Image Understanding via Continuous Questioning and Answering

Vatsal Mahajan (vmahaja1@asu.edu), and Saurabh Singh (ssing139@asu.edu), *ASU*

Abstract—This midterm report discusses the work done until now towards the completion of the project. In this project, we take inspiration from “Neural Self Talk: Image Understanding via Continuous Questioning and Answering” [Yang *et al.*, 2015] to work on the problem of continuously discovering image contents by actively asking image based questions and subsequently answering the questions being asked. We will create two modules, Visual Question Generation(VQG) and Visual Question Answering(VQA) module for which Convolution Neural Network and Recurrent Neural Network will be used. Both VQG and VQA are trained with different networks where, VQG uses image as input and corresponding question as output whereas VQA module uses images and questions as input and corresponding answer as output.

1 Problem Description

One of the important research work going in deep neural network is Image captioning and Visual Question Answering. This task is to include the semantic analysis of the image and answering the question referring to the image. The project proposes an end to end system that can continuously discover novel questions on an image, and then provide legitimate answers. This “self talk” approach for image understanding goes beyond as just a visual task, but to solve an interdisciplinary AI problem in vision & language.

2 Related Work

We are witnessing a renewed interest in interdisciplinary AI research in vision & language. The most established work in the vision & language community is ‘image captioning’, where the task is to produce a literal description of the image. It has been shown [Devlin *et al.*, 2015; Fang *et al.*, 2014] that a reasonable language modeling paired with deep visual features trained on large enough datasets promise a good performance on image captioning, making it a less challenging task from language learning perspective.

Previous approaches of question generation from natural language sentences are mainly through template matching [Brown *et al.*, 2005]. In [Yang *et al.*, 2015] they propose a visual question generation module through a technique directly adapted from image captioning system [Karpathy and Li, 2014], which is data driven and the potential output questions space is significantly larger than previous template based approaches, and the trained module only takes in image as input.

In the field of Visual Question Answering, very recently researchers spent a significant amount of efforts on both creating datasets and proposing new models [Antol *et al.*, 2015; Malinowski *et al.*, 2015; Gao *et al.*, 2015]. Interestingly both [Antol *et al.*, 2015] and [Gao *et al.*, 2015] adapted MS-COCO [Lin *et al.*, 2014] images and created an open domain dataset with human generated questions and answers. More recently, the work from [Ren *et al.*, 2015] reported state-of-the-art VQA performance using multiple benchmarks. The progress is mainly due to formulating the task as a classification problem and focusing on the domain of questions that can be answered with one word.

3 Approach

The paper [Yang *et al.*, 2015] describes an approach for Image Understanding via Continuous Questioning and Answering. (1) We can use the current state-of-the art image captioning systems to generate questions. (2) And then by using the previous output, a system could be trained from human like self-talk. The approach is focused on building 2 modules- (1) Question Generation Module, and (2) Question Answering Module. Fig. 1 shows the architecture described in the paper [Yang *et al.*, 2015].

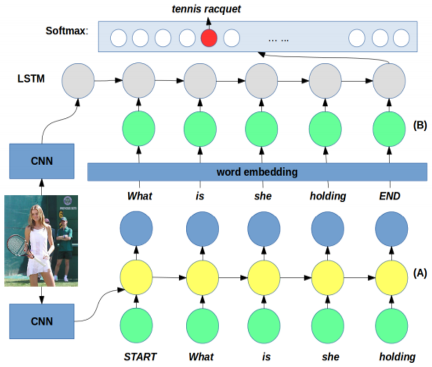
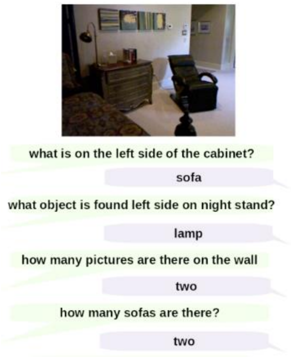


Fig. 1. The architecture of question generation module (part A) and question answering module (part B), and how they are connected as in ‘Neural Self Talk’.

Our work is related to three lines of research of natural image understanding: 1) question generation, 2) visual question answering.

