Visual Question Answering via Dense Captioning

Vinay Sriram
Department of Computer Science
Stanford University
vsriram@stanford.edu

Neel Ramachandran
Department of Computer Science
Stanford University
neelr@stanford.edu

Emmie Kehoe
Department of Electrical Engineering
Stanford University
emkehoe@stanford.edu

Abstract

In recent years,Visual Question Answering (VQA) (1) has become a prominent research problem at the intersection of natural language processing and computer vision. While historical approaches to visual question answering have relied on end-to-end models trained using a corpus of images with associated {Question, Answer} pairs (2), this paper analyzes an approach that performs answering in two stages. Our approach consists of independently training an image understanding task and a pure-language question answering task. This approach is motivated by a desire to provide a greater degree of interpretability to VQA, by effectively translating an image to natural language and then answering the question based on the translation, rather than simply outputting an answer. The first stage is a captioning architecture adapted from (7) that performs region-level captioning and consolidation of region captions into dense captions. The second stage uses the Bi-Directional Attention Flow (BiDAF) (4) model to perform question answering on the passage constructed from these region descriptions. We limit the scope of our experiments to question-answer pairs with one-word answers. Experiments show that the independent question-answering model achieves a 60% exact-match accuracy, and the best captioning architecture used in conjunction with this question-answering model achieves an 11% exact-match accuracy on a test set of question-answer pairs sampled from the Stanford Visual Genome (3) dataset.

1 Introduction

Visual Question Answering (VQA) is a task that relies on both a perceptually rich understanding of dense visual scenes, and an understanding of how natural language is used to communicate the relationships between objects in these scenes. In order to exploit this fact, we use the Stanford Visual Genome Dataset (3), which was designed with the broad objective of grounding visual concepts to language, thereby facilitating a deeper understanding of the visual world. The Visual Genome dataset was designed to provide a multi-layered understanding of pictures. Unlike other datasets such as Flickr30K and MS-COCO, Visual Genome offers complex information such as individual region descriptions. In this paper, we leverage these region descriptions to implement and analyze a multi-stage model for Visual Question Answering (VQA). Our VQA system takes as input an image and a natural language question about the image, and produces a single-word natural language answer as the output. Within the VQA system, there exist two stages. The first stage is a captioning architecture adapted from (7) that takes an input image, splits the image into regions, captions each region, and consolidates these captions into a dense caption.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Our extraction of region bounding boxes effectively mimics the concept of attention because it allows the captioning model to focus on a targeted area within an image that captures individual object relationships rather than the entire source image. An example image containing a person sitting on a couch, a dog on a rug, and an open door could have three individual region descriptions: {person on couch, dog on rug, open door} that will each be individually captioned. After all of the region captions are complied, they are then fed into the second stage of our system. The second stage uses the Bi-Directional Attention Flow (BiDAF) (4) model to perform question answering on the passage constructed from these region captions. This piece of the system takes as input a collection of phrases that describe an image and a question and outputs a natural language response to the question.

We use a dataset that contains 10,000 images with 200,000 region bounding boxes and corresponding labels from the Visual Genome dataset. Each of the images were paired with on average 30 {Question, Answer} pairs. Training both the captioning model and the question-answering model independently on such images, we are able to achieve reasonable accuracy. Specifically, the individual captioning model reaches an accuracy of 60% on visual genome data (answering single word queries from region descriptions). Furthermore, the captioning model produces captions that contain the correct answer 15% of the time. Chaining these two models together results in a one-guess question answering accuracy of 11%, a two-guess accuracy of 14%, and a three-guess accuracy of the upper bound of 15%. While these results do not represent state-of-the-art accuracy in visual question answering, the primary contribution is to show that, with more work in the area (in particular a more robust captioning model), a two-stage approach to answering questions from visually rich images can be a viable alternative to the end-to-end models that have been dominant in VQA to date.

