Question Answering with Deep Learning: A Survey

Conference Paper · May 2019

CITATIONS
2

READS
1,034

3 authors:

Martina Toshevska
Ss. Cyril and Methodius University in Skopje
14 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Georgina Mirceva
Ss. Cyril and Methodius University in Skopje
49 PUBLICATIONS **231** CITATIONS
SEE PROFILE

Mile Jovanov
Ss. Cyril and Methodius University in Skopje
63 PUBLICATIONS **153** CITATIONS
SEE PROFILE

Question Answering with Deep Learning: A Survey

Martina Toshevska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
martina.toshevska@finki.ukim.mk

Georgina Mirceva
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
georgina.mirceva@finki.ukim.mk

Mile Jovanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
mile.jovanov@finki.ukim.mk

***Abstract*—Automatically generating answer for a given question is a process in which the computer is supposed to answer a question in a natural language where the question itself is also provided in natural language. Deep learning techniques gained extensive research in both fields of computer vision and natural language processing. Therefore, they are extensively applied for the task of question answering using wide varieties of datasets.**

This survey aims to overview some of the latest algorithms and models proposed in the field, as well as datasets exploited for training and evaluating the models. In this survey, the models are presented as part of one of the following groups: classical deep neural networks, dynamic memory networks and relation networks. Several datasets have been proposed specifically for the research on automatic question answering. This survey briefly overviews datasets for two different categories of question answering: textual and visual. In the end, evaluation metrics utilized in the field are presented, grouped as: metrics for evaluation of an information retrieval system and metrics for evaluating automatically generated text.

***Keywords*—Question Answering, Visual Question Answering, Textual Question Answering, Natural Language Processing, Computer Vision, Deep Learning**

I. Introduction

The ability to provide an answer to a natural language question is known as question answering. Automatically generating answer for a given question is a process in which the computer is supposed to answer a question in a natural language. The question itself is also provided in natural language. Thus, the computer needs to understand the question too. Based on the domain, questions can be classified as questions referring to mathematical or logical tasks, questions asked in online communities (known as Community Question Answering), questions in the medical domain etc. In accordance with the type of information they refer to, questions can be separated into questions concerning text (known as Textual Question Answering), questions concerning images (known as Visual Question Answering), etc.

Textual Question Answering (TQA) is the task of extracting a text snippet from a passage which corresponds to a specific question. This task differs from classical information retrieval since the output is a particular piece of information rather than a collection of documents. Its purpose is to create textual answer for a specific question. The question is either from a specific domain (such as science, math, etc.) or from a general domain. With the advent on online communities, such as Quora[1], Stack Overflow[2] and Stack Exchange[3], a new type of textual question answering has increased popularity - Community Question Answering (CQA). The goal of this task is to resemble actions performed by users in the community such as ranking answers according to their relevance, selecting the best answer for a specific question, identifying duplicate questions etc.

Visual Question Answering (VQA) is the task where questions are asked about given image. It has received attention from researches in both natural language processing and computer vision communities. In the most common form of this task, the computer is given an image and a question about the image. It is supposed to create an answer for the question, which is typically a word or phrase. Images are either natural or synthetic i.e. computer generated. The latter are referred as abstract scenes. The idea behind creating such synthetic sets is to focus only on high-level reasoning rather than low-level image processing.

Deep learning techniques [1] gained extensive research in both fields of computer vision and natural language processing. Therefore, they are extensively applied for the task of question answering using wide varieties of datasets. This survey aims to overview some of the latest algorithms and models proposed in the field, as well as datasets exploited for training and evaluating the models.

The rest of this survey is organized as follows. Section 2 gives an overview of algorithms applied in the field of question answering. Section 3 explores some of the datasets being used. Evaluation metrics are presented in Section 4 and at the end, Section 5 concludes the survey.

II. Algorithms

Deep learning techniques gained extensive research in the domain of question answering, either visual or textual. The following subsections present some of the models utilized in the field. The models are grouped into three groups based on the architecture: deep neural networks, dynamic memory networks and relation networks.