4 Expected Result

Given an image the systems iteratively for N times (typically N=5) generates a question and passes it through the VQA system along with the image to achieve the answer. Fig. 2 shows the expected result. For this project, we will be using the MSCOCO-VQA dataset [Antol *et al.*, 2015].



what is on the left side of the cabinet?

sofa

what object is found left side on night stand?

lamp

how many pictures are there on the wall

two

how many sofas are there?

two

Fig. 2. Expected result of the system.

5 Method

5.1 Question Answering Module (VQA)

We implemented the approach from [Antol et al., 2016], which introduced a model builds directly on top of the long short-term memory (LSTM). We used the pretrained network parameters from the original paper for this model. This model combine features from an image using VGG Net [Simonyan et al., 2015] and embedding of the question using LSTM. Then runs a multi-layer perceptron on top these features. The last layer is a SoftMax over K possible outputs. Here K = 1000 most frequent answers from the training set. The Fig. 3 shows the best performing model [Antol et al., 2016].

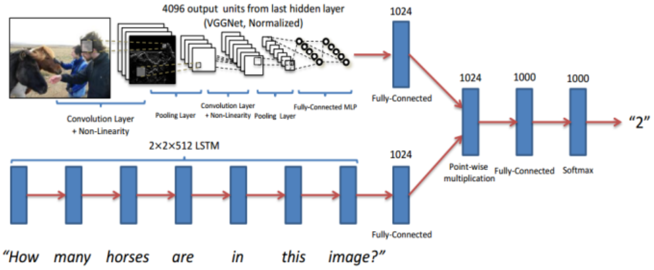


Fig. 3. The model uses a two-layer LSTM to encode the question and the last hidden layer of VGGNet for encoding the image. Both the question and image features are transformed to a common space and concatenated. Then send to a fully connected layer followed by a SoftMax layer to obtain a distribution over 1000 answers.

For image embedding the features form last hidden layer of VGGNet are used. This creates a 4096-dim image embedding. Further the activations of the last layer are l_2 normalized.

For question embedding an LSTM with two hidden layers is used. This creates a 2048-dim embedding for the question. The embedding is obtained by concatenating the last cell state and last hidden state representations of each the 2 hidden layers of the LSTM. Finally, a fully connected layer with a tanh non-linearity is used to o transform 2048-dim embedding to 1024-dim. The vocabulary contains all words seen in the training data.

The combined image channel (for image embedding) and the question channel (for question embedding) with the final fully connected layer, creates the end-to-end system for VQA model. The VGGNet parameters are frozen to those learned for ImageNet classification and not fine-tuned in the image channel.

Here the word embeddings are implicitly learned by the LSTM when the whole end to end system is trained together towards optimizing a common objective.

5.2 Question Generation Module (VQG)

We are implementing the method from [Karpathy and Li, 2014], where a simple but effective extension is introduced from previously developed Recurrent Neural Networks (RNNs) based language models to train image captioning model effectively. Here we take as input a set of images and questions corresponding to the images. The system uses an RNN [Karpathy and Li, 2014] to take as input a variable sized input, the question and an image. And learns to output a question given an image. The Fig. 4 shows the system design of the VQG module.

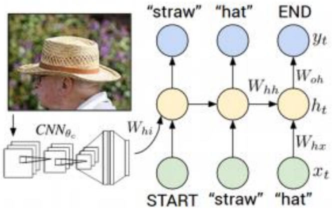


Fig. 4. Diagram of multimodal LSTM generative model. The LSTM defines a distribution over the next word in the question by taking in a word and context of the previous hidden layer. The image features are only given at t=0. START and END are special tokens.

The RNN takes as input the question $(x_1, \ldots, x_T)$ and image features for an I created by a VGG Net. The RNN then computes a sequence of hidden states $(h_1, \ldots, h_t)$ and a sequence of outputs $(y_1, \ldots, y_t)$ by going through the input sequence for t = 1 to T.

