ISSN 2395-1621

Question Generation from Images and Semantics Checker

Isha Pisal, Yash Oswal, Mehul Shah, Rushikesh Rote, Prof. Pranali Chavhan, Prof. Snehal Rathi

isha.pisal@viit.ac.in
yash.oswal@viit.ac.in
mehul.shah@viit.ac.in
rushikesh.rote@viit.ac.in
pranali.chavhan@viit.ac.in
snehal.rathi@viit.ac.in

Department of Computer Engineering, VIIT Pune

ABSTRACT

A goal to generate questions based on images which are given as input to the system. The web application will extract keywords from the image which will be used for question generation. For extraction of keywords first of all image recognition will be done using neural network algorithms. We are mainly focusing on the sentence generation based on image recognition for building up the questions. Image recognition is done using TensorFlow. Also, semantics of the words extracted are been checked for appropriate sentence formation. The image recognition is done using Tensor Flow and meaningful question formation is implemented using POS Tagging. There are basic three modules used to reach the goal:

- **Image to Text Analysis**
- **Text to Question Generation**
- **Semantics checking of generated questions**

The above three modules will work in co-ordination internally in the system. User will need to give the image input only to module 1 i.e; image to text analysis and wait for the resultant question to generate.

ARTICLE INFO

Article History

Received: 8th March 2020

Received in revised form :

8th March 2020

Accepted: 10th March 2020

Published online :

11th March 2020

Keywords— Image Processing, POS Tagging, Tensor Flow, Semantics.

I. INTRODUCTION

To generate question manually, professors/teachers require a lot of time and efforts to find right questions and its marks respectively. Hence, an approach to do this task in less time and efficiently is through Data Science, Deep Learning and Natural Language Processing. The question generation from text is already implemented but, using an image and extracting data from it to generate question is our key objective. We are focusing on image recognition to extract image description in sentences. Questions generated will be fully checked to approve the grammatical rules using a semantic checker. Our method is to extract objects present in the image and use it as keyword for generating questions. Our system contains 3 modules:

1. Image to Text
2. Text to Questions
3. Semantics checking of questions

The three modules of our system will take image as an input and give text as an output.

II. Research And Survey

We have done research on our key topics and identified the papers published in the similar field. We used Google for the research and PyCharm for the implementing the code written using python 3.6. An open source software library TensorFlow is used for identifying objects in the image. The dataset used is Coco developed by Microsoft. Rule based POS Tagging is implemented for generating rules and using them for tagging words according to their parts of speech and forming a question sentence.

III. RELATED WORK

Dense image annotations: Dense image annotations: We require this in our module to densely annotate the image

contents. Annotation is done to label the image using bounding boxes or other annotation types, we are using bounding boxes to obtain the image label these boxes are imaginary boxes. Bounding box is one of the annotation type other annotation types are Semantic Segmentation, 2D or 3D Cuboid, Polygons, Polylines or Splines, Point annotations are most commonly used annotations. Image segments were annotated by Socher et al.[22] and Barnard et al.[12] who studied multimodal correspondence between images and words. In our project image is given as an input to obatain labels of contents of image with main aim of correctly labeling scenes, regions, objects with fixed set of categories.[1]

Generating descriptions: Images can not only be labelled but also can be describe using sentences. Retrieval problem occurs as number of approaches pose the task, where most close compatible annotation from training set is given to the test image [6, 8, 23, 3, 24], or to generate descriptions for image the training annotations are broken up and put together to form an meaningful sentence. Few approaches based on fixed templates with generative grammar [11, 25] or with image content help generate image caption, but variety of possible outputs with this approach are limited. Full sentence descriptions for images with a log bilinear model developed by Kiros et al.[10] this model uses fixed window context. We are making use of Recurrent Neural Network(RNN) model whoes probability distribution depends over next word in a sentence on all previously generated words. RNN is simpler than other models and gives an average performance. We use RNN model to obtain image description in sentence format which is important part for question generation. [1]

