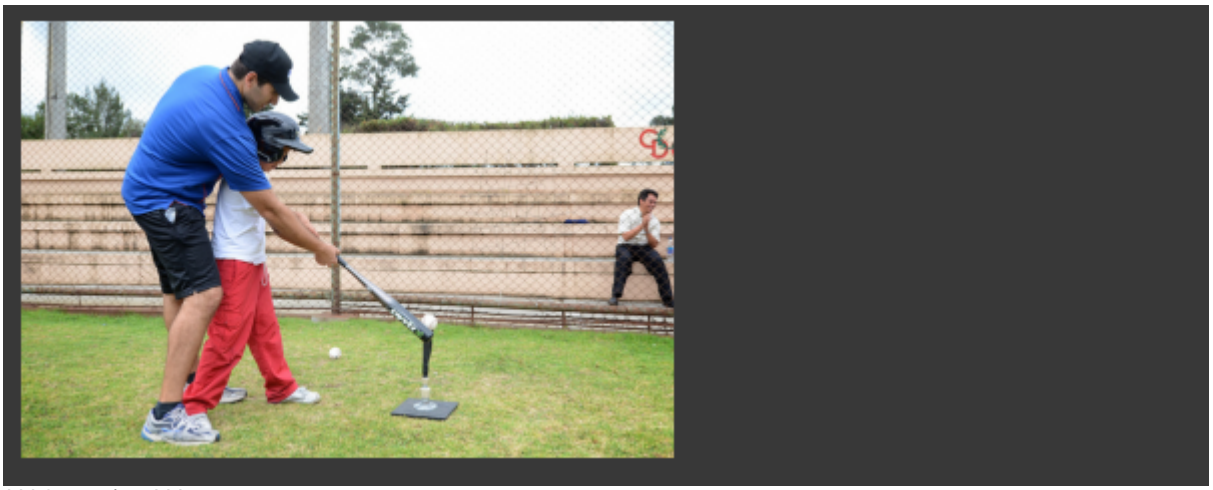


Attention Correlation Visualization

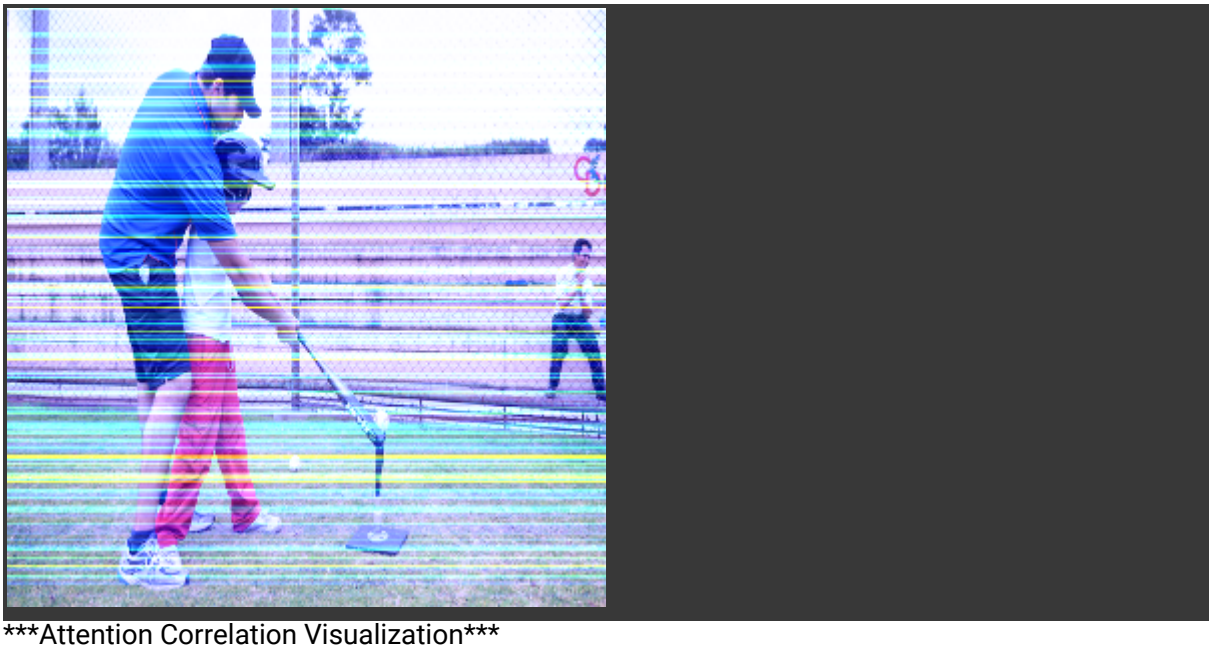


****

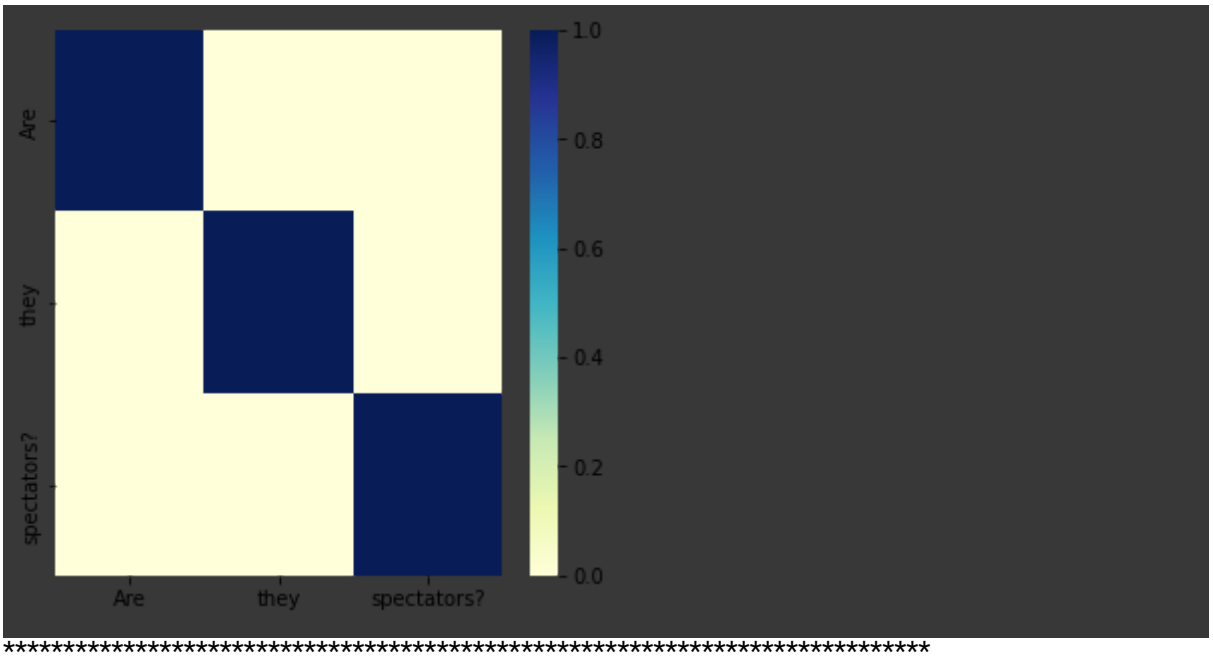
Image Index: 107
Plotting the Image



Question
Are they spectators?
1/1 [==============================] - 0s 22ms/step
Ground Truth Answer: yes
Predicted Answer: yes
***Confidence with which the Prediction is made**: 68.104126
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 30
Plotting the Image

Question

Is Lacoste a sponsor of this tennis match?

