A Comprehensive Survey on Automatic Image Captioning System

Chitra P
Pursuing M. Tech.
Computer Science and Engineering Department
N.S.S. College of Engineering
Palakkad, Kerala, India
chitra11393@gmail.com

Chitra S Nair
Assistant Professor
Computer Science and Engineering Department
N.S.S. College of Engineering
Palakkad, Kerala, India
chitranairis@gmail.com

***Abstract*—Automatic image captioning is a challenging problem that has recently gathered a large attention particularly in the computer vision and natural language processing domain. There is a heeding need for context specific natural language description of images. However, this may seem to be farfetched but the recent improvements in the fields like neural networks, computer vision and natural language processing has cleared way for the accurate description of images by representing their visual ground meaning. In this survey, we provide various approaches in automatic image captioning by highlighting the techniques present in the existing models and also listing their advantages and disadvantages. Finally, we also explore the future prospects in the domain of auto image annotation.**

***Index Terms*—Computer Vision, Machine Learning, Neural Networks, Image Processing, Natural Language Processing.**

I. Introduction

A quick look at an image is enough for a human being to point out and provide a large amount of information about the visuals presented in it. However, this remarkable capability has proven to be a difficult task. A majority of existing works in this domain has focused on providing images with labels from a pre-defined set of visual categories and great progress has been achieved in this field. However, these methodologies work for only a single image label at a time that provides only a numb definition of images. They are very restrictive when the huge amount of image descriptions which a human being can compose is considered.

Automatic image annotation or image captioning is an application that connects the two major fields of artificial intelligence, namely, computer vision and natural language processing. It is the process by which a computer system automatically assigns metadata in the form of captions or keywords to a digital image. Automatic generation of image captions requires a detailed understanding of an image and an ability to communicate that information via natural language. In other words, the task of automatic image description involves taking an image, analysing its visual content, and generating a textual description (typically a sentence) that verbalizes the most salient aspects of the image. This requires the joint use of both computer vision and natural language processing techniques. The main implication of auto image captioning is to automate the job of some person who interprets the image.

The initial step of an image captioning system is to understand the image by recognizing the objects present in it, justifying about the relationship among those objects and then focusing on the more salient parts in the image. The identified objects corresponds to the nouns in the caption to be generated [1]. The relationship between objects and their appearance corresponds to other linguistic constituents such as verbs and it also determines how to sequentially order the words into sentences. The main point to note here is that only the most important information is retained and the irrelevant secondary information is tuned out.

From a computer vision perspective, the description could in principle cover any visual aspect of the image. It can talk about objects and their attributes, features of the scene, or verbalize the interaction of the people and objects in the scene. More challenging, it can reference objects that are not depicted and provide background knowledge that cannot be derived directly from the image. In other words, image understanding, which essentially produces an unstructured list of object, scene and interaction labels, is necessary, but clearly not sufficient for producing a good description. A good description should be comprehensive but concise, while being formally correct, that is, consist of grammatically well-formed sentences [2].

From the perspective of natural language processing, the task of natural language generation takes a non-linguistic representation, that is, an image representation, and turns it into human-readable text. Generating text involves a series of steps: first we need to decide which aspects of the input to talk about (content selection), then we need to organize the content (text planning), and lastly, verbalize it (surface realization). Surface realization, in turn, requires choosing the right words (lexicalization), using pronouns whenever appropriate (referential expression generation), and grouping of related information (aggregation) [2]. In summary, automatic image description requires both image understanding and natural language generation, thus bridging the computer vision and the natural language processing communities.

The main aim of this survey is to give a comprehensive overview of the existing state-of-the-art models, datasets, and evaluation metrics that have been developed or adopted in this area of research. In this paper, Section II deals with various techniques of performing automatic image captioning. The findings and future prospects are presented in Section III and the conclusion is given in Section IV.

II. Image Captioning Models

The problem of automatic image captioning can be reviewed on the grounds of various methodologies available for the task. Some of those different types of image captioning models are reviewed herewith. Mainly, four categories of image captioning models are considered in this survey. First one is based on pre-defined templates where as second one wholly depends on retrieval based techniques. The third category explores image captioning from the neural network perspective while the last category is based on semantic attention.