[1]https://www.quora.com/, last visited: 24.01.2019
[2]https://stackoverflow.com/, last visited: 24.01.2019
[3]https://stackexchange.com/, last visited: 24.01.2019

A. Deep Neural Networks

There is a variety of deep neural models proposed in the field of question answering. The most common architectures consist of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The first two models described in this subsection address the problem of Community Question Answering (CQA), while the third model focuses on Visual Question Answering (VQA).

The first model, SwissAlps [2], utilizes CNN for computing similarity between question and its candidate answer. First, both the question and the answer sequences are processed by an embedding layer creating matrix representation for the sentences. Attention matrix is then calculated by computing pairwise similarity between the word embeddings of these matrices based on Euclidean distance. The attention matrix is multiplied by two weight matrices in order to generate attention features. These features are stacked on top of the sentence matrix creating three-dimensional array. The purpose is to give higher weight to the relevant part of the sentences. Such arrays are fed into a convolutional layer which creates a feature vector for the sentences by applying a set of convolutional filters. Attention values are generated by summing the attention matrix column-wise for the question and row-wise for the answer candidate. These values are used to weight the feature map matrices obtained by the convolutional layer. Standard max pooling is applied to the attention weighted feature map matrices. Finally, the map matrices are fed through a fully connected layer followed by a softmax regression layer.

The second model, FuRongWang [3], allows using either CNN or RNN for the processing of question and answer. In the same way as the previously described model, both question and answer are represented with a matrix composed of word embedding vectors. An augmented feature vector is added at the tail of these matrices. For each word in the question, the question augmented feature vector has value 1 at the corresponding position if it is present in the answer and value 0 otherwise. The same holds for the answer augmented feature vector. Next part of the model is a neural network which can be convolutional or recurrent. The convolutional network is represented as a convolutional layer consisted of several feature maps followed by max pooling. The recurrent network is represented as a bidirectional LSTM which processes the sentence in both directions. The output is a vector representation of the sentences. Another part is an interaction layer which calculates the relevance between question and answer by multiplying a weight matrix with the answer feature vector from right and with the transpose of the question feature vector from left. Features extracted by the neural networks altogether with extra features from the interaction layer and augmented features are concatenated and fed into a fully connected layer. In the end, softmax function is applied.

The SAN model [4] applies multi-step attention for the problem of VQA. A convolutional neural network produces an image feature map for the image regions and another convolutional network or LSTM encodes the question. The answer is then generated with an attention mechanism as described below. The image feature map and the question vector are combined, and then attention weights are produced with a softmax function. A weighted sum of region vectors is calculated based on the attention weights. This sum is then combined with the question vector forming a vector called refined query vector, which encodes information about the question and the relevant part of the image. For complicated questions that require more complex reasoning, single attention is not sufficient. Therefore, previously described attention mechanism is repeated multiple times before inferring the final answer by a softmax function.

B. Dynamic Memory Networks

Dynamic Memory Network (DMN) [5] is a framework based on neural networks capable of solving sequence tagging tasks, classification problems, sequence-to-sequence tasks and question answering tasks that require transitive reasoning. The DMN first computes a representation for all inputs and the question. An input module encodes input sequences into distributed vector representations. The raw text input is first transformed into word embedding vectors and is then fed through recurrent neural network that creates vector representation. End-of-sentence token is inserted after each sentence. The final vector representation is composed of the hidden state at each end-of-sequence token. A question module encodes the question into a distributed vector representation. Analogous as in the input module, the question is first converted into word embedding vectors and then fed into a recurrent network for creating the representation. Gated Recurrent Unit (GRU) is used in both input and question module as a recurrent neural network. The question representation then triggers an iterative attention process that searches the inputs and retrieves relevant facts. Given a collection of input representations, an episodic memory module chooses which parts of the inputs to focus on through the attention mechanism. It is a two-layer feed forward network that computes scalar score based on the input vector, previous memory and question vector. This score is used to weight the input sequence using a GRU in order to compute the episode. The episode memory module may pass over the input multiple times, updating episode memory after each pass. Each iteration provides the module with newly relevant information about the input and by the final iteration the episodic memory should contain all the information required to answer the question. Finally, an answer module generates the answer based on the final memory vector of the episodic memory module and the question itself.