1/1 [==============================] - 0s 22ms/step

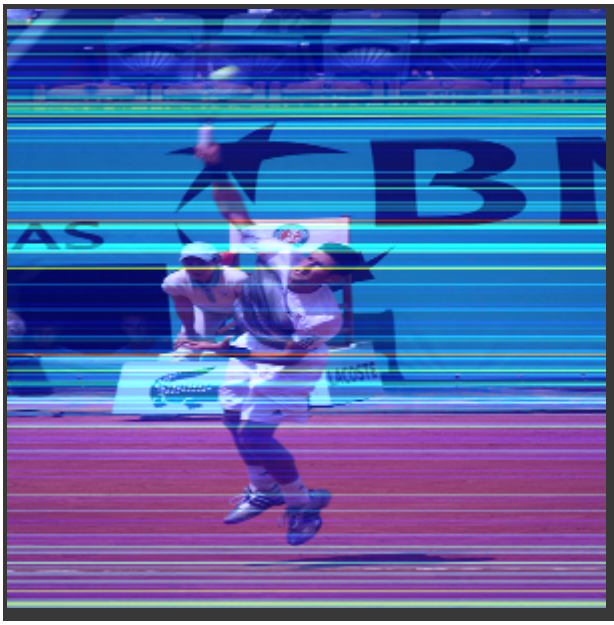
Ground Truth Answer: yes

Predicted Answer: yes

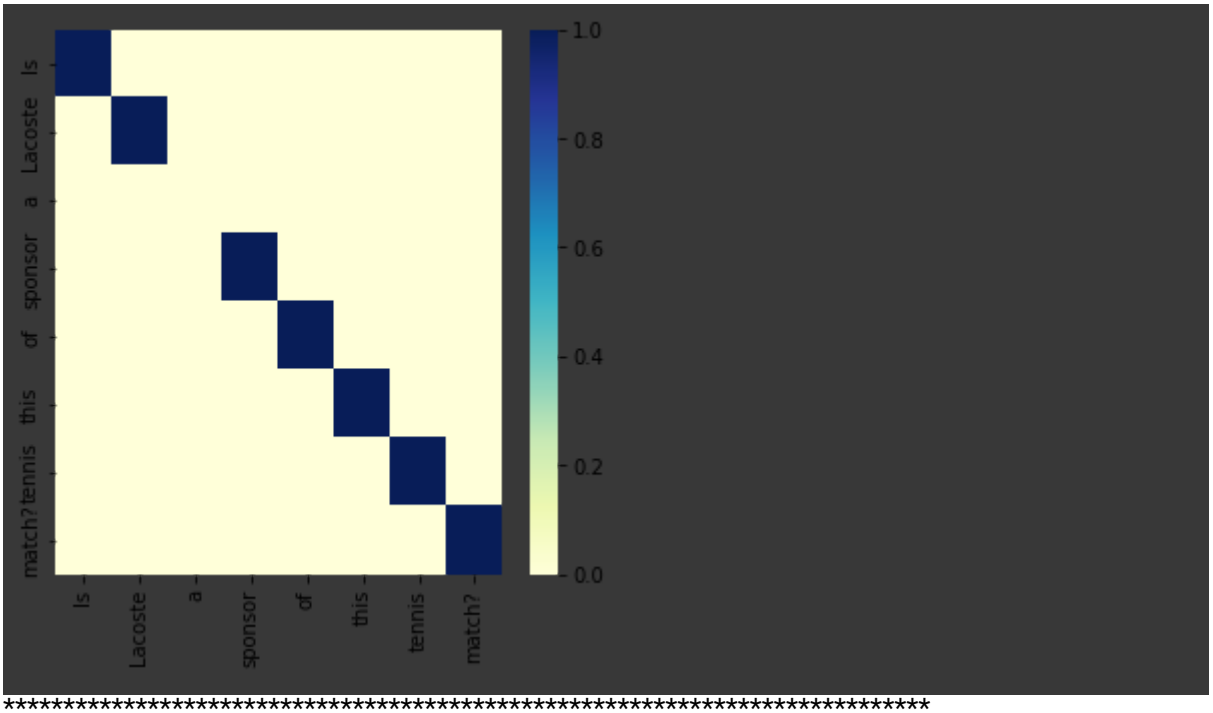
***Confidence with which the Prediction is made**: 68.478645