In the equations above, W_{hi}, W_{hx}, W_{hh}, W_{oh}, x_i and b_h, b_o are learnable parameters, and $CNN_{\theta c}$ (I) is the last layer of VGG Net. The output vector y_t gives the probabilities of words in the vocabulary; including a <START>, <END> and <UNK> symbol. The image context vector b_v was provided to RNN only at t=1. This is an empirical choice which gave a better performance. The size of the hidden layer of the RNN is 512 neurons. Instead of using RNNs to train the VQG module we used LSTMs. The above formulation remains the same. LSTMs were used over RNN because they are able to handle long-term dependencies as compared to RNNs.

$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t = 1) \odot b_v)$$
$$y_t = softmax(W_{oh}h_t + b_o).$$

The LSTM combines a word (x_t), with the previous context (h_{t-1}) to predict the next word (y_t). The LSTM's prediction is conditioned on the image context vector b_v on the first step. We use a pretrained VGG net with 16-layers for generating the image feature vector. We train on a specific

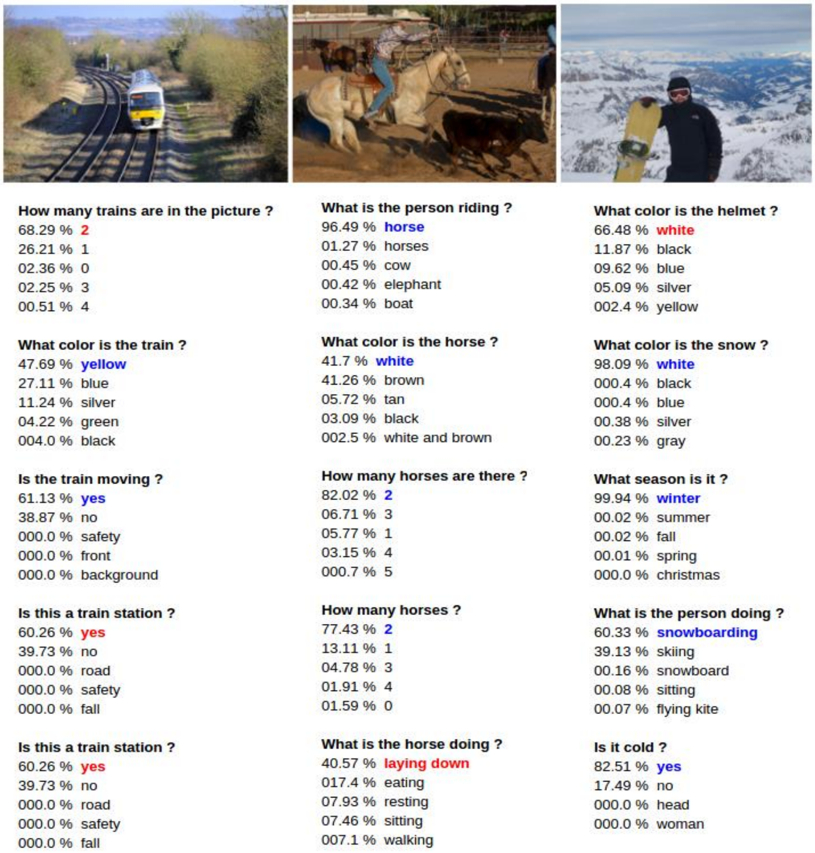


Fig. 5. This figure shows the results for the Continuous Question Answering Module. The images are from the MSCOCO dataset, the questions are generated by the VQG module and their respective answers given by the VQA module. The answers in blue are correct. Whereas, the answers in red are wrong.

input till the output word predicted by a sequence is a special <END> token. The cost function is to maximize the log probability assigned to the target labels (i.e. SoftMax classifier). At test time to predict a question, we compute the image representation b_v, set $h_0 = 0$, x_1 to the START vector and compute the distribution over the first word y1. We use beam search and argmax to sample 5 questions.

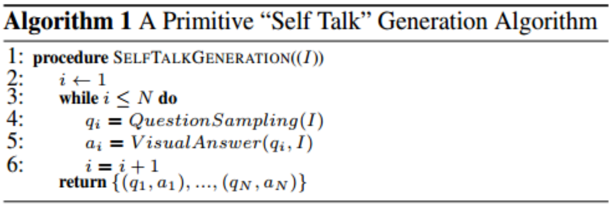
Here word embeddings used are separately learnt and not learned during the training of this module, unlike the previous VQA module.