2 Related Work

2.1 Captioning

Many end-to-end models have been proposed for VQA tasks, but few actually use the Visual Genome dataset. We focus in particular on a recent paper (7) by Anderson et al. which does uses the Visual Genome dataset for question answering and captioning. Both of the tasks are solved by Anderson through end-to-end systems. We focus first on Anderson's captioning system, which consists spatial region embeddings and a two-layer LSTM. The first LSTM layers offers 'soft' top-down attention on spatial region embeddings, weighting each spatial feature's importance towards predicting the next word in the caption at each time step. The second LSTM recieves this weighted sum as input, and performs the actual language generation to produce a caption. Mathematically, we can represent the input to the attention LSTM, x_1, as in equation 1. Here h_{t-1}^2 is the output of the language LSTM, $\bar{v}$ is the average spatial feature embedding, and W_t is the word-embedding of the previous time-step $(t-1)$ predicted word (which serves as input in the current timestep t). The language LSTM layer then has the following input, which consists of concatenating the attention LSTM output h_t^1 with the weighted region embeddings, as shown in equation 2.

$$x_1 = [h_{t-1}^2 \;\; \bar{v} \;\; W_t] \tag{1}$$

$$x_2 = [h_t^1 \;\; \sum_{i=1}^{K} w_i v_i] \tag{2}$$

Now we discuss several recent approaches to generating spatial embeddings of regions. Here, we describe two. A first approach by Jin et al. (8) uses selective search to identify salient image regions which are filtered with a classifier and then resized. The resulting images are encoded with a convolutional neural network (CNN) wherein the CNN encoding serves as input to an image captioning model with attention. A second approach proposed by Anderson uses Faster R-CNN in conjunction with ResNet-101 to identify salient image regions. ResNet-101 performs object detection, and faster R-CNN is used as region proposal network. Rather than using the final bounding boxes, an intermediate convolutional layer of faster R-CNN is used. This layer is of depth K (for K different embeddings, as in the above equations), and each embeeding is concatenated into a feature vector. Although the second method is effective for producing high quality captions, for the purpose of this project, we simplify this method by replacing Faster R-CNN with the pre-defined relationship

bounding boxes detailed in the Visual Genome dataset. This serves the purpose of allowing us to focus on the NLP components of the project, and reduce how dependent the project becomes on state-of-the art vision and region proposal systems.

2.2 Question-Answering

In the second stage of the model designed to answer a query, Anderson proposes a well-known joint multi-modal embedding of the question and image, followed by a prediction of regression scores over a set of candidate answers. Within the network, a special case of highway networks is implemented to learn non-linear transformations. This particular highway network uses gated hyperbolic tangent activations which have shown an empirical advantage over traditional ReLU or tanh layers (7).

Because our approach to question-answering is done on dense captions (pure text) rather than directly from images, we instead opt for the Bi-Directional Attention Flow (BiDAF) network (4). The BiDAF network defines a hierarchical multi-stage architecture for modeling a context paragraph at variable levels of specificity. The three main embedding types used in BiDAF are character level embeddings, word level embeddings, and contextual embeddings. Then, these embeddings are fed into a sequence of layers consisting of an attention flow layer, a modeling layer, and an output layer. At an abstract level, the network uses bi-directional attention over the embeddings to find an appropriate context representation to a given query (4).

3 Approach

3.1 Captioning Stage

Our approach to dense captioning is derived from the techniques used by Anderson et al. (7) on the Visual Genome dataset. The Anderson model uses object bounding boxes and spatial embeddings to produce captions with an average description length of 15 words per image. Because we wish to more densely caption our input images, we implement a region-specific variation of this model. The first phase of our model takes region proposals from Visual Genome (note that, in production systems, this would be done through a region proposal algorithm such as selective search or R-CNN, but the NLP focus of our project motivates the use of ground truths even at inference time).

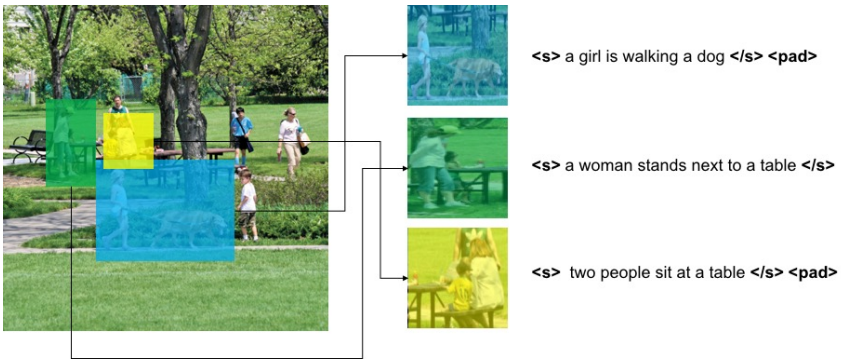


Figure 1: The original raw image before region pre-processing is given on the left, while the extracted region bounding boxes and associated captions are given on the right. Note that each caption is padded to max length, and each region is compressed to a square size accepted by ResNet.

