

# Visual Question Answering: A Comparative Study

\*Divya Jyothi Gaddipati  
Virginia Tech  
divyaj@vt.edu

\*Dhiraj Srivatsava  
Virginia Tech  
dhirajsrivastava@vt.edu

## Abstract

*Visual question answering (VQA) is a unique problem that involves both visual and natural language components. Many of the recently proposed VQA systems include attention or memory mechanisms designed to steer away from the black-box modeling. In this project, we present our exploration of networks that are attention-based and compare them with non-attention based baseline models that are built using pretrained CNNs and RNNs. We present our analysis along with quantitative and qualitative performance of the models on the VQA 2.0 dataset.*

## 1. Introduction

Visual Question Answering (VQA) has been a notoriously difficult problem to tackle because of its intersection of the two different tasks - computer vision and natural language processing. However, this has been an exciting area of research in the deep learning community in the recent years, as neural networks have made tremendous strides in image and text-based tasks, individually and combined. This is specifically due to the rise of convolution neural networks (CNNs) and recurrent neural networks (RNNs) and more recently the attention-based networks. These systems have several applications such as assisting the blind and visually-impaired in getting information from the images and also improving image-retrieval applications.

In this task, the input is an image and an open-ended question about the image, and the output is an open-ended answer to the question with respect to the image. The core challenge of this task lies in comprehending and correlating the information from two sources - image and text. In our project, we explore using CNN, LSTM and Attention-based architectures to fuse information from both the modalities. We also explore how attention in image alone contributes to the model's performance and how attention in text along contributes to the model's performance. The next sections of the report describe the methods/networks in detail, experiments performed, results, and analysis of the models.

## 2. Methods

In general, the outline of the approaches for VQA is as follows:

- Extract features from the image.
- Extract features from the question.
- Combine the features and generate the answer

We look at how to extract the individual component features using non-attention based networks and attention based networks.

### 2.1. Baseline models

We experimented with a set of baseline models by varying both the image and question feature extraction modules. The different configurations chosen are explained below.

#### 2.1.1 Image Feature Extraction

Transfer learning is a convenient way to use the existing pretrained networks. For our project, we use two pre-trained networks - VGG-19 and ResNet-50. The image features are extracted from the bottleneck layer of these pretrained networks are saved to file prior to training the VQA model. So, the input to the image-branch of the VQA model would be these extracted features. This input layer is followed by a dense layer of 1024 units with relu activation.

#### 2.1.2 Question Feature Extraction

The questions also need to be preprocessed and converted to a fixed-size representation. This is done in two steps - Tokenization and Sequence padding. All the questions that are fed as input to the model should first be integer-encoded and then undergo the preprocessing steps to convert them into a fixed-length sequence of length 24. This 24-length question input should be embedded in a latent dimension using the Embedding layer. Three different embedding techniques are used for comparison:

- Embedding layer (300-D) with random initialization.

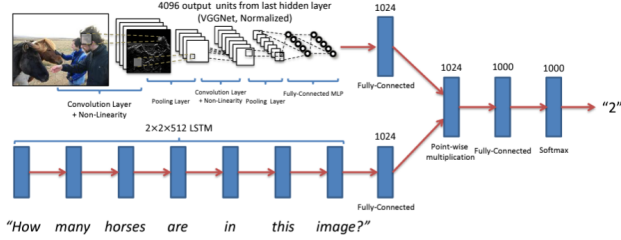


Figure 1. Baseline model structure. Source: [3]

- Embedding layer with pre-trained GloVe vector initialization. GloVe is a popular embedding technique based on factorizing a matrix of word co-occurrence statistics [7]. Specifically, we used the 300-dimensional GloVe embeddings that is publicly available.
- Embedding layer with pretrained BERT embeddings [?]. BERT [5] is a Transformer based model that produces word representations that are dynamically informed by the words around them. Aside from capturing obvious differences like polysemy, the context-informed word embeddings capture other forms of information that result in more accurate feature representations, which in turn results in better model performance [1]. The required preprocessing on the text is performed when BERT embeddings are used in the models.

Using such pretrained word embeddings can help in reducing the training time and also extract meaningful features. After the embedding layer, two more LSTM units are included. At the end, a Dense layer with 1024 units is used to make sure the features have the same shape as the image features.