DMN+ [6] is a modification of DMN that proposes modification of input representation, attention mechanism and memory update. The first modification is replacing the GRU in the input module with two different components: sentence reader (positional encoder adapted from [7]) and input fusion layer (bi-directional GRU). Beyond text, DMN+ can process image as input. This visual input module is composed of three parts: local region feature extraction (extracting features with convolutional neural network based on VGG-19 [8]),

visual feature embedding (linear layer with tanh activation that projects the local regional vectors to the textual feature space used by the question vector) and input fusion layer (bi-directional GRU). The second modification is updating the memory in the episodic memory module with Rectified Linear Unit (ReLU) layer instead of GRU. The last modification is the attention mechanism. The attention is implemented by associating a single scalar value called attention gate with each input fact. Two different mechanisms are proposed: soft attention, which produces a contextual vector through a weighted summation of the sorted list of input fact vectors and corresponding attention gates, and attention based GRU that is a modification of the standard GRU by incorporating attention gates.

C. Relation Networks

Relation Network (RN) [9] is a neural network module with a structure primed for relational reasoning. The main idea behind this network is the ability to compute relations without the need to be learned in a way in which recurrent neural networks learn to capture sequential dependencies and convolutional neural networks learn spatial dependencies. One model encompassing RNs is presented in [10]. In the first step, the image is embedded using a Faster R-CNN embedding method [11] creating a feature map for each region of interest, while the question is embedded using a GRU. Next step is applying visual attention to focus on important image regions. The attention mechanism takes as input the question embedding and embedded visual regions, and then weights the visual regions according to their relevance. An RN module performs pair-wise reasoning on objects. With the use of previously computed attention weights the most relevant regions of interest are selected. Then, for each pair of region embeddings, relational embedding is computed based on the region embeddings and the question. In the end, final relational embedding is produced by summing the embeddings for each pair. A joint embedding is computed with multimodal fusion by combining the question embedding, the attended image embedding and the relational embedding. These are combined using the Hadamard product. The final part of the model is a classifier that performs multi-label classification to infer the answer for the question according to the joint embedding.

III. Datasets

Several datasets have been proposed specifically for the research on automatic question answering. The following subsections briefly overview datasets for two different categories of question answering: textual and visual. Datasets for textual question answering typically comprise a set of questions with at least one answer, while datasets for visual question answering consist of a set of images, questions about them and their corresponding answers. Sometimes, datasets include additional information such as text articles, scene graph annotations, objects, object attributes, object relations etc.

A. Textual Question Answering

SemEval (International Workshop on Semantic Evaluation) is an ongoing series of evaluations of computational semantic analysis systems. SemEval-2017 is the eleventh workshop in the series. It comprises different tasks for semantic evaluation including Community Question Answering [12]. This task is divided into five subtasks each providing different data for ranking questions, question comments, question external comments, correct answers, as well as identifying duplicate questions.

SQuAD (Stanford Question Answering Dataset) [13] is a dataset consisted of question-answer pairs posed by crowd workers on Wikipedia articles. The dataset does not provide a list of answer choices for the question. On the contrary, the answer must be selected from all possible spans in the passage. It can be a segment of text, or span, from the corresponding reading passage. The second version of the dataset [14] introduces unanswerable questions. That is, the first version is extended with additional unanswerable questions. An additional challenge, besides determining the answer, when working with this dataset is determining when the answer is not available and abstain from answering.

QuAC (Question Answering in Context) [15] is a dataset comprising QA dialogues between two individuals. The first individual asks free-form questions, while the second individual gives an answer for the question. The goal is to predict text span which answers a question about Wikipedia article. The question can also be unanswerable.

The bAbI project is organized towards the goal of automatic text understanding and reasoning. It comprises different tasks where each task is associated with a specific dataset. The Simple Questions dataset [16] refers to open-domain question answering and is based on the Freebase knowledge database. It consists of a total of 108,442 questions written in natural language by human English-speaking annotators. The answer for each question is a fact formatted as tuple (subject, relationship, object) that also provides a complete explanation.

