Visual Q&A using Attention based Convolutional neural network

Aditya Raj Verma, Lenord Melvix, Srinath Narayanan
University of California, San Diego
{arv018, lmelvix, srn001}@ucsd.edu

Abstract

Visual Question-Answering has been a nucleus research problem in the field of computer vision. It had garnered a lot of interest due to its applications in natural scene embedding, automatic annotations, artificial conversational entities such as chat-bots etc. But, the field has not yet produced any realistically exciting results, since the state-of-the-art accuracy hovers around 60% only. There has been a lot of interest in attention networks, since psychophysics tells us that, intuitively, a human processes details of an image, by attending to its regions, rather than as a whole image. We take inspiration from Attention based Convolutional neural network (ABC-CNN) and propose our neural network architecture with some innovation to best suit our problem statement. We simplify the problem by attempting to solve only 'Yes/No', color based and numeral based questions. We innovate by using rgb-histogram features to assist in the color information. Our model is trains in around 20 hours, with an average accuracy of 47%.

1. Introduction

Computer vision is progressing beyond the holds of image recognition, object detection, and has forayed into solving inter-modal problems at the confluence of vision and related streams like natural language processing (NLP), semantic analysis, sound/music processing etc. Visual Question-Answering (VQA) is defined as the problem of answering a question, based on the semantic characteristics of the image. For a grown human, it is a pretty clear and simple task : Given an image, and a question based on its semantics or based on some hidden meaning drawn from the image, provide an answer to the question that follows a logical and intuitive explanation. The conversion or coordination between the semantic and natural language aspects of the question and visual non-lingual aspects of the image forms the biggest problem to be addressed. A human is able to access an image in remarkable detail and short-term recall. This allows the human to use these visual details in conjugation with the semantic structure of the question, to answer it. For example, a human baby is taught in its childhood, about the different real world objects repetitively. This forms a neural pathway, or a connection between the question and the image. Asking the same question with different models of the same object allows for invariance in translation of, affine and scale parameters. Hence, the confidence in detection of that particular object increases.

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?

Figure 1: The MS-COCO dataset

The visual Q&A can be described by 3 distinct modules : Image analysis, question analysis, and answer building. In this work, we focus on building models to generate one-word answers to questions regarding the semantics of the image, mainly Yes/No questions, color based questions and numeral based questions. This simplifies the modeling of the problem, and allows us to extract specific image features that might help in our quest.

In the question building module, there has been a lot of research in restricted vocabulary modeling of questions and its application to semantic analysis. Although this is not the ideal way to approach the problem, and is constrained in its description of the questions, it is a start. In this work we take a step forward, by using the GloVe embedding [24] of words to describe our question. GloVe is similar to Word2Vec [18] and provides semantically clustered vectors. As in, words that mean similarly are clustered together, and words that are different in meaning, and placed far way. The drawback of this method is that hidden meanings, dual meanings, and sarcasm cannot be extracted or well-modeled. However, we assume the questions in VQA to be simple and straightforward, and hence GloVe does a pretty good job for our quest.

The feature extraction is performed by a VGG-16 network [17] The vanilla model uses an architecture with very small (3×3) convolution filters with small (2×2) pooling layers, and pushes the depth to 16 weight layers. The VGG16 extracts feature information that is categorical of the image. It transforms an image, into a set of basis features, that try to explain different aspects of the image. This is ideal for our project, where we aim to extract useful features from the image, to be co-learned with the question embeddings. The kernels produced by the VGG-net are co-learned with the question embeddings, to provide a mapping from the questions space to the image space. In addition to the VGG features, we also innovate by introducing the histogram features, which we expect to be useful in extracting the color information of the image.

2. Prior Work

Earlier works in visual semantic analysis were focused on auto-annotations with object semantics, image captioning, and keyword-generation. VQA was a derivative of the above pursuits, and there has been constant interest in the field even before the advent of deep learning [1]. Works such as, Intelligent Virtual Agents, by Aylett et.al [2], Deployable Intelligent Systems by Lasecki [19], and Predicting motivations by text by Vondrick et.al [20], have tried to approach the problem via different vantages. But, most notably, in 2014, Microsoft released the MS COCO dataset [4] , with more than 300,000 images, captions, instances and key-words over 80 object categories, creating renewed interest in the field. Figure 1, shows some of the images from the MS-COCO dataset. Their works on Common Captions, and Common objects in context ([3] and [4]), provided insights into adapting the dataset for various vision tasks.