A. *Image Captioning Using Pre-defined Templates*

Here, some pre-defined templates are used to generate sentences by filling the detected visual elements such as objects. In [3], a simple yet effective approach to automatically compose image descriptions given computer vision based inputs and using web-scale n-grams is presented. This method composes fresh sentences from scratch, instead of retrieving or summarizing the existing text fragments associated with an image and the concept is shown in Fig. 1. Here, a surface realization technique based on web-scale n-gram data is used and it consists of two steps, namely, n-gram phrase selection and n-gram phrase fusion. The n-gram phrase selection collects the candidate phrases useful for generating image descriptions. The n-gram phrase fusion finds optimal compatible set of phrases using dynamic programming to compose new phrases. The produced captions are more human-like descriptions but the pattern matching is case sensitive, allowing only lower case letters.

Another approach towards image caption generation is a technique that describes images by predicting the core sentence components [4]. Here the input are initial noisy estimates of the objects and scenes detected in the image using the state-of-the-art trained detectors. This work explains the use of a large generic corpus such as English Gigaword as semantic grounding to predict and correct the initial visual detections of an image to produce reasonable sentence describing the image. A Hidden Markov Model inference scheme models sentence generation process with hidden nodes as sentence components and image detections as emissions. Based on purely summarization (unigram-overlap) point of view, it produces best descriptions. However, the detection of objects, actions and scenes from images is noisy and unreliable.

Another work is [6] in which a system is designed which is capable of producing natural-sounding descriptions from computer vision detections that are flexible enough to become more descriptive and poetic, or include likely information from a language model, or to be short and simple, but as true to the image as possible. Rather than using fixed template capable of generating one kind of utterance, [6] lies in generating syntactic trees. This follows a three-tiered generation process, utilizing content determination to first cluster and order the object nouns, create their local subtrees, and filter incorrect detections; microplanning to construct full syntactic trees around the noun clusters; and surface realization to order selected modifiers, realize them as postnominal or prenominal, and select final outputs. The system follows an overgenerate-and-select approach which allows different final trees to be selected with different settings. It also can automatically parse and train an unlimited amount of text. But it is unable to cluster similar nouns together and also on the computer vision side, incorrect objects are often detected and salient objects are often missed out.

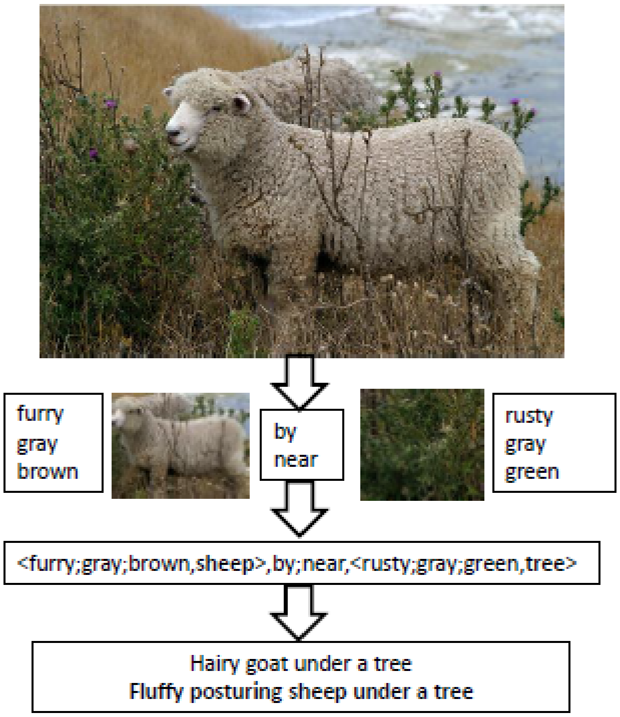
furry
gray
brown
by
near
rusty
gray
green
<furry;gray;brown,sheep>,by;near,<rusty;gray;green,tree>
Hairy goat under a tree
Fluffy posturing sheep under a tree

Fig. 1. Image captioning using web-scale n-grams [3]

BabyTalk [7] is a system to automatically generate natural language descriptions from images that exploits both statistics gleaned from parsing large quantities of text data and recognition algorithms from computer vision. Here, for an input image, detectors are used to detect objects and stuff. Each candidate object region is processed by a set of attribute classifiers. Each candidate region pair is processed by prepositional relationship functions. A conditional random field (CRF) is constructed that incorporates the unary image potentials with higher order text based potentials computed from large text corpora. Then, a labeling of the graph is predicted and the sentences are generated based on the labeling. The CRF labeling is used to predict the best labeling for an image. Nodes of CRF corresponds to several kinds of image contents like objects, attributes and prepositions. It automatically mines knowledge about textual representation and parses the image fully automatically. However, whole image features cannot be recognized.