B. Visual Question Answering

CLEVR (Compositional Language and Elementary Visual Reasoning) [17] provides a dataset that requires solving complex reasoning problems such as attribute identification, counting, comparison, spatial relationships, and logical operations. The dataset contains synthetic images generated by randomly sampling and rendering a scene graph. Scene graph is a kind of image scene representation in the form of a graph where nodes represents object and edges connect objects that are spatially related. CLEVR contains three object shapes: cube, sphere, and cylinder. They can come in two absolute sizes (i.e. small and large), two materials (i.e. shiny metal and matte rubber) and eight colors. The objects are spatially related via four relationships: left, right, behind and in front. Each question is associated with a functional program that can be executed on an image's scene graph, yielding the answer to the question.

VQA [18] is the most widely used dataset for the task of VQA. It is divided into two datasets according to the nature of

images: natural or abstract. The set of natural images (VQA-real) contains images from the MS COCO [19] dataset, while the set of abstract images (VQA-abstract) contains images with abstract scenes. The reason for creating the abstract scenes dataset is to avoid the low-level vision tasks and focus only on high-level reasoning. Each image has questions with both ground truth and plausible but likely incorrect answers. The questions are provided from human annotators. Answers are typically a word or a short phrase. However, for many of the questions, a yes or no answer is sufficient.

Visual Genome [20] is a dataset containing real-world images obtained as intersection of images in MS COCO [19] and YFCC100M [21]. It aims to connect structured image concepts to language. For each image, the dataset provides region descriptions, object instances, attributes, relations, region graphs, scene graphs and visual question answers. There are two types QA pairs associated with each image: based on the entire image (i.e. free-form) and based on specific image regions (i.e. region-based). Each image has at least one question of each type: what, where, how, when, who and why.

IV. Evaluation Metrics

Evaluating the quality of question answering systems is an important aspect of the problem. Their performance is measured with different types of evaluation metrics. According to their nature, the answers can be split into two groups: short answers composed of only one word and long answers composed of multiple words. Based on this division, we split the metrics into two groups: metrics for evaluation of an information retrieval system and metrics for evaluating automatically generated text. Creating answers from the first group, short answers, comes down to classification. Such answers are evaluated with classification and information retrieval metrics. Answers containing multiple words are evaluated with evaluation metrics for automatically generated text. Several evaluation metrics from both groups are described in the following subsections.

A. Metrics Based on Information Retrieval

The most common way for evaluating a machine learning model is to use metrics based on information retrieval. These metrics are applied when the purpose of the system is ranking answers, ranking similar questions or answers, or when the answer generation comes down to classification. Example evaluation metrics include accuracy, precision, recall, F1-measure, etc [22].

However, simple precision and recall do not apply for systems that rank the retrieved documents. That is, if we are comparing the performance of two ranked retrieval systems, we require a metric that will prefer the one that ranks the relevant documents higher [23]. One such metric is MAP (Mean Average Precision). It is calculated as follows. First, we descend through the ranked list of items and note the precision only at those points where a relevant document has been encountered. For a single query, these individual precision measurements are averaged over the return set up to some fixed cutoff. The final measure is the mean of such averages. Another metric assuming that the system retrieves relevant documents is MRR (Mean Reciprocal Rank). Each query is scored according to the reciprocal of the rank of the first correctly retrieved document. The final measure is the mean of such reciprocal ranks.

B. Metrics Based on Natural Language Generation

The evaluation of computer-generated natural language sentences is an inherently complex task. The most common way to assess the quality of automatically generated texts is the subjective evaluation by human experts. However, human evaluation is not always attainable. Another approach is to use automatic evaluation metrics. These metrics compute a score that indicates the similarity between generated and reference text. They are applied when the purpose of the system is to generate natural language phrase.

The METEOR [24] automatic evaluation metric is designed for evaluating machine translation. It is based on the harmonic mean of unigram precision and recall, where recall is weighted higher. It scores generated translations by aligning them to one or more reference translations. Alignments are based on exact, stem, synonym, and paraphrase match between words and phrases. BLEU [25] was also designed for automatic evaluation of machine translation. It measures how close a candidate sequence is to a reference sequence, i.e. the hits of n-grams of a candidate sequence to the reference. BLEU can be calculated with different length of n-grams. BLEU-N is the score where N is the maximum length of considered n-grams.

