INTERNATIONAL JOURNAL · ENGINEERING RESEARCH

International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

IJIEMR Transactions, online available on 1st May 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=ISSUE-05

DOI: 10.48047/IJIEMR/V11/I05/34

Title **Image Caption Generator Using CNN and LSTM**

Volume 11, Issue 05, Pages: 200-207

Paper Authors

Miss. K.Venkata Padmavathi, Miss. N.Sravani, Miss. P.anuradha, Miss.M.Swetha

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Image Caption Generator Using CNN and LSTM

Miss. K.Venkata Padmavathi, B.Tech, dept. of CSE in Tirumala Engineering College Jonnalagadda, Narasaraopet.
Miss. N.Sravani, B.Tech, dept. of CSE in Tirumala Engineering College Jonnalagadda, Narasaraopet.
Miss. P.anuradha, B.Tech, dept. of CSE in Tirumala Engineering College Jonnalagadda, Narasaraopet.
Miss.M.Swetha, B.Tech, dept. of CSE in Tirumala Engineering College, Jonnalagadda, Narasaraopet.

ABSTRACT

Automatically characterising what's in a picture or image has long been a research topic in Artificial Intelligence. The development of an Automatic Caption Generator employing CNN and LSTM models is described in this study. It integrates modern machine translation and computer vision research. Flickr8k was utilised as a dataset. We utilised BLEU scores to evaluate the performance of the stated model. The produced captions may be classified as excellent or terrible based on their scores. This model's main uses include virtual assistants, picture indexing, social networking, accessibility for visually impaired persons, modifying application suggestions, and much more.

Key Words: CNN, LSTM, BLEU, Deep Learning.

1. INTRODUCTION

The encoderdecoder architecture of Image Caption Generator models uses input vectors to generate valid and acceptable captions. This paradigm connects the worlds of natural language processing with computer vision. It's a job of identifying and evaluating the image's context before explaining everything in natural language like English. Our approach is based on two basic models: CNN (Convolutional Neural Network) and RNN-LSTM (Recurrent Neural Network-LSTM) (Recurrent Neural Networks- Long Short-Term Memory). CNN is utilised as an encoder in the derived application to extract features from the snapshot or image, and RNN-LSTM is used as a decoder to organise the words and generate captions. Self-driving cars, where it could describe the scene around the car; second, it could be an aid to the blind, guiding them in every way by converting scene to caption and then to audio; CCTV cameras, where alarms could be raised if any malicious activity is observed while describing the scene; and many more.

2. LITERATURE REVIEW

1.Topic modelling on Instagram hashtags: An alternative way to automatic image annotation

Authors:ArgyrisArgyrou ; StamatiosGi annoulakis ; Nicolas Tsapatsoulis

The practise of giving tags to digital photographs without the participation of humans is known as Automatic Image Annotation (AIA). The learning by example approach underpins the majority of recent automated picture annotation technologies. The first crucial step in such techniques is to create the training examples, which are pairs of photos with relevant tags. In earlier research, we've demonstrated that hashtags surrounding photos on social media, particularly Instagram, offer a reach source for AIA training sets. However, we discovered that only 20% of Instagram hashtags accurately represent the topic of the picture they accompany, necessitating a number of filtering procedures to find the suitable hashtags. We use topic modelling using Latent Dirichlet Allocation (LDA) on Instagram hashtags to predict the theme of linked photographs in this research. Because a topic is made up of a group of related phrases, identifying the visual topic of an Instagram picture using the suggested technique yields a reasonable collection of tags that can be utilised to train AIA algorithms.