Grounding natural language in images: Visual domain consists of number of approaches for grounding texts and sentences. Words and images were associated through a semantic embedding by Frome et al [14] who inspired our approach. Our work is more closely related to Karpathy et al [9] who decomposed sentences and images into portion and using ranking objective infer their inter – modal alignment. Our model aligns communicable section of sentences which are interpretable, defines certain meaning and have no fixed length rather than using grounding dependency tree relation based model. [1]

Neural networks in visual and language domains: Images and words can be represented at high – level representation using multiple approaches. We are using CNN and RNN for object detection [26] and image classification is done using Convolutional Neural Network (CNN) which is amongst the powerful model in neural networks. In our model CNN and RNN are used for descripting image through labels and sentences. We use these Neural Network models on images which were previously used for language modeling. [1]

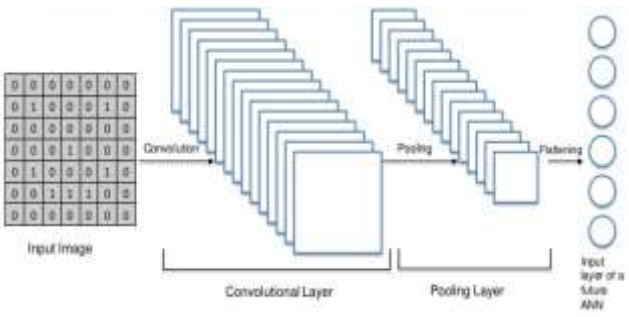
IV. OUR MODEL

Through previous published papers, we were able to learn the implemented and proposed methods for text and image processing. Initially we are using TensorFlow which is an open source software library for dataflow, it uses Coco dataset for identifying objects present in the image. The size of dataset is around 1GB but a bigger dataset can be used by using cloud computing. The keywords of the objects are extracted and semantics operation is been performed which will give an appropriate sentence. By using POS Tagging, the question is generated based on this keyword.

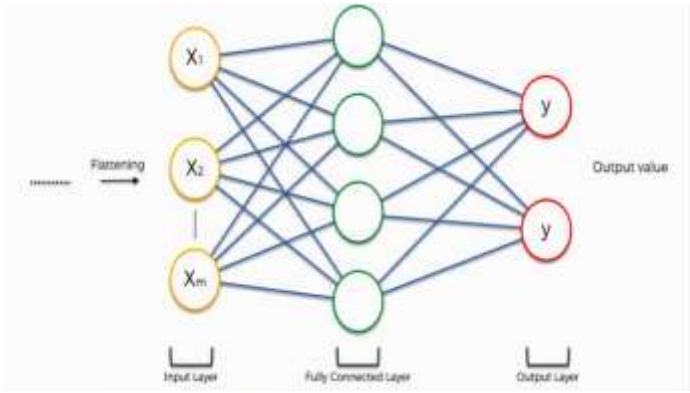
Our proposed system aims at generating the questions by analyzing vivid images. Soft-copy of the images is given to the system, and wh-question is the expected output. The system comprises of three modules, namely:

Image to Text Analysis: This is the first module of our proposed system, where the images are given as an input and the text is generated based on the analysis of the input image. There are various techniques that can be used for image classification and identification. Our proposed system uses multimodal RNN, Convolutional Neural Network. Our approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data.

Convolutional Neutral Network: Input image features are extracted by convolutional which then happens to be the first layer. image features are learnt with the help of small squares of input data which maintains and preserves the bond between pixels. image matrix and a filter or kernel are two inputs used in mathematical operations. convolutional neural networks class is one amongst the deep neural networks used for classifying and analyzing the images. CNN takes the input as an image and then further classifies it under certain categories. Internally the input image is seen as an array of pixels by computer (number of images) x (image width) x (image height) x (image depth) is a tensor shape depending on image resolution. this image input is further given for convolution within convolutional layer. After passing through convolutional the output obtained is an image abstracted to feature map with shape (number of images) x (feature map width) x (feature map height) x (feature map channels). [26] These feature maps are given as input to pooling layer which progressively reduces the spatial size of representation for network computation. Each feature map is operated independently in pooling layer. Max pooling is a commonly used technique in pooling. the result obtained from the input is further passed on to the next layer.