$$x = [h_{t-1} \;\; v \;\; e_t]^T \tag{3}$$

$$y_t = \textbf{softmax}(Wh_t + b) \tag{4}$$

After regions are extracted, the second phase of our model uses ResNet to produce an embedding of each region v. Let e_t denote the embedding of the word predicted at time t (e_0 is simply the embedding of the start token). Then, the input to our LSTM model at time t is given by equation 3.

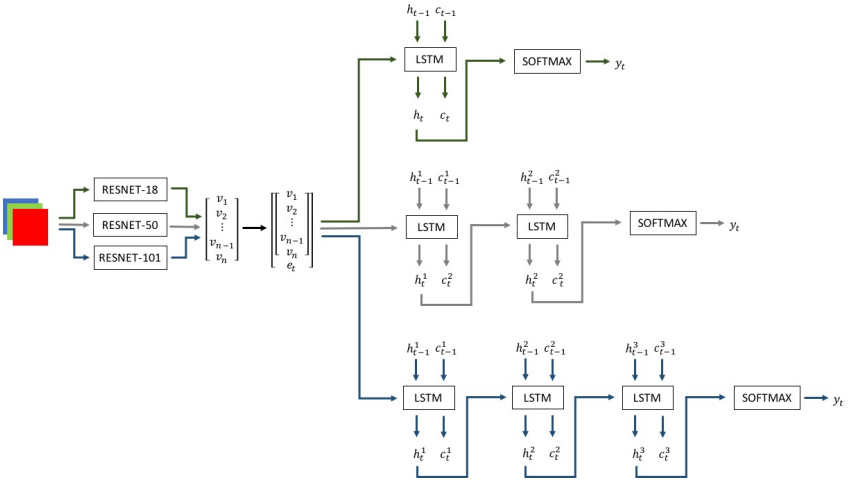


Figure 2: Various Captioning Models

The output probability distribution over the next word is given by equation 4. The most probable word at time t is fed as the next word embedding (at time $t + 1$) in greedy decoding at inference time.

At training time, we use teacher forcing and simply feed in the ground truth caption word at each time. As shown in figure 2, we evaluate three specific architectures for captioning. All three architectures use 300-dimensional word embeddings and 1000-dimensional image embeddings as inputs to the LSTM. Furthermore, each LSTM hiden layer has 1000 units, and the fully connected layer has $1000 \times |V| + |V|$ parameters for the weight and bias terms, where $|V|$ denotes the size of the vocabulary. This is typically on the order of tens of thousands in our experiments. The three architectures vary in the number of LSTM layers as well as the type of ResNet model used. Specifically, the simplest architecture uses ResNet-18 and consistes of 1 hidden LSTM layer, the medium complexity model uses ResNet-50 and consists of 2 hidden LSTM layers, and the most complex model uses ResNet-101 and consists of 3 hidden LSTM layers.

3.2 Question Answering

Once we have completed the captioning stage, we reduce the visual question answering task into a textual question answering task, similar to the SQuAD challenge. For a given image I, we aggregate the full set $R_I = r_1...r_n$ of region descriptions generated in the captioning stage, and treat R_I as the context for a question-answer pair associated with the image. At this stage, we assume that the answer to the posed question belongs to R_I, meaning that the generated region descriptions are rich enough to answer the question. We also limit the set of question-answer pairs to those where the ground-truth answer is one word. Thus, our task consists of selecting‘ a word from the set R_I to answer the posed question. We use an open-source PyTorch implementation (9) of implementation of BiDAF to learn the answers given a question and the region description context.