The system in [8] uses visual dependency representation (VDR) in order to represent the relationships between the objects in an image, and assumes that this representation can improve image description. This hypothesis is tested using a data set of region-annotated images associated with gold-standard descriptions. Visual dependency representations are used to represent the structure of images. This representation expresses the mathematical relationship between the regions of an image. This model also uses special grammar, namely, visual dependency grammar to describe the spatial relations between pairs of image regions. This work highlights four template-based models for generating image descriptions where the aim of each model is to determine what is happening in the image, which regions are important for describing it and how these regions relate to each other.

B. Image Captioning Based on Retrieval Methods

In this category of image captioning model, first a similar image is found out from the training data set and the new sentences are composed based on he retrieved images' sentences. In [5], a holistic data-driven retrieval-based approach to image description generation is presented. It exploits the vast amount of noisy parallel image data and associated natural language descriptions available on the web. It consists of two parts. For a query image, first retrieve candidate descriptive phrases from large image caption database using visual similarity measures. Then generate a coherent description from these candidates for content planning and surface realization. Image-level content planning deals with selecting the set of objects to describe and re-ordering sentences. Surface realization deals with selecting the set of phrases for each sentence and then re-ordering phrases within each sentence. Although it generates semantically correct and linguistically appealing descriptions the generated sentences are fixed and limited.

In [9], a tree based approach is presented in order to compose image descriptions that utilizes naturally available web images with captions. Two related tasks are considered here: image caption generalization and generation. The main idea of this work is to harvest expressive phrases as tree fragments from existing descriptions, then to form a new one by selectively combining the extracted or pruned fragments of tree. The key components are tree composition and tree compression, both combining the tree structure and sequence structure.

Given a query image, the images that are visually similar to the input image are retrieved. Then, phrases are extracted from their corresponding image descriptions and new captions are composed using these retrieved fragments. The tree composition is modeled as a constraint optimization whereas caption generalization is presented as sentence compression. It offers a higher level of linguistic expressiveness, and flexibility and also it has an improved image caption corpus with auto generalization. However, the task of sentence compression considers only deletion-only edit operation and this does not support complex edits like substitutions, insertions and so on.

C. Image Captioning Using Neural Network Models

The models based on neural network explores the various techniques of captioning an image by utilizing recent developments in the field of artificial intelligence and deep learning. The technique in [10] highlights a long-term recurrent convolutional network model which presents a novel architecture suitable for end-to-end trainable visual learning as illustrated in Fig. 2. This work combines a deep hierarchical visual feature extractor like a CNN with a model that will learn to recognize the temporal dynamics of sequential data.

The convolutional models can be compositional in spatial and temporal layers. Also learning long-term dependencies is possible when non-linearities are incorporated into network state updates. These models maps variable-length inputs, that is, video or still image frames, to variable-length outputs, that is, natural language text. These are optimized with back propagation. Using visual transformation has the important advantage of making inference and training parallelizable, facilitating independent batch processing and end-to-end optimization of visual and sequential sentence model. But this system is unable to handle time-varying visual input.

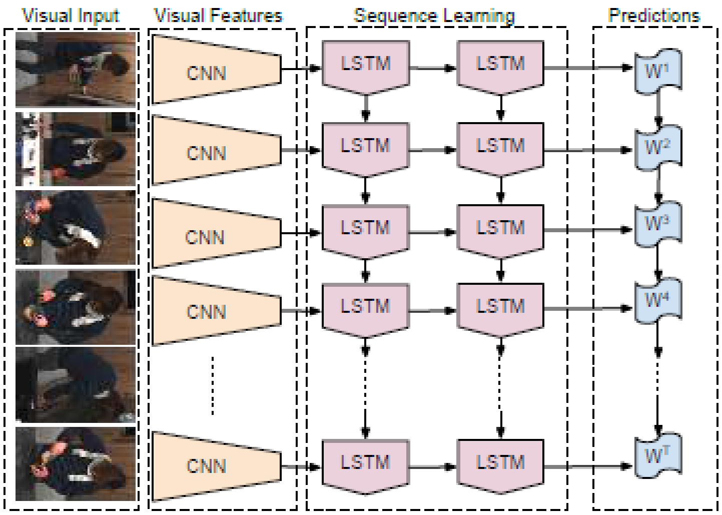

Fig. 2. CNN for image captioning [10]