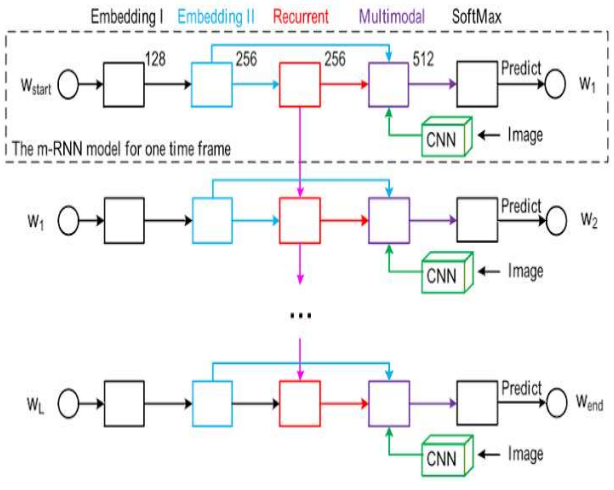


CNN layers and matrix

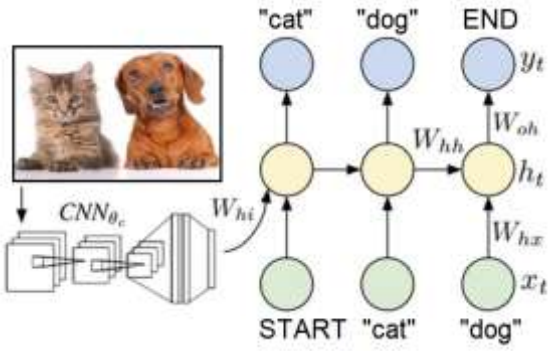


CNN fully connected model

Recurrent Neutral Network: a directed cycle is formed with connections in between its units which resembles a recurrent model; for example – future state is predicted by feedback connection from the current state. with this structure, network gets a complementary internal state. this state allows the network to show behavior which is dynamically temporal. tasks like sequence modeling, speech recognition, handwriting recognition and other natural language processing are effectively carried out by recurrent neural networks. vanishing gradient problem is one of the major drawbacks of RNN. this drawback features limitation of context range for input data as long dependencies are captured within the limited capacity of model. this problem was solved using long short term memory (lstm) which was proposed by hochreiter and schmidhuber. in this hidden layers were treated as multiple recurrently connected subnets, which were also called as memory blocks. these memory blocks were used to store and access information over long period of time via network. bidirectional network came into scenario when idea of unidirectional lstm network was extended by graves et al., bidirectional networks were preferred over unidirectional networks for their good improvement. multidimensional RNN came into existence by extending the one – dimensional RNN by grave et al. the idea is to replace 1d – RNN with multiple recurrent connections. graves et al also investigates deep structures of RNN.



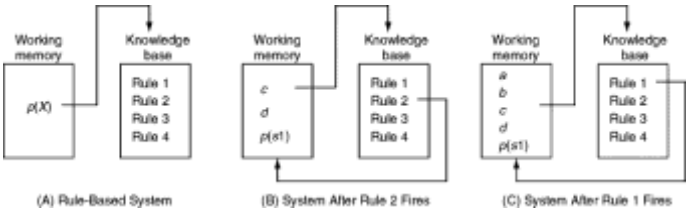
CNN flow diagram



CNN feature extraction

Text to Question Generation: After generating text from the image, the further task i.e. question generation is done using the text. Rule based techniques are used for this purpose where various question formation rules are defined. Context Free Grammar (CGF) rules are defined for formation of questions using text.

Rule-based Technique: By using a simple recognize-asset cycle, the inference cycle passes by it which is commonly known as forward chaining for data-driven reasoning, and backward chaining for goal-driven reasoning. When the premises of the rule are satisfied by the data, it concludes that the rules are true. Hence, this is the basic idea of forward chaining. The questions are framed as per the Context Free Grammar rules defined earlier.



Rule – based Technique

Semantics Checker for generated Questions: The questions that will be generated must be accurate grammatical format or in WH format. Therefore, the grammar of the generated questions must be checked by the third module. For this we propose to use Parts-of-Speech Tagging (POS) integrating with Markov Model.

