APPROVAL SHEET

Title of Thesis: Attention Correction mechanisms in Visual contexts in Visual Question answering mechanisms

Name of Candidate: Komal Sharan
M.S. in Computer Science,
2018

Thesis and Abstract Approved: __________________________
Dr. Tim Oates
Professor
Department of Computer Science and
Electrical Engineering

Date Approved: ________________________________

ABSTRACT

Title of Thesis:

Attention correction mechanism of visual contexts in visual Question answering mechanisms.

Komal Sharan, M.S. Computer Science, Dec 2018

Thesis directed by: Dr. Tim Oates, Professor
Department of Computer Science and
Electrical Engineering

To answer a question about an image or to merely describe an object in an image for answering, the current visual question answering systems have been augmented with attention mechanisms. The visual question answering mechanisms before the advent of attention mechanisms worked on the principle of training over a combination of image feature vectors and question and answer embeddings. Attention mechanisms like stacked attention networks and hierarchical co-attention attention mechanisms, help to figure out which parts of the image to attend but hardly emphasize on correcting attention. We propose a mechanism for correcting visual attention by using the concept of saliency of parts of the image being attended to. We primarily use a study of how the gaze of humans shifts over an image can help us improving the attention generated by introducing an auxiliary loss in a standard stacked attention network pipeline. For this mechanism, we use a dataset known as the VQA HAT dataset which is a large-scale collection of images containing regions explored by humans, and we use this dataset for further augmenting the work.

Keywords:Stacked Attention Networks, Human attention maps , VQA-HAT

Attention correction mechanisms in visual contexts in visual question answering

by

Komal Sharan

Thesis submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
M.S. Computer Science
2018

Dedicated to Dolly Sharan and Sanjay Sharan

ACKNOWLEDGMENTS

I would like to thank my advisor Dr.Tim Oates for his support and I am greatly indebted to his continuous encouragement and patiently mapping out my academic progress over the past year. I would like to thank Mr. AshwinKumar Ganesan for his incredible support and always guiding me towards solutions that led to this work. His ideas have always moved the thesis in a very positive direction and I will always be grateful to him for mentoring me. I would also like to thank Dr. Hamed Pirsiavash for giving his valuable suggestions and Dr. Tim Finin for agreeing to be a part of my committee.

I would also like to express my deepest gratitude to my mother Mrs. Dolly Sharan, my father Mr. Sanjay Sharan , my brother Rohit sharan,my grand parents for their unconditional support and my dearest friend Ajinkya Kolhe for his unfailing guidance during this journey of my masters program.

TABLE OF CONTENTS

LIST OF TABLES

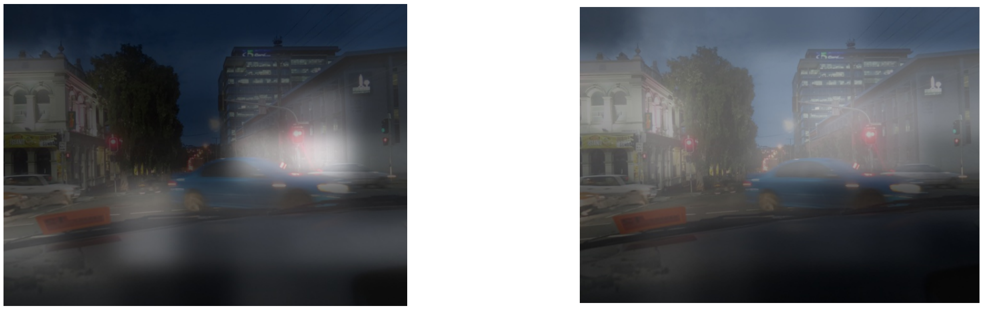
LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Overview and Motivation

Recently, a lot of impetus has been put on understanding scenes through contexts provided by natural language in AI systems. Attention mechanisms are an attempt to do this task. One may understand attention mechanisms as a replica of the steps taken by a human brain to reason about an image step by step. Attention mechanisms have played a role in improving AI tasks such as machine translation([Bahdanau, cho, & Bengio2016]), object detection ([Ba, Mnih, & Kavukcuoglu.2015]) and image captioning ([Xu *et al.*2015]) being a few of them. These advances however also bring up the differences between object detection and visual understanding. There has certainly been a marked improvement in these systems through attention mechanisms, but due to lack of enough contexts about the data for a system to learn the attention, the accuracies have a lot of scope of improvement. One of the famous statements by Des et al. is that ”machine-generated attention maps are either negatively correlated with human attention or have positive correlation worse than task independent saliency”([Jabri, Joulin, & der Maaten.2016]). The ability of a deep learning network to mimic human attention has been the most recent metric to improve attention mechanisms. This work further delves into perfecting such systems through learning dynamic weightings of input vectors using the human annotated attentions to act



(a) Attention generated attention

(b) Attention generated by the corrected model

FIG. 1.1: Attentions generated by two models

as an additional input to the attention mechanisms. In short, our work contributes to the correction of visual attention which can be used to improve the quality of the answers generated by the AI system for the given question about the image. To illustrate this by an example consider the figure 1.1 and the question asked about the given image being ”is it Sunny?”.Figure 1.1(a) shows attention generated by a standard attention network and figure 1.1(b) shows where the attention of the network should be if it were accurately generating attention maps.

1.2 Learning from human gaze data

Human gaze data has been collected in various ways to understand saliency, and there have been a lot of ways in which this data is being interpreted. The major challenge posed for a question answering system is when the question is based more on a semantic visual attribute rather than an object detection or a scene classification. There can be many interpretations about an image and recognizing such characteristics of data has been helped by collecting human gaze data. An example of this is when a human is asked to draw a boot, the general image would more or less have the same attributes, but when a human is asked to draw or pick an attractive person from a set of actors, it can vary across different people.([Murrugarra-Llerena & Kovashka2017]).

For our experiments, we have used a dataset annotated by humans which is a set of images emphasizing on the regions in the image important for answering a question about the image. This dataset is known as the VQA-HAT(Human Attention) dataset. To improve the currently existing attention networks for visual question answering systems, we use this dataset for a corrective mechanism used to improve the efficiency of the currently existing baseline models. There are various ways in which this dataset can be used to improve the existing model.

Chapter 2

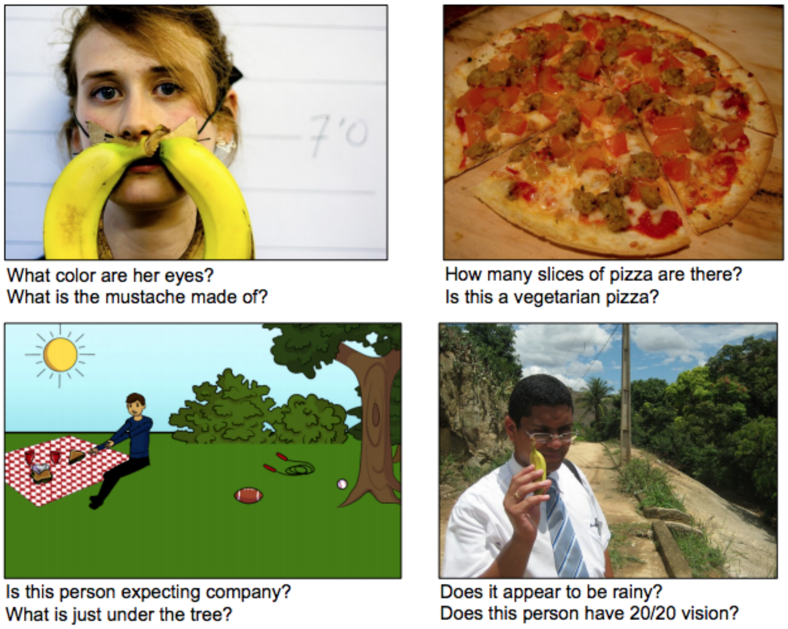
RELATED WORK

2.1 Visual Question Answering(VQA)

A visual question and answering system can be described as one which takes an image and a question as input and answers a natural language question about it. The very nature of such a system is multidisciplinary, as this system has to first understand the question and classify the type of the question and then answer it, for example, the answer to a ”how many” question has to be a number and the answer to an ”is this” question has to be a binary ”yes/no”([Jonathan Hui blog2017]). All of these questions query the image in a different manner for example for a how many question, there needs to be a object detection and for a binary ”yes/no” question needs to be a scene classification. This adds computer vision to the system.This reasoning about the image is the most critical part of such a system which adds knowledge representation reasoning (kr). A combination of computer vision, and knowledge representation reasoning makes it a multidisciplinary system. Shown below in Figure 2.1 ([Agrawal *et al.*2017]) are a few examples of visual question and answering.

2.1.1 Available datasets

Visual question answering datasets are hard to capture as there are many possibilities of types of questions and a large range of real-world scenarios to represent. A large number



What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

FIG. 2.1: Visual question and answering given images and questions

of datasets are a part of Microsoft common objects in context(MSCOCO) dataset.

One of the very first datasets released for visual question answering is the DAQUAR dataset which stands for ”DAtaset for Question Answering on Real-world images”. This dataset has 6794 training question-answer pairs and 5674 test question-answer pairs. The major drawback of this dataset is that it lacks the variety of different conditions of indoor and outdoor scenes. Hence it lacks the complexity required for capturing different scenarios. The second dataset largely used is known as COCO-VQA dataset. This dataset is significantly larger than the DAQUAR dataset as it contains 123,287 images from the COCO dataset. The uniqueness of this dataset comes from the fact that the creators of this dataset use an NLP algorithm to generate image captions. But just like any other NLP based dataset, this has drawbacks like grammatical errors. The dataset which is relatively larger than these is the VQA dataset. It has around 50,000 abstract images along with 204,721 images from the COCO dataset. This is the dataset we use for our experiments.

This dataset has 3 questions per image and ten answers per question.

2.2 Building blocks of VQA

2.2.1 Convolutional Neural Networks (CNN)

In the last few years, convolutional neural networks have been one of the most effective methods for visual recognition tasks like scene classification and object detection. There are many architectures based on CNNs which are used to extract features from images. VGG architecture is one such architecture which we use in VQA tasks. ([Singh2016]).

2.2.2 Long Short Term Memory Networks (LSTM)

LSTM's are a type of recurrent neural networks and facilitate in sequence based models by remembering the features inputs given in a sequence. Such models have been heavily used in machine translation and speech recognition activities. Particularly in VQA, they are used to comprehend the question. ([Singh2016]).

2.2.3 Word Embeddings

Word embeddings are dense word vectors used to represent words created by famous word2vec models like Google word2vec or Glove. Generally, the word vectors are of fixed length. Similar words have high cosine similarity.([Singh2016]).