ROUGE-L [26] is a recall-oriented metric developed for evaluation of text summarization. It applies the concept of Longest Common Subsequence (LCS). The intuition is that the longer the LCS between two summary sentences is, the more similar they are. The score is 1 when the two sequences are equal, and 0 when there is nothing in common between them.

CIDEr [27] was developed specifically for evaluation of image descriptions. The goal is to automatically evaluate how well a candidate sentence matches the consensus of a set of image descriptions, i.e. how often n-grams in the candidate sentence are present in the reference sentences. All words in the sentences (both candidate and references) are first mapped to their stem or root forms. SPICE [28] is another metric developed for evaluation of image captions. It measures how effectively image captions recover objects, attributes and the relations between them. It is based on the agreement of the scene-graph tuples of the candidate sentence and all reference sentences. Scene-graph is a semantic representation that parses the given sentence to semantic tokens. A set of tuples is formed by using the elements of the graph and their possible combinations. The score is defined as the F1-score based on the agreement between the candidate and reference caption tuples.

V. Conclusion

Automatically generating answer for a given question is a process in which the computer is supposed to answer a question in a natural language where the question itself is also provided in natural language. In accordance with the type of information the questions refer to, the task can be classified as Textual Question Answering, Visual Question Answering, etc. Textual Question Answering is the task of extracting a text snippet from a passage, which corresponds to a specific question. A specific type of textual question answering is Community Question Answering that refers to online communities. Visual Question Answering is the task about questions concerning images, either natural or abstract.

Deep learning techniques gained extensive research in the domain of question answering. There is a variety of deep neural models proposed in this domain. In this survey, the models are presented as part of one of the following groups: classical deep neural networks, dynamic memory networks and relation networks. Representative models of each group are considered.

Several datasets have been proposed specifically for the research on automatic question answering. Datasets for textual question answering are provided by SemEval, SQuAD, bAbI, etc. They typically comprise a set of questions with at least one answer. CLEVR, VQA, Visual Genome and others provide datasets for visual question answering. These datasets consist of a set of images, questions about them and their corresponding answers.

In the end, evaluation metrics utilized in the field are presented. They come as a part of one of the following groups: metrics for evaluation of an information retrieval system (such as accuracy, precision, recall, F1-measure, mean average precision and mean reciprocal rank) and metrics for evaluating automatically generated text (such as METEOR, BLEU, ROUGE-L, CIDEr and SPICE).

References

[1] J. Patterson and A. Gibson, *Deep Learning: A Practitioner's Approach*. O'Reilly Media, Inc., 2017.

[2] J. M. Deriu and M. Cieliebak, "Swissalps at semeval-2017 task 3: Attention-based convolutional neural network for community question answering," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 334–338, 2017.

[3] S. Zhang, J. Cheng, H. Wang, X. Zhang, P. Li, and Z. Ding, "Furonguang at semeval-2017 task 3: Deep neural networks for selecting relevant answers in community question answering," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 320–325, 2017.

[4] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21–29, 2016.

[5] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *International Conference on Machine Learning*, pp. 1378–1387, 2016.

[6] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *International conference on machine learning*, pp. 2397–2406, 2016.

[7] S. Sukhbaatar, J. Weston, R. Fergus, *et al.*, "End-to-end memory networks," in *Advances in neural information processing systems*, pp. 2440–2448, 2015.

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

[9] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in neural information processing systems*, pp. 4967–4976, 2017.

[10] L. J. Petersen, "Attended Relational Reasoning for Visual Question Answering," Master's thesis, Aalborg University, 2018.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[12] P. Nakov, D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, and K. Verspoor, "Semeval-2017 task 3: Community question answering," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 27–48, 2017.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[14] P. Rajpurkar, R. Jia, and P. Liang, "Know what you dont know: Unanswerable questions for squad," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 784–789, 2018.

[15] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer, "Quac: Question answering in context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2174–2184, 2018.

[16] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," 2015.

[17] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

[18] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[21] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "The new data and new challenges in multimedia research," 2015.

[22] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," *Waltham: Elsevier*, 2012.

[23] J. H. Martin and D. Jurafsky, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River, 2009.

[24] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, pp. 376–380, 2014.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.

[26] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[27] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

[28] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*, pp. 382–398, Springer, 2016.