Parts-of-Speech Tagging: In Parts of Speech Tagging, a sentence is converted to a form, list of words, and list of tuples. The form of tuple is (*word, tag)*. The tag describes whether the word is a noun, adverb, adjective, verb, prepositions, pronouns, etc. in the part of speech.

Part of Speech	Tag
Noun	n
Adjective	adj
Verb	vb
Adverb	ad
Pronoun	pn
Preposition	pre
Conjunction	con
Interjection	in

POS Table

The two stage architecture of Rule based POS Tagging are -

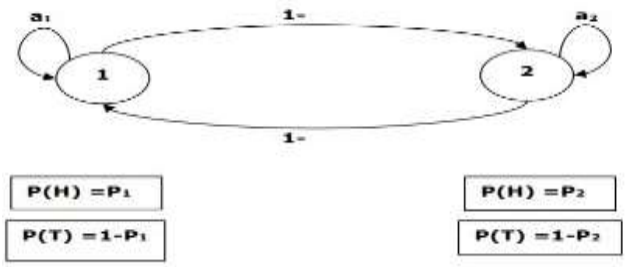
- **First stage** − part of speech for each word is assigned using a dictionary.
- **Second stage** − to sort the list of words to a single part of speech the hand written disambiguation rules are used.

Hidden Markov Model (HMM) POS Tagging:
Before digging deep into HMM POS tagging, we must understand the concept of Hidden Markov Model (HMM).

Hidden Markov Model:
An HMM model, may be defined as the doubly-embedded stochastic model, where the underlying stochastic process is hidden. The sequence of observations that produces is the hidden stochastic process that can only be observed through another set of stochastic processes.

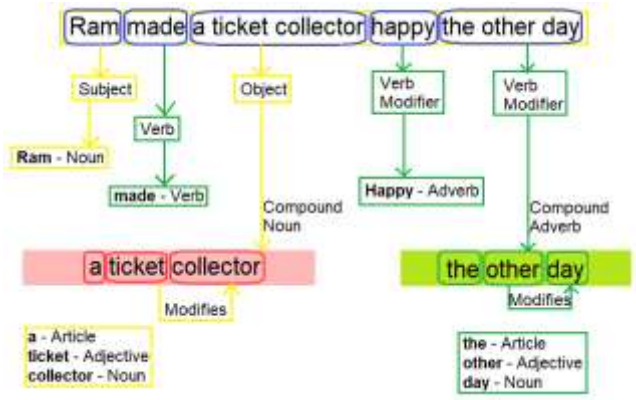
Example

For example, a sequence of hidden coin tossing experiments is done and we see only the observation sequence consisting of heads and tails. The actual details of the process - how many coins used, the order in which they are selected - are hidden from us. By observing this sequence of heads and tails, we can build several HMMs to explain the sequence.
Following is one form of Hidden Markov Model for this problem –
This way, we can characterize HMM by the following elements –



HMM example diagram

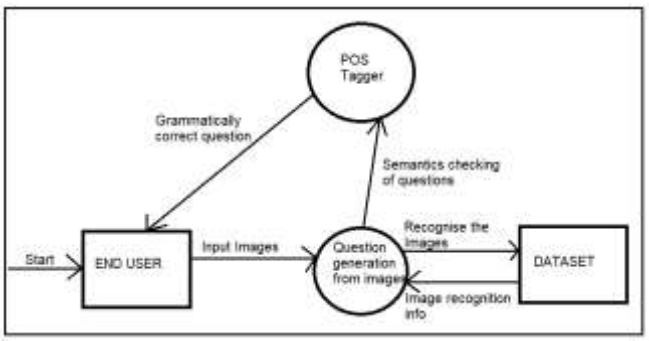
- N, the number of states in the model (in the above example N =2, only two states).
- M, the number of distinct observations that can appear with each state in the above example M = 2, i.e., H or T).
- A, the state transition probability distribution − the matrix A in the above example.
- P, the probability distribution of the observable symbols in each state (in our example P1 and P2).
- I, the initial state distribution.