2.3 Famous VQA pipeline architectures

2.3.1 Convolutional Neural Networks and bag of words model (BOW)

In this model, the image is passed through the VGGnet and the extracted features which are a 4096-dimensional vector are concatenated with the question vector which is a summation of the vectors of all individual words in the question. A softmax layer is attached at the end, and it gives us a probability distribution over the entire answer space.([Singh2016]).

Fig 2.2 shows the architecture of this system.

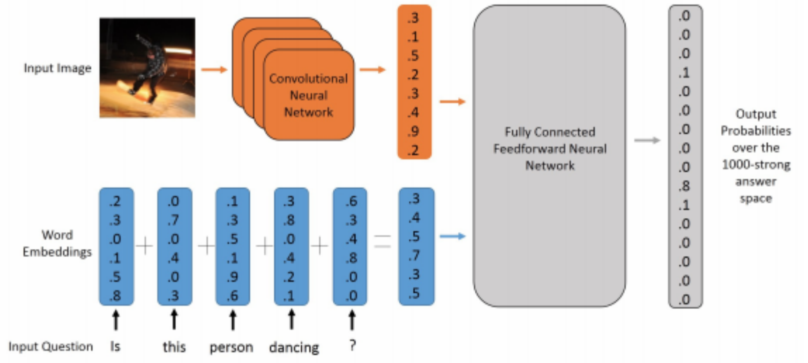


FIG. 2.2: VQA bag of words model

2.3.2 Convolutional Neural Network and Long Short Term Memory Network

The BOW model ignores the sequence in which the words are occurring. An LSTM-based question model captures the sequence by first converting the words into an embedding and then passing them into an LSTM. The final output of this LSTM is used as the question embedding which is concatenated with the image vector. Figure 2.3 shows how the different states of the LSTMs are generated according to the word sequence in a sentence to get one final word embedding to represent the sentence.([Singh2016]).

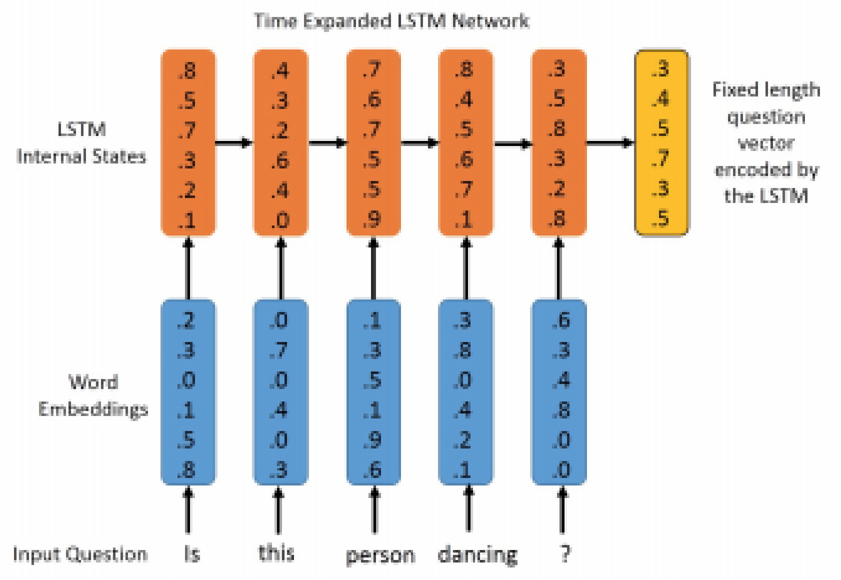


FIG. 2.3: Word embedding generated by LSTM

Overall these baseline models consider a combination of image features and word embeddings of questions and answers. The second category of baseline models includes the contextual and spatial understanding of the image features as training features for the VQA model that is spatial attention is taken into consideration when using image features for training. For example, if the question is ”what is the color of the table” this mechanism considers the visual elements around the table and suppresses others. These models are called attention based models.

2.4 Attention in VQA systems

2.4.1 What is attention?

In the previous section, attention-based VQA models were explained. The goal of attention mechanisms is to focus on relevant areas of the image to answer a question. In the last few years a number of attention models have been proposed for visual question and answering that generate attention maps along with an answer to highlight the part of the image that is relevant for answering the question about the image. Figure 2.4 ([Yang *et al.*2016]) shows an example of how attention looks in a VQA set up.

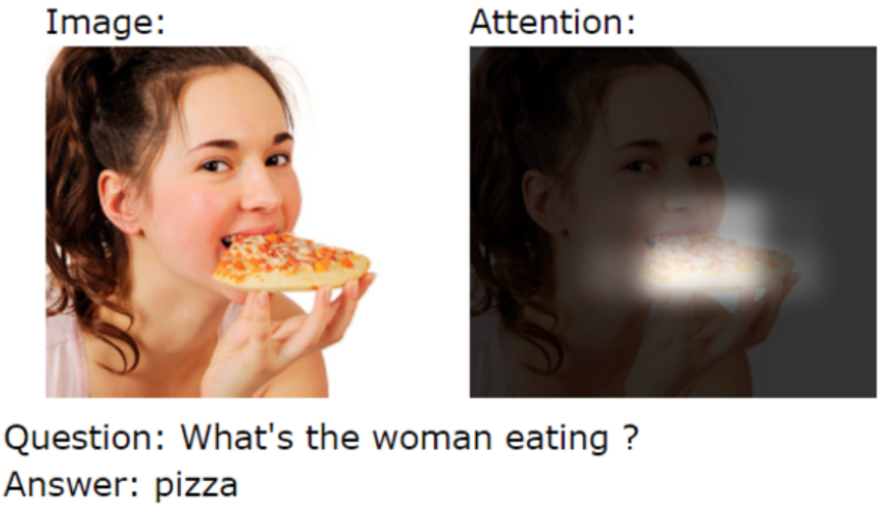


FIG. 2.4: Example of attention in a VQA system

To further understand attention we take an example of an image captioning system. As described earlier, attention is a method to weigh spatial locations according to their perceived importance. From the image in Figure 2.5 we eventually generate the caption. A man holding a couple plastic containers is walking down an intersection towards me. ([Jonathan Hui blog2017])

The first word predicted for this image is 'A' which is updated as a context.After that

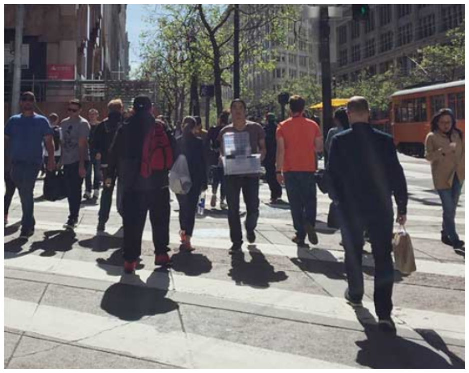
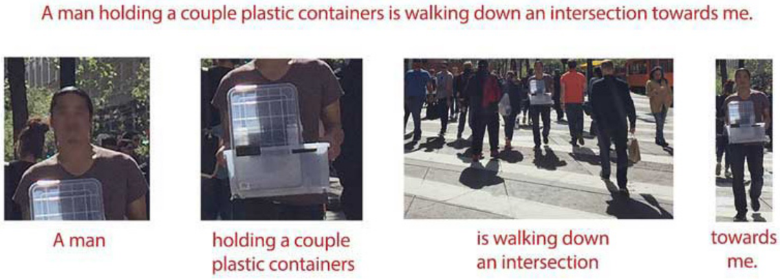


FIG. 2.5: Image for generating a caption

the word ’man’ is updated as the new context.”For the next prediction, our attention shifts to what he is holding in his hands.



A man holding a couple plastic containers is walking down an intersection towards me.
A man
holding a couple plastic containers
is walking down an intersection
towards me.

FIG. 2.6: Attention in an image captioning system

Hence, after a series of updated attentions on the image we are able to generate the caption. Mathematically, the system is performing the following function.

$$nextword = f(image, lastword) \tag{2.1}$$

where image is given as input ,the last word is the word predicted at step time step t-1 and next word is the word predicted at time step t.

This which can be translated as

$$h_t = f(x, h_{t-1}) \tag{2.2}$$

where x is the image and h_t is the hidden RNN state at time step t..

To introduce attention we are trying to replace the image x in LSTM model, with an attention module:

$$h_t = f(attention(x, h_{t-1}), h_{t-1}) \tag{2.3}$$

The attention module generates the areas to importance in an image x.

Figure 2.7 shows how attention works in a caption generating system.

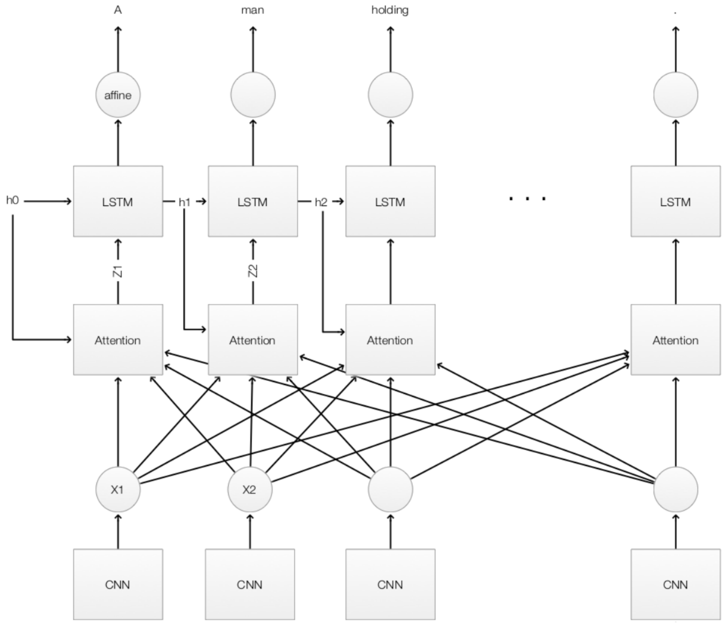


FIG. 2.7: Generating caption using attention module

The attention module has two inputs

- context : Context implies a hidden state h_t from the previous time step t. In an LSTM the input is the feature of a fully connected layer of the CNN applied to an image. But in an attention mechanism this we need spatial features too, so we choose the features of a convolution layer as input.
- Image features in each localized areas.