Image HeatMap

1/1 [==============================] - 0s 31ms/step

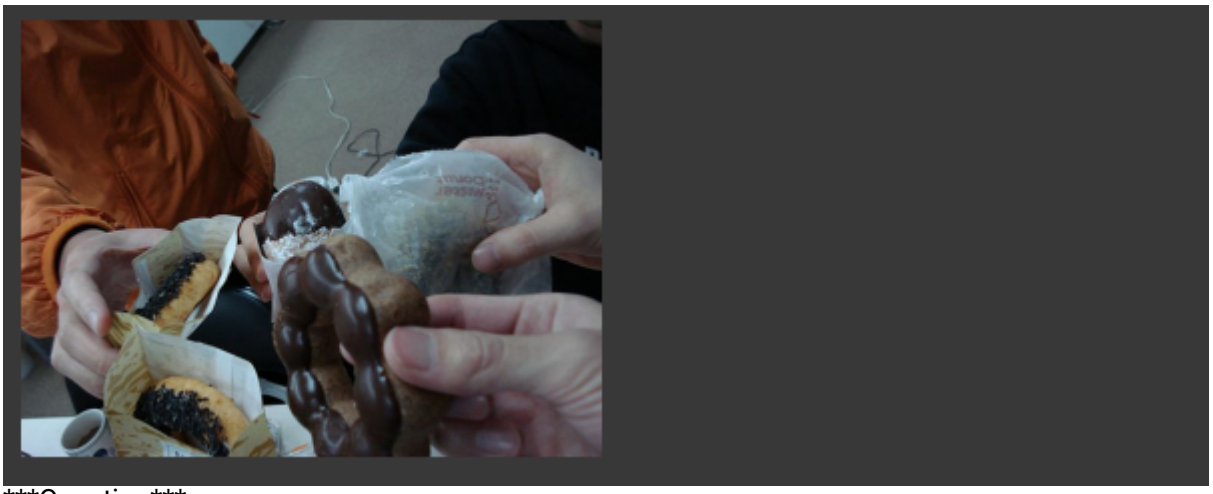


Attention Correlation Visualization

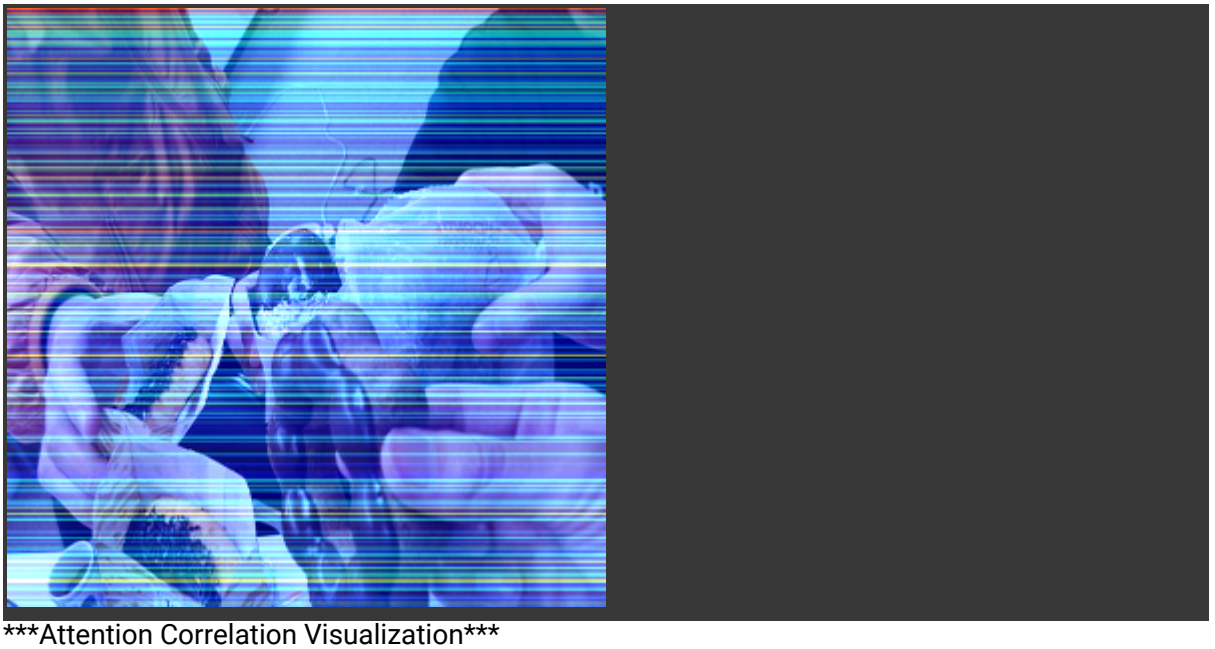


****

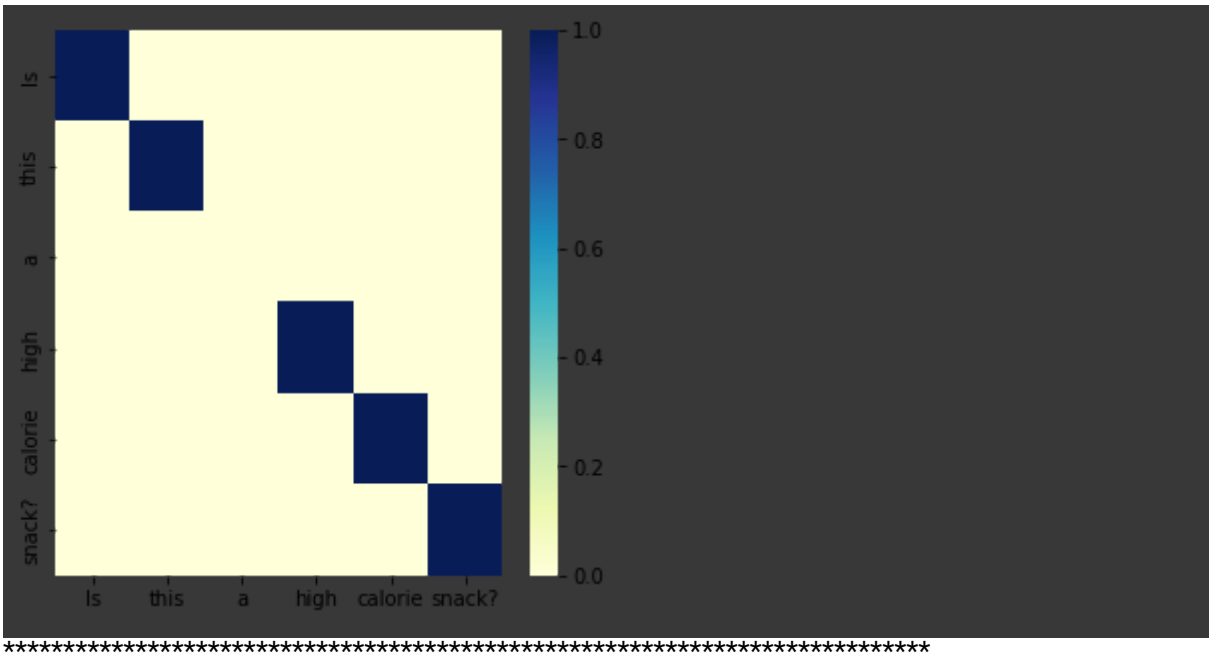
Image Index: 295
Plotting the Image



Question
Is this a high calorie snack?
1/1 [==============================] - 0s 22ms/step
Ground Truth Answer: yes
Predicted Answer: yes
***Confidence with which the Prediction is made**: 66.70954
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 102
Plotting the Image

Question

Are there a lot of people on the mountain?

1/1 [==============================] - 0s 23ms/step

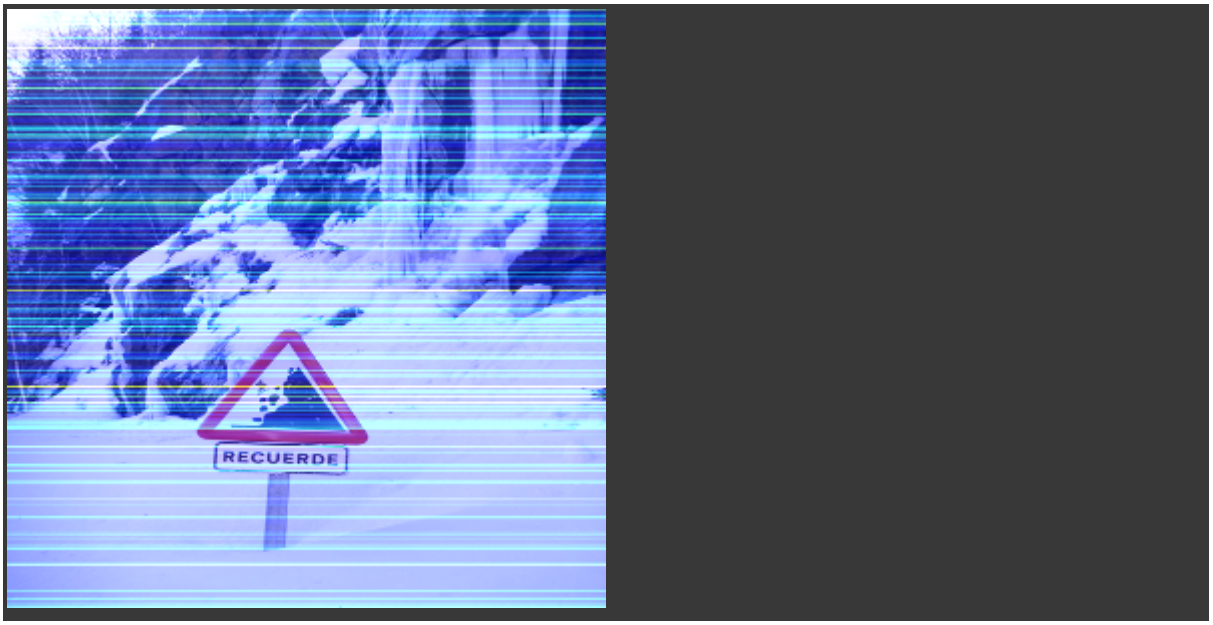
Ground Truth Answer: no

Predicted Answer: yes

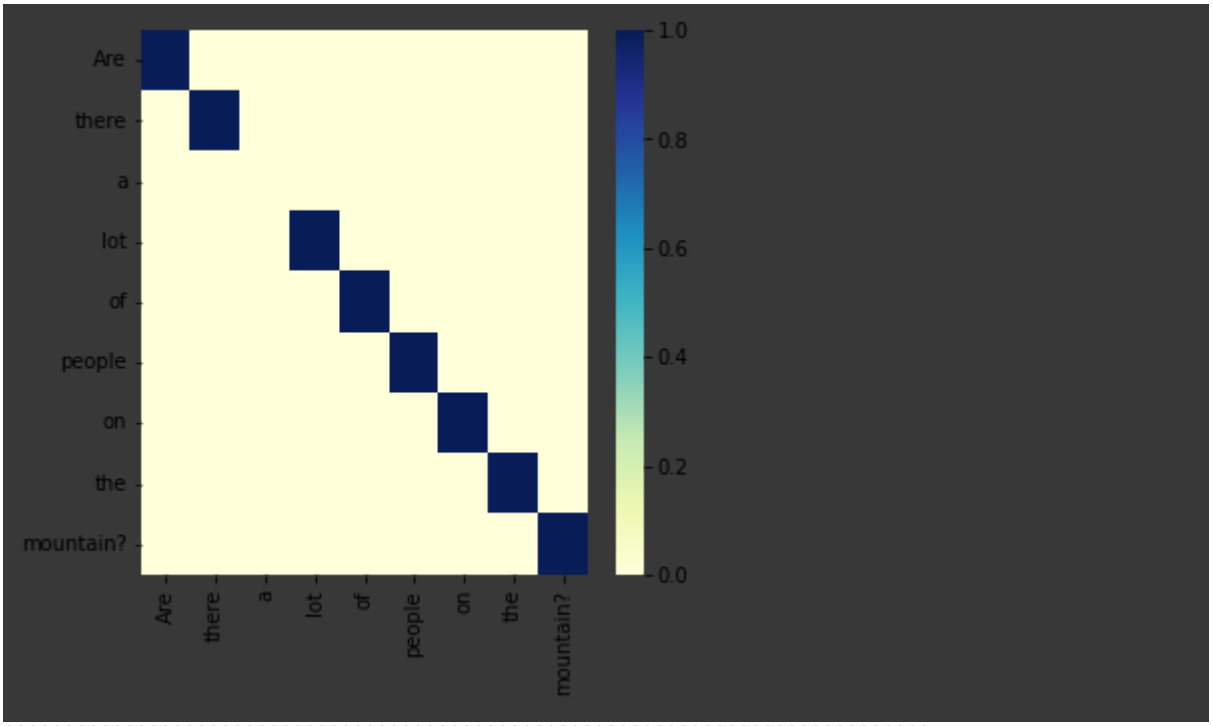
***Confidence with which the Prediction is made**: 68.737076

