

Image Captioning using Machine Learning

R. Malar kodi

Dr.G. Mathivanan

Department of computer science Engineering

Sathyabama Institute of Science and Technology

Abstract

Description of an image generated spontaneously in accordance with the computer vision and natural language processing. This model unites computer vision and machine translation using the mechanism of deep recurrent architecture. Based on the image it generates the natural sentences which describes the image. In order to improvise the likelihood of the aimed description, the model is trained accordingly. Precision and adaptability of the language is experimented on various datasets of the target image description. The Proposed model is accurate one in a way we verified qualitatively and quantitatively. Analysis on various improvements applied and tested with the performance.

Keywords: Image captioning, Recurrent Neural Network (RNN), Convolution Neural Network(CNN), language model.

1. INTRODUCTION

Having the capacity to consequently depict the substance of picture utilizing properly formed English sentences is an extremely difficult task, yet it could have incredible effect, for example by helping visually impaired person to get the content of pictures on the web. An Object recognition and image classifications are main focus of computer vision community. The Description of an image should explain the objects present in image as well as how the objects relate to each other based on its attributes and activities. For visual understanding, the natural language English needs an additional model. Instead of stitching sub-problems into one to produce outcome, in our model single joint proposed scheme which can take input image Im as input and model is trained to increase the likelihood $p(S|Im)$ in order to produce sequence of words $S=\{S_1, S_2, \dots\}$. Where each word S_i comes from the dictionary from that an image can be described adequately.

This main motivation of our work comes from recent growth in machine translation field. Which intend to give the task where modify the sentence form from sentence written in source language S into target language T . It is done by

Maximizing $(T|S)$. For a long period, machine translation is tough task with sequence of distinct tasks (interpreting words separately, arrangement of word in order, reordering, etc.). Complexities of the model can be rectified by using [11] Recurrent Neural Network (RNNs). An “encoder” RNN function is to read the source sentence and converts into fixed length vector form and “decoder” RNN is that to produce the target sequence accordingly.

In our proposed work we are replacing the existing RNN encoder by a deep Convolution Neural Network (CNN). For the past few years CNN producing the better result than RNN and its presentation of the input image which is embedded it with the fixed length vector form. This approach is used for different types of vision tasks. So it is obvious to use this model as an image “encoder” and pre training it to classify the image using the last hidden layer as an input to RNN decoder which generates the sentences. This model is termed as the Neural Image Caption (NIC).

The Proposed work contains two proposals. They are presenting end to end system for problem and it is a neural net that can fully trainable using stochastic gradient descent. Other one is the mode which combines the state of art sub-networks for vision and language models. It has advantage is it can be pre trained on larger corpora and added advantage over additional data.

It produces better functionality in the measure of performance over other comparative approaches.

Convolutional Neural Network

The Design of Convolutional Neural Network (CNN) is to take advantage of two dimensional structure of given input image. It is achieved with the help of local connections and tied weights which is continued by certain form of pooling. This results the translation invariant specification. Also CNN has added advantage in the factor of fewer parameters and it can be trained easily. Input and Output specification for CNN network is always fixed size. It is the type of feed forward artificial neural network which is designed to use less amount of preprocessing.

Recurrent Neural Networks

RNN is an artificial network where its connections between the units are in the form of directed graph along a sequence. The Processing of the sequence in RNN is done with the internal state (memory). Finite and infinite impulse recurrent models have an additional stored state and it is directly controlled by the neural network. These controlled states are termed as gated memory which is a part of Long Short Term Memory (LSTM). The Storage of data can be replaced by any other graph or network model if it is involved with any time delays or any feedback loops. RNN can handle dynamic input, output sequence lengths and uses time series information in order to create text sequentially. It is ideal for text and speech analysis.