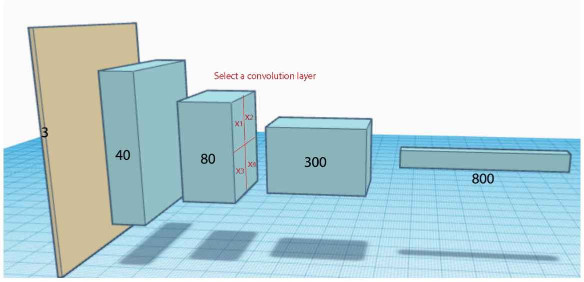


FIG. 2.8: Spatial features in CNN model

As we can see in Figure 2.8 the feature maps in the second convolution layers are divided into four regions which resemble the top right and left and the bottom right and left of the original pictures. Replacing the ’x’ in the LSTM input with these four spatial regions and the context h_{t-1} makes the attention module. Figure 2.9 shows in more detail the inputs to the attention module which are different regions of the image.

2.4.2 Types of attention mechanisms

There are two kinds of attention mechanisms-

Soft attention: As mentioned in equation 2.3 the attention, the term x can represent an image. In soft attention, each state is a representation of what is learned about the

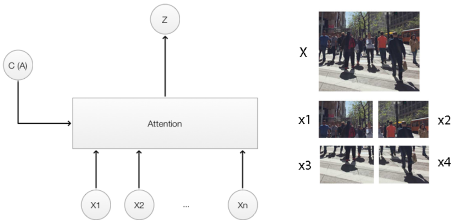


FIG. 2.9: attention module with spatial inputs

image from the previous state in terms of image features. If we condense this idea in terms of inputs, soft attention marks down the irrelevant image region features by getting them closer to zero. A new feature map is created every time by darkening the irrelevant image features. Hence in soft attention, instead of using the image x as an input to the LSTM, we input weighted image features accounted for attention.

Hard attention: As discussed in the above section soft attention modulates the value of the image features based on the importance of the corresponding image region. For this a term representing a probability α_i is multiplied to the corresponding image feature vector x_i, for the ith image region feature. In hard attention only the probability value α_i is used as an input for the LSTM.Hence in hard attention we use a probability instead of the weighted vectors of images as the input for attention. In our model we use the concept of hard attention that is we generate a probability distribution to generate the new attention mask for the image.

2.4.3 Current attention based VQA models

Stacked attention models (SAN) : Given the image feature matrix and the question feature vector SAN predicts the answer using multi-step reasoning. In many cases, an answer only related to a small region of an image. For example, in Fig.

2.10([Yang *et al.*2016]), although there are multiple objects in the image namely bicycles, baskets, window, street and dogs and the answer to the question only relates to dogs.

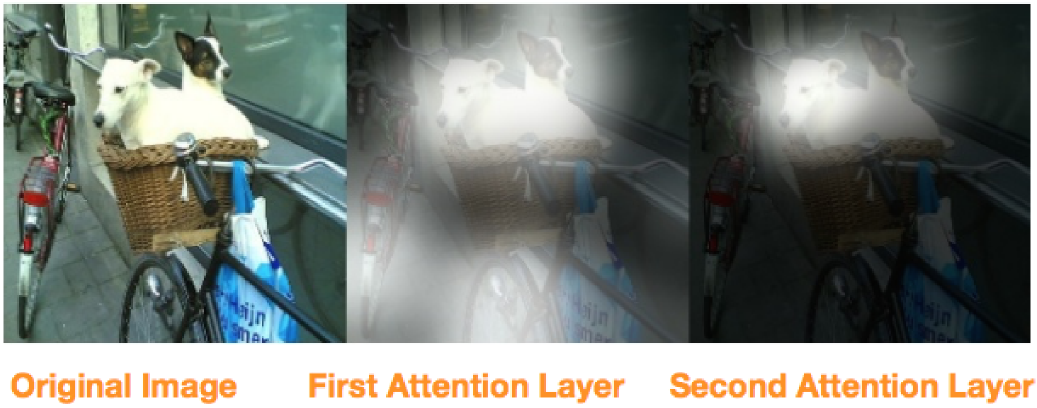


FIG. 2.10: Visualization of attention layers

Therefore, using the one global image feature vector to predict the answer could lead to sub-optimal results due to the noise introduced from regions that are irrelevant to the potential answer. Instead, reasoning via multiple attention layers progressively, the SAN is able to gradually filter out noises and pinpoint regions that are highly relevant to the answer. Given the image feature matrix v_i and the question vector v_Q, we first feed them through a single layer neural network and then a softmax function to generate the attention distribution over the regions of the image. The stacked attention network architecture is shown in Figure 2.11 ([Yang *et al.*2016]).

Hierarchical Co attention model : This mechanism uses the fact that to correctly answer a question about an image we need to look at both the question and the image. Hence it uses both the question attention and the image attention in parallel. For example give the following image if the question asked is ”how many apples are there in the image” or ”how many apples can you see in the image” , the model should consider the first three

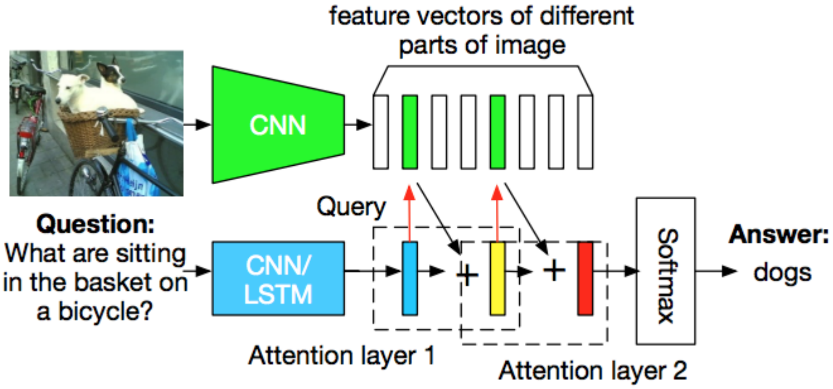

feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
+
+
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

FIG. 2.11: Stacked attention network

words in the question.

Hence a Co attention model jointly reasons about the visual attention and the question attention.There are two co attention mechanisms namely Parallel co-attention and Alternating co-attention.

2.5 Saliency

2.5.1 What is saliency?

To make progress in the field of attention mechanisms, there is a need to understand how to look at an image given some information about it. Capturing human gaze on an image largely fulfills this requirement. With the advent of neural networks and large annotated datasets, saliency prediction techniques have successfully created feature maps very close to the way a human would look at an image but mimicking the human gaze completely has not been achieved yet.

2.5.2 How has saliency been explored till now?

Saliency has been explored by carrying out various studies on the shifting of the human gaze shifts over images and videos to capture context. The VQA-HAT (Human attention) dataset.This dataset which consists of a total of 62,597 images which have been explored by people to answer a question. They were given a blurred image to mimic the intermediate layers generated by an attention mechanism and were asked to deblur the portion of the image which, according to them, were relevant to answer the question([Qiao, Dong, & Xu2017]).

These human-generated attention maps were used to conduct experiments which could use these maps instead of an attention mechanism as a corrective mechanism. One such experiment was based on the idea that these maps be directly given as the pool5 layer output in the VQA pipeline.

Furthermore, to correct visual attention the VQA-HAT dataset was used to add supervision to the VQA models. The human attention network first trains on the VQA-HAT dataset. This trained model is then applied on the VQA version 2.0 dataset to produce the human-like attention dataset(HLAT). As a final step, this new dataset is used the attention networks to show improvement in the attention. The overall accuracy is said to have improved by 0.15% utilizing this mechanism which further validates the intuition behind our experiments([Qiao, Dong, & Xu2017]). The architecture for a human attention network trained using the VQA-HAT dataset is given in Figure 2.12.

After creating the human attention network based on VQA-HAT data, this work generated human-like attention maps for the VQA dataset. This dramatically augments the dataset which acts as ground truth for further training. Because there is a large base for human attention maps or similar data this work further incorporates human in the loop mechanism by adding a loss in a standard Multilinear Low-Rank Bilinear Attention Net-

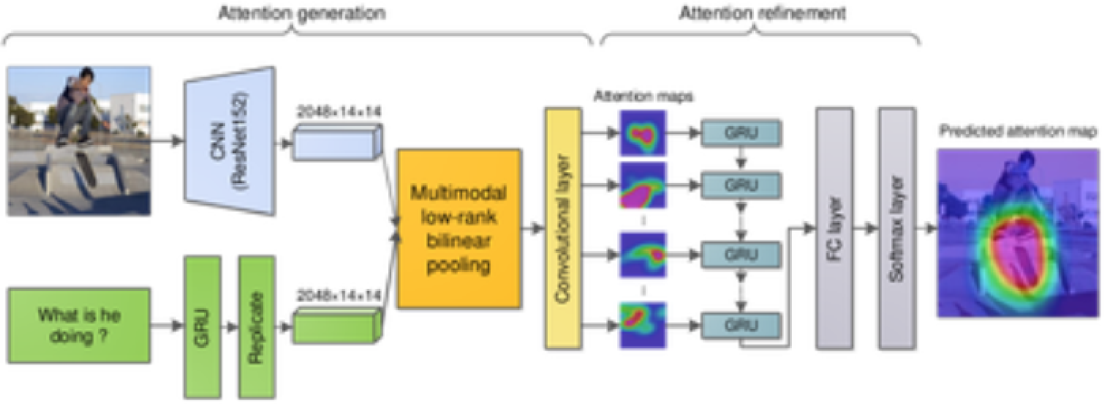


FIG. 2.12: Human attention network

work.

The complete pipeline of this work is given in Figure 2.13. The attention generated by a standard attention network is compared with the human-like attention map for the same image and question pair, and the model learns from the difference. This makes this process close to a supervised learning process.

The interesting part of observing human gaze has been the fact that human gaze does not focus on a particular object but also includes peripheral vision and a context to create a sentence or a word for an object.

2.5.3 Systems using the concept of saliency

As mentioned in section 2.5.2, the concept of saliency has been used by many other systems. One interesting use of the concept of saliency was done in image captioning systems. Image captioning models attend to different areas of the image to generate words.

