EXTENSION TO VISUAL INFORMATION NARRATOR USING NEURAL NETWORK

Shantanu S Suryawanshi, Rushikesh M Vaidya, Akshay V Patil, Prashant A Kale(Guide)

Students of BE Computer, Late.G.N.Sapkal College of Engineering, Trimbakeshwar, Nashik

---***---

Abstract - *Their were many advancements in Deep Learning based Machine Translation and Computer Vision which have led to excellent Image Captioning models using advanced techniques like Deep Reinforcement Learning and many more. While these models are very accurate, but these models need an expensive high configuration hardware which make it difficult to apply these models in real time scenarios, So we have developed an Android application which can run reliably on an inexpensive hardware i.e the low end hardwares such as hand-held devices like smartphones which operate on Android Operating System which will be helpful for Visually Impaired Peoples to visualize their surrounding using our application .We have also compared our results evaluated using various hardware systems such as CPU and GPU with our models and analyze the difference between computation speed and quality. Our model is trained on FLICKR dataset using an TensorFlow framework by Google, we have implemented this Android application to demonstrate the real time applicability and optimizations for the sake of Visually Impaired Peoples.*

***Keywords*: Tensorflow, CNN, Text-to-speech, LSTM, Natural Language Generation**

1. INTRODUCTION

Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task, but it could have great impact, for instance by helping visually impaired people better understand the content of images on the web. This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding.

2. LITERATURE SURVEY

2.1 Generating image descriptions using dependency relational patterns

Published year : 2016

Author : Ahmet Aker

This paper presents a novel approach to automatic captioning of geotagged images by summarizing multiple web documents that contain information related to an images location. The summarizer is biased by dependency pattern models towards sentences which contain features typically provided for different scene types such as those of churches, bridges, etc. Our results show that summaries biased by dependency pattern models lead to significantly higher ROUGE scores than both n-gram language models reported in previous work and also Wikipedia baseline summaries. Summaries generated using dependency patterns also lead to more readable summaries than those generated without dependency patterns.

2.2 Neural machine translation by jointly learning to align and translate

Published year : 2015

Author : Dzmitry Bahdanau

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

2.3 Learning phrase representations using RNN encoder-decoder for statistical machine translation

Published year : 2018

Author : Kyunghyun Cho

In this paper, we propose a novel neural network model called RNN Encoder Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder Decoder as an additional feature in the existing log-linear model. Qualitatively, we show that the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases.

3. SYSTEM ARCHITECTURE

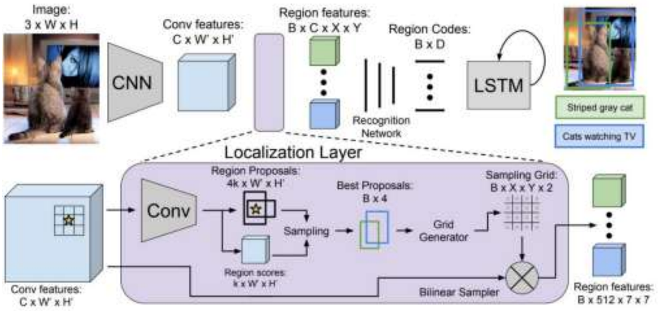

Fig-1: General System Architecture

• **Input Image**: It is the actual nature which we want to view.

• **Convolutional Feature Extraction**:In CNN model the features are extracted from the input image.

• **RNN**: In RNN model the extracted features from CNN we generate actual textual output.

4. METHODOLOGY

4.1 Tensorflow

TensorFlow is an open-source software library for dataflow programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.

4.2 Imagenet inception-v4

Very deep convolutional networks have been central to the largest advances in image recognition performance in recent years. One example is the Inception architecture that has been shown to achieve very good performance at relatively low computational cost. Recently, the introduction of residual connections in conjunction with a more traditional architecture has yielded state-of-the-art performance in the 2015 ILSVRC challenge; its performance was similar to the latest generation Inceptionv3 network. This raises the question of whether there are any benefit in combining the Inception architecture with residual connections.

4.3 VGG

VGG Netis a neural network that performed very well in the Imagenet Large Scale Visual Recognition Challenge(ILSVRC) in 2014. It scored first place on the image localization task and second place on the image classification task. Localization is finding wherein the image a certain object is, described by a bounding box. Classification is describing what the object in the image is. This predicts a category label, such as cat or bookcase. Image Netis a huge database of images for academic researchers. Every year the people who run ImageNet host an image recognition competition. The goal is to write a piece of software these days usually a neural network of some kind that can correctly predict the category for a set of test images. Of course, the correct categories are known only to the contest organizers.