Following the dataset, a huge wave of interest led to a sizable number of papers being published on VQA. Vondrick et.al [20] studied the problem of inference from images, by creating a new dataset with images annotated with likely motivations, and asking the why people in a image perform a particular task. The aim was to establish that any semantic understanding of the image required knowledge and natural understanding of massive amount of texts. VQA by Antol et.al [23] redefined the problem and established an online judge and an annual competition *(VQA 1.0)* tasked to solve the problem. It was one of the first papers to provide insights into approaching the problem, and is one of the widely used datasets. It used a CNN embedding of an image, a LSTM embedding of the question, and directly combined the two via a point-wise multiplication technique. This linearly embedded image-question pair, was passed on to a perceptron, with a soft-max layer at the output, to classify the answer. It was a novel idea and performed better than the benchmark at that time. Hierarchical Co-Attention model (HieCoAtt) [25], modeled the problem in a hierarchical fashion, by attending the question in word-level, phrase-level and sentence level and combined the levels recursively to find the best possible answer to the question.

Chen et.al [4] proposed a recurrent visual representation of the image, by generating features at multiple scales to solve the problem of image captioning. Deep visual semantics by Karpathy et.al [26], works on problem by employing a convolutional neural network (CNN) over the images, and a bidirectional recurrent neural network (B-RNN) over the questions. They provide a structured objective by aligning the modalities, through a multi-modal embedding. Duygulu et al. [5] and Socher et al. [6] studied the complex mapping between words and images to annotate segments of images. They studied the problem of holistic understanding of the question-image pair, and their inter-dependencies, leading to a multi-modal correspondence between scenes and words. The show-attend-tell concept builds on the show-tell concept, and was introduced by Mao et.al [7]. They train a discriminative model using standard back-propagation, and stochastically maximizing a variational lower bound. The algorithm learns to search, attend and fix its focus on salient objects in the image, and generate captions by stitching together the words with syntactic meaning. Cho et.al [8] proposed a phrase-translation representation approach using RNN encoder-decoder for statistical machine translation. Chen et.al [4] proposed a recurrent image representation for visual image captioning. These works inspire the use of recurrent machines for applications in image captioning, VQA and machine translation, since they are interconnected. Donahue et.al [9] proposed a LSTM based recurrent network for visual recognition and description, that performs well in normal descriptive scenario with well-modeled vocabulary.

Kong et. al's paper on text-to-image co-reference [10], talks about the different methods of co-analyzing textual and visual information. It provides a bridge between text-text translation and text-image translation. Fukui et. al [22] discusses about bilinear pooling of images and text for complex visual grounding analysis and VQA. Bilinear pooling is an effective method to efficiently combine multi-modal features to express the different aspects of the image. They apply the pooling twice, once for predicting attention over spatial features and again to combine the attended representation with the question representation, which leads to better embedding in the question-image space.

However, the problem is not completely understood yet. The complex compositional structure of language and its semantic and syntactic content makes problems at the intersection of vision and language challenging, but the language with its diverse and distinct structure also provides a stronger prior. Works by Devlin et. al [11] and [13] points to the fact that the language provides a stronger prior to the embedding than the image. Zhang et.al [14] talks about a 'visual priming bias' due to the fact that subjects see an image, before annotating it. This leads to a inherent bias in constructing the n-grams as image captions, or as answers

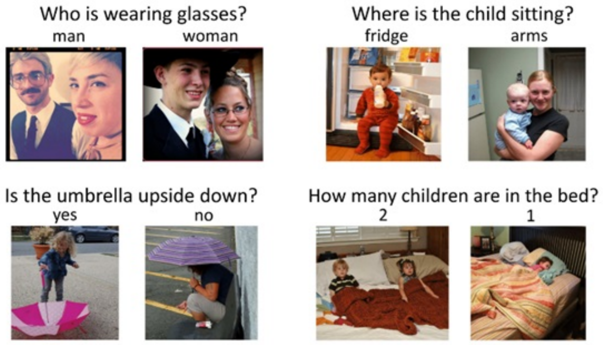

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

Figure 2: Examples of single-word answer question-image pair

to a question based on the image. Works by Torralba et.al [12] and Zhang et.al [14] try to take an unbiased look into these image biases and ways to overcome them.

Although multiple works have established the interdependencies between image captioning, image retrieval, scene understanding and VQA, we are yet to fully conquer the domain. The accuracy and recall of these algorithms vastly depend on the dataset used, the validation set used and the methods of retrieval or sentence building structure. Thus, it is difficult to find an universal algorithm that could work perfectly in all situations. In fact, humans themselves have retrieval issues with age and stress. Thus, psychophysics and neurology may also provide us with new opportunities of research.