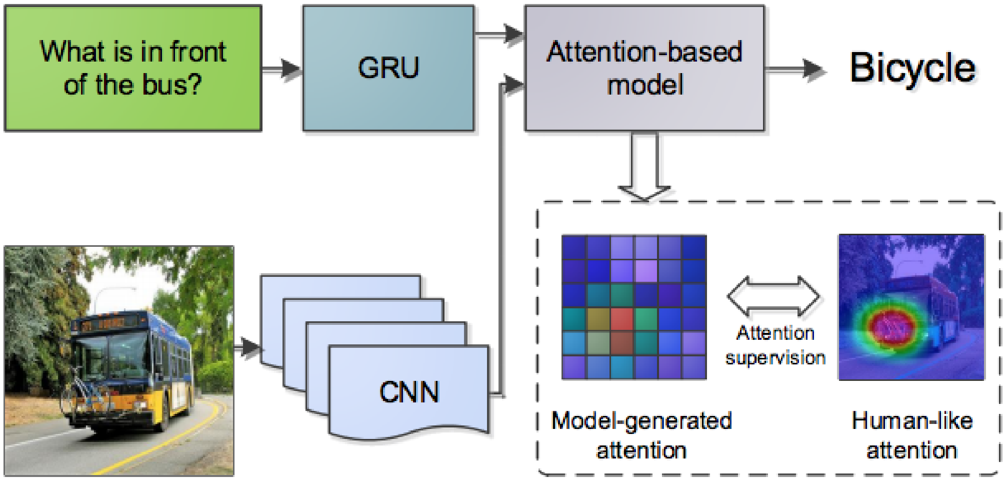


FIG. 2.13: supervised attention model

To correct the generated maps human supervision and annotations were used as ground truth to optimize the image captioning models.

The concept of attention correctness using saliency is also used with video captioning. This dataset is known as VAS and has a set of clips of video with captions along with human gaze data to generate those captions. This network is known as the Gaze Encoding Attention Network which uses the human gaze tracking data for generating spatial and temporal information([Kim2017]).

One very interesting work is based on evaluating a model based on the explanations generated by it for answering a question. ([Park *et al.*2017]).

There has been work in the direction of correcting attention not only using visual attention correction but also by evaluating the natural language explanations generated by a system. Figure 2.14 shows the working of such a system. The AI system also generated explanations or reasons about why it answered a particular question about the image. This helps as a mechanism to adjust the attention according to the direct comparison that can be made with the understanding of the system and the ground truth.

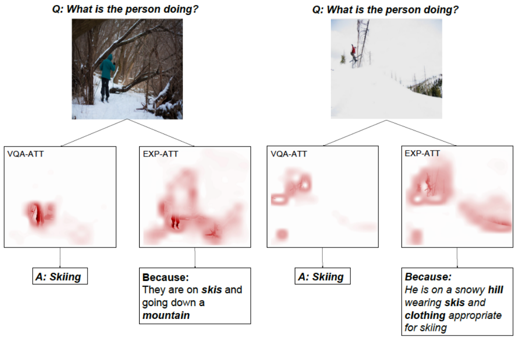

Q: What is the person doing?
Q: What is the person doing?
VQA-ATT
EXP-ATT
VQA-ATT
EXP-ATT
A: Skiing
Because:
They are on skis and going down a mountain
A: Skiing
Because:
He is on a snowy hill wearing skis and clothing appropriate for skiing

FIG. 2.14: Attentive Explanations

2.5.4 Multi-task Learning

To get optimal results, various approaches are used to train, either a single model or an ensemble of models. After training a model sufficiently, we either fine-tune a model or tweak its hyperparameters [Ruder2017].To understand multi-task learning we have to understand the concept of related tasks. For example, two single tasks of detecting the color of a flower and another of detecting the type of a flower can use the same network. Two related tasks have a slight difference but basically, use the same subset of lower level features to train on for different tasks. Based on the parameter sharing between two models, the multi-task learning can be divided into two types. The first is hard parameter sharing and the second is soft parameter sharing.

Hard parameter sharing: As the name suggests in hard parameter sharing the hidden layers are shared between tasks. Figure 2.15([BAXTER1997]), depicts hard parameter sharing in a deep neural network between three tasks A,B, and C.Hard parameter intuitively also has an advantage of reducing the risk of overfitting([Ruder2017]). It is proved that the risk of overfitting is in the order of N, where N is the number of tasks([BAXTER1997]). In this work, we have used hard parameter sharing as we use commonly hidden layers for both of our tasks.

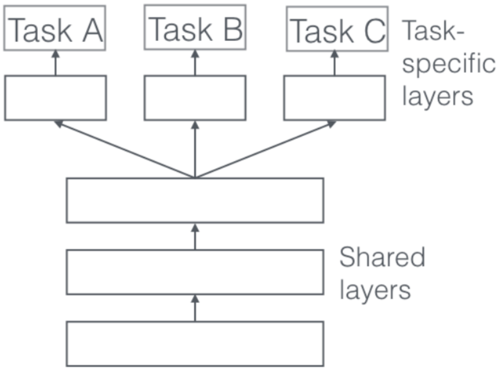


FIG. 2.15: Hard parameter sharing(MTL)

Soft parameter sharing: In soft parameter sharing all tasks have separate models. All the tasks have their own set of parameters, and the distance between them is regularized to keep them similar. Figure 2.16 ([BAXTER1997]) shows such a system with each task having its model but with regularizing the parameters as a form of sharing.

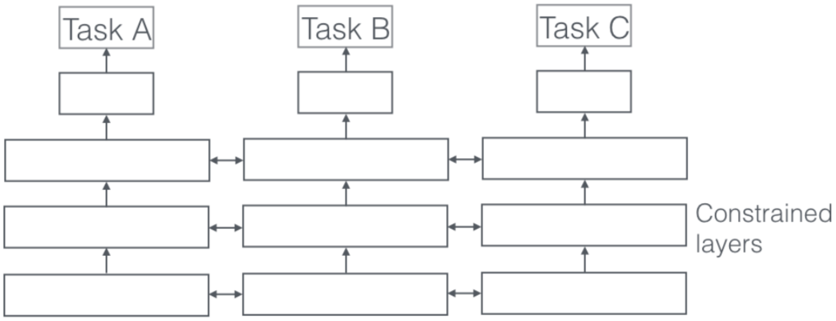


FIG. 2.16: Soft parameter sharing(MTL)

Hence we can use a neural network which was trained for some other related task in conjunction with one trained for a separate task, and this is a form of inductive transfer which helps in improving generalization. Another example is a network trained to recognize a specific object can be used for a task that doesn't require us to categorize the objects in the same way as the previous task but can still be very useful. For example, if we have an object detection task in Figure 2.15,we can treat the detection of different objects as separate tasks and use multi-task learning.

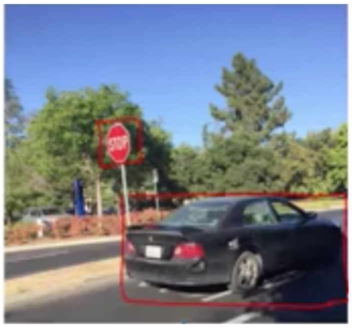


FIG. 2.17: Multitask Learning in object detection

In this image in the scenario of multitask learning instead of providing one label we can have multiple labels for the same scenario.

In our work, we use the concept of multitask learning by learning from two tasks. The first task is training on an { I, q, a } tuple where I is the image feature vector, q is the question embedding and a is the answer embedding. The second task is training over an { I, q , a , I'} where I, q and a mean the same as before with addition of I' which is the corresponding human attention map. This enables us not to lose information and helps the model to generalize better at the original task. Multi-Task learning can be done in two ways. To define the loss function for multi-task learning, we can use two mechanisms. We can either define the two tasks and optimize them separately, or we can optimize them jointly. In this work, we have optimized them individually by training them in alternate

objects	y^i Binary (0/ 1)
Pedestrians	0
Cars	1
Stop Signs	1
Traffic Lights	0

Table 2.1: Multitask Learning in object detection

batches with two separate losses.

Chapter 3

DATASET

3.1 VQA

The Microsoft COCO dataset is a dataset commonly used for visual question answering and contains human annotated questions and answers. The dataset contains 265,016 images. (COCO and abstract scenes)([Agrawal *et al.*2017]).

3.1.1 Dataset Analysis

To further analyze the dataset we describe each component of the dataset namely, images, questions, and answers. The images are categorized in two sets which are real images and abstract scenes. The real images are images of multiple objects containing rich contextual information. This accounts for a total of 123,287 training and validation images and 81,434 test images. Next, we have a collection of 50k images which are known as abstract scenes. These scenes were generated to help the model train for high-level reasoning about the image.

As described earlier the dataset has 614,163 questions and 7,984,119 answers. The questions can be clustered by the words with which they start. The figure below depicts the distribution of the questions based on the first four words for both real images and abstract scenes([Agrawal *et al.*2017]).

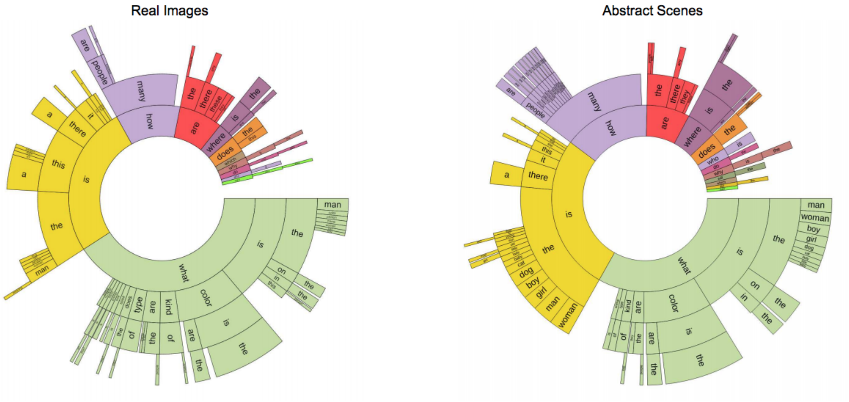


FIG. 3.1: Clustering of questions

3.1.2 Data split

There are 248,349 training questions, 121,512 validation questions, 244,302 testing questions, and a total of 6,141,630 question-answers pairs. There are three subcategories according to answer types including yes/no, number, and other. Each question has 10 free-response answers([Agrawal *et al.*2017]).

3.2 VQA-HAT

Many studies on ’human attention’ in Visual Question Answering (VQA) try to understand where humans choose to look to answer questions about images. The study required the subjects to sharpen regions of a blurred image to answer a question. Thus this dataset is called the VQA-HAT (Human Attention) data set. The images used for this study are a subset of the COCO data set. The training set has 58,475 attention maps, and the validation set has 4,122 attention maps with multiple attention maps for a single image [**?**].

3.2.1 Dataset Analysis

This VQA-HAT dataset was collected in three scenarios :

In the first scenario, the human subjects were given a blurred image and question. They were then asked to sharpen the regions they felt were useful for answering the question about the image. In this scenario, the humans were not given the answer. This helped in the collection of exploratory attention on the images the relevant regions in the image were sharpened by the subjects ([Das *et al.*2016]). Figure 3.2 ([Das *et al.*2016]) below shows this experiment where the subject was given the question ”How many players are visible in the image?”.