The Neural Image Caption (NIC) model in [11] deals with a generative technique based on a deep recurrent architecture. It is illustrated in Fig. 3. It is trained in order to increase the likelihood of target sentence given the training image. A CNN is used as the image encoder by implementing it as an image classification task and the last hidden layers are used as input to RNN decoder that generates sentences. Here, a neural and probabilistic approach is used to generate image descriptions. The model uses a RNN to encode variable length input to fixed dimensional vector and thereby uses it to decode to desired output. An LSTM-based sentence generator is used for the tasks of translation and sequence generation. The method of Beam Search is used as a part of sentence generation. The implementation of batch normalization in this model considers the fact that each layer is normalized with respect to the current batch of examples. But the model fails to answer user specified questions.

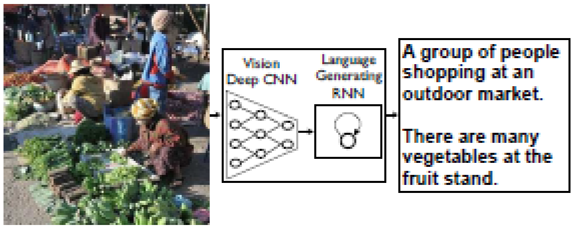
Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Fig. 3. NIC model [11]

The work in [13] deals with an extension to the LSTM model to tackle the problem of image caption generation, called as g-LSTM model. The model adds semantic information which is extracted from the image as input to each unit of LSTM block so that the solutions are more tightly coupled with image content. The cream idea here is the sentence normalization. The semantic information being extracted here serves as a guide in the procedure of word sequence generation.

In this system, as shown in Fig. 4, the generation of words is carried out under the guidance of global semantic information, which is, in fact a biggest advantage of this model. When compared to the standard LSTM architecture, here a new term is appended to the computation of each gate and cell state in LSTM. This new term represents the semantic information which acts as a bridge between visual and textual domains. There exists three ways to extract the semantic information. First is a cross-modal retrieval task and it is simply used to retrieve sentences. Second one is representation as embedding in semantic space where visual and textual representations are equivalent. Last one is to use the image itself as guidance.

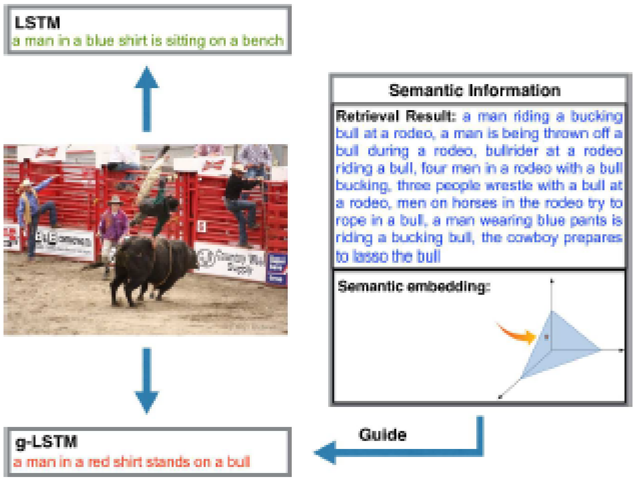
LSTM
a man in a blue shirt is sitting on a bench
Semantic Information
Retrieval Result: a man riding a bucking bull at a rodeo, a man is being thrown off a bull during a rodeo, bullrider at a rodeo riding a bull, four men in a rodeo with a bull bucking, three people wrestle with a bull at a rodeo, men on horses in the rodeo try to rope in a bull, a man wearing blue pants is riding a bucking bull, the cowboy prepares to lasso the bull
Semantic embedding:
Guide
g-LSTM
a man in a red shirt stands on a bull

Fig. 4. g-LSTM model [13]

In [14], an encoder-decoder pipeline is introduced to learn multimodal joint embedding space with images and text and provides a language model for decoding distributed representations from the space. The pipeline unifies the joint image-text embedding models along with multimodal neural language models. A structure-content neural language model is considered in this work in order to unravel the structure of a sentence to its content, by making it conditioned on representations produced by the encoder. The encoder allows ranking of images and sentences while the decoder generates novel descriptions from scratch. However, in this approach only a small region is relevant at any given time.