3. Our Approach

In this work, we try to focus on 'single-word answer' questions based on an image, by using attention layers and extracting relevant features to build the model. Some of the examples of single-word answers, are shown in Figure 2. The dataset is subcategorized into 3 parts : Questions with answers based on color, numeral and other single-word nouns. This simplification allows us to model the data well, and helps us to figure out the feature extraction techniques that could be employed to assist in the different sets of questions. For example, for the color based questions, we employ the histogram of the image as a feature in addition to the features extracted by the CNN, as it intutively will provide a good understanding of the color space of the image. Similarly for numeral-answer based questions we could employ a pre-trained CNN model from the MNIST-dataset, as a feature extractor. This is a very simple model, but it allows good ablation studies and helps us decide the advantages and disadvantages of a particular feature.

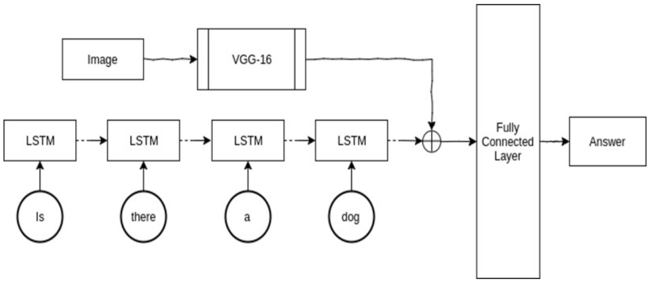


Figure 3: Examples of single-word answer question-image pair

3.1. Blind-Deaf model

We start with a blind-deaf model (Figure 3), which is our basline. It consists of a LSTM network to process the temporal or syntactic representation of a question, and a CNN to process the corresponding image and extract its relevant regional features. The LSTM is the default recurrent network that is employed in temporal domain, since we require the network to efficiently remember and recall the corresponding image features relating to the question. We one-hot encode the question as a fixed length vector to train the LSTM. Our vocabulary is extracted from the training set itself, and it might fail when it encounters a new word in the test set. Hence, the blind-deaf model is not robust. The hidden node of the LSTM provides the encoded question vector.

For the CNN, we use a vanilla VGG16 network [17]. that extracts the visually dominant features in an image. We input a pre-processed 14×14 image, and use the pre-trained VGG model as a black-box. The VGG16 is a very deep convoutional neural network with small convolutional filters (3×3) and max-pooling layers (2×2). This allows good local embedding of the image features. The Figure 4 shows the features extracted from a VGG16 network. The output is a $14 \times 14 \times 512$ vector. The 512 feature matrices could be described as a decomposition of the image features that the network deems important.

The image features and the one-hot encoded question vector is linearly combined and passed on to a fully-connected layer. It computes the joint-probabilities of the question-answer pairs, and performs a soft-max at the output layer, to extract the best possible single-word answer.

3.2. Attention based CNN

Our model improves upon the blind-deaf model, by using an attention layer based CNN (ABC-CNN) [15], to focus on specific aspects of the image that contribute to the answer selection. This improves the focus of the network to search and confidently reply about specific regions of the image, rather than search the whole image space. It is a way of question-guided attention that is focused on the image-question pair. The ABC-CNN is composed of four compo-

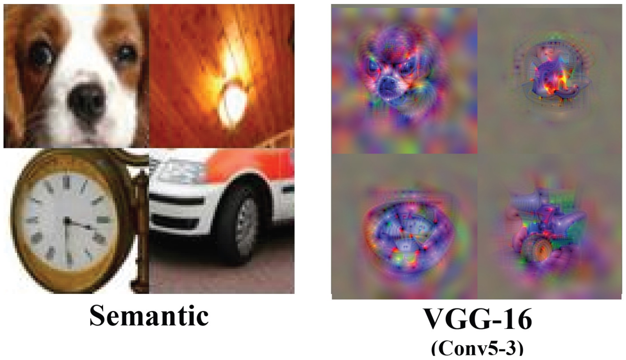


Figure 4: VGG16 features

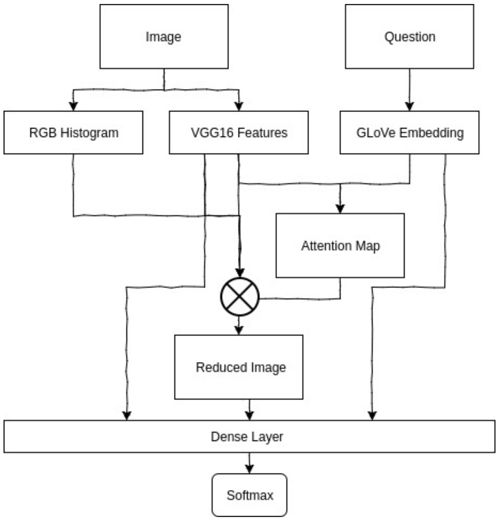


Figure 5: Our model

nents:

1. The Image feature extraction part
2. The Question understanding part
3. The Attention extraction part
4. The Answer generation part

The image extraction is done by splitting the image as cell grids and extracting a feature vector for each cell in the grid. This is done very similar to a VGG16 module, where the features are averaged across the grid for each cell. The final feature map is a concatenation of the grid maps.