2.Crowdsourcing for multiple-choice question answering

Authors:Bahadir Ismail Aydin, Yavuz Selim Yilmaz, Yaliang Li, Qi Li, Jing Gao and Murat Demirbas

We use crowd wisdom to answer multiple-choice questions, and we use lightweight machine learning approaches to increase the aggregate accuracy of the crowdsourced responses. We designed and implemented a crowdsourced system for playing the "Who Wants to Be a Millionaire?" quiz show in order to explore more effective aggregation algorithms and statistically assess them. After analysing our data (which includes over 200,000 responses), we discovered that by just selecting the most popular answer in the aggregate, we can correctly answer over 90% of the questions, but the success rate drops to 60% for the later/harder questions in the quiz show. We examine unique weighted aggregation strategies for aggregating the crowd's replies to increase the success rates of these later/harder queries. We demonstrate that by utilising weights adjusted for participant dependability (derived from the participants' confidence), we can increase the accuracy rate for the tougher questions by 15%, and the total accuracy rate to 95%. Our findings support the use of machine learning methods in the development of more accurate crowdsourced question answering systems.

3.Validity and reliability of naturalistic driving scene categorization judgments from crowdsourcing

Humans may need to classify large amounts of recorded visual information, which is a typical difficulty when analysing naturalistic driving data. We studied the possibility of crowdsourcing to characterise driving scene elements (such as the presence of other road users, straight road segments, etc.) at a larger scale than a single individual or a small team of academics could do using the internet platform CrowdFlower. In all, 200 professionals from 46 nations took part in the 1.5-day event. Validity and reliability were investigated using the CrowdFlower technique known as Gold Test Questions, both with and without incorporating researcher-generated control questions (GTQs). External employees' identification of driving scene objects was much more valid (correct) and dependable (constant) when using GTQs. In a CrowdFlower Job of 48 three-second video clips, GTQs were shown to have a 91 percent accuracy (i.e., relative to the evaluations of a confederate researcher) on items, compared to 78 percent without. There was a difference in bias, with external employees returning more false positives without GTQs than with GTQs. At a higher scale, a CrowdFlower Job using just GTQs provided 12,862 three-second video segments for annotation. Because checking the correctness of each at this scale was impossible (and self-defeating), a random selection of 1012 categorizations was verified and yielded comparable levels of accuracy (95 percent).