Image HeatMap

1/1 [==============================] - 0s 31ms/step



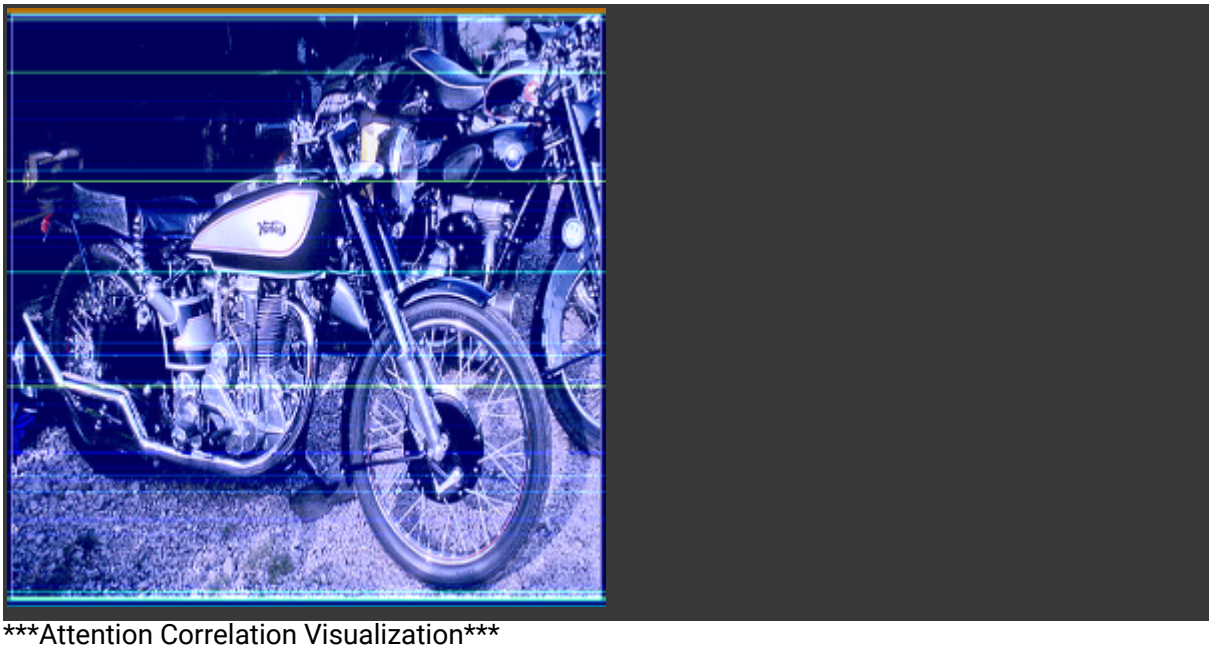
Attention Correlation Visualization



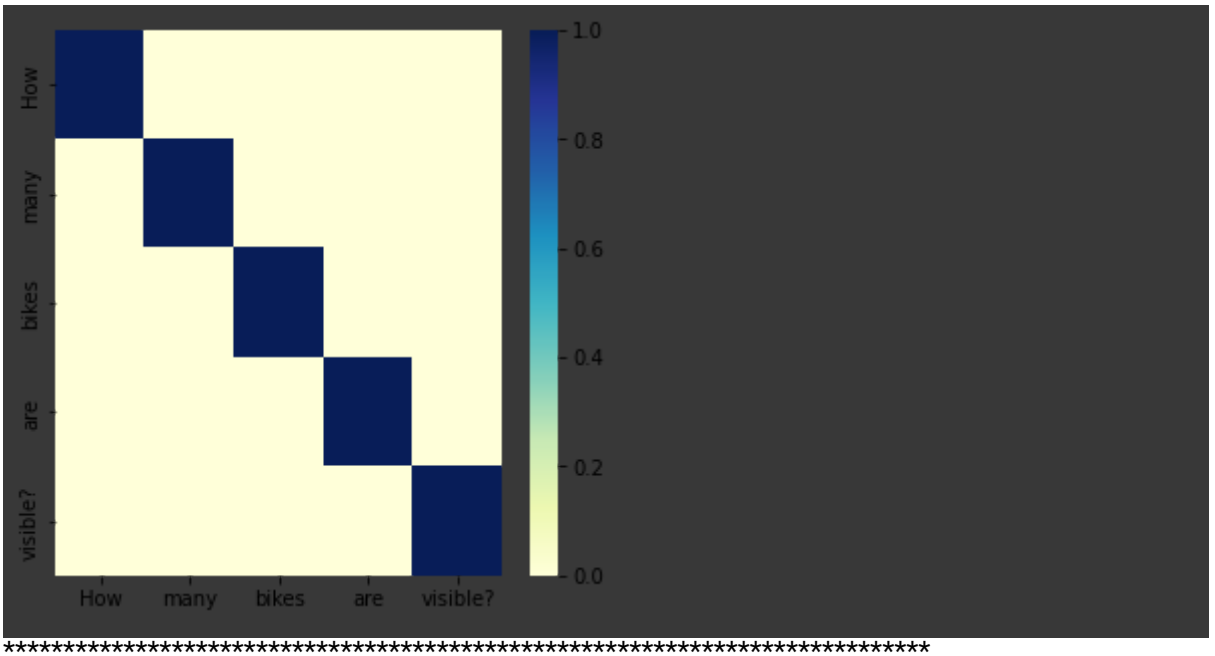
****

Image Index: 38
Plotting the Image

Question
How many bikes are visible?
1/1 [==============================] - 0s 21ms/step
Ground Truth Answer: 2
Predicted Answer: 2
***Confidence with which the Prediction is made**: 18.485092
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 262

Plotting the Image

Question

Is the rider wearing a fluorescent vest?