4.3 ConvNets

Convolutional networks (ConvNets) currently set the state of the art in visual recognition. The aim of this project is to investigate how the ConvNet depth affects their accuracy in the large-scale image recognition setting.

5 CNN

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery. CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually refer to fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. Typical ways of regularization includes adding some form of magnitude measurement of weights to the loss function. However, CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns.

CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage.

5. ARCHITECTURE

Our Model works on encoder-decoder neural network. It works by first "encoding" an image into a fixed-length vector representation, and then "decoding" the representation into a natural language description. The image encoder is a deep convolutional neural network. This type of network is widely used for image tasks and is currently state-of-the-art for object recognition and detection. Our particular choice of network is the Inception v4 image recognition pretrained model.
The decoder is a long short-term memory (LSTM) network. This type of network is commonly used for sequence modeling tasks such as language modeling and machine translation. In the Show and Tell model, the LSTM network is trained as a language model conditioned on the image encoding. Words in the generated text are represented with an embedding model. Each word in the vocabulary is associated with a fixed-length vector representation that is learned during training.

The following diagram illustrates the model architecture.

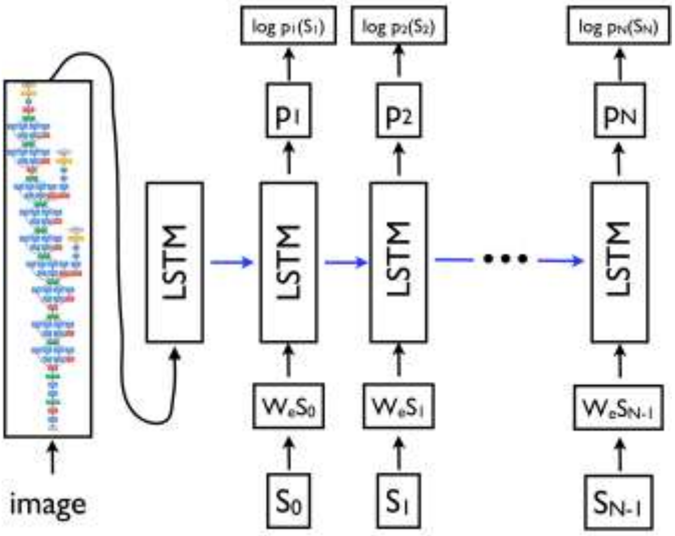

Fig-2: System Architecture

In this diagram, {s0, s1, ..., sN-1} are the words of the text and {wes0, wes1, ..., wesN-1} are their corresponding word embedding vectors. The outputs {p1, p2, ..., pN} of the LSTM are probability distributions generated by the model for the next word in the sentence. The terms {log p1(s1), log p2(s2), ..., log pN(sN)} are the log-likelihoods of the correct word at each step; the negated sum of these terms is the minimization objective of the model.

During the first phase of training the parameters of the Inception v4 model are kept fixed: it is simply a static image encoder function. A single trainable layer is added on top of the Inception v4 model to transform the image embedding into the word embedding vector space. The model is trained with respect to the parameters of the word embeddings, the parameters of the layer on top of Inception v4 and the parameters of the LSTM. In the second phase of training, all parameters - including the parameters of Inception v4 - are trained to jointly fine-tune the image encoder and the LSTM.

Given a trained model and an image we use beam search to generate text for that image. Text are generated word-by-word, where at each step t we use the set of sentences already generated with length t - 1 to generate a new set of sentences with length t.
This way we generate our text.

6. Conclusion

From this project we have made an Android application using Convolutional Neural Network and Natural Language Processing (Text-to-speech) for the sake of Visually Impaired People in order to visualise them their surroundings just by pointing their mobile device to get information using Narration Form

REFERENCES

[1] A. Aker and R. Gaizauskas. Generating image descriptions using dependency relational patterns. In ACL, 2010.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, 2014.

[3] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP, 2014

[4] Pranay Mathur, Aman Gill, Ayush Yadav, Anurag Mishra and Nand Kumar Bansode. A real time image to caption generator, 2017

[5] D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP, 2013.