Segmentation guided attention networks for Visual Question Answering

Vasu Sharma
Indian Institute of Technology, Kanpur
sharma.vasu55@gmail.com

Labhesh Patel
Abzooba Inc.
labhesh@gmail.com

Ankita Bishnu
Indian Institute of Technology, Kanpur
ankitab.iitk@gmail.com

Abstract

In this paper we propose to solve the problem of Visual Question Answering by using a novel segmentation guided attention based network which we call **SegAttendNet**. We use image segmentation maps, generated by a Fully Convolutional Deep Neural Network to refine our attention maps and use these refined attention maps to make the model focus on the relevant parts of the image to answer a question. The refined attention maps are used by the LSTM network to learn to produce the answer. We presently train our model on the visual7W dataset and do a category wise evaluation of the 7 question categories. We achieve state of the art results on this dataset and beat the previous benchmark on this dataset by a 1.5% margin improving the question answering accuracy from 54.1% to 55.6% and demonstrate improvements in each of the question categories. We also visualize our generated attention maps and note their improvement over the attention maps generated by the previous best approach.

1 Introduction

Visual Question Answering (VQA) is a recent problem in the intersection of the fields of Computer Vision and Natural Language Processing, where a system is required to answer arbitrary questions about the images, which may require reasoning about the relationships of objects with each other and the overall scene.

There are many potential applications for VQA. The most immediate is as an aid to blind and visually impaired individuals, enabling them to get information about images both on the web and in the real world.

The task of Image Question answering has received a lot of traction from the research community of late (Ren et al. (2015), Gao et al. (2015), Antol et al. (2015a), Malinowski et al. (2015)) due to the inherent challenging nature of the problem which involves combining question understanding in context of the image, scene understanding and common sense reasoning to be able to answer the question effectively. The problem is much more complicated than the purely text based Question answering problem which has been extensively studied in the past (Berant and Liang (2014), Kumar et al. (2015), Bordes et al. (2014), Weston et al. (2014)) and needs the model to be able to combine information from multiple sources and reason about them together.

Most recent approaches are based on Neural Networks, where a Convolutional Neural is first used to extract out image features and then these image features are used along with some RNN model to understand the question and generate an answer. However the problem with such approaches is that they do not know where to look. Recent approaches solve this problem by calculating an attention over the image by using the question embeddings to try and guide the model where to look, however such attention maps are still not very precise and not grounded at the image level. Moreover, there is no way to explicitly train these attention maps and the hope is that the model will implicitly learn them during training. In this paper we propose an approach which tries to guide these attention maps to learn to focus on the right regions in this image by giving them pixel level grounded annotations in the form of segmentation maps which we generate using a Fully Convolutional Deep Neural Network.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 43–48
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-3008

The rest of the paper is organized as follows. The existing literature on this problem is presented in Section 2 followed by a description of the datasets we used in Section 3. Section 4 introduces our approach and gives a detailed explanation of how we generate the segment maps and use them to guide our model to learn better attention maps which are subsequently used to perform the task of visual question answering. Finally we present the results in Section 5 and outline the papers conclusions and directions for future research in Section 6.

2 Literature Review

VQA is a fairly recent problem and was proposed by Antol et al. (2015b). Despite being a recent problem, several researchers from across the world have attempted to solve it. However, the performance still remains a long way off from the human performance which means there is still scope for improvement.

One of the early neural network based model for this problem proposed by Malinowski et al. (2015) combines a CNN and a LSTM into an end-to-end architecture that predict answers conditioning on a question and an image. In this model at each time step the LSTM is fed with a vector which is an one hot vector encoding of word in the question and the CNN encoding of the whole image. In Ren et al. (2015), a similar kind of approach was employed, with the main differnce that CNN features was fed to LSTM only once for each question; either before the question or after the last word of the question. This model achieved better accuracy than Malinowski et al. (2015).