POS extraction

The object's location with respect to some other object or the object's behavior is used for question generation based on the object identified. Semantics checker is applied to check the question's accuracy with respect to language, Part of Speech tagging, also called grammatical tagging, is the process of marking up a word in a text [corpus] or as a tag, which corresponds to a particular part of speech such as noun, adjective, verb, etc., based on both its definition and its context i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. POS tagging is done using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags. There are two distinctive groups under POS-tagging algorithm: rule-based and stochastic.

V. IMPLEMENTATION

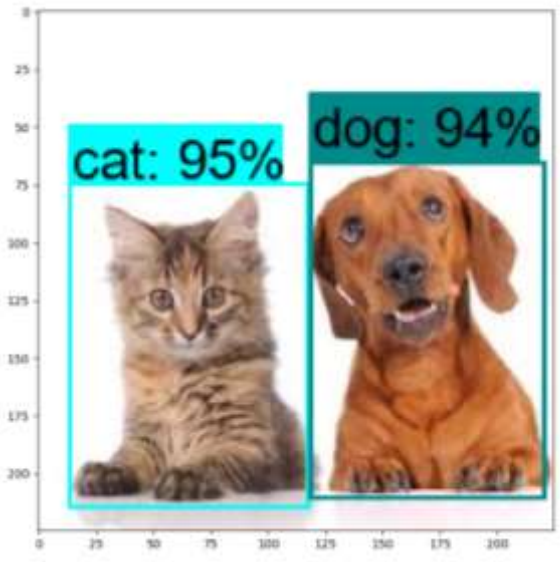


Data Flow Diagram

TensorFlow: TensorFlow is an open source software library. It is developed by Google Brain and is used for dataflow and machine learning applications such as neural

networks. TensorFlow has a flexible architecture which allows for the easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of servers to mobile and edge devices. TensorFlow computations are expressed as stateful dataflow graphs. It works on multidimensional arrays known as tensors, hence is named as TensorFlow.

TensorFlow is supported by various operating systems like Linux, Windows, macOS and also mobile computing platforms like Android and iOS. TensorFlow has its major application in image – captioning, speech recognition, self driving cars, sentiment analysis, text summarization, video tagging, etc.



Output of CNN using TensorFlow

Steps for generation of Image captions is as follows:

1) Data Preparation

Firstly, collect your data and put it in a form that the network can train which involves collecting images and labeling them. Even if the dataset is downloaded which someone else has prepared, you must preprocess and prepare the data before you use it for training. Data preparation involves dealing with things like missing values, corrupted data, data in the wrong format, incorrect labels, etc.

2) Creating the Model

You have to make various choices about parameters and hyper parameters while creating neural network model. Decisions such as number of layers to use in your model, what the input and output sizes of the layers will be, what kind of activation functions you will use, whether or not you will use dropout, etc have to taken by you.
Learning which parameters and hyper parameters to use will come with time (and a lot of studying), but right out of the gate there are some heuristics you can use to get you running.

3) Training the Model

Create an instance of the model and fit it with your training data, once your model is created. The time consumed by the model while training is the biggest parameter to be taken into consideration. You can specify number of epochs to train to manually specify the length of training for a network. The longer you train a model, the greater its performance will improve, but too many training epochs and you risk over fitting.
Choosing the number of epochs to train for is something you will get a feel for, and it is customary to save the weights of a network in between training sessions so that you need not start over once you have made some progress training the network.

4) Model Evaluation

Model evaluation can be done in multiple steps. The first step in evaluating the model is comparing the model's performance against a validation dataset, a data set that the model hasn't been trained on. You can analyse the performance of the model through different metrics by comparing the model's performance against this validation set.

Dataset: We use MSCOCO dataset, which occupies memory of 19GB consisting of total 164K images. These images are further classified as 118K training images, 5K validation images, 20K test-dev images, 20K test-challenge images.
We can also use Flikr8K and Flickr30K which consists of 8000 and 30000 images respectively.

VI. SUMMARY

Our proposed system is to solve the problems and reduce the efforts of professors in question formation for an examination. Many software are already developed for question generation from textual paragraphs and text format. Our software system leverages this concept of automatic question generation to a next level. We propose to generate accurate questions from the images as an input rather than any textual paragraphs or text format as an input.