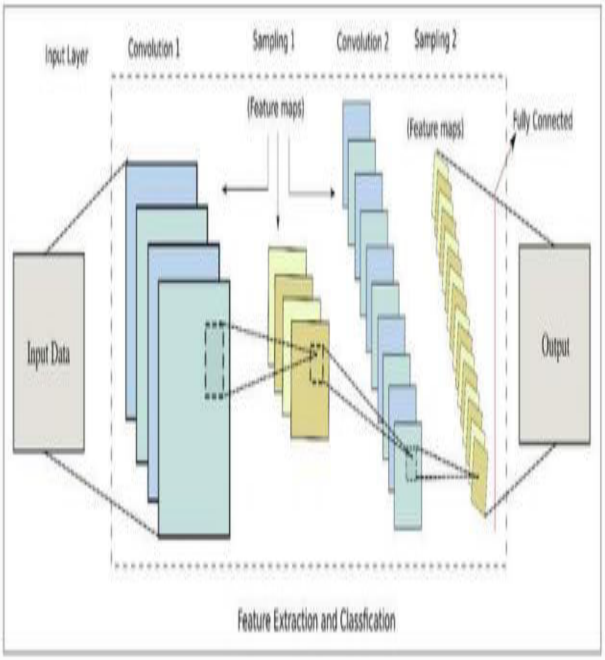
3.SYSSTEM ANALYSIS

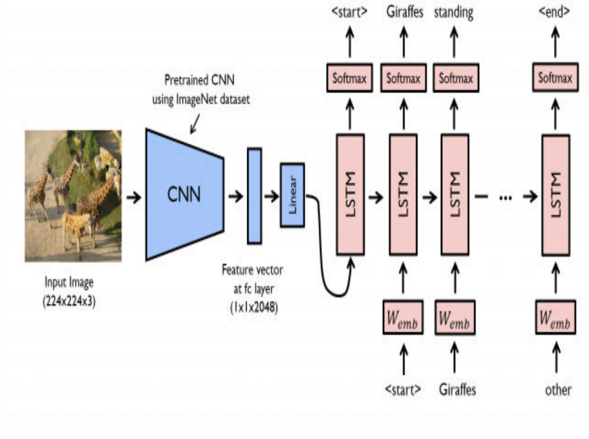
EXISTING SYSTEM

We will discuss the experimental findings obtained using the MSCOCO dataset. In their suggested work, they have incorporated a feature called guiding network to the encoder/decoder architecture. The guiding network technique primarily deals with learning the vector using a neural network v=g(A), where A is the collection of annotation vectors. The difficulty of creating natural language descriptions from visual data is a significant one. It has long been researched in the field of computer vision. As a result, elaborate systems based on visual basic recognizers and structured formal languages such as And-Or Graphs or logic systems have emerged. The topic of describing still images with natural words has recently attracted a lot of attention.

PROPOSED SYSTEM

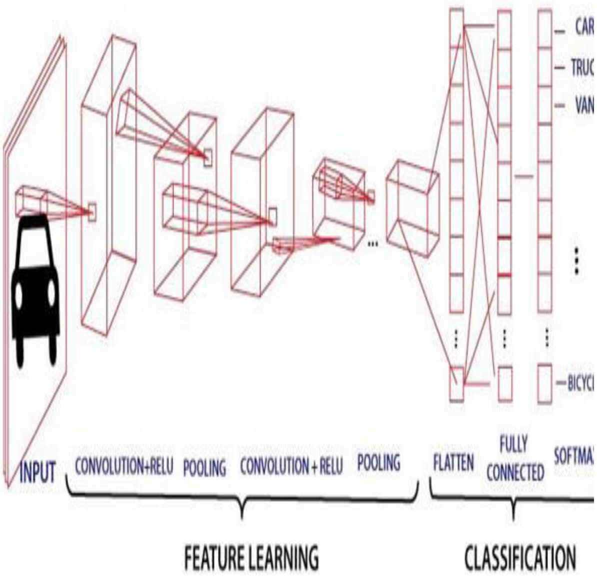
Here To achieve our aim (picture caption generator), we utilise CNN and LSTM. We begin by learning about CNN and how it might help us with our challenge. A convolutional neural network is a kind of deep learning neural network that is created artificially. Picture classifications, computer vision, image recognition, and object identification are all possible with it. CNN image classifications takes an input picture, processes it, and categorises it into several groups (Eg., Dog, Cat,etc). It scans photos from left to right and top to bottom to extract significant elements before combining them to categorise them. Second, define LSTM. Long short-term memory (LSTM) is a form of RNN (recurrent neural network) that is particularly well adapted to sequence prediction challenges. We can guess what the following word will be based on the preceding paragraph. It has outperformed regular RNNs in terms of overcoming the constraints of RNNs with short term memory. The LSTM may carry out important information throughout the processing of inputs, and it can discard non-related information using a forget gate. We combined these two models into a single CNN-RNN model. generally The success of the top-down image generating models outlined above informs our approach. The visual picture characteristics are retrieved using a deep convolutional neural network, while semantic information are extracted using the semantic tagging model. The visual information from the CNN and the semantic features from the tagging model are combined and fed into an LSTM network, which subsequently creates captions.





4. ALGORITHM CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Networks are one of the most common types of neural networks used for image categorization and recognition. Convolutional neural networks are commonly utilised in domains such as scene labelling, object identification, and facial recognition, among others.