In Agrawal et al. (2015) the best model model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet Simonyan and Zisserman (2014) to encode the images. Both the question and image features are then transformed to a common space and fused by a hadamard product and passed through a fully connected layer followed by a softmax layer to obtain a score over 1000 most frequent answers. The model proposed in Gao et al. (2015) had four components: Two separate LSTM modules for question representation and context of answer generated so far with a shared word embedding layer, a CNN to extract the image representation and a fusing component to fuse the information from other three components and generate the answer. All of these models look at the CNN feature of the whole image whereas to answer the real word questions concentrating to parts of the image is more useful in most of the cases. Many of the proposed VQA systems afterwards have incorporated spatial attention to CNN features, instead of using global features from the entire image. Both Shih et al. (2016); Ilievski et al. (2016) used Edge Boxes Zitnick and Dollr (2014) to generate Bounding Box proposals in the image. In Shih et al. (2016) a CNN was used for local features extraction of the images from each of these boxes. The input to their model was consisting of these CNN features, question features and one of the multiple choice answer. Weighted average score for each of the proposed region's features was used to calculate the score for an answer. In Ilievski et al. (2016) the authors use region proposals for the objects present in the question. At training time the objects labels and bounding boxes are taken from the annotation of COCO dataset and at test time bounding box proposals are classified using ResnetHe et al. (2015). Word2vecMikolov et al. (2013) is used to get a similarity between bounding box labels and objects present in question. Any bounding box with a similarity score greater than 0.5 is successively fed to an LSTM and at last time step the global CNN features for the image is also fed to the LSTM. A separate LSTM was used to represent the question. The output of these two LSTMs are then fed to a fully connected layer to predict the question. In Zhu et al. (2015) the model actually learns which region of the image to attend rather than feeding the model any specific region of the image. Here the LSTM is fed with the CNN feature of the whole image and the question word by word. Based on the image features and hidden state, the model actually learns which part of the image it should look at and generates an attention vector. This attention vector is operated on the CNN feature of the whole image resulting in some focused parts of the image. The model computes the log-likelihood of an answer by a dot product between CNN features of the image and the last LSTM hidden state.

We build on this model by proposing how to generate better attention maps and use them to

improve the performance on the VQA task.
Several newer approaches also propose novel methods of computing these attention maps. Notable among these are Z. Yang and Smola. (2015) and J. Lu and Parikh (2016). The former among these uses the question's semantic representation to search for the regions in an image that are related to the answer and used a multilayer approach to attend important parts of the image. In each layer of the attention it actually refines where to look at in the image.

3 Dataset

We did our experimentation on the Visual7W Dataset which was introduced by Zhu et al. (2015). Visual7W is named after the seven categories of questions it contains: What, Where, How, When, Who, Why, and Which. The dataset also provides object level groundings in the form of bounding boxes for the objects occuring in the question. The Visual7W dataset is collected on 47,300 COCO images. In total, it has 327,939 QA pairs, together with 1,311,756 human-generated multiple-choices and 561,459 object groundings from 36,579 categories. In addition, it also provides complete grounding annotations that link the object mentioned in the QA sentences to their bounding boxes in the images and therefore introduce a new QA type with image regions as the visually grounded answers.

We use this dataset for our task as we wanted to study how having pixel level groundings in form of segmentation maps affect each particular question type among how, when, where, why etc. We expect the improvement to be substantial for questions like 'how many' and 'where' which intuitively should benefit most from such pixel level groundings. This study allows us to validate this. We can also compare how these segmentation maps correspond with the provided object level groundings. Hence this dataset is our dataset of choice for this study.

4 Approach

We now present the approach we used to solve the problem of Visual Question Answering. A complete diagrammatic representation of our **SegAttendNet** model is presented in Figure 2. Each component of this model is explained in the subsequent subsections.