4 Experiments

4.1 Data

4.1.1 Captioning Model

From the the Visual Genome Dataset, we extracted a subset of raw images and their corresponding region descriptions and bounding boxes. In total, we extracted 200,000 bounding box region pairs and on average, there are about 200 region bounding boxes per raw input image. The specific partitioning we used was a train:val:test ratio of 160,000:20,000:20,000, though for periodic validation we randomly sampled the validation set by 10x for faster evaluation speed. All of the extraction, preprocessing, and data subset generation code (from the Visual Genome dataset) was written by us.

4.1.2 Question-Answering Model

To train the question-answering model, we consider the full dataset of region descriptions and question-answer pairs on the 108,077 images and approximately 1.4 million question-answer pairs in the dataset. After extracting the question-answer pairs with one word answers that are present in the ground-truth region descriptions, we are left with 504,671 question-answer pairs, from which we choose a random subset of 120,000 question-answer pairs that form our dataset. Finally, we use a 100,000:20,000 train:val ratio. Note that the 20,000 region descriptions and their associated question-answer pairs from the captioning model are also used here to test the multi-stage model.

4.2 Evaluation Methods

4.2.1 Captioning Model

To evaluate the captioning model, we compute BLEU scores of the predicted captions with respect to reference translations on our validation set. We restrict the BLEU score computation to 1-grams and 2-grams, with equal weighting $\lambda_1 = 0.5, \lambda_2 = 0.5$. We also note that the BLEU scores represent an intermediate evaluation of our multi-stage VQA model. As such, the evaluation is not necessarily indicative of our ultimate results; our captioning model could successfully capture the salient features of the image for the question-answering model without achieving a high BLEU score. Thus, using generated captions on a test set, we compute (as an upper-bound for question-answering accuracy) the proportion of captions for which the answer word truly exists in the caption. This serves as a more clear metric for how effective the captions are at the task of answering the question.

4.2.2 Question-Answering Model

For this paper, we evaluate question answering using top-k accuracy. Specifically, we evaluate the percent of questions for which the top guess is the correct answer ($k = 1$), the percentage for which the correct answer is in the top two guesses ($k = 2$), and the correct answer is in the top three guesses ($k = 3$). The model's answer $\hat{y}$ is said to be "correct" if it exactly matches one of the ground-truth answers y; note that when $k = 1$, the metric is equivalent to exact-match accuracy. This is the evaluation metric used in the Ranjay et. al baseline (3). In this paper, we do not consider the commonly-used F1 measure, since we our limiting ourselves to one-word question answers.

4.3 Experimental Details

4.3.1 Captioning Model

In order to develop a robust captioning model, we first perform hyper-parameter tuning. We evaluated the various architectures at different learning rates, specifically considering the 9 combinations associated with training each of the three architectures at each learning rate in the set $\{10^{-2}, 10^{-3}, 10^{-4}\}$. While training, we periodically computed the validation loss after every 50 iterations. we consider a limited training regime of 3000 iterations, and tracked the total validation set loss over this span. The validation loss was used to determine which model to deploy for wider training. While certain functional elements were adapted from (7) and from assignments 4 and 5, we wrote the vast majority of the code for the caption model ourselves (using torchvision.models for ResNet).

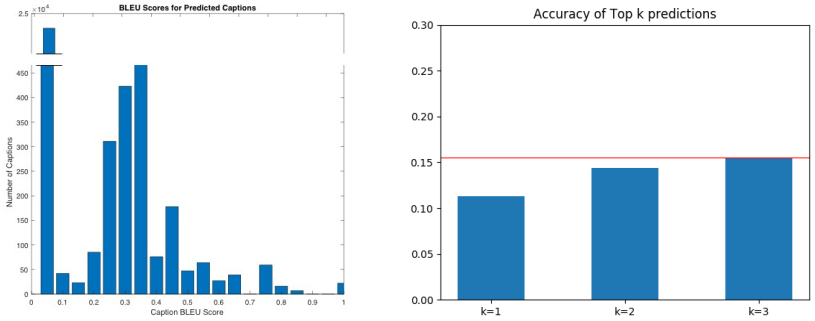
4.3.2 Question-Answering Model

We run BiDAF using character-level embeddings of size 10 and pre-trained GloVe embeddings of size 100. As we only consider question-answer pairs with one-word answers, we make slight modifications to the output layer of BiDAF. Instead of outputting probability distributions p_1 and p_2 over the starting and ending indexes of the answer, we limit the output to a single probability distribution p_1.