### 2.1.3 Combining Image and Question Features

As mentioned in the sections 2.1.1 and 2.1.2, the features at the last layer from both the branches are made sure to have the same shape. This is done by projecting them into the same dimension length space using dense layers. There are several ways to combine these image and textual features. We used a point-wise multiplication approach. Then the features are fed into the final stage of the model, which is a multi-layer perceptron with a softmax non-linear layer at the end that outputs the score distribution as a 1000-dimensional vector.

## 2.2. Attention models

Attention mechanism has become one of the prominent aspect in the field of deep learning and has become widely

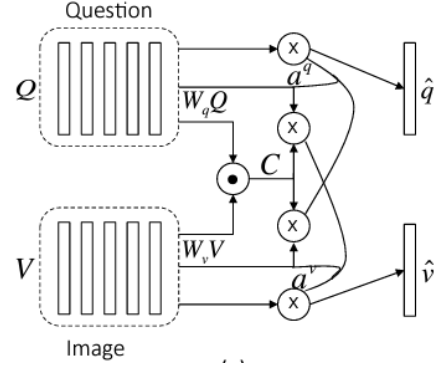


Figure 2. Question and Image attention maps joint mechanism. Source: [6]

popular in many areas of the field. Memory and attention-based models rely on the hypothesis that in order to answer a question correctly the model has to understand where in the image to “look” or which words of the questions to “listen”. This is what is required for the problem of Visual Question Answering as well - the machine needs to understand both the image (“where to look”) and also the question (“which words to listen to”). Both are equally important for an accurate model. To this respect, [6] proposed a novel model based on both visual and question attention. HiCoAtt model proposed an hierarchical architecture that co-attends to the image and question at the word, phrase, and question level.

### 2.2.1 Co-attention Model

The main idea proposed by [6] is based on the symmetrical mechanism that the image representation should be guiding the question attention and the question representation(s) should be guiding the image attention. The paper also proposes a hierarchical architecture that co-attends to the image and question at three levels: (1) word-level, (2) phrase level, and (3) question level. At the word-level, the words are embedded to a vector space through an embedding matrix. At the phrase level, 1-dimensional convolution neural networks are used to capture the information contained in unigrams, bigrams, and trigrams. Specifically, the word representations are convolved with temporal filters of varying support and then the various n-gram responses are pooled into a single phrase-level representation. At the question level, recurrent neural networks are used to encode the entire question. As shown in Figure 2, for each level of the question representation in the hierarchy, the joint question and image co-attention maps are constructed.

In this way, co-attention attends to the image and question simultaneously. The image and question are connected by calculating the similarity between image and ques-

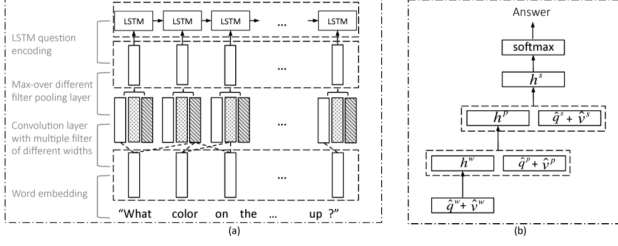


Figure 3. (a) Hierarchical question encoding at word, phrase and sentence level. (b) Encoding for predicting answers. Source: [6]

tion features at all pairs of image-locations and question-locations [2]. The image and question attention vectors are calculated as follows:

$$C = \tanh(Q^T W_b V) \quad (1)$$

$$H^v = \tanh(W_v V + (W_q Q)C) \quad (2)$$

$$a^v = \text{softmax}(w_{hv}^T H^v) \quad (3)$$

$$H^q = \tanh(W_q Q + (W_v V)C^T) \quad (4)$$

$$a^q = \text{softmax}(w_{hq}^T H^q) \quad (5)$$

$$\hat{v} = \sum_{n=1}^N a_n^v v_n, \quad \hat{q} = \sum_{t=1}^T a_t^q q_t \quad (6)$$

$a^q$  and  $a^v$  are the question-attention and image-attention maps. The affinity matrix  $C$  transforms the question attention space to image attention space and vice versa.  $\hat{q}$  and  $\hat{v}$  are the respective question and image attention vectors that are calculated as the weighted sum of image features and question features. These vectors are obtained at each level of the question hierarchy as shown in Figure 3.

Finally, a distribution over the answers is predicted based on the co-attended image and question features from all three levels. A multi-layer perceptron (MLP) is used to recursively encode the attention features as show in Figure 3

To get a deeper understanding of how the image attention and question attention are contributing to the co-attention model, we perform the following ablation study to help quantify the contribution of each of the components.