Sampling. The output of an LSTM is a probability distribution over the vocabulary instead of just one word. While generating text we choose only one of the words ourselves given the probabilities and feed that back into the network. We explored the following sampling strategies: 1) Greedy Search or argmax - here at each iteration t we pick the word which has the highest probability. This strategy only generates one sample. 2) Beam Search – here instead of selecting the best next word it selects the best k words. k is called the beam width of this method. This strategy can generate multiple samples, and the best of those are often better than samples created using greedy search. 3) Random Sampling – This can generate multiple samples but the results are not usually comprehensible. In our experiments, we make use of greedy search and beam search.

5.3 Continuous Question Answering Module (VQG + VQA)

The two above described modules put together create the continuous question answering modules. Given an image the system samples a question from the VQG module and uses the VQA system to get the answer for the question. The system only samples five questions per image, but this can be extended to any arbitrary number. The Algorithm 1 [Yang *et al.*, 2015] further explains this module.

ALGORITHM 1 (VQG + VQA)



Algorithm 1 A Primitive "Self Talk" Generation Algorithm

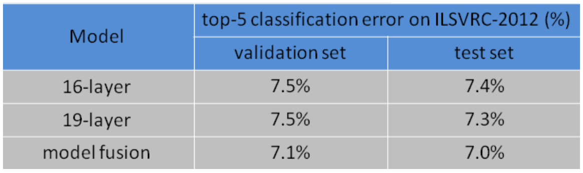
```
procedure SELFTALKGENERATION((I))
    $i \leftarrow 1$
    while $i \leq N$ do
        $q_i = QuestionSampling(I)$
        $a_i = VisualAnswer(q_i, I)$
        $i = i + 1$
    return $\{(q_1, a_1), ..., (q_N, a_N)\}$
```

6 RESULTS

Dataset. MSCOCO-VQA [Antol et al., 2015] is VQA dataset that contains open-ended questions about arbitrary images collected from the Internet. This dataset contains 443,757 questions and 4,437,570 ground truth answers based on 82,783 MSCOCO images. For training the VQA module we used 82,783 images and for training the VQG module we used 40,504 images. We made a 70-30 training and validation split. The variation of the images in this dataset is large and till now it is considered as the largest general domain VQA dataset. The effort of collecting this dataset cost over 20 people year working time using Amazon Mechanical Turk interface.

The Fig. 5 shows the results for the Continuous Question Answering Module (VQG + VQA). The results show images from the MSCOCO data-set along with questions generated by the VQG module, and answers generated by the VQA module. The first question is the argmax output and the other four are sampled using beam search of size 10. The answers in blue are correct. Whereas, the answers in red are wrong.

Also, we used a pretrained VGG net with 16 layers for extracting image features. VGG Net is very versatile, simple, relatively small and more importantly portable to use. The Fig. 6 the VGG 16 performance on ILSVRC-2012



Model	top-5 classification error on ILSVRC-2012 (%)	
	validation set	test set
16-layer	7.5%	7.4%
19-layer	7.5%	7.3%
model fusion	7.1%	7.0%

Fig. 6. The VGG 16 performance on ILSVRC-2012

7 Evaluation

For the generated question answer pairs, since there are no ground truth annotations that could be used for automatic evaluation, we would be using human evaluation. For evaluation, we will use these five metrics: 1) Readability Score: measures how readable is the generated conversation (range from 1 to 5). 2) Correctness Score: measures how correctly the content of the generated QA pairs describes the image content (range from 1 to 5); 3) Human-likeness Score: measures how human-like does the robot perform (range from 1 to 5). 4) Repetitiveness Score: measure similarity between the questions generated, we want the generated question not to be centered around a single mode (range from 1 to 5). 5) Generality: measures how generic is the conversion, if talks about a range of different objects in the mage or just a particular object (range from 1 to 5).

For human evaluation, we sample 100 images from the MSCOCO-VQA dataset. We generated the questions and answers for these images. We sampled 5 question, and used argmax to sample the first question and beam search for rest 4 questions. Finally, we both evaluated the results based on the above-mentioned parameters.