1/1 [==============================] - 0s 21ms/step

Ground Truth Answer: yes

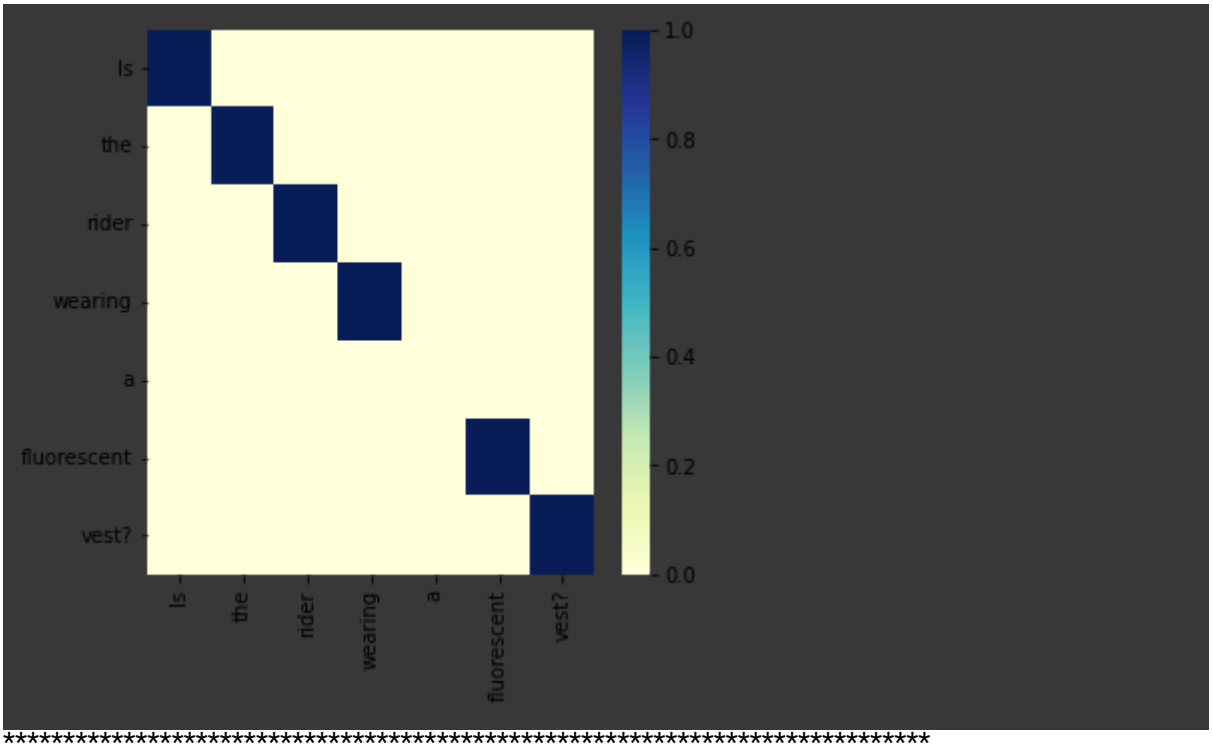
Predicted Answer: yes

***Confidence with which the Prediction is made**: 63.384968

Image HeatMap

1/1 [==============================] - 0s 32ms/step

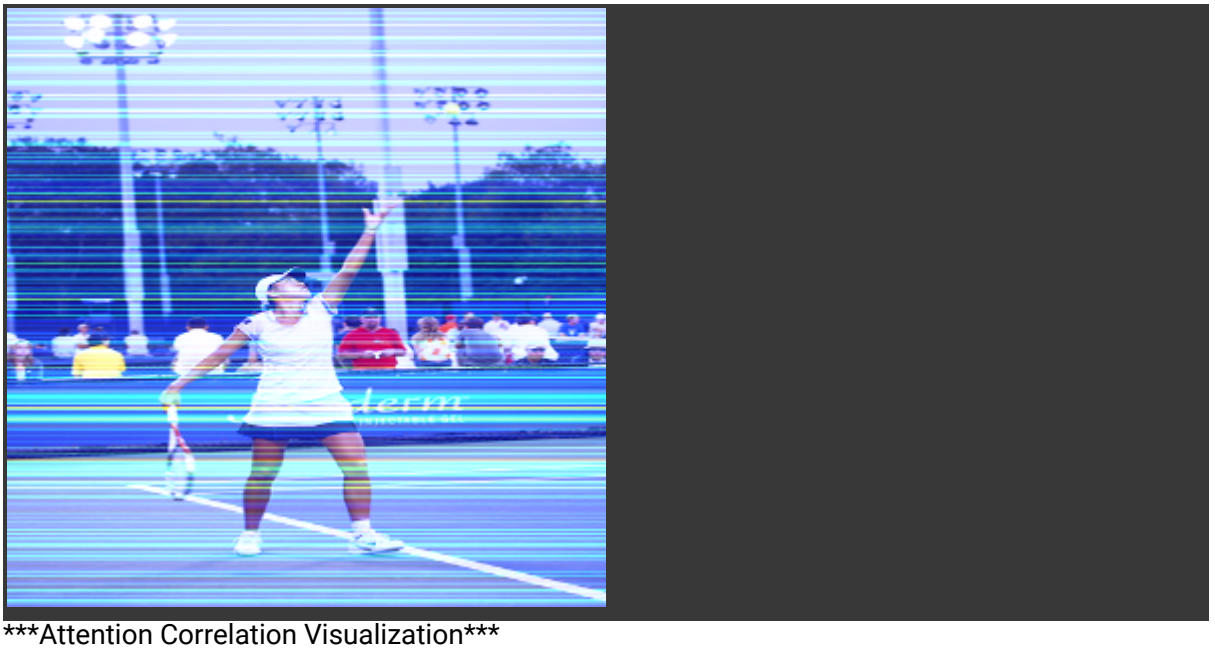
Attention Correlation Visualization



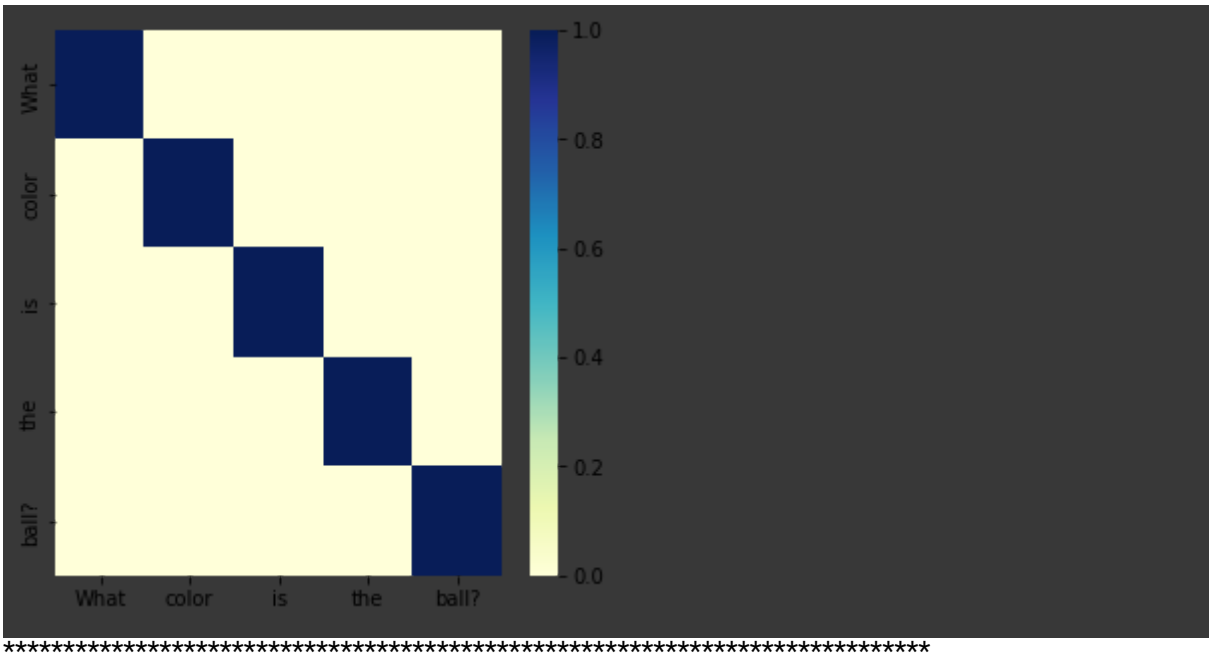
****

Image Index: 172
Plotting the Image

Question
What color is the ball?
1/1 [==============================] - 0s 21ms/step
Ground Truth Answer: yellow
Predicted Answer: 2
***Confidence with which the Prediction is made**: 20.911087
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

Image Index: 207

Plotting the Image

Question

Is this an expensive vehicle?