Convolution Layer

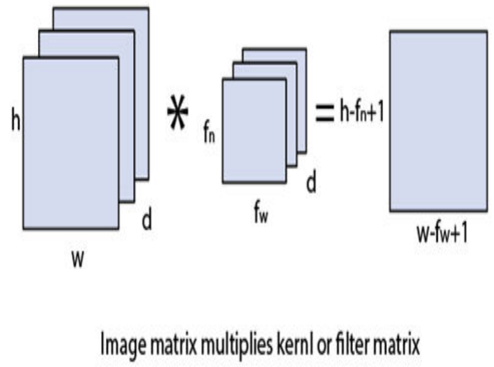
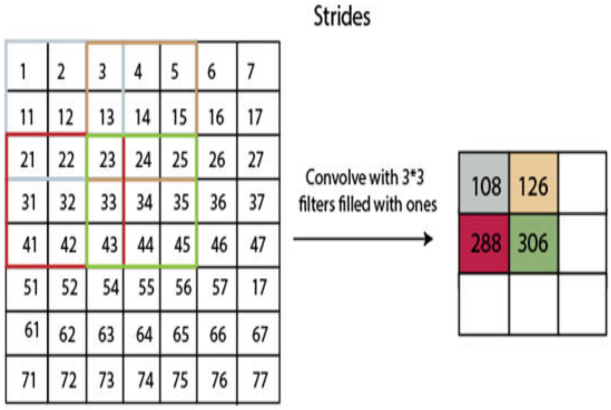


Image matrix multiplies kernl or filter matrix

Strides

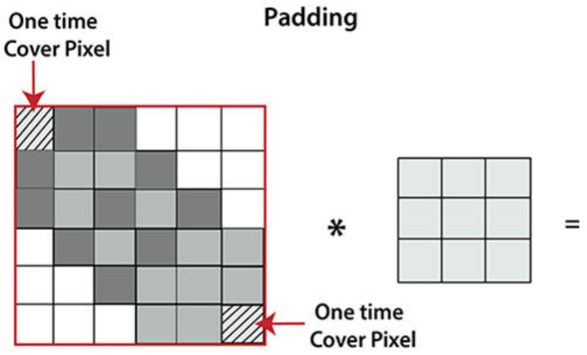


1	2	3	4	5	6	7
11	12	13	14	15	16	17
21	22	23	24	25	26	27
31	32	33	34	35	36	37
41	42	43	44	45	46	47
51	52	54	55	56	57	17
61	62	63	64	65	66	67
71	72	73	74	75	76	77

108	126	
288	306	

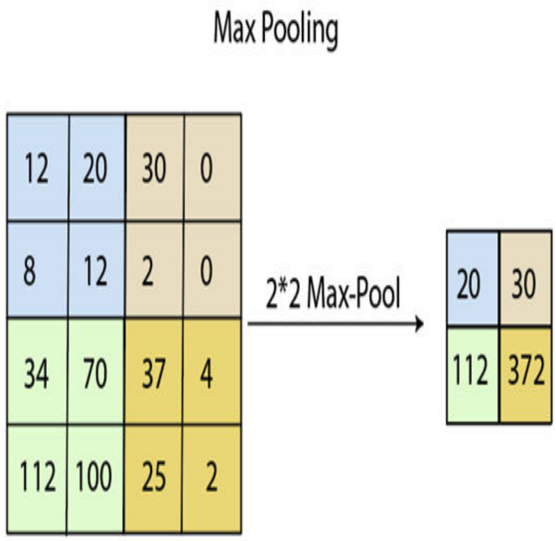
Padding

Padding plays a crucial role in building the convolutional neural network.



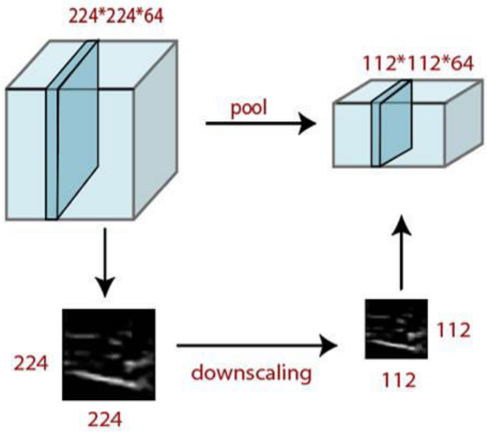
Pooling Layer

Pooling layer plays an important role in pre-processing of an image. Pooling layer reduces the number of parameters when the images are too large. Pooling is "**downscaling**" of the image obtained from the previous layers.