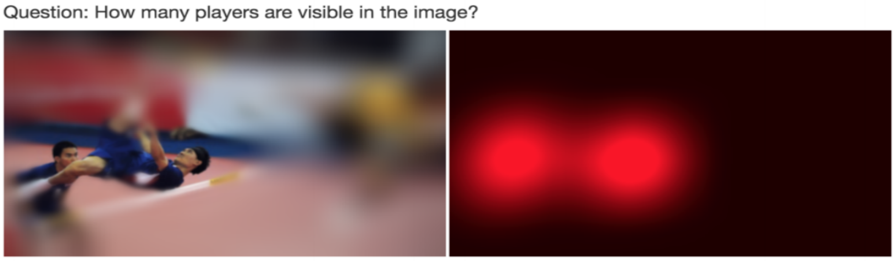
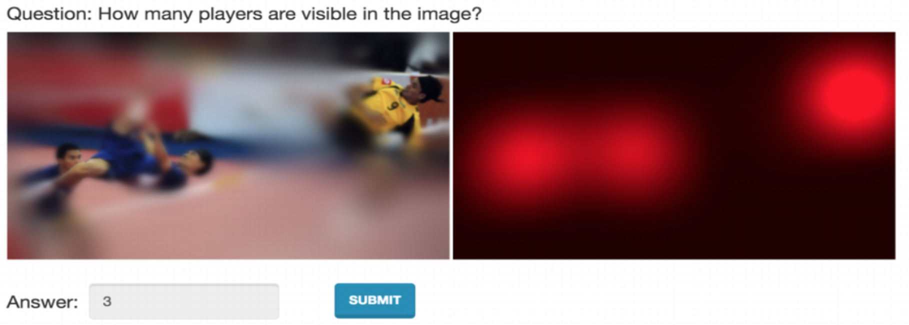


FIG. 3.2: First scenario

In the second scenario the human subjects were given a blurred image, a question and also an answer in addition to the first scenario. They were then asked to sharpen precisely enough regions needed to answer the question by someone who is shown the blurred image with the question[Das *et al.*2016]. Figure 3.2 ([Das *et al.*2016]) below shows this experiment where the subject was given the question ”How many players are visible in the image?” with the answer ’3’.

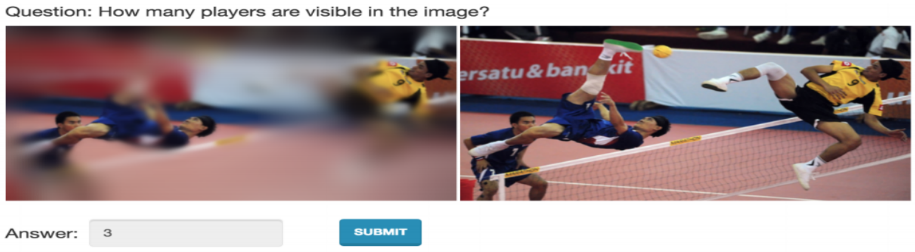
In the third scenario, the human subjects were given the question, the answer, and the high-resolution image. The subjects were then asked to imagine a scenario where someone has to answer the question without looking at the original image. This enables them to provide accurate attention maps[Das *et al.*2016]. Figure 3.3 (citevqahat) below shows this


Question: How many players are visible in the image?

Answer: 3

FIG. 3.3: Second scenario

experiment where the subject was given the question ”How many players are visible in the image?” with the answer ’3’.


Question: How many players are visible in the image?

Answer: 3

FIG. 3.4: Third scenario

It was observed that the VQA accuracies of the answers given by human subjects under the three scenarios had strikingly different values and the accuracy of the second scenario was the best ([Das *et al.*2016]).

In order to map the questions with images, the names of the images include the question ID they belong to in the VQA dataset for version 1. As described above there are three scenarios for each image. The first scenario is denoted by adding a numeric 1 2, or 3 depending on the scenario they belong to. Hence the image names are in the format quesid_1.png/ quesid_2.png /quesid_3.png ([Das *et al.*2016]).

Interface Type	Human Accuracy
Blurred Image without Answer	75.2
Blurred Image with Answer	78.7
Blurred and Original Image with Answer	71.2
Original Image	80.0

Table 3.1: Accuracies given different scenarios

Just like in the VQA dataset the VQA-HAT datset questions have been clustered into 6 categories. The Figure 3.5 shows the clusters and examples of types of questions falling into each category.

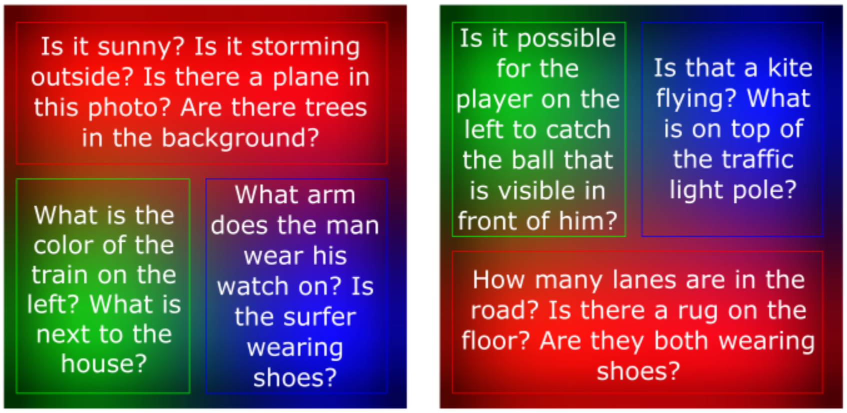

Is it sunny? Is it storming outside? Is there a plane in this photo? Are there trees in the background?
What is the color of the train on the left? What is next to the house?
What arm does the man wear his watch on? Is the surfer wearing shoes?
Is it possible for the player on the left to catch the ball that is visible in front of him?
Is that a kite flying? What is on top of the traffic light pole?
How many lanes are in the road? Is there a rug on the floor? Are they both wearing shoes?

FIG. 3.5: Question Clustering

3.2.2 Data split

The dataset had human attention maps for 58475 training instances (out of 248349 total) and 1374 validation instances (out of 121512 total) question-image pairs in the VQA dataset ([Das *et al.*2016]).

Chapter 4

ARCHITECTURE AND METHODOLOGY

This section explains the architecture and proposed methodology for improving attention in stacked attention network for a visual question answering system. Before we elaborate on the design and architecture of the proposed methodology, we review a few concepts to understand their usage in our method.

4.1 Preliminaries

4.1.1 Kullback - leibler divergence

Kullback-leiber divergence, which is also known as relative entropy, is a measure of difference between two probability distributions. Entropy for a probability distribution 'p' is defined as

$$H(p) = -\sum_{i=1}^{N} p_i.log(p_i) \tag{4.1}$$

where H denotes the entropy for discrete distributions over N outcomes.

If we have a second distribution q , then we can define KL divergence as: Let $D(p||q)$ denote the difference between the entropies of p then,

$$D(p||q) = H(p,q) - H(p) \tag{4.2}$$

$$D(p||q) = -\sum_{i=1}^{N} p_i.(log(p_i) - log(q_i)) \tag{4.3}$$

or

$$D(p||q) = -\sum_{i=1}^{N} p_i.(log((p_i)/(q_i))) \tag{4.4}$$

since log(a)-log(b)=log(a/b)

4.1.2 Normalization methods

Two common methods used to normalize a feature vector are the L1 norm and L2 norm. After applying L1, or least absolute deviation, the sum of the of the normalized vector equal to 1.

It is basically minimizing the sum of absolute differences between the target value y and the estimated value $f(x_i)$. In the equation below is represented by S:

$$S = \sum_{n=1}^{n} |y - f(x_i)| \tag{4.5}$$

Whereas, L2 or least squares, is basically minimizing the sum of square of differences of the target value y and the estimated value $f(x_i)$. In the equation below is is represented by S:

$$S = \sum_{n=1}^{n} (y - f(x_i))^2 \tag{4.6}$$

Out of these two, L1 is more robust because it is affected less by outliers.

4.2 Proposed Architecture

Any attention mechanism outputs the attention on an image in terms of a probability distribution. The current base model used for this experiment is a stacked attention network architecture which outputs a probability distribution. A typical attention network architecture works on the principle of creating a probability distribution over the image to indicate the relevance of the part of the image being queried. This distribution is scaled up to the image size and then treated as a mask on the original image. The attention mechanism creates attention weights over the image. As we can see that different parts of the image feature vectors are being queried sequentially instead of having one global image feature vector. This architecture is shown in the figure 4.1([Yang *et al.*2016]):

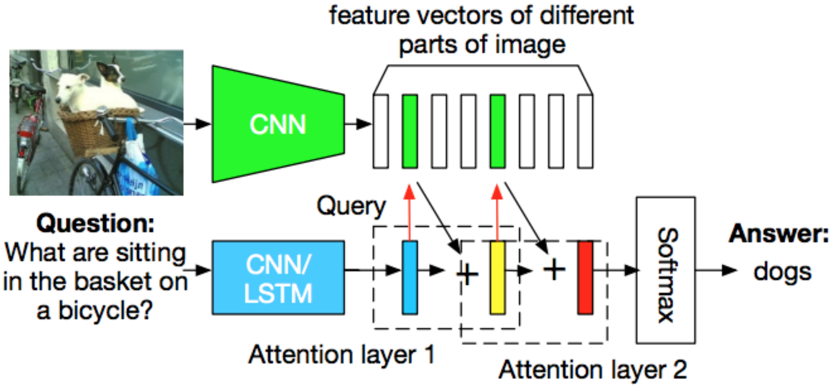

feature vectors of different parts of image
CNN
Question:
What are sitting in the basket on a bicycle?
Query
CNN/ LSTM
+
+
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

FIG. 4.1: Stacked attention network

We propose modifying this mechanism by modifying attention weights more effectively using the VQA-HAT dataset. The VQA-HAT dataset is a set of 3 channel png images having attention in the form of pixel values of the brighter areas which are converted into a probability distribution to denote attention.

To get the probability distributions from these values we use L1 normalization. After getting a probability distribution of the map we now have two probability distributions. The first distribution is the probability generated by the second attention layer in our pipeline and the second distribution which we get by normalizing the VQA-HAT images. As we have two distributions we use a popular method know as the Kullback-leibler divergence to take a difference between these two distributions to add as a new loss to the standard cross-entropy. This mechanism is shown in the flowchart in figure 4.2 shows how the KL divergence loss is being used to update the model.