Question understanding is crucial for visual question answering. The semantic meaning of questions not only provides the most important clue for answer generation, but also determines the convolutional kernels to generate attention maps. We employ an LSTM model to generate a dense question embedding to characterize the semantic meaning of questions.

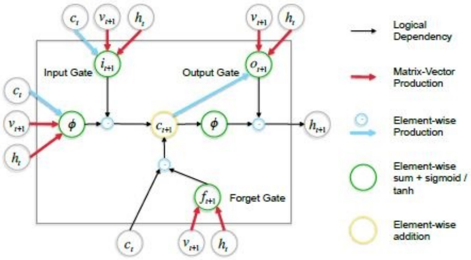


Figure 6: LSTM structure [15]

For embedding thee question, to preserve the semantic information provided by the text, we use GloVe embedding. It was developed by the stanford NLP group. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. It is similar to Word2Vec, and preserves the similarity between words in higher dimensional embedding space. The categorization of a word depends upon the kulbeck-leiber divergence of its conditional probability across all the words. This allows similar words to be clustered together, and dissimilar words to be seperated far from each other. It runs in $O(n^2)$, and hence is efficient in accomodating new vocabularies. But, in this project we use a pre-trained GloVe model. According to a question dictionary, each word is represented as a dense word embedding vector, which is learned in an end-to-end way.This type of embedding has intuitive advantages over one-hot encoding and provides better results.

An LSTM is applied to the word embedding sequence to generate a state h_t from each question vector v_t, using memory gate c_t and forget gate f_t, as shown below.

$$\begin{aligned}
i_t &= \sigma(W_{vi}v_t + W_{hi}h_{t-1} + b_i) \\
f_t &= \sigma(W_{vf}v_t + W_{hf}h_{t-1} + b_f) \\
o_t &= \sigma(W_{vo}v_t + W_{hi}h_{t-1} + b_o) \\
g_t &= \sigma(W_{vg}v_t + W_{hg}h_{t-1} + b_g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \phi(c_t)
\end{aligned} \tag{1}$$

where ϕ is a hyperbolic tangent function and $\odot$ is the element-wise product between two vectors. The basic LSTM structure in shown in Figure 6

The attention extraction is the novel feature that ABC-CNN provides. We use the VGG16 features to build a overall high-level understanding of the image, and we provide histogram features to answer color based questions. We extract attention layers based on a element-wise product between the question and the image features. The configurable convolution operation can be thought of as searching spatial image feature maps for specific visual features that correspond to the questions intent. The dense embedding of the convolution features encodes the semantic information asked in the question, on the image. Thus as the network trains, the attention layer develops and tries to look for those regions in the image, that respond to a particular aspect of the question. The more general the question is, the weaker or more spread out is the attention. The narrower the question, the strong is the attention, but has more chance to fit into a local maximum. This question embedded allows us to use a reduced image, as shown in figure 5, as input to the dense layer. some of the outputs of the attention layer is shown below.

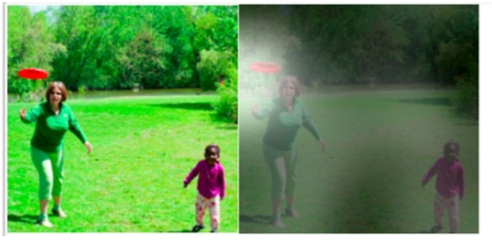


Figure 7: Attention outputs

The answer generation part is a multi-class classifier based on the original image feature map, the dense question embedding, and the attention weighted feature map. We also introduce the concept of histograms, as an additional feature vector as shown in Figure 5. This allows us to better model the question-answer pairs relating to the hue of the objects in the image. We later propose to do ablation studies to test the effectiveness of this designed feature. We employ the attention map to spatially weight the image feature map. The weighted image feature map focuses on the objects asked in the question. The spatial weighting is achieved by the element-wise production between each channel of the image feature map and the attention map. The attention weighted feature map lowers the weights of the regions that are irrelevant to the meaning of question.

A dense fully-connected layer with softmax activation, which is trained on the final projected features, predicts the index of an answer word specified in an answer dictionary. The answer generated by ABC-CNN is the word with the maximum probability. The full approach is described as a flowchart in Figure 8. The tensor-board flowchart of the algorithm is given in Figure 9

4. Results

We train the algorithm on the MS-COCO dataset. We initially extract the questions which have a single-word answer, and are covering the topics of numerals, color or other distinct field nouns. This simplifies the working of the algorithm and allows us to extract newer domain features designed for specific question, like the histogram for the color.