2. RELATED WORKS

Benjamin Z. Yao et al.'s work[2] explains Image to text description (I2T) is discussed and the generation of image and video content with respect to the understanding of an image. These are divided into three major steps which include decomposition of input image and parsing sentences in natural language then outcomes are transformed into semantic representation. These representations are finally converted into meaningful and human readable form. Ali Farhadi et al.'s work[3] The Intermediate form is specified in this model and it is oversimplified. For evaluating this model quantitative and qualitative analysis are listed. First phase is to map image space to meaning space and

second phase qualitative result is images for sentences.

Jacob Devlin et al.'s work[4] gives the clear idea over the result improvement over other techniques are discussed clearly. Multimodal Recurrent Neural Network (MRNN) possesses better BLEU performance than ME LM or LSTM conditioned on set discrete functions. ME LM + DMSM methods outperforms the Multimodal Recurrent Neural Network in basis of human judgements. A k-nearest neighbour retrieval algorithm is introduced and which performs similarly to the MRNN in terms of both automatic metrics and human judgements. Usage of MRNN system with DMSM to provide additional scores in MERT re-ranking of the n-best produced by the image conditioned MELM is given.

Ryan Kiros et al.'s work[5] the technique of an encoder-decoder model that learns two strategies is discussed. Namely multimodal joint embedding space with images and text, unique language model for decoding distributed representations. This pipeline concept unifies the image-text embedding models with multi model neural language models. The Author presented the structure content neural language model that unravels the structure of a sentence to its substance, modelled on portrayals created by the encoder. The Encoder enables one to rank pictures and sentences while the decoder can create novel depictions starting with no help from outside.

Ralf Gerber et al.'s work[6] he describes the image sequence interpretation process which gives the descriptions behaviour of real world traffic scenes. It is treated with only a single vehicle or single agent and simultaneously the behaviour of other vehicles in the scene and those vehicles are considered to be relative agent. This type of approach is used to quantify the occurrences so that it can create more global conceptual descriptions using natural language quantifier. These quantified occurrences are presented using the special logic structures. Descriptions of the natural language are drawn from internal knowledge specifications.

3. MODEL

The Proposed model is categorised into four modules based on the procedures. Each one

has different requirements and functioning mechanisms. First one and basic one is data collection for further processing and the data is collected from MSCOCO Dataset. Second function is pre-processing in which the data are processed to TFRecord format. As third function training phase carried over by using the Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). Finally implementation process takes place which is connecting the trained model with input i.e, user given image to give caption output.

3.1 MODEL APPROACH

The Proposed model for image captioning is based on the neural network which can generate image descriptions. Advancements in the field of statistical machine translation give the robust sequence model. The possibility to attain state of art results is higher by directly increasing the probability of the matching translation for given input sentence in end to end fashion which is common for both interface and training.

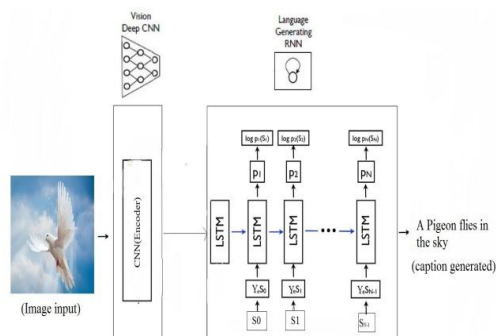


Fig.1. NIC model is based complete neural network consisting of a vision CNN followed by a language generating RNN with LSTM model.