1/1 [==============================] - 0s 22ms/step

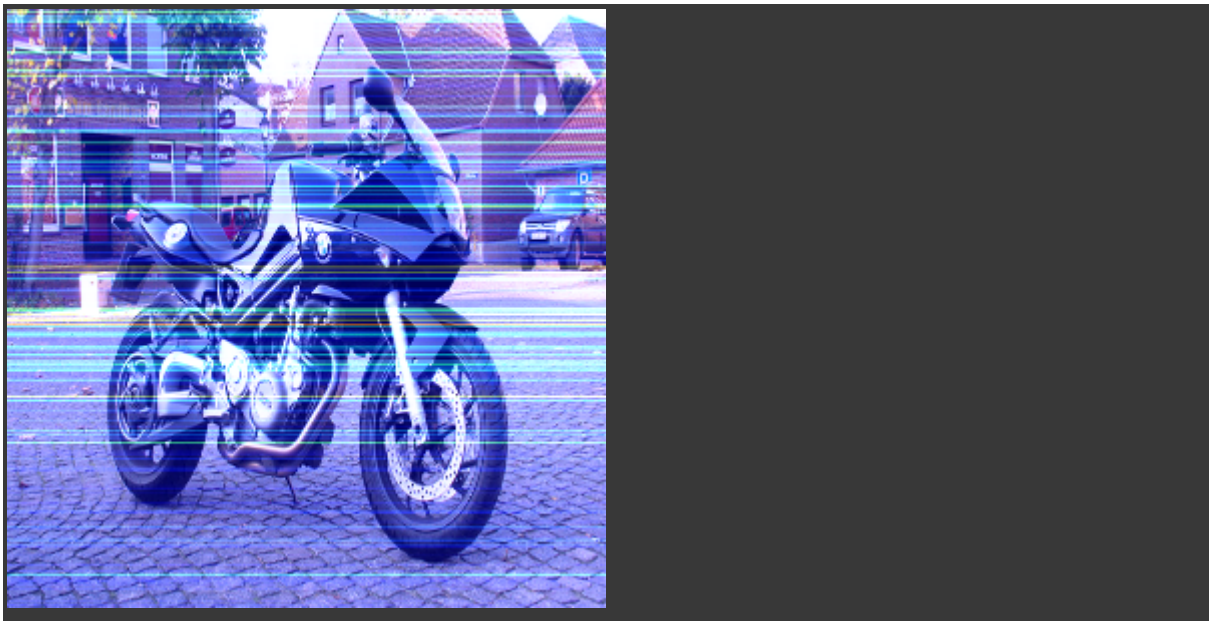
Ground Truth Answer: yes

Predicted Answer: yes

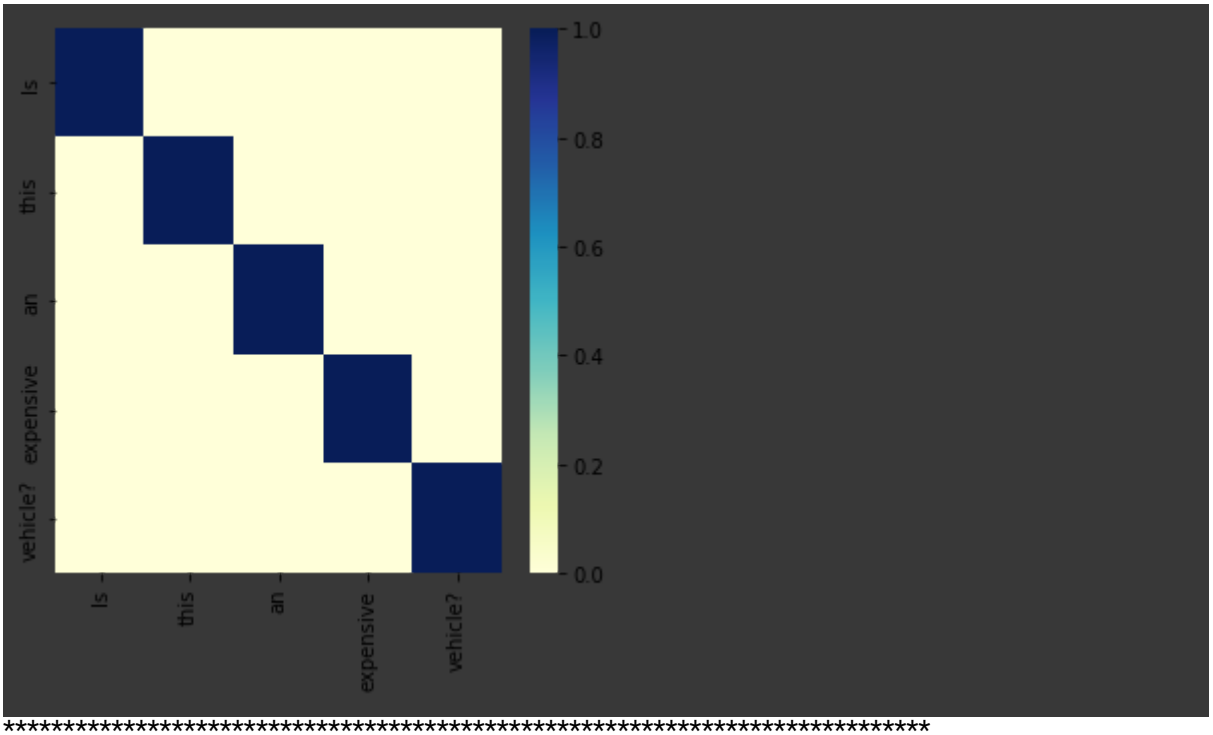
***Confidence with which the Prediction is made**: 69.04924