## 2.2.2 Image-only Attention

Attention mechanism is only performed on the image and we do not use any question attention. Hence, the attention mechanism in the question component is replaced by a single layer perceptron (dense layer) to transform the question feature vector into a new vector that has the same dimension as the image-attention vector.

$$\hat{v} = \tanh(W_q f_q + b_q) \quad (7)$$

The question corresponding attention variables,  $H^q$  and  $a^q$  are accordingly not calculated for this image-only attention model.  $H^v$  and  $a^v$  are still calculated as we are retaining the attention mechanism in the image component.

### 2.2.3 Question-only Attention

Attention mechanism is only performed on the question and we do not use any image attention. For this purpose of replacing the attention mechanism in the image component, a single layer perceptron (dense layer) is used to transform the image feature vector into a new vector that has the same dimension as the question-attention vector.

$$\hat{v} = \tanh(W_i f_i + b_i) \quad (8)$$

The image corresponding attention variables,  $H^v$  and  $a^v$  are accordingly not calculated for this question-only attention model.  $H^q$  and  $a^q$  are still calculated as we are retaining the attention mechanism in the question component.

## 3. Experimental Results

### 3.1. Dataset

We used the VQA 2.0 dataset for all the experiments. It is one of the largest and earliest datasets containing human annotated questions and answered on the MSCOCO dataset. The dataset contains 443,757 training questions, 214,354 validation questions and 447,793 testing questions. There are 4,437,570 training answers and 2,143,540 validation answers. The testing answers have not been provided. The answer-types include yes/no, number, and other open-ended. Each question has 10 free-response answers. We used the top 1000 most frequent answers as the possible outputs. For quantitative metrics, we used the same accuracy metric proposed by the VQA paper:

$$\min\left(\frac{\# \text{human labels that match that answer}}{3}, 1\right) \quad (9)$$

i.e, the answer is deemed 100% accurate when three or more of the ten human labels match the answer, otherwise partial credit is given. Following [3], we haven't used the evaluation metrics like BLEU because they are typically applicable for sentences containing multiple words. Moreover, such metrics were found to poorly correlate with human judgement for similar such tasks [4].

### 3.2. Model configuration and Setup

We used Tensorflow Keras to build our models. We used the Adam optimizer with a base learning rate of 0.0001 and

Method	Accuracy
VGG + LSTM	43.99
VGG + GloVe	44.05
VGG + BERT	42.81
ResNet + LSTM	44.34
ResNet + GloVe	43.89
ResNet + BERT	42.81
Question-only Attn (VGG)	42.95
Question-only Attn (ResNet)	43.42
Image-only Attn (VGG)	46.88
Image-only Attn (ResNet)	43.73
<b>Co-attention (VGG)</b>	<b>47.34</b>
Co-attention (ResNet)	40.95

Table 1. Performance of the models

set the batch size to 64 for VGG19-based models and 16 for ResNet50-based models. All the images are rescaled to 224 x 224 and the questions are integer encoded and converted to a fixed 24-length sequences. The models are trained for up to 50 epochs with early stopping if the validation loss has not decreased in the last 5 epochs.

For the BERT embeddings, model *uncased\_L12\_H-768\_A12* is imported from the Tensorflow library which has 12-layer, 768-hidden, 12-heads and 110M parameters.

### 3.3. Quantitative Results

Table 1 shows the accuracy results (calculated as eq - 9) on all the models. The results shown are on the combined set of open-ended and multiple-choice question-answer types. The corresponding training and validation loss curves are shown in Figures 4. For better visualization and comparison of the model training and validation, VGG-based and ResNet-based models are separately plotted in figures 5, 6

We performed hyperparameter tuning for the co-attention model. Different parameters for optimizers, dropout, LSTM units and the results are shown in the Table 2. However, these did not perform well.

We also observed that changing the input image size to 448x448 during the initial feature extraction step also didn't have much effect on the model's performance. In fact, the model trained on 224x224 image-extracted-features performed slightly better.

Model parameter	Accuracy
SGD	31.04
(best = Adam)	
LSTM units = 256	38.91
(best = 512)	
Dropout = 0.5	31.40
(best = 0.3)	

Table 2. Hyperparameter tuning

### 3.4. Qualitative Results

We show the predictions made by the models along with the confidence of the prediction on some of the inputs. The predictions on the test set are shown in Figure 3. The VQA 2.0 dataset doesn't provide the groundtruth answers for the test set, so there was no way to validate the performance other than to check the predicted answers manually. The predictions on the validation set are shown in Figure 4. We show the groundtruth answer for the images in the validation set to compare against the predicted answers.