One challenge of treating the region descriptions in the Visual Genome dataset as our ground-truth context pagragraph in the question-answering model is that, unlike data in the SQuAD challenge, for example, we are not given the index of the ground-truth answer in our context paragraph. Further, because the Visual Genome region descriptions are rich and thorough, our dataset often contains examples where the answer word appears frequently in the context paragraph. For example, a context paragraph might contain the phrase *"The grass is green, The girl's shirt is green,...,There are green*

trees" and ask the question *"What color are the trees?"* In assigning the index of the answer, we simply search the context paragraph and choose the index at which the answer word first appears. This introduces challenges for training our question-answering model effectively, because we cannot be certain that the assigned index is in fact the true index of the answer. In the example above, for example, we would choose an answer index that is not associated with the true context of the answer, causing the model to learn false associations. To address this issue, we clip the context paragraphs to length 100 from an average length of 250. This strategy alleviates the issue by reducing the number of times the answer word appears, but does not eliminate it entirely.

4.4 Results



(a) Captions Test Set BLEU Scores

(b) Question Answering Accuracy

Figure 3: Quantitative Results on Captioning and Question Answering

4.4.1 Captioning Model

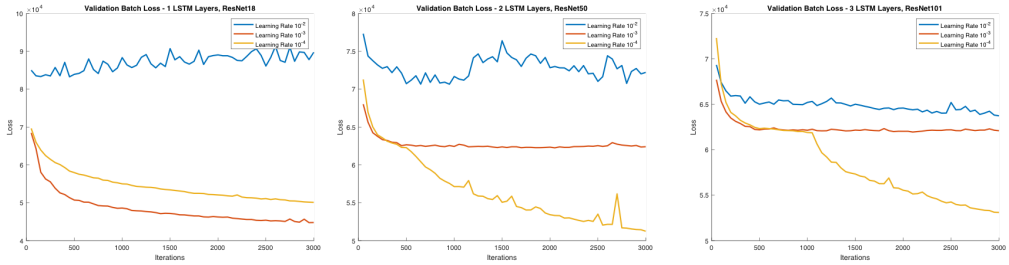


Figure 4: Validation Loss for Various Experiments

Figure 4 illustrates the validation loss graphs for the three models at different learning rates. We train over a batch size of 32. Evidently, we see that the loss drop is the most desireable for the slowest learning rate, as the other two are prone to stagnate sub-optimally. Furthermore, we see that the rate of decline for the loss in the most complex model (3 LSTM layers with ResNet-101) is the fastest and most consistent, and so this model will likely do the best to prevent overfitting on the training set. Therefore, we select the 3-layer ResNet-101 model at a learning rate of 10^{-4} to train for an extended period of time. After training this best model for 50,000 iterations, we evaluated the model on the test set regions set aside in the original data partitioning. We evaluated the BLEU score of each caption using the ground-truth captions as references. We noted in particular that the BLEU scores were very low, as indicated in figure 3a. Specifically, we found that the mean BLEU score was 0.01, and the lowest BLEU score bin [0, 0.5] contained the vast majority (92%) of the captions. Upon visually inspecting the results, we discovered that this was because many of the captions were describing different elements of the region than the ground truth. As a concrete example, consider the following three region descriptions {the background sky is blue, the boat color is green, the ocean color is blue}. The region crops associated with these three descriptions could be almost identical, each depicting a boat on an ocean with a sky in the background. Despite the fact that each of the captions is a valid

caption, only one will be the ground truth for a given region crop. This phenomenon results in low BLEU scores on average. However, it also directly implies that the BLEU score is not necessarily an accurate metric for the validity of a caption. Therefore, as explained in the subsequent section, we use the metric of answer presence in the caption to evaluate the question-answering potential.