The Fig. 7 shows the average scores of these 100 test images for all the above five metrics. Here, we can see the average readability is high at a value of 3.65. This indicates that the VQG modules can learn a very good distribution over all questions to generate for a given image. The generality score of 3.32 is a very good measure that VQG generates questions that relate to different objects in the image rather a particular object. Thus, the VQG is able to generate varied questions. The corrected score falls around 2.76. This reflects that the VQA module is not performing close to human standards. Hence the correctness of the conversation drops. This is also reflected the human-likeness sore. We see a repetitiveness of almost 1. The does not have much semantic understanding when it generates two questions which are similar. The future work will focus to reduce this score.

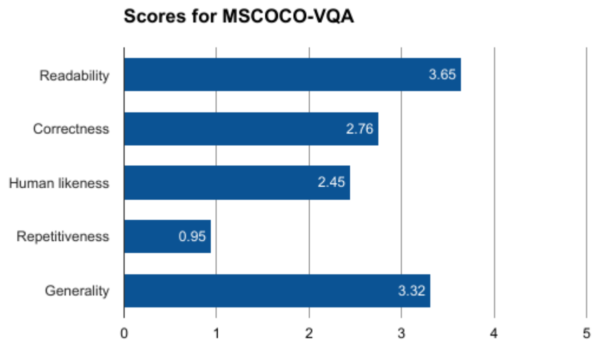


Fig. 7. The average scores of 100 test images for all the mentioned five metrics.

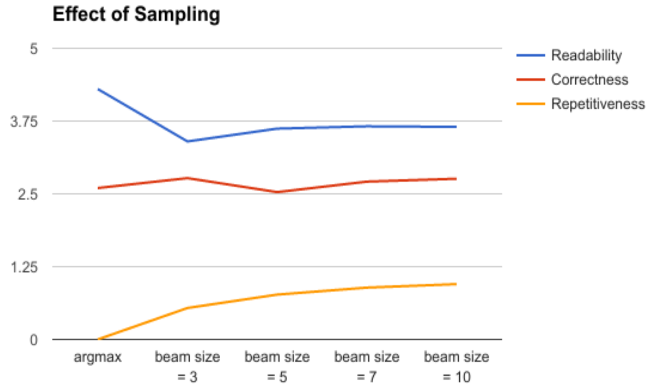
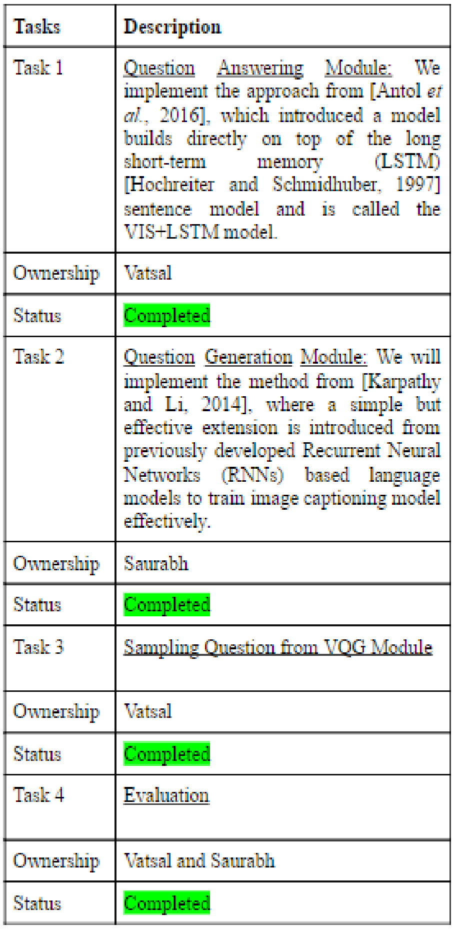


Fig. 8. This shows how different sampling methods affect readability, correctness and repetitiveness.