Image HeatMap

1/1 [==============================] - 0s 35ms/step

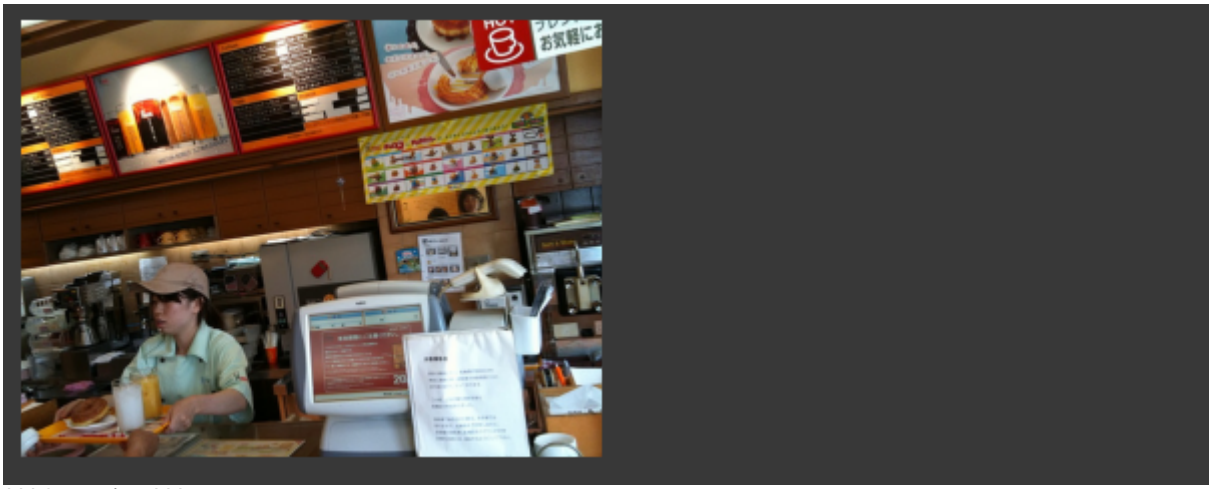


Attention Correlation Visualization



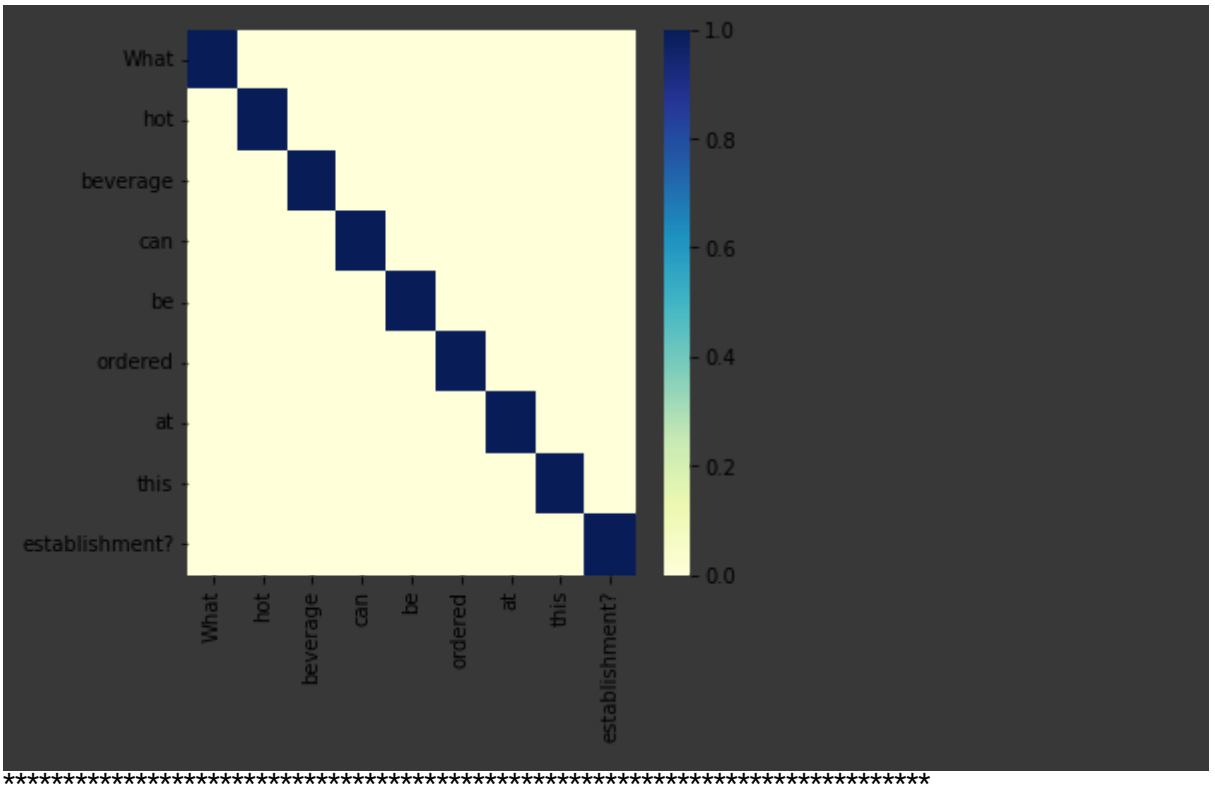
****

Image Index: 44
Plotting the Image



Question
What hot beverage can be ordered at this establishment?
1/1 [==============================] - 0s 23ms/step
Ground Truth Answer: coffee
Predicted Answer: no
***Confidence with which the Prediction is made**: 4.132215
Image HeatMap
1/1 [==============================] - 0s 30ms/step

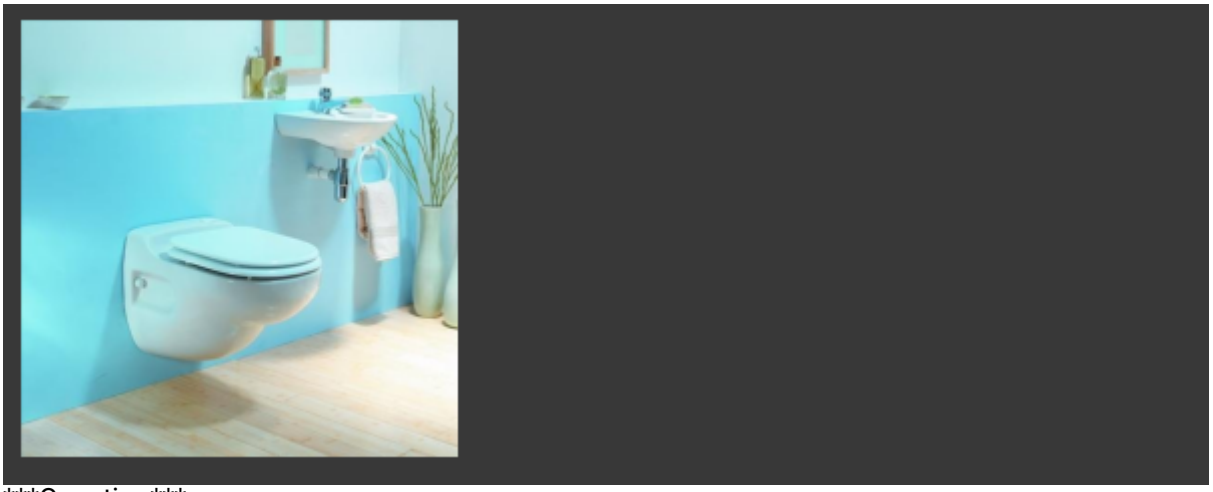
Attention Correlation Visualization



****

Image Index: 9

Plotting the Image



Question

What is the main color in this room?