## 4. Discussion

Table 1 shows the comparison of all the models. It is clear from the results that VGG based models performed better than ResNet based models on a majority. In case of the attention based models, evidently, the Co-attention model has performed better than all other models as expected. This clearly demonstrated the positive impact of attending to different regions of the image as well as different fragments of the question. However, it is interesting to note that the Image-only Attention model has only slightly less performance, whereas the Question-only Attention model performed worse than both. This shows that image-attention has a greater contribution than question-attention.

From the training plots in Figures 4 - 6, we can see that the attention based have a steady learning curve during training and also during validation. It is interesting to note that the Image-attention has lower validation loss than Co-attention model, although the image-attention saturated around 15-20 epochs.

As the test answers from the VQA 2.0 dataset were not accessible, we show the qualitative results on some of the inputs in Table 3. In the figure of the bear (question: what is the bear doing?), the attention models gave the more appropriate answers (sleeping, sitting, standing), whereas it is interesting to see some of the baseline models had "snowboarding" and "skiing". It is funny that most of the models predicted that the teddy can read in real (image-3). This could be because of the text written on the image (which says "even teddy bears love a great book").

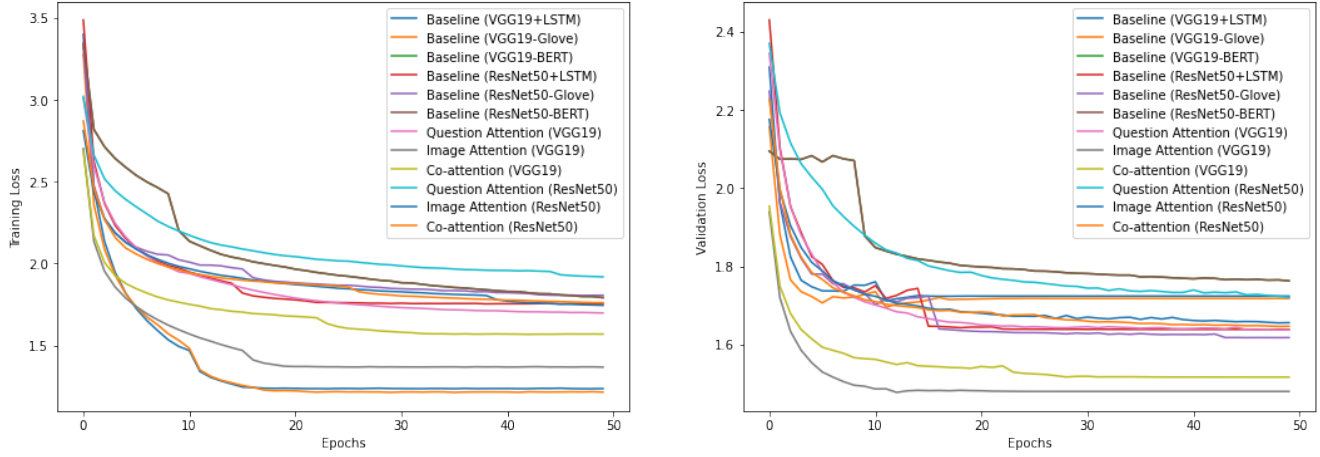


Figure 4. Training and Validation loss plots for the models corresponding to Table 1