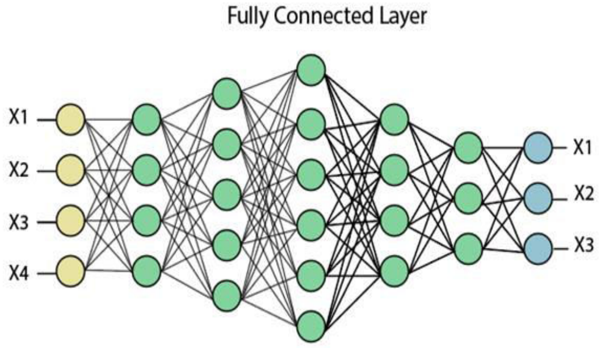
12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	2

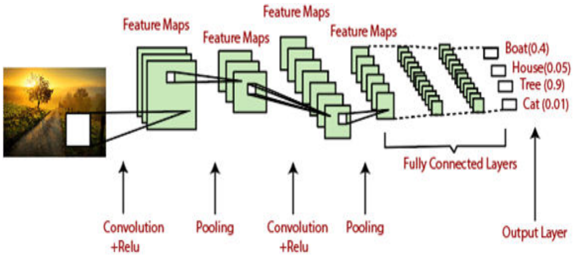
20	30
112	372



Fully Connected Layer

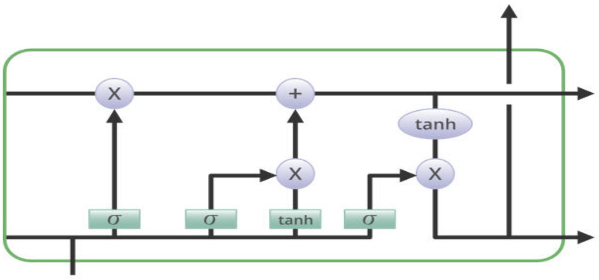
The fully connected layer is a layer in which the input from the other layers will be flattened into a vector and sent. It will transform the output into the desired number of classes by the network.





LSTM

Structure Of LSTM:



5.RESULTS

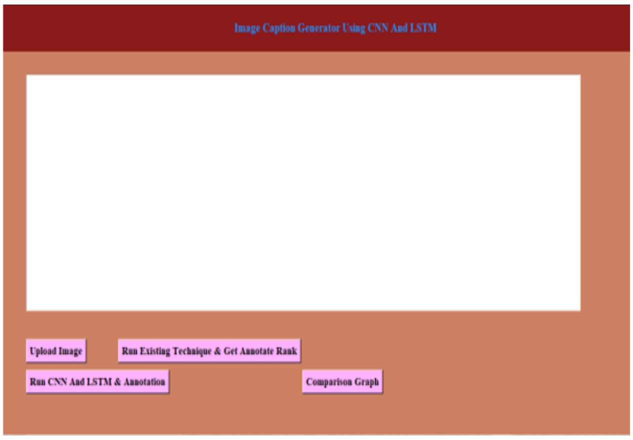


Image Caption Generator Using CNN And LSTM
Upload Image
Run Existing Technique & Get Annotate Rank
Run CNN And LSTM & Annotation
Comparison Graph