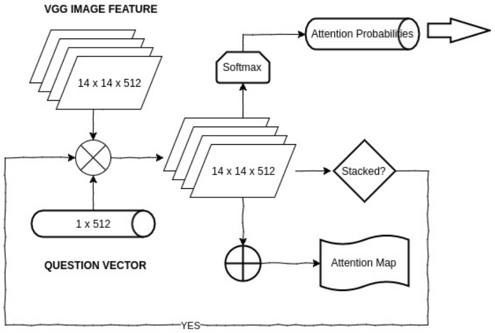


Figure 8: Our Approach - Flowchart

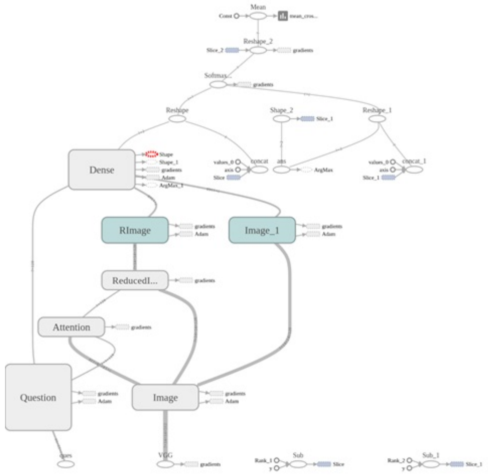


Figure 9: Tensorboard flowchart

Our whole framework is trained in an end-to-end way with **Adam** gradient descent algorithm. Each batch of the stochastic gradient descent randomly samples 32 image question pairs independently, and back propagation is applied to learn all the weights of the ABC-CNN architecture. We randomly adjust the initialization weights of all the layers to ensure that each dimension of the activations in all layers has zero mean and one standard variation. The initial learning rate is set to be 0.1. In our experiments, the weights in image feature extraction part are fixed to allow faster training speed, although it is possible to train all the weights in ABC-CNN in an end-to-end way. During the testing stage, an image feature map is extracted for each image. Given a question, we can produce its dense question embedding, and utilize the question embedding to generate the attention map. The multi-class classifier generates the answer using the original feature map including histograms, the question embedding, and the attention weighted feature map.

Table 1: Parameter configuration

Parameter	Configuration
Framework	Tensorflow
Cost Function	Cross-Entropy
Optimizer	Adam
Word Embedding	GloVe
Image Embedding	VGG-16
Batch Size	32
Epochs	32
Training Time	20 hours

4.1. Implementation details

The algorithm was implemented in the Tensorflow framework. We used a $8GB, i7$ processor for local implementation. We ran the algorithm in the AWS machine with an Amazon Linux AMI, $8GB$ GPU. The parameter configurations is shown in Table 1

The LSTM is performed as a hidden-box computation with default weights given by the Tensorflow module. The weights of the forget and allow gates are initialized randomly. The image input size is pre-processed to 14×14 and the resolution of the image map is 3×3. Each image cell generates a $14 \times 14 \times 512$ dimensional image feature vector using a pre-trained VGG network. We append the histogram of the image to it, as a priori information regarding the image.

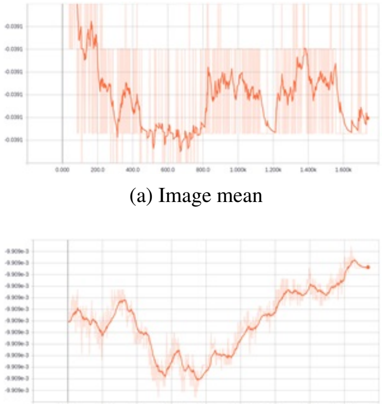
The GloVe embedding is directly performed as a dictionary lookup from the downloaded GloVe dictionary. We can re-train the dictionary, if needed, but since the questions were simple and rudimentary the pre-trained dictionary works well for the task.

The dense layer is designed as a 4096×1 network, or as a double layered perceptron. This allows the fully connected component in the network, to map the question-image pair to the respective answer word. This mapping is trained using back-propagation.

The performance plots from the reduced images and attention layers are shown in Figure 11. The saturation at each epoch shows that the network is training. The average value of the reduced image and attention map increases and saturates, indicating that the maps are being built and learnt. But, we still do not have a validation of the attention maps. The image and question variation vs Epochs is shown in Figure 10. This shows the diversity of image-question pairs in the training stage to generalize the learning.