The m-RNN model in [15] is used for generating sentence descriptions to explain the image content. It directly represents the probability distribution of generating a word when the previous words and the image are already given. Image descriptions are generated by sampling from this distribution. The model consists of a deep recurrent neural network for sentences and a deep convolutional network for images. These two sub-networks interact with each other in a multimodal layer to form the whole model. The m-RNN model is much deeper than the simple RNN model. It has six layers in each time frame, namely, the input word layer, two word embedding layers, the recurrent layer, the multimodal layer, and the softmax layer. Despite the merits of this system, it is difficult to handle large datasets with m-RNN model.

The approach in [16] highlights a deep multimodal similarity model for caption generation of images. Here, multiple instance learning approach is used to train the visual detectors for words that commonly occur in captions. These include many parts of speech such as nouns, verbs and adjectives. The outputs from the word detector serves as conditional inputs to a maximum-entropy language model. The task of language generation searches for the most likely sentence which is conditioned on a set of visually detected words.

The language model in [16] defines the probability distribution over word sequences. It performs a left-to-right Beam Search operation. This model also employs an advantageous sentence re-ranking to find out the best sentences using the techniques of deep multimodal similarity. It learns two neural networks that map images and text fragments to common vector representation. It will also measure the similarity between images and text by taking the cosine similarity measure between corresponding vectors. This cosine similarity measure is used to re-rank sentences effectively. However, the supervised learning techniques are not supported here.

D. Image Captioning Based on Semantic Attention

The existing approaches to automatically generating a natural language description of an image are either top-down, which start from a gist of an image and convert it into words, or bottom-up, which come up with words describing various aspects of an image and then combine them. This category is an extension to those that used neural network models but an added feature called semantic attention is used.

One such model is a NIC with visual attention given in [12]. It is an attention based method which automatically learns to describe the image contents. It incorporates attention with two different variants: a hard stochastic attention mechanism and a soft deterministic attention mechanism. The main advantage of including attention is to bring out the ability to visualize what the model sees. Soft attention mechanism is trainable by standard back propagation methods whereas hard attention mechanism is trainable by maximizing the approximate variational bound. However as the implementation needs time proportional to length of longest sentence, this model always requires a dictionary while preprocessing.

In [17], the model combines both top-down and bottom-up approaches through semantic attention as illustrated in Fig. 5. The algorithm learns to selectively address the semantic concept and combine them into hidden states and outputs of recurrent neural networks. The selection and fusion form a feedback connecting the top-down and bottom-up computation. Semantic attention is the ability to provide a detailed, coherent description of semantically important objects. The semantic concepts or attributes are recognized as candidates for bottom-up approach, and utilize a top-down approach to guide the direction and location of attention.

The model in [17] is built on top of RNN, whose inceptive state captures global information from the top-down feature. As the RNN state transits, it receives feedback and interaction from the bottom-up attributes. This feedback allows the algorithm to not only predict more accurately new words, but also lead to stronger inference of the semantic gap between existing predictions and image content. Visual attribute prediction is a key part in both training and testing.

Two approaches exist for predicting attributes from an input image. First, a non-parametric method based on nearest neighbour image retrieval from a large collection of images with rich and unstructured textual metadata such as tags and captions. The attributes for a query image can be obtained by transferring the text information from the retrieved images with similar visual appearances. The second approach is to directly speculate visual attributes from the input image using a parametric model. But this model do not support phrase-based attribute prediction.

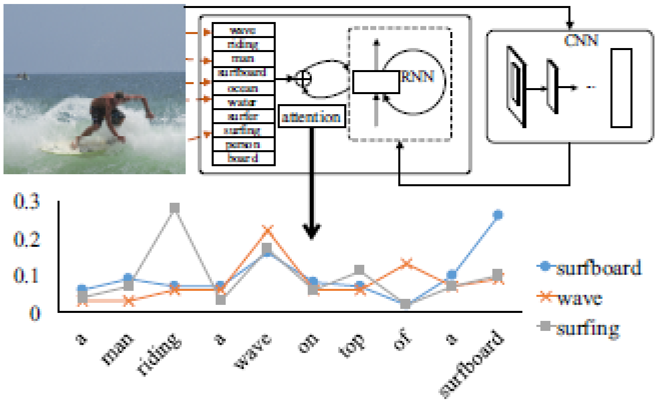

Fig. 5. Semantic attention model [17]