Recurrent Neural Network (RNN) is utilised to encode the variable length input sequence into a fixed dimensional vector. Also this representation is used to decode the encoded information and desired output sequence. The Same criteria followed in image captioning model where instead of input sequence an image is given. Same principle is followed for translating it into description. Thus to attain correct description of given input image is attained directly maximizing the probability of correct description.

$$\theta^* = \arg \max_{\theta} \sum_{(Im, S)} \log p(S|Im; \theta), \quad (1)$$

Where θ and Im are parameters of our model, Im indicates an image and S its correct transcription. S represents any sentence and its length meant to be unbounded. Where N is the length of this particular example as

$$\log p(S|Im) = \sum_{t=0}^N \log p(S_t|Im, S_1, S_2, \dots, S_{t-1}) \quad (2)$$

The Step by step functioning is categorized into four main blocks. They are given below,

- [1] The Show and tell model works by first "encoding" an image into a fixed-length vector representation, and then "decoding" the representation into a natural language description.
- [2] The image encoder is a deep convolutional neural network.
- [3] The decoder is a long short-term memory (LSTM) network.
- [4] Words in the captions are represented with an embedding model. Each word in the vocabulary is associated with a fixed-length vector representation that is learned during training.

4. LSTM BASED SENTENCE GENERATOR

Long Short Term Memory (LSTM) is a RNN with the inbuilt memory cell and it stores information and exploits long range context. Gating units surrounds the LSTM memory cells and the task of these units includes the writing, reading and resetting information. LSTMs have been used to achieve acceptable performance in several tasks such as recognition of speech and hand writing, sequence generation and machine translation among others.

$$ig_t = \sigma(Y_{ix}x_t + Y_{im}m_{t-1}) \quad (3)$$

$$fg_t = \sigma(Y_{fx}x_t + Y_{fm}m_{t-1}) \quad (4)$$

$$og_t = \sigma(Y_{ox}x_t + Y_{om}m_{t-1}) \quad (5)$$

$$cg_t = f_t \odot c_{t-1} + i_t \odot h(Y_{cx}x_t + Y_{cm}m_{t-1}) \quad (6)$$

$$m_t = o_t \odot c_t \quad (7)$$

$$p_{t+1} = \text{Softmax}(m_t) \quad (8)$$

Where ig (input gate), fg (forget gate), og (output gate), $\odot$ indicates product with gate value and Y matrices are trained parameters. Equation of mt indicates the feed for the Softmax module. The term p_t means that the probability distribution of overall words.

4.1 PREPROCESSING

Preprocessing is the stage where the training of the model is taken care. To train the model, data needs to be fed in TFRecord Format. TFRecord Format consists of set of shared files containing serialized ``tf.SequenceExample`` protocol buffers. The ``tf.SequenceExample`` proto contains an image (JPEG format), a caption and metadata such as the image id. During preprocessing, a dictionary is created that assigns each word in the vocabulary to an integer-valued id. The caption is encoded as a list of integer word ids in the ``tf.SequenceExample`` protocol buffer.

4.2 TRAINING

The Functioning of LSTM is controlled by three gates and different colours are used to represent different actions in its structure. Recurrent connections are indicated in blue colour. This means to be the output 'm' at time (t-1) is fed back to the memory at time 't' through gates. The value of the cell is given back to the forget gate and the word predicted at time of (t-1) is fed back in addition to the memory output.

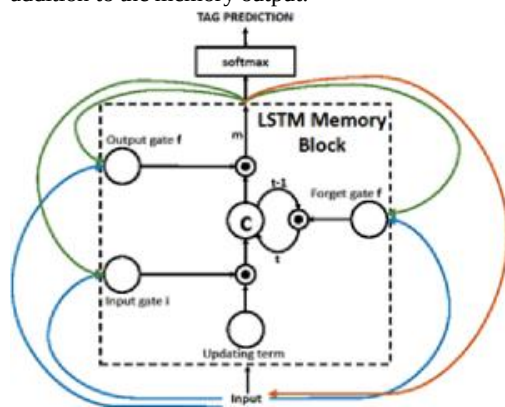


Fig.2. LSTM sequence generator

The Block Softmax which gets the input from 'm' and it generates tag prediction. In training phase the LSTM model is trained to predict top tags for all image inputs. The Features extracted from input images is said to be ACS DCNN feature and it is produced using GoogLeNet. The Training of the LSTM is done with the combination of

DCNN features and labels for pre-processed images.