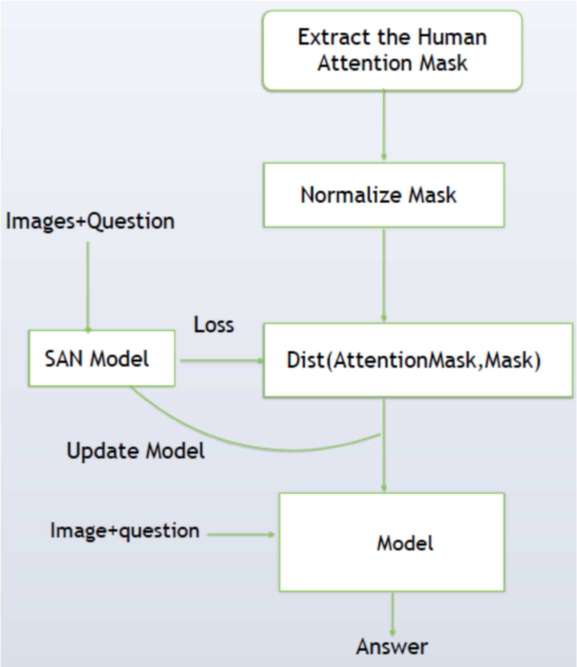


FIG. 4.2: Workflow : illustrates how the KL loss is added to the standard architecture as a new loss

4.2.1 Using multi - task learning for VQA task and VQA-HAT task

As described in section 2.5.4, for the model to generalize better it can learn from two tasks. If we only use one of the models, there is a possibility of ignoring the information learned from a large set of images and questions while refining the attentions using the model using the new loss. Hence to optimize the model we train on the VQA images on the oiginal stacked attention model and the new model which includes the loss calculated from the VQA-HAT images as ground truth in alternates batches.

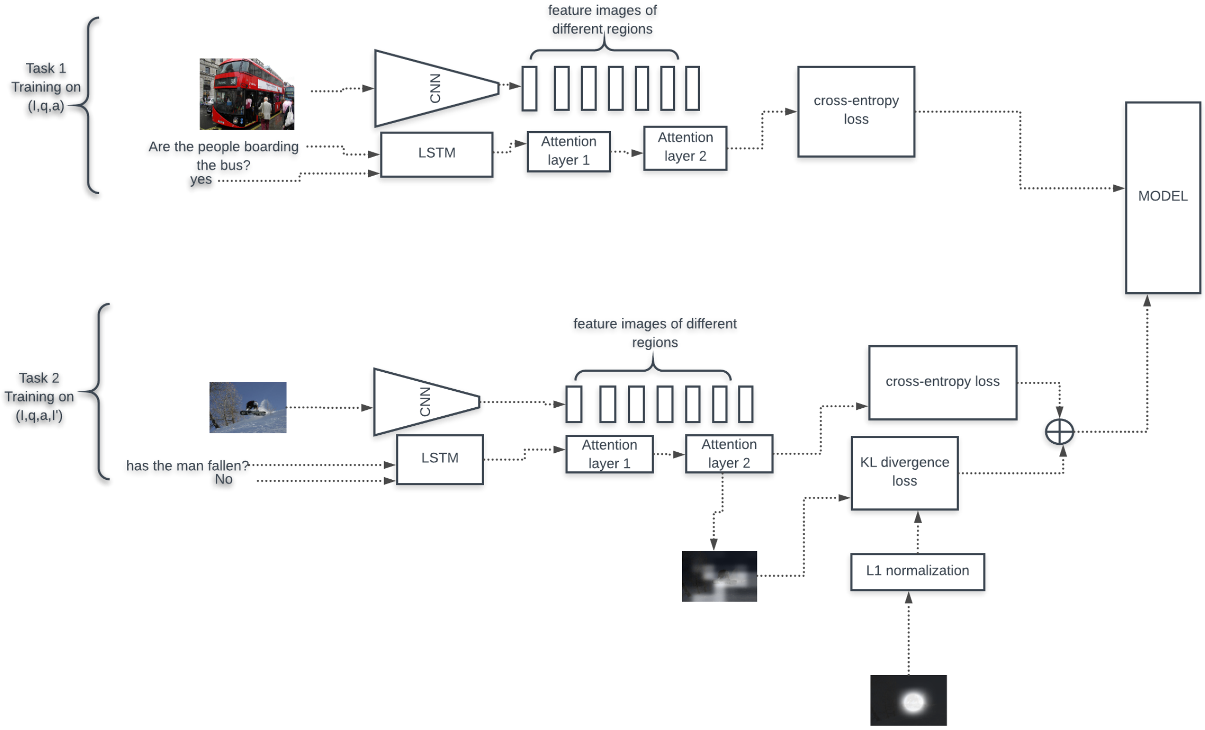


FIG. 4.3: Multitask Learning in an attention network:In the above diagram two tasks are shown of a MTL pipeline. The first task learns from the (I,q,a) tuple which denotes the image(I) , question(q) and answer(a) and the second task learns from the (I,q,a,I') which denotes the Image(I), question(q), answer(a) and the human annotation(I')

As we can see in the figure 4.3 the two models are being used in alternate batches. The dataset used by the first model is VQA version 1 and the dataset used by the second model is VQA-HAT, dataset which is a subset of the VQA version 1 dataset along with the corresponding human annotations. Figure 4.3 also shows the difference in losses between the two tasks. The task which uses only the VQA data has a only a cross entropy loss. Let p' be the predicted probablity and p be the probability of the class. The cross entropy can be mathematically written as:

$$H(p,p') = -\sum_{i=1}^{n} p_i.log(p'_i) \tag{4.7}$$

and the KL loss when q is the second probability which is calculated from the VQA-HAT dataset for the corresponding question ID is denoted by equation 4.4. We also multiply a parameter to boost the KL loss for our experiments. We have chosen this number to be a positive number and a randomply chosen number for this experiment is 25.

Hence for the first task where we only have the VQA dataset, the loss is described by equation 4.6 and for the second task the loss is given as :

$$D(p||q) + H(p, p') \tag{4.8}$$

or

$$(\alpha * (-\sum_{i=1}^{N} p_i.(log(p_i/q_i))) + (-\sum_{i=1}^{N} p_i.log(p_i')) \tag{4.9}$$

As we discussed in section 2.5.4 we can multi-task in two ways, hard parameter sharing and soft parameter sharing. In this work we are basically implementing hard parameter sharing which means that most of the layers of the model are being shared and the two types of tasks are shown in Figure 4.4.

4.2.2 Preprocessing

The VQA-HAT images are converted into probability distributions using L1 normalization as explained in section 4.1.2. For each question ID that is present in the VQA-HAT dataset, we create the dataset which contains the image feature , question and answer vector along with the VQA-HAT image converted into probability distribution. THe VQA-HAT image is first averaged across the third channel as there are three channels , then we resize it into a (7,7) numpy array. Then we reshape in into a one dimensional array of size 49 before we carry out the L1 normalization. This process is done because these steps are done in the opposite manner when the network generates the attention. So to get an equivalent probability distribution we carry out these steps.

4.2.3 Postprocessing

After getting the probability distribution we need to process the result to create a mask that can be overlayed on the original image to show the attention areas on the image. There are two attention layers in the pipeline and the attention calculated is in the form of a list of size 49, which is then reshaped into a (7,7) numpy array, this is further resized into the original image dimensions and overlayed on the original image to get the an image which depicts the attention regions.

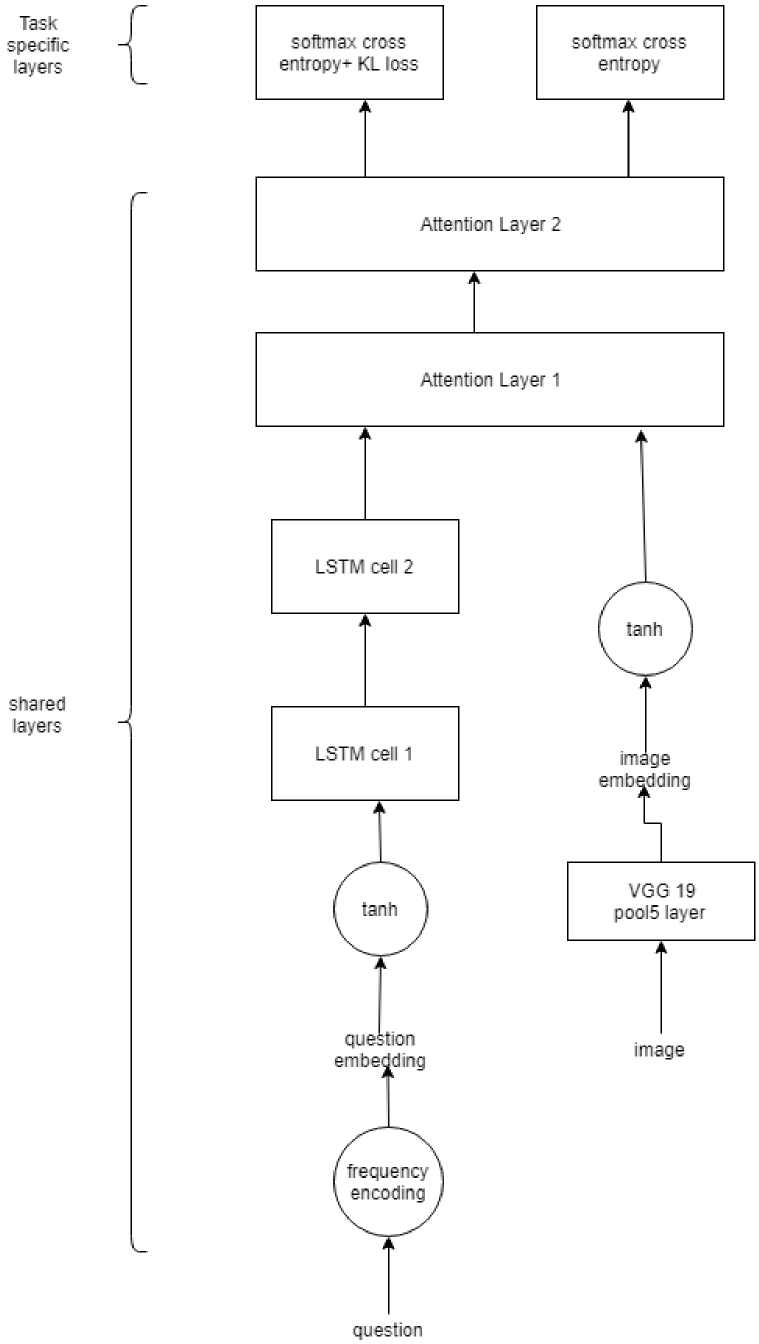


FIG. 4.4: Hard parameter sharing in multi-task in a stacked attention network pipeline: In learning.Two attention layers and an RNN layer are shared by the two tasks with only the output of attentions as the difference in the two tasks.