4.1 Generating segmentation masks for the image using the question

We first use the question to determine the objects whose segmentation maps we need to extract. This is done by using a POS tagging of the question to determine the nouns occurring in the question. After pre-processing these nouns, we match them to the 60 object categories from the Pascal context dataset Mottaghi et al. (2014) to know which of these objects might occur in the image. We then generate the segmentation maps from the question using the following steps:

- The Image is then fed to a Fully Convolutional Neural Network (FCN) Long et al. (2015), trained on the Pascal Context dataset to perform semantic segmentation on it based on the 60 classes of PASCAL Context dataset

- The FCN-16 feature map is generated using the architecture described in Figure 1. The lower resolution segment map (16X lower spatial resolution than the original image) is obtained from the fuse pooled layer, which combines both local features from lower layers and global features from higher layers to generate a segmentation map. We take a softmax over the 60 channels (corresponding to the 60 object categories) to obtain a probability map over the various classes.

- Now we extract the channels from this segmentation map which correspond to the nouns occurring in the question. We sum the segmentation map probabilities for these channels to obtain a single channel combined segmentation map. The intuition behind summing these channels is that, a particular pixel location in the image can have any of the objects occurring in the question with a probability which is the sum of the probability of each individual object occurring at that location.

- This map is further used in the attention network to refine the attention maps as described in the next subsection.

4.2 Using segmentation maps to guide the attention network for VQA

Once we have generated the segment maps and combined them into a single map based on the objects occuring in the question, we use this map to