Further the Fig. 8 shows how different sampling methods affect readability, correctness and repetitiveness. Greedy search or argmax gives only one output, hence we have repetitiveness score of zero for argmax. Here, we can see a trend in repetitiveness, as we increase the beam search width the model tends to sample more questions that have semantically similar meanings; i.e. their answers will be same. The readability score is maximum in case of argmax as it generated the most probable output from the distribution. The readable tends to reach a constant value as we increase the beam size. It takes a constant value of approximately 3.5. Correctness score does not seem to follow any trend. Which mean sampling different questions does not help the VQA module to increase this correctness. VAQ gives almost the same performance for different questions sampled. Finally, the results we presented in Fig. 5 correspond to the first question being sampled using argmax and rest four samples using beam search of width 10.

TABLE 1
TASKS AND OWNERSHIP



Tasks	Description
Task 1	Question Answering Module: We implement the approach from [Antol *et al.*, 2016], which introduced a model builds directly on top of the long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] sentence model and is called the VIS+LSTM model.
Ownership	Vatsal
Status	Completed
Task 2	Question Generation Module: We will implement the method from [Karpathy and Li, 2014], where a simple but effective extension is introduced from previously developed Recurrent Neural Networks (RNNs) based language models to train image captioning model effectively.
Ownership	Saurabh
Status	Completed
Task 3	Sampling Question from VQG Module
Ownership	Vatsal
Status	Completed
Task 4	Evaluation
Ownership	Vatsal and Saurabh
Status	Completed

8 TASKS AND OWNERSHIP

The Table 1 describes the tasks along completed during this project. Also, every task has the ownership assigned for each of the team members.

8 FUTURE WORK

Currently the answers generated by VQA model are one word. For future work, We will extend this module for multiple words [Malinowski *et al.*, 2015] answers. Also, Common-sense knowledge has a crucial role in question raising and answering process for human beings. To improve the readability of the conversation, we propose and are working towards adding knowledge constraint embedding to accommodate for common sense reasoning [Liu *et al.*, 2015].

REFERENCES

[1] [Yang *et al.*, 2015] Yezhou Yang, Yi Li, Cornelia Fermuller, Yiannis Aloimonos. Neural Self Talk: Image Understanding via Continuous Questioning and Answering. arXiv: 1512.03460, 2015.

[2] [Devlin *et al.*, 2015] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 100–105, Beijing, China, July. Association for Computational Linguistics.

[3] [Brown *et al.*, 2005] Jonathan C Brown, Gwen A Frishkoff, and Maxine Eskenazi. Automatic question generation for vocabulary assessment. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 819–826. Association for Computational Linguistics, 2005.

[4] [Karpathy and Li, 2014] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. arXiv preprint arXiv:1412.2306, 2014.

[5] [Antol *et al.*, 2015; Antol *et al.*, 2016] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In International Conference on Computer Vision (ICCV), 2015 and arXiv:1505.00468, 2016.

[6] [Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Com- ´ mon objects in context. In Computer Vision–ECCV 2014, pages 740–755. Springer, 2014.

[7] [Ren *et al.*, 2015] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. arXiv preprint arXiv:1505.02074, 2015.

[8] [Fang *et al.*, 2014] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng ´ Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2014. From captions to visual concepts and back. CoRR, abs/1411.4952.

[9] [Malinowski *et al.*, 2015] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neuralbased approach to answering questions about images. arXiv preprint arXiv:1505.01121, 2015.

[10] [Gao *et al.*, 2015] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. arXiv preprint arXiv:1505.05612, 2015.

[11] [Malinowski and Fritz, 2014] Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge. arXiv preprint arXiv:1410.8027, 2014.

[12] [Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[13] [Liu *et al.*, 2015] Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, Yu Hu. Learning Semantic Word Embeddings based on Ordinal Knowledge Constraints. Annual Meeting of the Association for Computational Linguistics (ACL) 2015, Beijing, China, July 2015.

[14] [Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524v5, 2014.

[15] [Mostafazadeh *et al.*, 2016] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, Lucy Vanderwende. Generating Natural Questions About an Image. arXiv preprint arXiv:1603.06059v3, 2016.

[16] [Simonyan *et al.*, 2015] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image

Recognition. arXiv preprint arXiv:1409.1556v6, 2015.

[17] [Uijlings *et al.*, 2013] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. IJCV, 2013