Chapter 5

EXPERIMENTS AND ANALYSIS

In the previous chapter, we described our system architecture to generate new attention maps from an ensemble of models using a multi-task learning pipeline. Here we discuss evaluation metrics and results. Evaluation of this pipeline is a complicated process as in many cases the quality of answers is greatly affected by the category of questions and images used in the VQA-HAT images. The test set of the VQA version 1 dataset has a variety of images and questions. The questions are categorized into 64 categories based on the words they start with. The codebase for the base model has been referred from ([TingAnChien2016]).

5.1 Processing before Evaluation

Before we evaluate we need to do some processing of the questions and answers. In order to recognize all of the characters regardless of their case we make them all lower case. Then all periods are removed except when they are decimals. All the number words are converted into digits, all the articles such as 'a', 'an' and 'the' are removed.Numbers like 100,978 are converted to 100978 . After this processing we use the evaluation tools provided by the VQA website which provides the original dataset. The VQA evaluation tool accepts the results in following form,

$results = [result]$

$result\{$

$''question\_id'' : int,$

$''answer'' : str$

$\}$

where ’int’ stands for integer and ’str’ stands for string format([TingAnChien2016]).

To decide the ground truth answer, the hypothesis used is that a minimum of 3 people should have answered the same answer for it to be considered as the ground truth.Hence the equation takes the minimum of 1 and an average of total number of humans saying an answer as the accuracy of that answer to be correct. The following equation depicts the described equation:

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

The first set of experiments are done based on the multi-task model. There are two sets of experiments, quantitative and qualitative. The quantitative experiment is done on the validation set of VQA for to determine overall changes in language attention and to test the visual attention we take a set of 3000 question ids which are a subset of VQA-HAT dataset used as the test set.

5.2 Model trained with multi-task learning

5.3 Quantitative Evaluation

The quantitative evaluation was done on the complete validation set of the VQA 1.0 for three models. The first model is model based on a stacked attention network, the second model is the model trained with multitask learning pipeline and the third model being the finetuned model with VQA-HAT dataset. There are two parts to this analysis.

5.3.1 Visual attention

The training with VQA-HAT dataset was carried out excluding a set of 3000 images which we use in the analysis on the improvement of visual attention. The mechanism used to evaluate the improvement in visual attention is called intersection over union (IOU). Since we have the 3000 VQA-HAT images which are human annotations, we treat them as the ground truth. We also have the generated attention maps from both the initial base model and the new model.

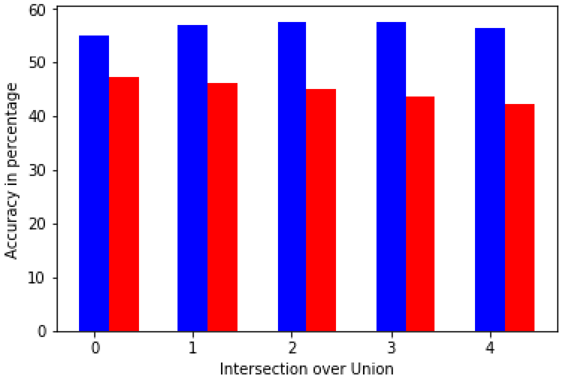
Let ’X’ be the attention mask generated by any of the models, and the ground truth is ’G’ , then the measure of performance of the map generated the model can be written as:

$$IOU = X \cap G / X \cup G \tag{5.1}$$

Mathematically the intersection operation is a logical ’and’ operation, and the union is a logical ’or’ operation. To be able to use these operations the attention mask which is a probability distribution has to be converted into 0 or 1 depending on a threshold.Let the attention generated by the base model be ’A’, the attention generated by the new model be ’B’.

Then the IOU for the base model and the ground truth can be written as :

$$A \cap G / A \cup G \tag{5.2}$$



X-axis: Thresholds in probability distribution for binary format conversion.

0: 6.10258372e-06
1: 7.10258372e-06
2: 8.10258372e-06
3:9.10258372e-06
4: 10.10258372e-06

Y-axis :% of intersection over union

FIG. 5.1: The X- axis depicts the various thresholds taken in order to convert the probability distribution into a binary format from the generated attention masks in increasing order and the Y-axis denotes the accuracy in percentage of the intersection over union for the models. The color red represents the base model and the color blue represents the new model

and the IOU of the multi-task model and the ground truth can be written as:

$$B \cap G / B \cup G \tag{5.3}$$

The results from equations 5.2 and 5.3 have been plotted in the bar graph given in the figure 5.4 for 5 different thresholds used to convert probability distribution into a binary format.

As we can see in figure 5.1 there is a clear difference between the IOU calculated using the attention maps generated by the multi-task model(denoted by blue color) and the IOU calculated using the attention maps generated by the base-task model(denoted by red color) . The difference also increases as we increase the threshold. The overlap of the generated probability with the ground truth generated by the multi-task model is better than the overlap of the generated probability with the ground truth generated by the base model. Hence there is a clear improvement in the visual attention.

Model Type	Binary (Yes/ No) Questions	Number Questions	Other	Overall
SAN Multitask model	78.08 %	32.56%	40.03%	53.33%
SAN Base Model	78.38%	33.46%	41.70%	54.39%

Table 5.1: Accuracies of models based on question types and Overall

5.3.2 Overall language attention

The next part is the overall quantitative evaluation over the complete VQA test set with the two models shown in table 5.2. There are three categories of questions for which the accuracy has been calculated namely ”Binary(Yes/No)” , the ”number questions” and the ”other” category which are basically the Multiple choice answers.

These accuracies are taken over the test set of the VQA dataset and consider the ground truth as explained in section 5.1 to find the confidence in a candidate answer.

5.3.3 Qualitative Evaluation

The multitask trained model has been evaluated based on the accuracy in percentage of the type of question being answered based on the first 3-4 words in the questions. Fig 5.4 shows a plot of the such accuracies where the labels on the x-axis denote a number that represents a specific type of a question. The table 5.2 below gives a mapping of the numbers and the actual question type to give further clarity.

We can see in figure 5.2 the accuracies of the two models where the black color represents the accuracy of the new model and the orange color represents the accuracy of the base model.

Figure 5.3 shows the categories of questions where our model has performed better which are [0, 3, 13, 26, 28, 29, 34, 40, 41, 47, 48, 50, 53 and 61 in the table 5.2. Here also the orange color represents the base accuracies and the black color represents the new

X-LABEL	QUESTION	X-LABEL	QUESTION
0	are there	27	what is on the
1	what brand	28	has
2	what room is	29	was
3	what color is	30	what type of
4	is	31	is this an
5	are they	32	do
6	what number is	33	what is the man
7	what sport is	34	which
8	are	35	are these
9	is the	36	what are
10	what is the person	37	what is the
11	how many	38	where are the
12	does this	39	is this a
13	is there a	40	can you
14	is he	41	what time
15	what	42	what are the
16	does the	43	are there any
17	is the person	44	what color are the
18	where is the	45	why
19	what animal is	46	what is this
20	how	47	how many people are in
21	what is the woman	48	do you
22	none of the above	49	is this
23	who is	50	why is the
24	is the woman	51	what is the color of the
25	are the	52	what is
26	how many people are	53	could
54	is that a	55	what is in the
56	what does the	57	what kind of
58	is it	59	is the man
60	what is the name	61	is there
62	what color is the	63	what color
64	is this person		

Table 5.2: Mapping of x- labels and question types

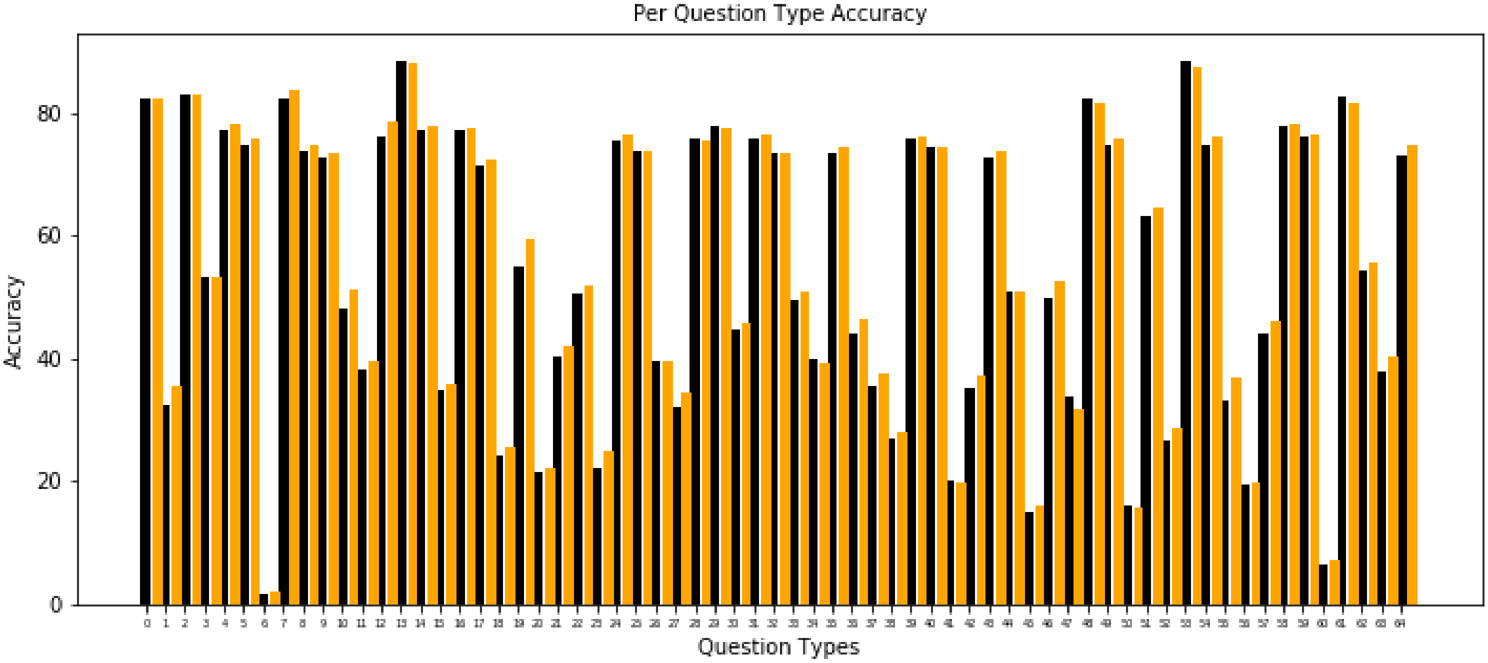


FIG. 5.2: Question vs Accuracy

model accuracies.