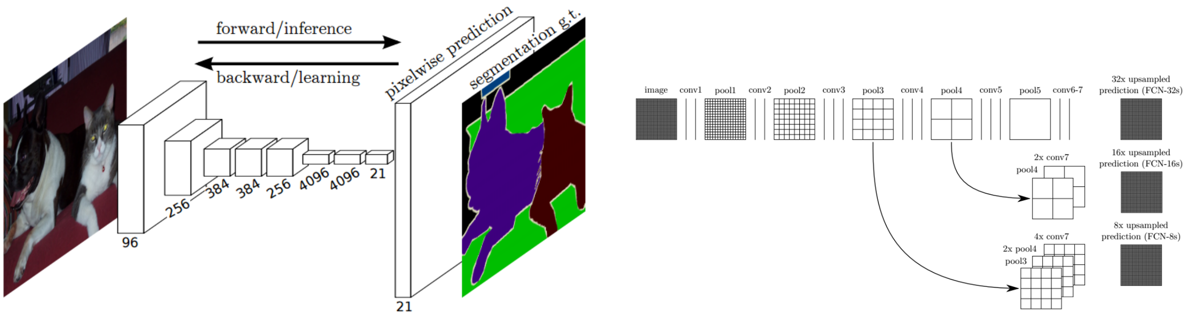

Figure 1: Fully Convolutional Neural Networks for Semantic segmentation

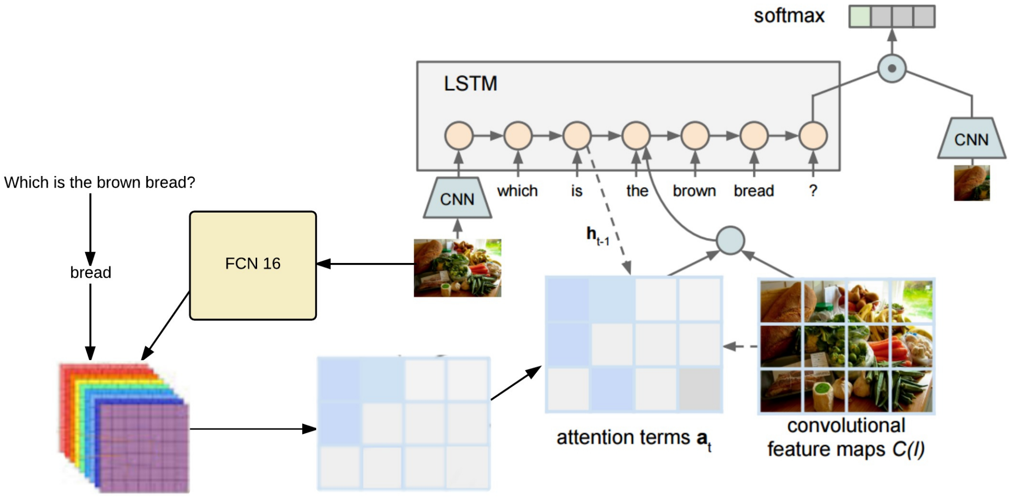

Figure 2: Our SegAttendNet for Visual Question Answering

Model	What	Where	When	Who	Why	How	Overall
Human(Question only)	0.356	0.322	0.393	0.342	0.439	0.337	0.353
Human(Question + Image)	0.965	0.957	0.944	0.965	0.927	0.942	0.964
Logistic Regression (Ques + Image)	0.429	0.454	0.621	0.501	0.343	0.356	0.359
LSTM (Question + Image)	0.489	0.544	0.713	0.581	0.513	0.503	0.521
Visual7W, LSTM-Attn(Ques+Image)	0.529	0.560	0.743	0.602	0.522	0.466	0.541
SegAttendNet(Ours)(Ques+Image)	**0.539**	**0.581**	**0.754**	**0.611**	**0.542**	**0.494**	**0.556**

Table 1: Comparison of results of our model against some existing approaches on the VQA task

guide our attention model to help it know where to look. We use the following steps to combine our segmentation maps with the attention based VQA network:

- The image is first passed through a VGG 16 network Simonyan and Zisserman (2014) in a feed forward manner and the fc7 features are extracted from the VGG network giving us a 4096 dimensional vector. These image features are fed as input to the LSTM at $t = 0$ and forms an initializing mechanism for the LSTM network.

- The question is passed through an LSTM network word by word, with a one hot word embedding being fed to the network at each time step. We also record the LSTM state at each time step. Lets say the previous such state was $h(t-1)$. The LSTM's ability to remember temporal context allows the network to understand the question with reference to the

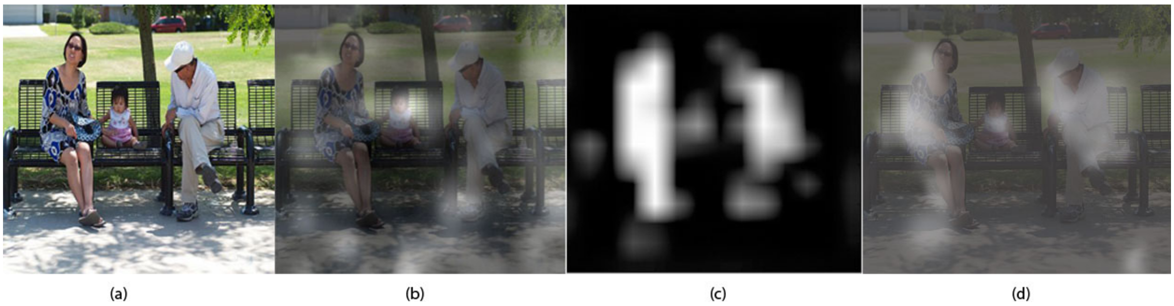

Figure 3: Question: "How many people are in the image?" Answer: "three"
a) Original image b) Attention map generated by previous state of the art approach c) Our low resolution segmentation map guidance d) Attention map generated by our SegAttendnet Model

input image and to subsequently refine it's internal representation at each time step based on the new input it receives.

- The above steps can be represented by the equations:

$$\begin{aligned} v_0 &= W_i[F(I)] + b_i, \\ v_i &= W_w[OH(t_i)], i = 1, ..., m \end{aligned}$$

Here F is the transformation function which uses the VGG's fc7 layer to convert an image into a 4096 dimensional embedding. $OH(.)$ represents he one-hot encoding for the word t_i. The weight matrices W_i and W_w embed the image and word embeddings into d_i and d_w dimensional embedding spaces such that d_i and d_w are both 512. The embedded image vector is used as the initial input to the LSTM network.