The accuracy and cross-entropy plots are shown in Figure 12

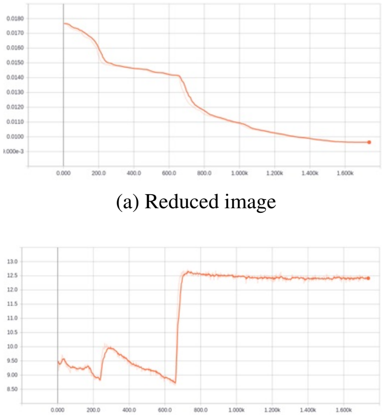
Our network performs far better than the baseline model that had an average accuracy of around 33%. We can see that our network performs on par with the ABC-CNN model, and that better configurations, architecture and feature designs would greatly improve the performance of the model.



(a) Image mean

(b) Question mean

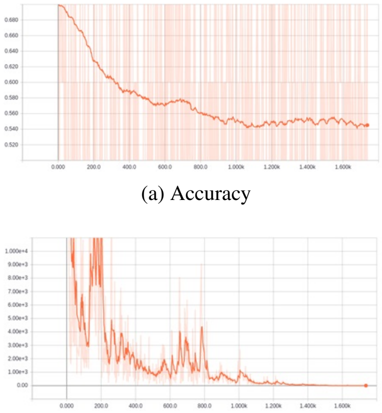
Figure 10: Image-Question pair variation per epoch



(a) Reduced image

(b) Attention map

Figure 11: Performance plots



(a) Accuracy

(b) Cross-Entropy loss

Figure 12: Accuracy plots

The figures $[13, 14, and 15]$, show the correct and incorrect predictions by the network. Also, Figure 16 shows the attention maps for a given image-question pair. we can clearly see that the network identifies specific aspects of the image, hidden or otherwise, like shadows, occluded benches etc. Also it is able to differentiate between a numeral question and a Yes/No question in Figure 13. The network is also able to differentiate between the white color of the shirt of the person, and the white background in the image. It is also able to identify the person in the foreground and avoid the people in the background while answering the question, showing the development of the attention layer.

Also in Figure 14, the network is able to correctly identify the Frisbee even when it is heavily reflecting sunlight. This means the network is trained to identify the frisbee with illumination invariance. Also the position of the person throwing it, might be a contributing factor. This image proves the advantage of attention layers, since normal CNNs would not be able to identify these features separately, due to a very dense green background. The other answers are also in the same domain; basketball, ball etc. proving the effectiveness of the attention layer. The out-of-the-domain answer 'black' could be due to the addition of the histogram as a feature. Similarly in the second image, it is able to identify the surfer from the background, and is also identifying that the surfer has fallen from the surfboard, which is a great bit of detail to be learnt. This proves that the network is able to not only learn syntactic information, but also is able to generalize, infer and learn semantic truth about the image, like a human does. Again, the introduction of colors in the top-5 answers, could be due to the histogram feature.

In Figure 15, we can see the incorrect choices made by the network. Although they are incorrect, we can see that the network has learnt well, and mostly the ground-truth answer lies in the top-5 answers, or the network atleast identifies the domain correctly. In the last image, the directions are listed, and it is a ill-made question, since the answer is subjective. Still the network learns to correctly list out directions, from the word 'facing' in the question, againg proving the learning of semantic information.

Figure 16 shows the attention maps embedded on the image. We can clearly see that the hot-spot in the attention map is the exact region we as humans focus on, when we are asked the same question. This proves the learning of the attention layers, and its effectiveness to reduce the image, and train the dense layer properly.

The validation accuracy is given in Table 2.

5. Conclusion

In this paper, we approached the problem of Visual Q&A, and proposed a unified attention based convolutional neural network, inspired from ABC-CNN. Our model unifies the visual feature extraction and semantic question understanding via question-guided attention map. ABC-CNN significantly improves both visual question answering performance and the understanding of the integration of question semantics and image contents. We stack the attention layers along with histogram features, and VGG16 features, and calculate a reduced image, that is passed on to a dense fully-connected stage that performs soft-max to classify the output. We are interested only in single-word answers such as Yes/No questions, questions based on color or single-word nouns etc. We proposed that inclusion of histogram features would force the network to identify colors better, and we can see significant improvement in the results.

Table 2: Validation Accuracy

Question Class	Accuracy
Baseline average	**33.68%**
Yes/No	62.04%
Numbers	30.23%
Color	35.41%
Single-word Noun	22.12%
Total	**47.01%**

6. Future Work

We propose to work more on the feature extraction of the image. We believe that question-specific semantic information in the image would help in the domain knowledge about the question. We would like to perform ablation studies and verify that the inclusion of these features do improve the results. Also, we would like to propose stacked architectures of the attention maps, that focus on parts of image in iterations, like a sieve, to get a particular answer, which is higher in confidence.