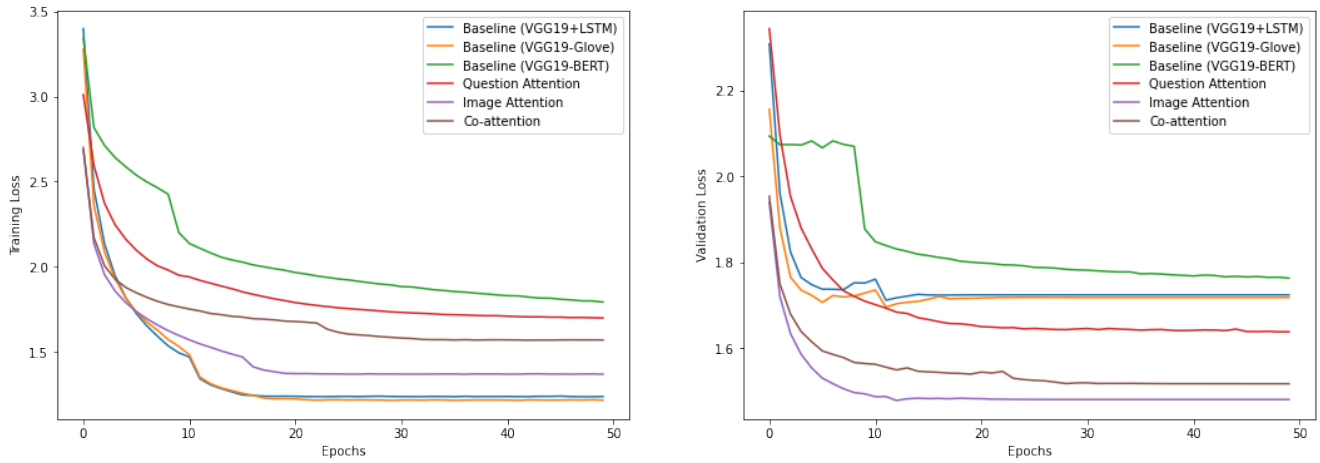


Figure 5. Training and Validation loss plots for the models with VGG features from Table 1

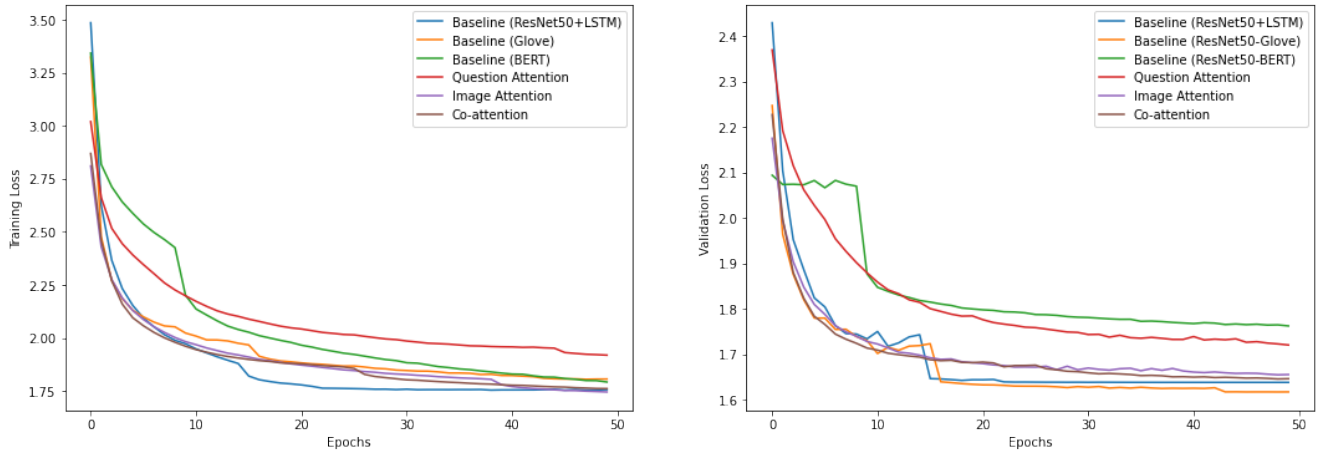


Figure 6. Training and Validation loss plots for the models with ResNet features from Table 1






Model					
Question	Is that a normal sized skateboard?	What is the bear doing?	Can teddy bears really read?	How many levels does this bus have?	How many compartments are in the container?
VGG + LSTM	yes (13.18)	snowboarding (5.83)	no (25.66)	1 (60.5)	1 (33.68)
VGG + GloVe	yes (60.84)	sitting (19.45)	no (54.24)	1 (36.97)	1 (43.76)
VGG + BERT	yes (55.78)	no (54.30)	yes (58.85)	yes (58.69)	yes (59.48)
ResNet + LSTM	skiing (19.54)	skiing (21.56)	right (14.68)	1 (50.80)	1 (24.74)
ResNet + GloVe	yes (54.02)	standing (18.23)	yes (62.27)	1 (27.62)	1 (28.91)
ResNet + BERT	yes (55.78)	no (54.30)	yes (58.85)	yes (58.69)	yes (59.49)
Question-only Attn (VGG)	no (50.71)	eating (20.95)	yes (57.19)	2 (43.31)	1 (17.22)
Question-only Attn (ResNet)	no (51.59)	standing (15.68)	yes (50)	1 (41.03)	2 (12.39)
Image-only Attn (VGG)	yes (58.52)	sleeping (38.10)	yes (41.02)	2 (44.86)	1 (33.25)
Image-only Attn (ResNet)	no (56.39)	sitting (22.21)	yes (49.97)	2 (51.02)	2 (21.31)
Co-attention (VGG)	fence (5.01)	sleeping (38.47)	yes (60.67)	2 (56.80)	1 (36.51)
Co-attention (ResNet)	no (52.21)	standing (20.34)	yes (8.97)	1 (41.97)	2 (23.19)

Table 3. Prediction of the models on the test set. The table shows the predicted answer along with the confidence of the prediction by each model.