4.3 INFERENCE

There is an availability of multiple techniques used to bring about sentence for a given image. This involves with two major steps and they are Sampling and Beam Search approach. In sampling process the sample of first word based on p_1 , then providing corresponding embedding as input and sample p_2 . This process is carried over until the sampling reaches the token which indicates the end of the token or maximum length. Second process is Beam Search and it consider the set of best sentences up to the duration of as the participants to create sentences with size $t+1$ and it keeps only the best k of it. By varying the beam size the experiments are conducted. According to the beam size BLEU points are varied.

5. EXPERIMENTS

To confirm the worthiness of our model various set of experiments are done using number of metrics, data sources and different model architectures in order to compare and make sure our model provides better performance.

5.1 EVALUATION METRICS

Even though there is a lack of clarity in the description of images, there is proposed evaluation metrics. The Best way to ask score on the usefulness of the image description is that asking raters to give score. But it is a time consuming process so the reinforcement of automatic metrics are correlated with attained scores and the procedures are listed in [12]. The Created sentences are analyzed by graders in the scale of 1 to 4. Each and every image is evaluated by two workers. Basic level of coordination between the workers range is given as 65 percent. When the problem arises on agreement then average of scores are calculated and it is stored for further process.

Other type of metrics can be calculated automatically assuming that one has access to the ground truth (descriptions created by humans).

The Literature on image descriptions is done using the BLEU score and it is the most used

one. The perplexity is the geometric mean of the inverse probability for each predicted word. The

choices are made among different parameters for selecting model but it's not reported because BLEU is always preferred. This metrics has certain drawbacks but it is correlated with the human observations. METEOR and ROGUE metrics are discussed and used for result comparison.

Transformation of description process to ranking process is hard and the complexity of the model increases as the size of the sentences probability grows exponentially with its dictionary. The Proposed model gives the descriptions with acceptable quality. We trust it is far better to concentrating on evaluation metrics for generation task rather than for ranking.

5.2 DATASETS

The Datasets are the one which contains images and sentences with English description. Kinds of datasets used for evaluation and comparison are SBU, Pascal VOC 2008, Flickr8k, Flickr30k and MSCOCO.

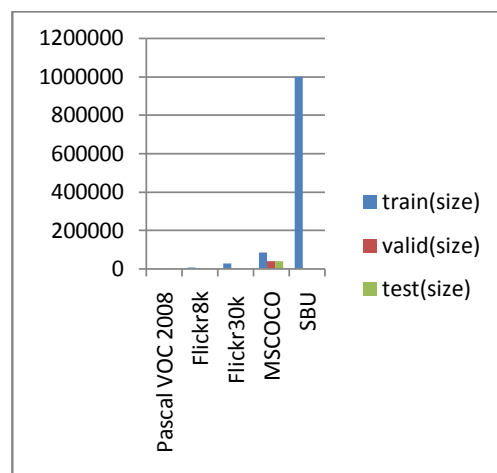


Fig.3. Different dataset with train, valid and test values

In specific Pascal dataset is used for testing only after system has been trained on various data such as any of the four dataset. For testing and training of SBU set 1000 images are reserved. Likewise same kind of reservation of images takes place for MSCOCO set also called COCO-4 and it is the one used to report results.

6. RESULTS

To justify the model proposed in this paper we need to answer certain queries regarding

impact of data size on generalization, the kinds of learning able to achieve with this model and how our model deals with weakly labeled examples.