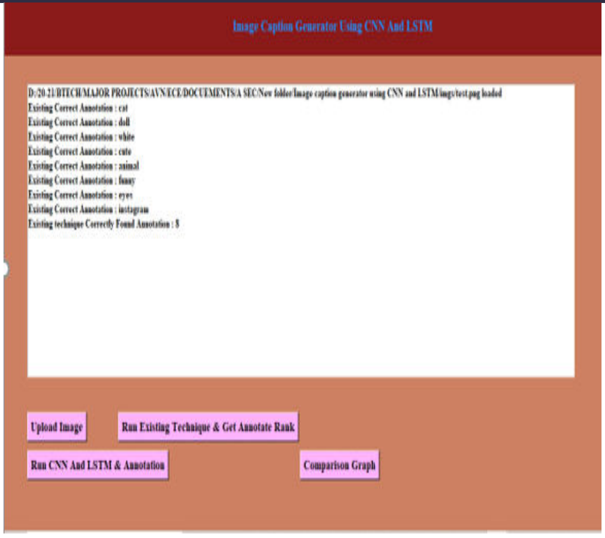


Image Caption Generator Using CNN And LSTM

```
D:/20-21/BTECH/MAJOR PROJECTS/AVN/ECE/DOCUEMENTS/A SEC/New folder/Image caption generator using CNN and LSTM/imgs/test.png loaded
Existing Correct Annotation : cat
Existing Correct Annotation : doll
Existing Correct Annotation : white
Existing Correct Annotation : cute
Existing Correct Annotation : animal
Existing Correct Annotation : funny
Existing Correct Annotation : eyes
Existing Correct Annotation : instagram
Existing technique Correctly Found Annotation : 8
```

Upload Image
Run Existing Technique & Get Annotate Rank
Run CNN And LSTM & Annotation
Comparison Graph

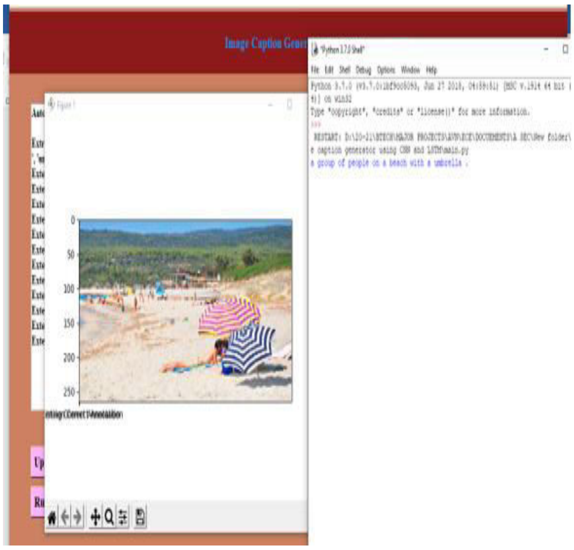


Image Caption Gener

Python 3.7.0 Shell

```
Python 3.7.0 (v3.7.0:1bf9cc5093, Jun 27 2018, 04:59:51) [MSC v.1914 64 bit (
4)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
 RESTART: D:\20-21\BTECH\MAJOR PROJECTS\AVN\ECE\DOCUMENTS\A SEC\New folder\
e caption generator using CNN and LSTM\main.py
a group of people on a beach with a umbrella .
```

Figure 1
File Edit Shell Debug Options Window Help
0
50
100
150
200
250

CONCLUSION

Using a keep probability of 75 percent for dropout and two layers for our decoder LSTM network, we conducted an extensive hyperparameter search over the CNN-LSTM model architecture, producing a best model that achieves results that are 3.3 BLEU-4 points and 3.8 CIDEr points behind the state-of-the-art. The model seems to be capable of correctly captioning a broad range of photos from the MSCOCO dataset, according to a detailed quantitative and qualitative study of the output metrics. Owing to a lack of attention to precise characteristics in photos, partial mistakes are common (for example, mislabeling an image of elephants wandering in an enclosure as 'elephants in a field' due to being distracted by trees in the background). This implies that the attention processes investigated in recent research might help with this job. The influence of emitted words on hidden states in the LSTM that were previously viewed as black boxes is our key innovative addition to the field. We showed that semantically near emitted words (e.g. 'plate' and 'bowl') cause identical hidden state movements despite differing preceding context, and that divergences in hidden state occur only when semantically distant words (e.g. 'vase' and 'meal') are emitted.

REFERENCES

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV,

ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.

[2] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. [4] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014. [5] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). CoRR, abs/1412.6632, 2014. [6] Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan. Show and tell: A neural image caption generator. CoRR, abs/1411.4555, 2014.

[7] Quanzeng You, HailinJin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. CoRR, abs/1603.03925, 2016.