7. Contributions

We confirm that this paper was a team effort, both in its idea and execution. We would like to credit the contributions of our team members. Aditya Raj Verma, worked on synthesizing the idea and tried out simple models to solve the problem and additional network structures such as implementing stacked networks of the model. Lenord Melvix, worked on the Tensorflow implementation of the ABC network. Srinath Narayanan worked on the literature review, implementing Protocol Buffer [16] for the input pipeline, and the writing of this paper.

8. Acknowledgements

We would like to thank our Professor, Dr. Zhuowen Tu, for his sustained advice and motivation throughout the course. We would also like to thank our TA, Mr. Saining

Xie, for his thoughts on our project. We would finally like to thank UCSD, for providing AWS credits, enabling us to work on our research.

References

[1] W. S. Lasecki, *Crowdsourcing for Deployable Intelligent Systems*. AAAI, 2013.

[2] R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, "Intelligent virtual agents," vol. 13, pp. 29–31, August 2013.

[3] T. zy. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014*, pp. 740–755, 2014.

[4] X. Chen, H. Fang, T. y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server." arXiv preprint, 2015.

[5] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Heyden A* (G. Sparr, M. Nielsen, and P. Johansen, eds.), vol 2353. Springer, Berlin, Heidelberg: Computer Vision — ECCV 2002. ECCV 2002. Lecture Notes in Computer Science, 2002.

[6] R. Socher, J. Pennington, E. Huang, A. Ng, and C. Manning, *Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions*. In Proceedings of EMNLP 2011, 2011.

[7] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks." arXiv preprint, 2014.

[8] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. In EMNLP, 2014.

[9] J. Donahue, L. A. Hendrikcs, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description." November 2014.

[10] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler, *What are you talking about? text-to-image coreference*. In CVPR, 2014.

[11] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning: The quirks and what works." arXiv preprint, 2015.

[12] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, (Washington, DC, USA), pp. 1521–1528, IEEE Computer Society, 2011.

[13] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring nearest neighbor approaches for image captioning." arXiv preprint.

[14] 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, 2015.

[15] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[16] "Protocol buffer by google,
https://www.overleaf.com/9983350ryjycysjvksf#/36684264/,"

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." arXiv preprint, 2014.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, *Distributed Representations of Words and Phrases and their Compositionality*. NIPS, 2013.

[19] W. Lasecki, "Crowdsourcing for deployable intelligent systems," 2013.

[20] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba, "Predicting motivations of actions by leveraging text," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[21] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba, "Predicting motivations of actions by leveraging text," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pp. 2997–3005, 2016.

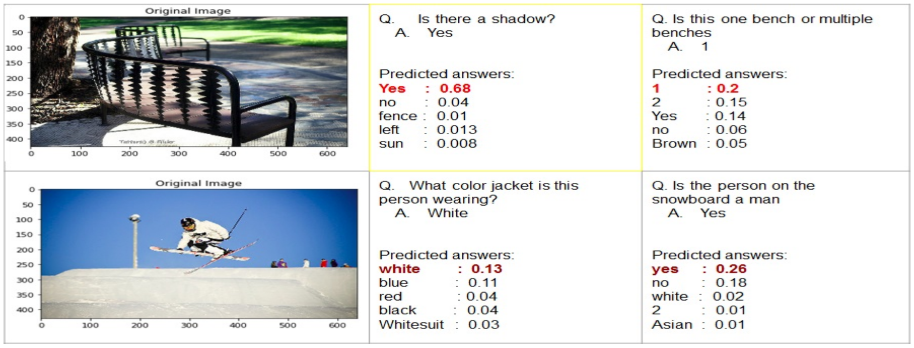
[22] A. Fukui, D. Huk Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," *ArXiv e-prints*, June 2016.

[23] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.

[24] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[25] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, pp. 289–297, 2016.

[26] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.