Statistical significance In order to see how well the new model has performed over some categories we take confidence scores for answers generated by the base model and the new model and perform a randomization test. In this test we take lists of scores per question for each category and calculate the difference in the confidence values in noth the lists and sum those values. After this we calculate the difference 100000 times in a randomized manner assuming that score is equally likely to come from either models and create another list of differences. Then we basically calculate number of elements in the list which are greater than the sum of differences of the confidence scores from the new and the original list.We are essentially trying to calculate if we can reject the null hypothesis that there is no difference between the algorithms. The result is the percentage of samples for which the difference of confidence values is smaller as compared to the sum of all the

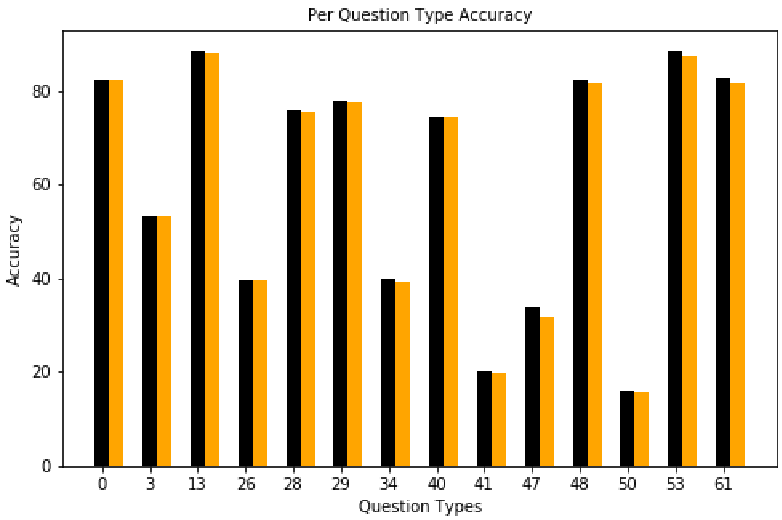


FIG. 5.3: Question vs Accuracy

differences for that category.

Figures 5.4 and 5.5 show some examples of the generated attention for images and questions from the VQA test set.

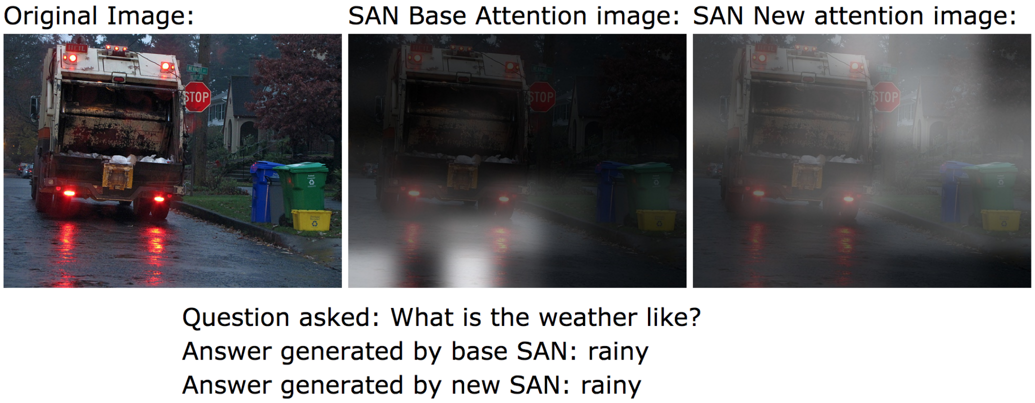
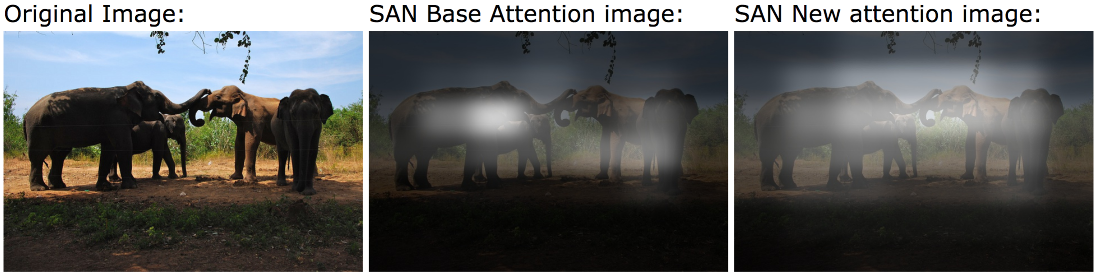

Original Image:
SAN Base Attention image:
SAN New attention image:
Question asked: What is the weather like?
Answer generated by base SAN: rainy
Answer generated by new SAN: rainy

FIG. 5.4

Fig 5.4 shows that there is a marked improvement in attentions although the quality of the answer remains unaffected. The model visualizations on samples shows improvement

Category	Accuracy(Base model)	Accuracy(New model)	%samples
are there	82.23%	82.422%	33.000%
what color is	53.291%	53.319%	49.000%
is there a	88.133%	88.468%	8.000%
how many people are in	39.435%	39.566%	43.000%
has	75.388%	75.939%	29.000%
was	77.515%	77.912%	32.000%
which	39.408%	39.824%	26.000%
can you	74.371%	74.556%	43.000%
what time	19.712%	20.069%	19.000%
how many people are in	31.577%	33.901%	5.000%
do you	81.777%	82.444%	27.000%
why is the	15.734%	16.109%	33.0000%
is there	81.751%	82.782%	0.0%
could	87.448%	88.486%	11.0%

Table 5.3: Percentage of samples as a result of randomization test



Question asked: How many elephants are there?
Answer generated by base SAN: 2
Answer generated by new SAN: 2

FIG. 5.5

in attentions but the quality of answers doesn't improve overall because of the following reasons. There is an evident change in the visual attentions but the overall accuracy of answering doesn't improve, hence it shows that improvement of attentions in images does not create a difference in reasoning by the model. The model learns to attend to more accurate regions in the image but the same shift of focus is not observed in the answering hence pointing to a hypothesis that the reasoning capability of a model is not directly correlated with the improvement in visual attention.

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

In this paper, we explore new methods of using a human annotated dataset for improving a stacked attention network used as a baseline for visual question answering systems. We first experimented with training an attention networks with only the human annotations. This experiment showed us that a relatively smaller dataset with human annotations as the ground truth gives accuracies for different categories close to the dataset that is thrice as large hence validating the idea of human annotations being a very effective mechanism. Deriving from this idea we try to perform the second experiment which is fine-tuning the pre-trained stacked attention model trained on VQA dataset version 1. The fine-tuned model gave new insights about using human attention maps. Because of fine-tuning on the previous data, the new model's performance deprecated a little bit because of losing the generality. In short, we end up ignoring information which might help us do better.

After the above experiments, we come to the conclusion of preferring an ensemble learning method called multitask learning. We share the weights/information between the two models and train on the models and we observe an improvement in the category of ”other” type questions which brings us to the conclusion that multi-task method of training was effective in bringing up the performance of the model for questions which were largely uncategorized and helped in overall contextual understanding of the scene by the model.

We accomplish this with human attention maps for 1/3rd of the total vqa dataset as compared to the human-like attention model which utilized the human annotations to bring up the performance by making it a supervised task.

We can augment this work by experimentation with other baseline models and creating a similar pipeline for them. There a lot of scope of using similar mechanisms in other type of AI tasks also which can help avoid the need of a large number of human attention maps [Ren, Kiros, & Zeme2015].

REFERENCES

[Agrawal *et al.*2017] Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Batra, D.; and Parikh, D. 2017. Vqa: Visual question answering. In *Facebook AI Research, Menlo Park, CA 94025, USA*, 1–6.

[Ba, Mnih, & Kavukcuoglu.2015] Ba, J.; Mnih, V.; and Kavukcuoglu., K. 2015. Multiple object recognition with visual attention. In *In ICLR*, 1–8.

[Bahdanau, cho, & Bengio2016] Bahdanau, D.; cho, K.; and Bengio, Y. 2016. Neural machine translation by jointly learning to align and translate. In *Accepted at ICLR 2015 as oral presentation*, 1–9.

[BAXTER1997] BAXTER, J. 1997. A bayesian/information theoretic model of learning to learn via multiple task sampling. In *Department of Mathematics, London School of Economics and Department of Computer Science, Royal Holloway College, University of London*.

[Das *et al.*2016] Das, A.; Agrawal, H.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Facebook AI Research, Menlo Park, CA 94025, USA*, 1–9.

[Jabri, Joulin, & der Maaten.2016] Jabri, A.; Joulin, A.; and der Maaten., L. v. 2016. Revisiting visual question answering baselines. In *Facebook AI Research*, 1–11.

[Jonathan Hui blog2017] Jonathan Hui blog. 2017. Soft hard attention. `https://jhui.github.io/2017/03/15/Soft-and-hard-attention/`.

[Kim2017] Kim, Y. Y. J. C. Y. K. K. Y. S.-H. L. G. 2017. Supervising neural attention models for video captioning by human gaze data. In *Facebook AI Research, Menlo Park, CA 94025, USA*, 1–26.

[Murrugarra-Llerena & Kovashka2017] Murrugarra-Llerena, N., and Kovashka, A. 2017. Learning attributes from human gaze. In *Department of Computer Science University of Pittsburgh*.

[Park *et al.*2017] Park, D. H.; Hendricks, L. A.; Akata, Z.; Schiele, B.; Darrell, T.; and Rohrbach, M. 2017. Attentive explanations: Justifying decisions and pointing to the evidence. In *UC Berkeley EECS, CA, United States 2Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrucken, Germany*, 1–16.

[Qiao, Dong, & Xu2017] Qiao, T.; Dong, J.; and Xu, D. 2017. Exploring human-like attention supervision in visual question answering. In *arXiv:1709.06308v1 [cs.CV]*, 1–10.

[Ren, Kiros, & Zeme2015] Ren, M.; Kiros, R.; and Zeme, R. S. 2015. Exploring models and data for image question answering. In *University of Toronto, Canadian Institute for Advanced Research*.

[Ruder2017] Ruder, S. 2017. An overview of multi-task learning in deep neural networks. In *arXiv:1706.05098 [cs.LG]*, 1–14.

[Singh2016] Singh, A. 2016. Deep learning for visual question answering. In *Facebook AI Research, Menlo Park, CA 94025, USA*, 1–6.

[TingAnChien2016] TingAnChien. 2016. san-vqa-tensorflow. `https://github.com/TingAnChien/san-vqa-tensorflow`.

[Xu *et al.*2015] Xu, K.; Ba, J.; Kiros, R.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and

Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *In ICM*, 1–8.

[Yang *et al.*2016] Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.