Image Captioning with Recurrent Neural Networks

Research proposal

B.Sc. Jakub Kvita

kvitajakub@gmail.com

Department of Computer Graphics and Multimedia
Faculty of Information Technology
Brno University of Technology
Brno, CZ
2015

Abstract

This research is trying to create image captioning system with good performance using combination of convolutional and recurrent neural networks. Solution can be very useful for both users and companies focused on image processing and searching, and researchers creating AI understanding images.

Data necessary for this task are already collected in form of COCO and Flickr datasets, which are publicly available. Code will be implemented with Torch, framework for machine learning, which is one of the most popular and developed among community. There has been created several metrics and benchmarks for machine translation, which has been adapted to captioning and are usually used. From those - BLEU, METEOR and CIDEr metrics will be used to evaluate created system.

Investigation of this problem is my Masters thesis, with scope appropriate to its level. It should be completed by May 2016.

Table of Contents

Introduction

Scene understanding is one of the fundamental and most difficult tasks of computer vision. In form of automatically generating image description using proper English sentences, it can have great effect in both real world applications and scientific research. Currently, there is no standardized solution for this problem, but with recent advancements of machine learning, researchers all around the world are trying to their best to tackle this task, using new methods for image processing and machine translation.

With solution that has sufficient performance, we can for example help visually impaired people understand images on web or to orient in range of different locations. Other possibility is to automatically search through images with keywords. In research this can be used as a foundation for artificial intelligence which has its input based on cameras or is processing large amount of visual data.

Since around 2012 when renaissance of neural networks in form of deep learning started, they have been by far the best performing tool for solving this problem, with no competitor in sight. Differences in solutions occur on level of deciding which types of neural networks researchers use and how they are connected together. This is not as simple as it looks like, because after several decades of neural networks research there are a lot of options to pick from and to combine to accomplish the task.

Background and Literature Review

There are different kinds of neural nets for processing different kinds of data. Convolutional forward nets are used for images as they can process them with output invariant to positions of features. Some ideas of its usage goes back to 1980s, but shape they have today has been formed in 2003 by Behnke in [1] and Microsoft team in [2].These networks usually have different kinds of layers, convolutions being most distinctive, among others max pooling which is looking for maximum value in a window of values, or layer providing nonlinear function, usually in form of sigmoid function or arc-tangent.

For sequences in form of text or sound waves (speech), recurrent neural nets turned out to be particularly useful, as they model sequential dependencies directly in them. At the beginning there was a problem with modeling and learning long term dependencies with vanilla RNNs, which was described as vanishing gradient problem. This has been overcome by work of Hochreiter and Schmidhuber [3] in 1997 with creation of new unit called Long Short Term Memory (LSTM). Since then this is one of the most often used units for RNN. This research has its successors and many variations of the basic unit has been created. Most notably, from last year, Gated Recurrent Unit (GRU) was proposed in [4] and was evaluated as simplified LSTM, which still has same power. General architecture for sequence to sequence learning was proposed by Sutskever last year in [5], which shown great results in translating English to French.

Nowadays as deep convolutional neural nets trained on GPUs became the state-of-the-art and can evaluate images in blink of an eye, researchers are trying to create nets with attention, which can focus on different parts of image during each forward evaluation. Since people and animals benefit greatly from this feature, this can prove very useful in progress of CNNs. In paper by Stollenga, Masci et al. [6],"dasNet" Deep Attention Selective network architecture has been created. It is using learned feedback connections and proved itself very effective. Several attention mechanisms were introduced in work [7] of people from universities in Toronto and Montreal. This work also focused on generating captions to images and finished among the top 5 models in last year MS COCO Challenge [8].

Most of the best models for captioning are competing in annual MS COCO Captioning Challenge, which is using their own data set. Besides paper mentioned in previous paragraph, notable is work of team from Google [9], which is trying to combine convolutional nets to generate vector representing image and then use it to start recurrent network to generate description. Worth mentioning is also model with similar results from Microsoft team [10], which creates set of words in the picture (nouns, verbs, adjectives) and then generating sentences with these words.

Most of the research in deep learning area is performed with the frameworks created to speed up development and testing new ideas. The biggest three are Caffe [11], working primarily in C/C++, Theano [12] with its huge ecosystem of packages implemented in Python, and Torch [13] for Lua. There are a lot of others, which are not as widely used. Ranging from Keras [14] and Lasagne [15] based on Theano to recently released TensorFlow [16] from Google or Microsoft and Samsung

tools. Most of these frameworks are open source, with development based on GitHub. It is very useful from user point of view to monitor what is implemented and how the authors plan to move forward, as this field is very live and new things are discovered every day. This also means that not everything is implemented in every framework. For example: Caffe has been known for not very good support of recurrent neural nets or some tools like Keras are not yet supporting import of Caffe models. Speed can also be limiting factor, as Keras is much slower than Torch, but usually researchers are limited by access to computing resources.