Q. Is there a shadow?
A. Yes

Predicted answers:
Yes : 0.68
no : 0.04
fence : 0.01
left : 0.013
sun : 0.008

Q. Is this one bench or multiple benches
A. 1

Predicted answers:
1 : 0.2
2 : 0.15
Yes : 0.14
no : 0.06
Brown : 0.05

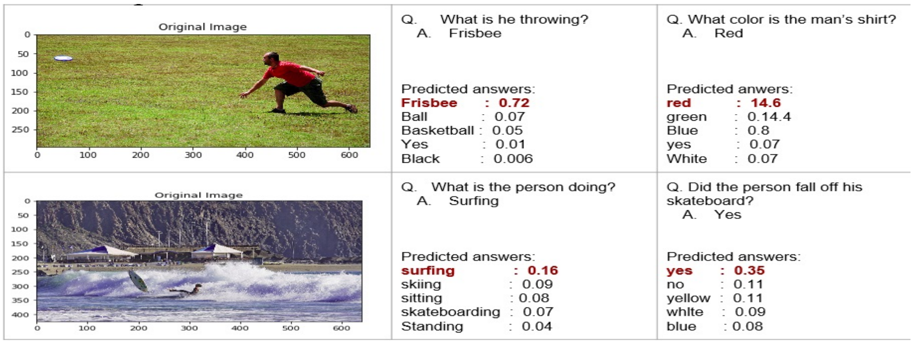
Q. What color jacket is this person wearing?
A. White

Predicted answers:
white : 0.13
blue : 0.11
red : 0.04
black : 0.04
Whitesuit : 0.03

Q. Is the person on the snowboard a man
A. Yes

Predicted answers:
yes : 0.26
no : 0.18
white : 0.02
2 : 0.01
Asian : 0.01

Figure 13: Correct Prediction-2



Q. What is he throwing?
A. Frisbee

Predicted answers:
Frisbee : 0.72
Ball : 0.07
Basketball : 0.05
Yes : 0.01
Black : 0.006

Q. What color is the man's shirt?
A. Red

Predicted anwers:
red : 14.6
green : 0.14.4
Blue : 0.8
yes : 0.07
White : 0.07

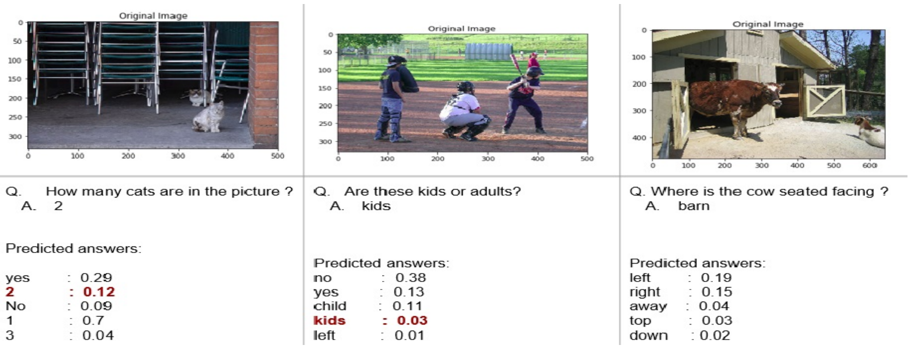
Q. What is the person doing?
A. Surfing

Predicted answers:
surfing : 0.16
skiing : 0.09
sitting : 0.08
skateboarding : 0.07
Standing : 0.04

Q. Did the person fall off his skateboard?
A. Yes

Predicted answers:
yes : 0.35
no : 0.11
yellow : 0.11
whlte : 0.09
blue : 0.08

Figure 14: Correct Prediction-1



Q. How many cats are in the picture ?
A. 2

Predicted answers:

yes : 0.29
2 : 0.12
No : 0.09
1 : 0.7
3 : 0.04

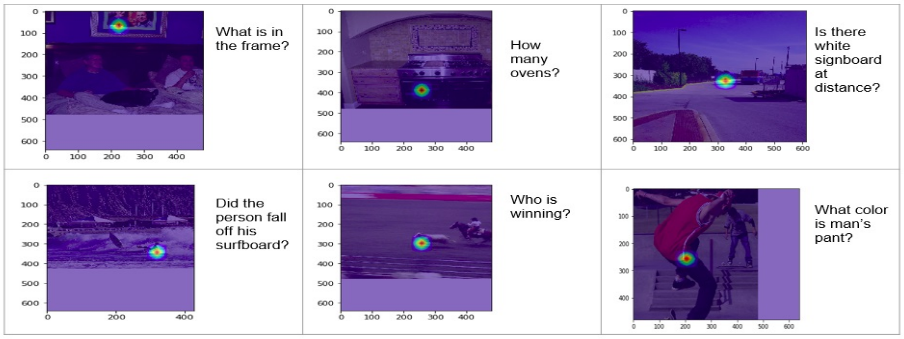
Q. Are these kids or adults?
A. kids

Predicted answers:
no : 0.38
yes : 0.13
child : 0.11
kids : 0.03
left : 0.01

Q. Where is the cow seated facing ?
A. barn

Predicted answers:
left : 0.19
right : 0.15
away : 0.04
top : 0.03
down : 0.02

Figure 15: Incorrect prediction



What is in the frame?

How many ovens?

Is there white signboard at distance?

Did the person fall off his surfboard?

Who is winning?

What color is man's pant?

Figure 16: Attention heat-map