III. Findings and Future Prospects

The problem of automatic image captioning is quite daring. Various methodologies of providing image captions have been discussed here. While earlier works focused mainly on generating captions based on pre-defined templates, there existed retrieval based methods which found the similar image in training set and composed descriptions based on the retrieved image. Then the works focusing on automatic image annotation based on language models arrived which included multimodal RNN, deep CNN and systems using LSTM models. Several other methods which used lexical representations instead of visual representations also evolved. Later, in recent times, the captioning task is supported by attention based models. Table I illustrates a brief summary of all these image captioning techniques.

TABLE I
Image Captioning Models

S. No.	***Technique***	***Merits***	***Demerits***
1	Pre-defined Templates	Human-like image descriptions produced	Do not perform sentence selection and often salient objects are missed out
2	Retrieval Models	Linguistic expressiveness and flexibility	Generated sentences are fixed and complex sentence compression tasks not considered
3	Neural Network Models	End-to-end optimization and batch processing	Fails to answer user specified questions
4	Semantic Attention Models	Visual attribute prediction	Background scene analysis is not done

The computer vision and natural language processing communities have witnessed an upsurge in interest in automatic image description systems. With the help of recent advances in deep learning models for images and text, substantial improvements in the quality of automatically generated descriptions has been registered. In spite of these recent advancements, a series of challenges still remain. The supervised algorithms takes advantage of carefully collected large datasets, and hence, lowering the amount of supervision in exchange of access to larger unsupervised data can be an interesting avenue for future research.

Leveraging unsupervised data for building richer representations and description models is another open research challenge in this context. Having multilingual repositories for image description is an interesting direction to explore. Future work should investigate whether transferring multimodal features between monolingual description models results in improved descriptions compared to monolingual baselines. Overall, image understanding is the ultimate goal of computer vision and natural language generation is one of the ultimate goals of natural language processing. Image description or image captioning is where both these goals are interconnected and this research domain is therefore likely to benefit from individual advances in each of these two fields.

IV. Conclusion

In this survey, we discussed recent advances in automatic image description generation. We reviewed a large body of the existing work, highlighting common characteristics and differences between existing researches. Compared to traditional keyword-based image annotation automatic image description systems produce more human-like explanations of visual content, providing a more complete picture of the scene. Advancements in this field could lead to more intelligent artificial vision systems, which can make inferences about the scenes through the generated grounded image descriptions and therefore interact with their environments in a more natural manner. They could also have a direct impact on technological applications from which visually impaired people can benefit through more accessible interfaces.

References

[1] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, “Aligning Where to See and What to Tell: Image Captioning with Region-based Attention and Scene-specific Contexts,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, December 2016.

[2] A. Karpathy and L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in CVPR, 2015.

[3] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, “Composing simple image descriptions using web-scale n-grams,” in CoNLL, 2011.

[4] Y. Yang, C. L. Teo, H. Daume III, and Y. Aloimonos, “Corpus guided sentence generation of natural images,” in EMNLP, 2011.

[5] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, “Collective generation of natural image descriptions,” in ACL, 2012.

[6] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daume III, “Midge: Generating image descriptions from computer vision detections,” in EACL, 2012.

[7] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Babytalk: Understanding and generating simple image descriptions,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 12, pp. 28912903, 2013.

[8] D. Elliott and F. Keller, “Image description using visual dependency representations.” in EMNLP, 2013.

[9] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, “Treetalk: Composition and compression of trees for image descriptions,” Transactions of the Association for Computational Linguistics, vol. 2, no. 10, pp. 351362, 2014.

[10] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in CVPR, 2014.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in CVPR, 2015.

[12] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in ICML, 2015.

[13] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, “Guiding long short term memory for image caption generation,” in ICCV, 2015

[14] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual semantic embeddings with multimodal neural language models,” Transactions of the Association for Computational Linguistics, 2015.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, “Explain images with multimodal recurrent neural networks,” in ICLR, 2015.

[16] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt et al., “From captions to visual concepts and back,” in CVPR, 2015.

[17] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in CVPR, 2016.