So the experiment on five different dataset is carried over and presented to make sure the depth of our model. The pictorial ratio is given in Fig.4

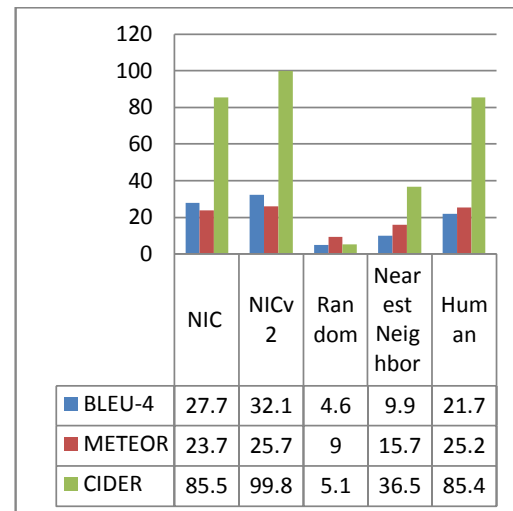


Fig.4. Points gained in MSCOCO improvement Setfordouble Models

6.1 TRAINING DETAILS

Number of difficulties faced during the training of our model is based on over fitting. Direct access needs huge amount of data, but datasets which has high aspect contains less than 100000 images. Referring description of an image is very tough work and it is harder than object classification. We trust that the attained results are absolutely good and the dominance of our method versus current human engineered approaches will get increases as the training set size increases.

The Exploration of the various methodologies is tried with our model in order to get rid of the problem over fitting. Familiar way to notoverfit is to start up the weights of the CNN component. This is done with all experiments and it helps in basis of generalization. All weights of trained sets are arbitrarily initialized except for CNN weights. Dimension used for embeddings 512 and size of the LSTM memory.

6.2 GENERATION RESULTS

Among two datasets PASCAL and MSCOCO, second one is preferred for training of model since the PASCAL does not contain any training set. The State of art results of PASCAL and SBU is not using the features of the images using deep learning. Human scores were compared with the other four and it is tabulated. Evaluation of our model captions by human rates results that our model needs work in order to achieve better result. From this section the improvisation factors are discussed with respect to original model.

6.3 TRANSFER LEARNING, DATA SIZE AND LABEL QUALITY

Among various model trainings and training sets we analysed the possibility of transferring model to an alternate dataset and the mismatch compensation capability also reviewed in terms of higher quality labels or increased training data. While considering transfer learning and data size between Flickr30k and Flickr8k, those were generated by same group. Flickr30k type of dataset produces the result with increased four BLEU points. MSCOCO set has 5 times bigger data training than with collection process varies with present one Flickr30k. It produces greater mismatch and more variations in vocabulary. Even though BLEU score of type lacks by 10 points it produces the reasonable descriptions.

There is no proper training set for PASCAL type and it is trained using the transfer learning from MSCOCO and Flickr. But it produces the degraded result with BLEU-1 score. Labelling ability of the SBU has no strong effect since the captions and the labels are not generated by humans. It is very tough task based on the larger data and noisier vocabulary. The Data available for training this model is greater than available model and which produces the poor performance. In terms of points it degrades from point 28 to 16.

6.4 GENERATION DIVERSITY DISCUSSION

A Trained model which gives the $p(S/Im)$ and the question arises regarding this is whether model generates novel captions and whether captions produced are diverse and high quality. N-best list is generated with the beam search decoder rather than go for best hypothesis. BLEU score for generated 15 sequences of sentences is 58 and it is more similar with humans. Among this generated model

contains few sentences which are not present in the training sequence. Even though the training data set is quite small it is easy to find the exemplar sentences. Analysis on top 15 generated sentences gives the result of novelty in description 50 per cent of the time. According to BLEU score generated sequences has enough quality and provides acceptable diversity.

6.5 RANKING RESULTS

It is not an effective way to represent result using the ranging scores. But in most of the papers they used ranging score to test training set of an image. The Technique function better on other parameters is MNLM (ranking aware loss). NIC can do a better job in both functioning (Ranks images for given description and ranks description for given image). Normalisation of the scores has been carried over in the image annotation work.