4.4.2 Question-Answering Model

We first run BiDAF given the Visual Genome ground-truth region descriptions to train the model over 20 epochs, achieving an exact-match accuracy of 60.33% on the validation set. We then evaluate the multi-stage model, feeding the region descriptions generated by the captioning model on the test set to the trained BiDAF model. First, we gauge how effective our captions could be at question-answering. We find that 15.55% of the generated test captions have the answer word within them, representing reasonable captioning accuracy, and providing us with an upper bound on the overall possible accuracy of our multi-stage model. We then evaluate our multi-stage model using top-k accuracy, and find exact-match scores of **11.3%** ($k = 1$), **14.4%** ($k = 1$), and **15.4%** ($k = 3$), as shown in figure 3b. Thus, the model is effective when allowed one prediction, and reaches the upper accuracy bound when allowed three predictions.

5 Analysis

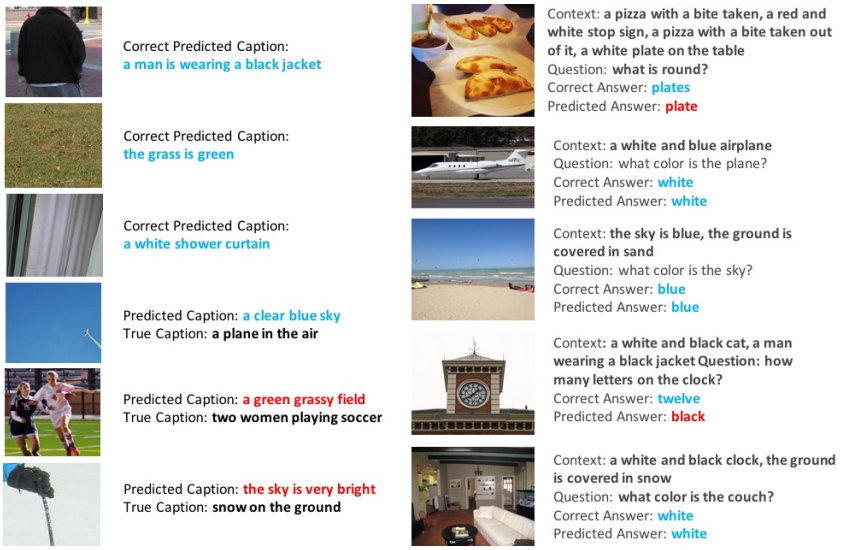


Figure 5: Qualitative Results for Image Captioning (Left) and Visual Question Answering (Right)

In qualitatively analyzing results, we found that the captioning model primarily produces captions in the following three categories: (1) Low BLEU score and invalid, (2) Low BLEU score but valid, and (3) high BLEU score and valid. It is clear that the majority of the captions fall into the first two classes. However, it is difficult to isolate the percentage that fall into the first class vs. the second. Examples are presented in figure 6. The top three examples are of perfectly predicted captions. These clearly have high BLEU scores, and the captions accurately match the contents of the regions. The next two examples are of captions in which the BLEU scores are low (near 0), but the captions may be considered valid. For example, case 4 depicts a plane flying through a clear blue sky, but the ground truth caption is “a plane in the air." Thus, when the model predicts “a clear blue sky" this caption has a low BLEU score. Another example is of the two women playing soccer in a grassy field. Once again, the predicted caption is valid but not close in terms of n-grams to the ground truth. Finally, there are cases of poor captions that also have low BLEU scores. The final example in the left column of figure 6 depicts this. Humans, upon looking at this picture, can infer that snow is on the ground because of the skier’s arm and ski pole, but the model is absent this contextual intuition, and so there is no way for it to distinguish the snow from a bright sky. Thus, qualitatively the captions overall appear reasonable, even though this does not manifest quantitatively.

In analyzing the results of the multi-stage VQA model, we find that it is adept at answering "What color is...?" questions, as our captioning model is effective at extracting colors from the regions. We achieve 25% accuracy – more than twice our overall accuracy – on this sub-category of question. We also find that by considering one-word question-answer pairs, many images contain the same abstract question. For example, the question "When was this photo taken?" with answer "Daytime" represents nearly 10% of our test set, but the word "Daytime" is never represented in our test captions, and thus we achieve 0% accuracy on this sub-category of question. This type of question illustrates one major shortcoming of our multi-stage approach, which is that abstract questions such as the above break our assumption that we can extract an answer from region descriptions. Though the time of day is clearly a salient feature of any image, it is not well-captured by any one region and we would not necessarily expect a more sophisticated image captioning model to produce a caption including that word. Thus, even simple abstract questions pose a large challenge for the multi-stage model.