Question generation from images and semantic checker helps to generate wh-questions by taking images as an input. It helps in checking the grammar of the generated questions as well. Our system follows a certain number of steps to generate the questions. Firstly, the image will be extracted from the document and stored in a file for input to the software system. These input images will be recognized using the MS COCO dataset with the help of TensorFlow python library. Once the images are recognized and the objects in the images is are known, a sentence will be formed w.r.t. to object positions in the images with the help of preposition identification.

POS tagging is done on these formed sentences for generation of syntactically and semantically correct and accurate questions.

VII. CONCLUSION

The system successfully identifies objects present in the image by using dataset. Word extracted by identifying object is used to generate question after successfully checking its semantics. Further recognizing complex images such as graphs will be implemented.

VII. ACKNOWLEDGEMENT

I would like to take this opportunity to thank my internal guide Prof. Pranali Chavhan for giving me all the help and guidance I needed. I am really grateful to them for their kind support. Their valuable suggestions were very helpful.

I am also grateful to Prof. Sachin Sakare, Head of Computer Engineering Department, VIIT for his indispensable support, suggestions.

At last we must express our sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped me directly or indirectly during this course of work.

VIII. REFRENCES

[1] Andrej Karpathy, Li Fei-Fei - "Deep Visual-Semantic Alignments for Generating Image Descriptions." in CVPR 2015.

[2] SanjaFilder, Raquel Urtasun - "A Sentence Is Worth a Thousand Pixels" Conference Paper in Proceedings/CVPR, IEEE Computer Society Conference on Compuuter Vision and Pattern Recognition. June 013.

[3] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In NIPS, 2011.

[4] Notes of Pushpak Bhattacharyya, CSE Dept, IT Patna.

[5] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava –" From Captions to Visual Concepts and Back" in CVRP 2015 Paper.

[6] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013.

[7] Kobus Barnard, Pinar Duygulu, David Forsyth, NandodeFreitas. Matching Words and Pictures. Journal of Machine Learning Research 3 (2003) 1107–1135

[8] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV. 2010.

[9] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. arXiv preprint arXiv:1406.5679, 2014.

[10] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539, 2014.

[11] M. Yatskar, L. Vanderwende, and L. Zettlemoyer. See no evil, say no evil: Description generation from densely labeled images. Lexical and Computational Semantics, 2014.

[12] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei,andM.I.Jordan. Matchingwordsandpictures. JMLR, 2003.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[14] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In NIPS, 2013.

[15] M.SchusterandK.K.Paliwal.Bidirectionalrecurrentneural networks. Signal Processing, IEEE Transactions on, 1997.

[16] Itzair Aldabe and Montse Maritxalar. Semantic Similarity Measures for the Generation of Science Tests in Basque. At IEEE Transactions on Learning Technologies.

[17] Dhawaleswar Rao Ch and Sujan Kumar Saha. Automatic Multiple Choice Question Generation from Text : A Survey. At IEEE Transactions on Learning Technologies.

[18] ShengfengHe , Chu Han, Guoqiang Han , and Jing Qin. Exploring Duality in Visual Question-Driven Top-Down Saliency. At IEEE Transactions on Neural Networks and Learning Systems.

[19] Junqiang Liu, Ke Wang, and Benjamin C.M. Fung. Mining High Utility Patterns in One Phase without Generating Candidates. At IEEE Transactions on Knowledge and Data Engineering.

[20] Junwei Bao , Yeyun Gong, Nan Duan, Ming Zhou, and Tiejun Zhao. Question Generation With Doubly Adversarial Nets. At IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[21] Kai Yu, Zijian Zhao, Xueyang Wu, Hongtao Lin and Xuan Liu. Rich Short Text Conversation Using Semantic Key Controlled Sequence Generation. At IEEE Transaction on ASL.

[22] R. Socher and L. Fei-Fei. Connecting modalities: Semisupervised segmentation and annotation of images using unaligned text corpora. In CVPR, 2010.

[23] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. TACL, 2014.

[24] Y.Jia,M.Salzmann,andT.Darrell. Learningcross-modality similarity for multinomial data. In ICCV, 2011.

[25] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daum´e, III. Midge: Generating image descriptions from computer vision detections. In EACL, 2012.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C.Berg, and L.Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