6.6 ANALYSIS OF EMBEDDINGS

The Usage of word embedding is meant for LSTM decoder which produces the S_t and it produced in order to represent the previous word S_{t-1} . This one is quite opposite to one hot encoding method. By learning the relationships between the words it helps the CNN to extract from the sequence generated.

6.7 IMPROVEMENTS OVER PRESENT MODEL

There are functions carried over to improvise the model ability to produce image caption accurately. Image model is improvised by normalization with batch normalization. In which each layer of neural network batch examples are normalised. Next process involves with image model fine tuning and that helps to neglect the over fitting problem. LSTM block alone specified in training with MSCOCO set and which improvises the performance of image captioning task. LSTM predict words by using the current caption of an image. Then to improve the performance of machine learning, resembling process takes place. Beam size reductionism process of maintaining most hit n sequences and keep track of it. For improved model beam size is decreased which generated the result with increased CIDER and the

score was high. It also gained the BLEU score with increment by four.

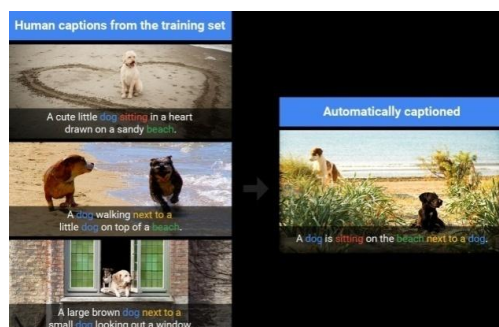


Fig.5. Model Image caption generated

7. CONCLUSION

Discussed model in this paper produces the best result and it is justified in terms of analysis reports and scores of the metrics or BLEU. Model is designed in way that complete package which can procreate moderate caption in understandable English. In Given model input image is encoded by CNN (Convolution Neural Network) and it produces concise representation. Sentence generation is carried over by the RNN (Recurrent Neural Network). The Model is testified with different dataset models and its outcomes were analysed accordingly. So the proposed model has the ability to produce image description based on properties of the image in different aspects. The Major requirement of this kind of model is in the field of robotics. It is used to make the robot analyse the image and produce the caption in order to functioning on command of the user.

REFERENCES

- [1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 39, NO. 4, APRIL 2017 PP., 652-663.
- [2] By Benjamin Z. Yao, Xiong Yang, Liang Lin, MunWai Lee, and Song-Chun Zhu, "I2T: Image Parsing to Text Description", Proceedings of the IEEE, Vol. 98, No. 8, August 2010 PP., 1485-1506
- [3] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, "David Forsyth1 Every Picture Tells a Story: Generating Sentences from Images", PP., 1-14.
- [4] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, "Language Models for Image Captioning: The Quirks and What Works", Association for Computational Linguistics, China, July 26-31, 2015, pp., 100-105.
- [5] Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models", Nov 2014, vol., 1, pp., 1-13.
- [6] Ralf Gerber and Hans-Hellmut Nagel, "KNOWLEDGE REPRESENTATION FOR THE GENERATION OF QUANTIFIED NATURAL LANGUAGE DESCRIPTIONS OF VEHICLE TRAFFIC IN IMAGE SEQUENCES", IEEE 1996, pp., 805-808.
- [7] Alex Graves, "Generating Sequences with Recurrent Neural Networks", June 2014, vol., 5, pp., 1-43.
- [8] Tsung-Yi Lin, Michael Maire, "Microsoft COCO: Common Objects in Context", Feb 2015 v3 PP., 1-15.
- [9] Junhua Mao, Wei Xu & Yi Yang & Jiang Wang & Zhiheng Huang, Alan Yuille, "DEEP CAPTIONING WITH MULTIMODAL RECURRENT NEURAL NETWORKS (M-RNN)", Published as a conference paper at ICLR 2015, pp., 1-17
- [10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.
- [11] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. Empirical Methods Natural Lang. Process., 2014.
- [12] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," J. Artif. Intell. Res., vol. 47, pp. 853-899, 2013.