- Now lets call our segmentation map obtained from the FCN-16 as $S(I)$. Also let's call the pool5 features extracted from the VGG network as $C(I)$ Now we compute the attention by the following set of equations:

$$\begin{aligned} e_t = {W_a}^T \cdot tanh(W_{he}h(t-1) + W_{ce}C(I) \\ + W_{se}S(I)) + b_a \end{aligned}$$

$$a_t = softmax(e_t)$$
$$r_t = {a^T}_t \cdot C(I)$$

Here a_t is the generated attention map which helps the model decide how much attention to pay to various parts of the image by taking a dot product with the convolutional feature map of the image to generate r_t.

- Now this computed attention weighted convolution map is fed back to the LSTM network and the whole process repeats till the whole question is exhausted.

- In the end, the final state of the LSTM network and the pool 5 convolutional features are used to generate the final answer to the question. The end of the question is denoted by the question mark token.

- A decoder LSTM is used for open ended question and a softmax for multiple choice questions. In case of open ended questions, the previous word output is fed back to the LSTM network as input for generating the next answer word.

- A cross entropy loss is used to train the model using Backpropagation using Adam update rule. Hyperparameter tuning is done on the validation set and the results are reported after testing on a held out test set. The train, val and test sets are kept exactly the same as the original Visual7W paper to allow for a fair comparison. We also compare our approach with the human performance on this task.

5 Results

We evaluated our model for the telling questions in the Visual7W dataset using the approach we described in the previous section. The results of the same are presented in Table 1.

We note that our model outperforms the existing best reported result on this dataset by close to 1.5% margin. We also notice that we achieve substantial improvements in all the question categories. A closer observation of Figure 3 also reveals that our intuition that the model will perform substantially better on 'how many' and 'where' kind of questions does seem to be empirically justified as we can see a 3% improvement in the 'how' questions and a 2.1% improvement in the 'where' questions. Visualizing the attention maps also tells us that our attention maps are much more refined than the ones produced by the older approaches.

6 Conclusion and Future Work

In this paper we presented our model SegAttendNet to use segmentation maps to guide our attention model to focus on the right parts of an image to answer a question. We demonstrate that our model outperforms all other approaches on this dataset and attains superior performance in all question categories.

Right now we haven't tried combining our approach with more complicated attention mechanisms like the Stacked Attention Networks and Hierarchical Co-Attention networks. Our approach can easily be extended to the same and can help us achieve even better performances. We also plan to experiment with other much larger datasets which too can let our model train much better.

References

A. Agrawal, S. Antol, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. 2015. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468* .

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. 2015a. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468* .

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015b. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

J. Berant and P. Liang. 2014. Semantic parsing via paraphrasing. *In Proceedings of ACL* 2.

A. Bordes, S. Chopra, and J. Weston. 2014. Question answering with subgraph embeddings. *arXiv preprint arXiv:1406.3676* .

H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. *NIPS* .

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR* abs/1512.03385. http://arxiv.org/abs/1512.03385.

I. Ilievski, S. Yan, and J. Feng. 2016. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485* .

J. Yang D. Batra J. Lu and D. Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *CoRR.abs/1606.00061* .

A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. 2015. Ask me anything: Dynamic memory networks for natural language processing. *arXiv preprint arXiv:1506.07285* .

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. *CVPR* .

M. Malinowski, M. Rohrbach, and M. Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. *arXiv preprint arXiv:1505.01121* .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

M. Ren, R. Kiros, , and R. Zemel. 2015. Exploring models and data for image question answering. *arXiv preprint arXiv:1505.02074* .

K. J. Shih, S. Singh, and D. Hoiem. 2016. Where to look: Focus regions for visual question answering. *CVPR* .

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556. http://arxiv.org/abs/1409.1556.

J. Weston, S. Chopra, and A. Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* .

X. He J. Gao L. Deng Z. Yang and A. J. Smola. 2015. Stacked attention networks for image question answering. *CoRR, abs/1511.02274* .

Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei. 2015. Visual7w: Grounded question answering in images. *CoRR abs/1511.03416* .

C. L. Zitnick and P. Dollr. 2014. Locating object proposals from edges. *ECCV* .