The qualitative results on the inputs from validation set are shown in Figure 4.

From the examples, it looks like most of the models were able to correctly answer the yes/no type of questions, countable questions, color-related questions as well. However, attention models could not predict correctly when asked about more than one color (image-3: what colors are the train?). They were only able to identify one color, either blue or white. It is surprising to see that for the image-4 where a woman with short hair is carrying a bouquet, most models predicted that it was a man. Could this be a bias in the data where short hair is stereotyped as man? Another kind of bias can also be observed from the last image of the bird. When asked where the bird is, most models predicted as "branch". This could come from the fact that most birds live on trees, so the models are biasing birds to trees/branches.

There is still scope of improving and better capturing

the interaction between query phrasing representation and visual representations. Datasets should also be balanced to ensure the models are not inherently learning the biases in the data. Visualizing the attention maps of image and question would have been helpful in understanding what the models are actually focusing on, but due to time constraints we could not perform that analysis. The implementation, trained models and more results will be made available at [https://github.com/divyayjyothig/VisualQuestionAnswering\\_ComparativeStudy](https://github.com/divyayjyothig/VisualQuestionAnswering_ComparativeStudy).

## 5. Contributions

- Baseline models with VGG-19 and ResNet-50
- Co-attention on ResNet-50 image features
- Image-only attention model on VGG-19 and ResNet-50 image features





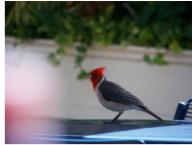
Model					
Question	what is the man holding up to his head?	what color is the horse?	what colors are the train?	which person is carrying a bouquet?	what is the bird perched on?
Actual answer	umbrella	brown	blue and gray	bride	table
VGG + LSTM	umbrella (28.78)	brown (85.63)	blue and yellow (17.0)	woman (21.99)	water (11.69)
VGG + GloVe	umbrella (20.62)	brown (73.02)	blue and white (19.36)	woman (33.79)	water (19.27)
VGG + BERT	yes (59.0)	yes ( 58.54)	yes (58.77)	yes (59.53)	yes (58.25)
ResNet + LSTM	nothing (15.37)	white (47.61)	red and white (10.12)	boy (26.16)	branch (20.37)
ResNet + GloVe	hat (12.03)	white (40.30)	white (11.63)	man (40.83)	branch (37.06)
ResNet + BERT	yes (59.0)	yes ( 58.54)	yes (58.77)	yes (59.53)	yes (58.25)
Question-only Attn (VGG)	umbrella (17.99)	brown (48.57)	red and white (9.32)	man (32.07)	branch (25.77)
Question-only Attn (ResNet)	phone (8.96)	brown (45.41)	white (11.2)	man (31.39)	grass (4.26)
Image-only Attn (VGG)	umbrella (16.95)	brown (59.99)	blue (9.52)	woman (36.25)	branch (9.05)
Image-only Attn (ResNet)	phone (8.46)	brown (49.04)	white (8.7)	man (31.87)	branch (11.41)
Co-attention (VGG)	umbrella (16.05)	brown (60.91)	blue (7.83)	man (30.48)	branch (19.07)
Co-attention (ResNet)	umbrella (12.20)	brown (54.92)	red (12.90)	man (19.46)	branch (28.19)

Table 4. Prediction of the models on the validation set. The table shows the predicted answer along with the confidence of the prediction by each model.

- Question-only attention model on VGG-19 and ResNet-50 image features
- Evaluation and visualization (equal contribution)

## References

- Question-only attention model on VGG-19 and ResNet-50 image features
  - Evaluation and visualization (equal contribution)

## References

  - [1] <https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/#what-is-bert>. [Online; accessed 6-May-2022].
  - [2] <https://medium.com/@harshareddykancharla/visual-question-answering-with-hierarchical-co-attention-augmented-transformer-5d6c2e2f1061>. [Online; accessed 6-May-2022].
  - [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.
  - [4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.
  - [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
  - [6] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061, 2016.
  - [7] Jeffrey Pennington, Richard Socher, and Christopher D. Man-

ning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.