6 Conclusions

The conclusions which may be drawn from this project are several. First and foremost, we have demonstrated that multi-stage VQA can be a viable technique (if improved upon) to perform question answering on visually dense scenes. We argue that the most notable advantage of multi-stage VQA over traditional end-to-end models is that it provides a greater degree of interpretability. Specifically, the multi-stage model allows us to better analyze sources of error in question-answering by translating an image to natural language before making a prediction, as opposed to an end-to-end model which outputs an prediction, and is difficult to interpret if it makes a mistake. Meanwhile, the multi-stage approach has its own set of challenges, including answering abstract questions that are not well-defined by image regions or captions, as described above. Second, we have shown that, due to high flexibility in image captioning, captions with low BLEU scores relative to ground truths can still perform effectively in question-answering. Specifically, such captions do in fact contain the correct answer word for a significant proportion of the questions in the dataset. Finally, we have shown that within three guesses it is possible to reach the accuracy upper bound for the question answering stage, and within one guess to reach a substantial fraction of this upper bound.

6.1 Captioning Model Future Work

There are two primary directions for future work that we could potentially explore. The first direction involves the input data for the captioning model. We noticed that our current dataset contains regions with over-extended captions. For instance, an image displaying a sidewalk made of bricks would be accompanied by the caption “a sidewalk made of bricks next to a street." Using Visual Genome’s relationship data (rather than just region descriptions) could solve this problem because it provides two bounding boxes, one for each item mentioned in the relationship. We could use both of these bounding boxes, passing them through a ResNet architecture to obtain a pair of encodings to feed into later stages of the model. The second direction for improvement involves further segmenting the regions. Rather than attempting to learn region-level embeddings directly, we could attempt instead to learn region-based object-level embeddings and intelligently consolidate such embeddings into a more intelligent region embedding. This would also likely improve captioning substantially.

6.1.1 Question-Answering Model Future Work

There are two directions for future work with our question-answering model. One aforementioned challenge with our approach to the VQA task is that we inherently do not have ground-truth indices of the answer in our context paragraphs at training time. Improving our naive pre-processing step of choosing a the first index that matches the answer might lead to better overall performance. The most effective – albeit labor-intensive and time consuming – solution to this issue would be manually labeling the starting indices. Other computational strategies might involve using heuristics such as matching term frequencies in the question to subsections of the context paragraph in order to more accurately identify the correct indices. Secondly, the model could easily be extended to consider question-answer pairs with multi-word answers by re-introducing the distribution p_2 of the ending index in the output layer of our BiDAF model. Pursuing this avenue may in fact help solve the first mentioned issue, as multi-word answers are less likely to appear multiple times in the context paragraph, while also making the overall task more difficult.

7 Additional Information

Our mentor is Sahil Chopra.

References

[1] Antol, Stanislaw, et al. “VQA: Visual Question Answering." 2015 IEEE International Conference on Computer Vision (ICCV), 2015, doi:10.1109/iccv.2015.279.

[2] Goyal, Yash, et al. VQA Dataset. VQA: Visual Question Answering, Virginia Tech, 2019, https://visualqa.org/download.html.

[3] Krishna, Ranjay, et al. “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations." International Journal of Computer Vision, vol. 123, no. 1, 2017, pp. 32–73., doi:10.1007/s11263-016-0981-7.

[4] Seo, Minjoon, et al. “Bidirectional Attention Flow for Machine Comprehension." International Conference on Learning Representations, 2017.

[5] Karpathy, Andrej, and Li Fei-Fei. “Deep Visual-Semantic Alignments for Generating Image Descriptions." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, doi:10.1109/cvpr.2015.7298932.

[6] Xiong, Caiming et al. “Dynamic Coattention Networks for Question Answering." 2017 International Conference on Learning Representations, 2017.

[7] Anderson, Peter et al. “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[8] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv preprint arXiv:1506.06272, 2015.

[9] galsang. Re-implementation of BiDAF(Bidirectional Attention Flow for Machine Comprehension, Minjoon Seo et al., ICLR 2017) on PyTorch. https://github.com/galsang/BiDAF-pytorch, June 2017.