1/1 [==============================] - 0s 26ms/step

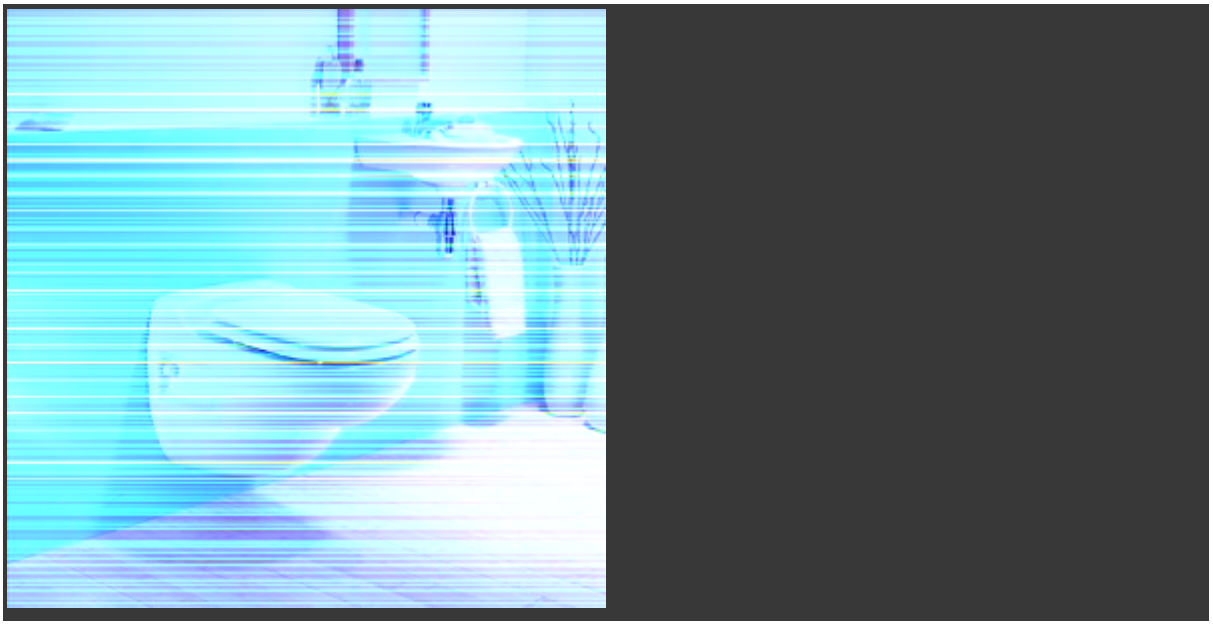
Ground Truth Answer: blue

Predicted Answer: white

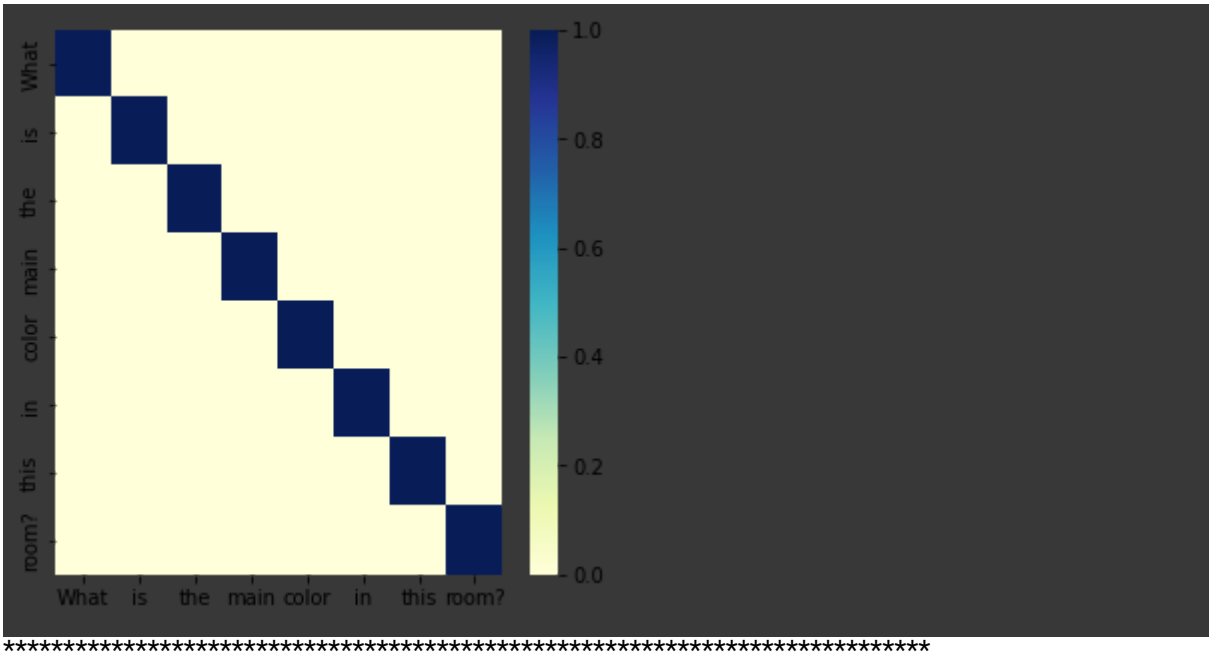
***Confidence with which the Prediction is made**: 13.1088

Image HeatMap

1/1 [==============================] - 0s 32ms/step



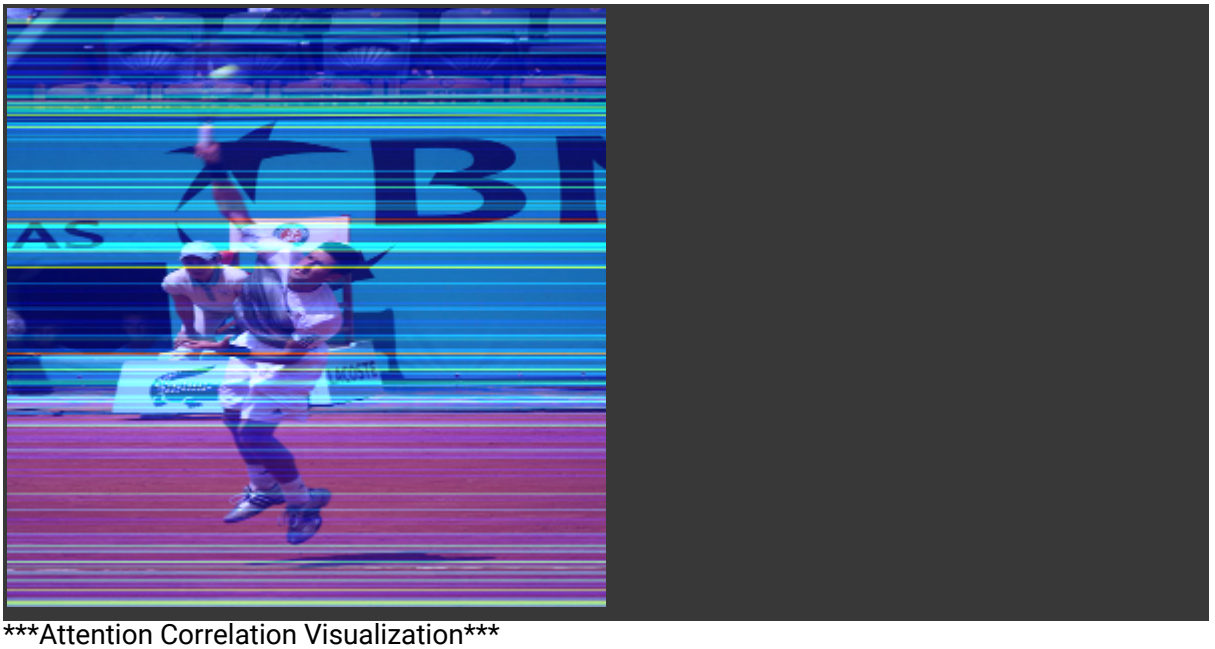
Attention Correlation Visualization



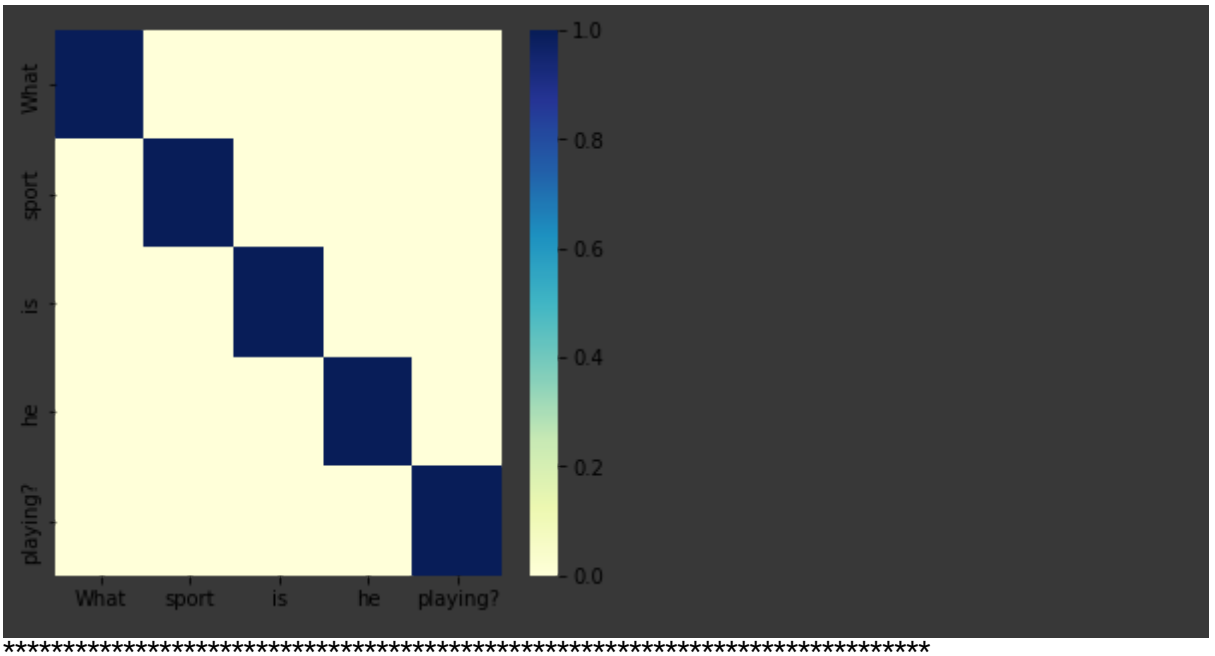
****

Image Index: 31
Plotting the Image

Question
What sport is he playing?
1/1 [==============================] - 0s 22ms/step
Ground Truth Answer: tennis
Predicted Answer: no
***Confidence with which the Prediction is made**: 4.7054033
Image HeatMap
1/1 [==============================] - 0s 31ms/step



Attention Correlation Visualization



****

< ——----------END>

References used for implementing this project :

- https://github.com/VedantYadav/VQA
- https://github.com/iamaaditya/VQA_Demo
- https://github.com/harsha977/Visual-Question-Answering-With-Hierarchical-Question-Image-Co-Attention
- https://github.com/gradient-ai/Seq-to-Seq-Machine-Translation