Until now, I have studied neural networks from basics to state-of-the-art models and algorithms. Among others I learned from very popular Geoffrey Hinton Coursera course [17], which is known for being very precise. I also studied papers which focus on similar topics to mine. Then I learned how is Theano, Keras and Torch working and got familiar with them. Currently I am training smaller models locally on my machine and bigger ones on Salomon, one of European supercomputers, or on Czech academic grid with powerful GPUs.

Research design and methodology

General idea for my research is to create two-part model. First part will be convolutional feedforward network which will be remotely based on some of the ImageNet [18] trained networks for image classification, with slight changes in last several layers. This network will be producing fixed length vector representing image which will be used as input for the next part.

Second part will be recurrent network suitable for text generation. There are multiple options how to use input from previous layer. At the beginning input can be fed to network or be used as initialization values for first layer. Second possibility is that it can be static input which will remain there for all the time slices until the end of generation. To tackle this problem it will require more research and simulations to prove which is more useful. Recurrent part of the network could be initialized with one of the networks used for machine translation to English, which will alleviate the need to teach the network from scratch as it will know basic principles and words in English. Some teams were very successful with this approach, for example team from Google in this [9] paper.

For training, there are three main datasets, I would like to use. First MS COCO [19] with over 90 000 images and several captions for each image. This will be the main dataset which will be divided into training, testing and evaluation part. Second and third are from Flickr (30k, 8k) [20][21]. Datasets are similar, based on images uploaded to Flickr. These two will be used to validate model on completely different data than training dataset.

As far as evaluation metrics are concerned, there are multiple options available. Many researchers are using BLEU [22] metric which is used to measure quality of translations between natural languages and has multiple versions, but recently was marked as flawed and obsolete. Therefore improved metrics were introduced – METEOR [23], which was designed directly to fix some of the problems of BLEU, and CIDEr [24]. I want to use all three metrics to evaluate my research, because they are commonly used and they are also included in COCO evaluation server, which is trying to standardize evaluation of captioning.

This research should be finished by May 2016. Torch will be framework of my choice, mostly because of the issues I mentioned earlier and that it is one of the most commonly used frameworks. Also because installation is done in user space and there is no need for superuser rights, which might not be available in some environments.

According to similar studies, it is expected to get to fairly good success rate reasonably quick. Improvements will then diminish and it will be much more complex to reach same level of increase. This captioning solution can be then applied in range of different problems, for example processing images on web – searching through images, processing uploaded galleries, generation of descriptions for the visually impaired. It can be also used in creation of general intelligence as it requires detailed understanding of picture and interpretation of this understanding via natural language.

Conclusion

Image captioning problem is not a new one, but with recent progress in deep learning field, new state-of-the-art solutions are emerging. Research is very live with no standard solutions, only preferred methods. As mentioned earlier, image captioning can be applied in wide range of web applications, image searching, in robotic vision systems or as a subsystem of artificial intelligence.

Nowadays all the necessary equipment such as tools and datasets, has been created and only thing left is to work on the best solution possible.

List of References

1: Sven Behnke, Hierarchical Neural Networks for Image Interpretation, 2003
2: Patrice Simard, David Steinkraus and John C. Platt, Best Practices for Convolutional Neural NetworksApplied to Visual Document Analysis, 2003
3: Sepp Hochreiter and Jürgen Schmidhuber, Long short-term memory, 1997
4: Kyunghyun Cho, et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014
5: Ilya Sutskever, Oriol Vinyals and Quoc V. Le, Sequence to Sequence Learning with Neural Networks, 2014
6: Marijn Stollenga, Jonathan Masci, Faustino Gomez and Juergen Schmidhuber, Deep Networks with Internal Selective Attention through Feedback Connections, 2014
7: Kelvin Xu, et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2015
8: , http://mscoco.org/dataset/#captions-challenge2015, 2015
9: Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell: A Neural Image Caption Generator,
10: Hao Fang, et al., From Captions to Visual Concepts and Back,
11: , Caffe | Deep Learning Framework, , http://caffe.berkeleyvision.org/
12: , Theano Documentation, , http://deeplearning.net/software/theano/
13: , Torch | Scientific computing for LuaJIT., , http://torch.ch/
14: , Keras Documentation, , http://keras.io/
15: , Welcome to Lasagne Documentation, , http://lasagne.readthedocs.org/en/latest/
16: , Home - TensorFlow, , http://www.tensorflow.org/
17: , Neural networks for Machine Learning - University of Toronto | Coursera, , https://www.coursera.org/course/neuralnets
18: , ImageNet, , http://image-net.org/
19: , MSCOCO Dataset, , http://mscoco.org/dataset/
20: , Flickr 8K Data, , https://illinois.edu/fb/sec/1713398
21: , Flickr30K & Denotation Graph data, , https://illinois.edu/fb/sec/229675
22: Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, BLEU : a Method for Automatic Evaluation of Machine Translation, 2002
23: Satanjeev Banerjee and Alon Lavie, METEOR: An Automatic Metric for MT Evaluation withImproved Correlation with Human Judgments, 2005
24: Ramakrishna Vedantam, C. Lawrence Zitnick and Devi Parikh, CIDEr: Consensus-based Image Description Evaluation, 2015