CS 441/541: Artificial Intelligence

Image Captioning using CNN and RNN

Abstract:

Image Captioning is the process of generating a textual description of an image. It uses both Natural Language Processing and Computer Vision to generate the captions. Creating a web application that will extract features from an image and automatically annotate them and generate the output in the form of audio. Visually challenged people will not be able to interpret what is being shown on the screen. In response to the problem, our study proposes a model which will be able to extract features from an image and automatically annotate them and generate the output in the form of audio. We will be using Convolutional Neural Networks (CNN) for separate image classification and Recurrent Neural Networks (RNN) for the purpose of caption generation and converting them into audio format.The goal of this project is to develop a visual system that generates a textual description of the object in the image. Given an image, break it down to extract features and finally generate a meaningful caption for the image. Convert the generated caption into voice output.

Introduction:

Image Captioning is the process of generating a textual description of an image. It uses both Natural Language Processing and Computer Vision to generate captions for visually challenged people who will not be able to interpret what is shown on the screen. In response to the problem, our study proposes a model which will extract features from an image and automatically annotate them and generate output in the form of audio.

We will be using convolutional neural networks for separate image classification and recurrent neural networks for the purpose of caption/description generation. The image captioning model can be used for helping visually impaired people, telling stories using albums, etc. Given an image, the model is able to describe the image. In order to achieve this, our model includes an encoder which is a CNN (Convolutional Neural Networks), and a

decoder which is an RNN (Recurrent Neural Network).The CNN encoder gets an image as an input for a classification task and its output is fed into the RNN decoder which in turn outputs the audio about the image.

The main objective of the project is to develop a visual system that generates a textual description of the image. Given an image, break it down to extract features and finally generate a meaningful caption for the image. Convert the generated caption into voice output. Visually challenged people will not be able to interpret what is shown on the screen. In response to the problem, our study proposes a model which will extract features from an image and automatically annotate them, and generates output in the form of audio. We will be using Convolutional Neural Networks (CNN) for image classification and Recurrent Neural Networks (RNN) for the purpose of caption/description generation and then convert it into audio. The image captioning model can be used for helping visually impaired people, telling stories using albums, etc.

This paper discusses the pattern recognition system by relying more on automatic learning and character recognition, that trains all the modules to optimize a global performance criterion[1]. LSTM networks are well-suited for classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. [2] This paper discusses the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. The evaluation of networks of increasing depth using an architecture with very small convolution filters and it proposes a modified convolutional network architecture by increasing the depth, using smaller filters, data augmentation and a bunch of engineering tricks.[3] This paper discusses the existence of interpretable cells that keep track of long-range dependencies such as line lengths, quotes and brackets. Moreover, the comparative analysis with finite horizon gram models traces the source of the LSTM improvements to long-range structural dependencies. Finally, provides analysis of the remaining errors and suggests areas for further study. [4]

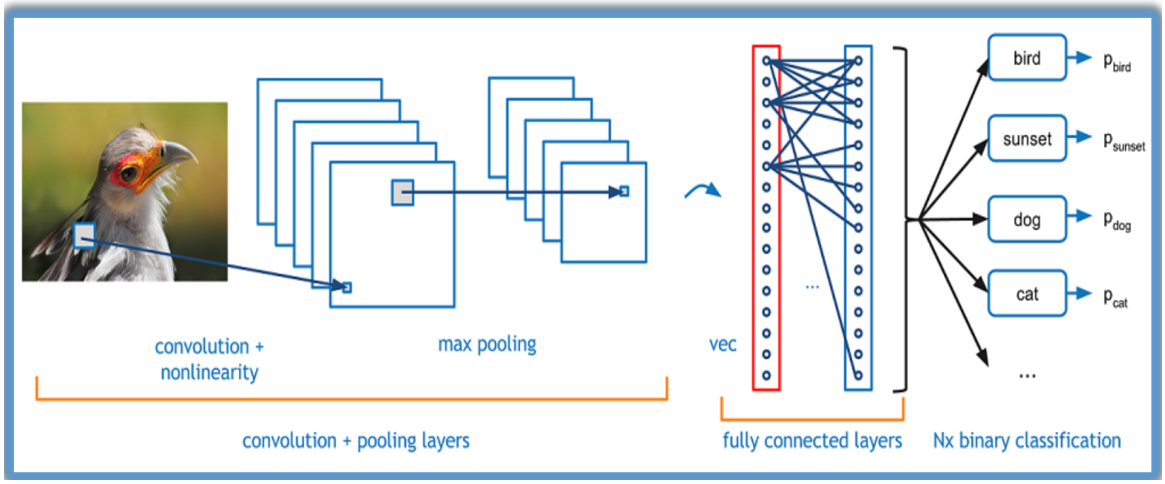
Methods:

CONVOLUTIONAL NEURAL NETWORKS (CNN):

Convolutional Neural Networks are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product, and optionally follows it with a non-linearity. Convolutional neural networks are deep artificial neural networks that are used to classify images, cluster them by similarity and perform object recognition within scenes.

Image classification is the task of considering an input image and finding a class or a probability of classes that best describes the image. For humans, this task of recognition is one of the first skills we learn from the moment we are born and is one that comes naturally and effortlessly as adults. Without even thinking twice, we're able to quickly and seamlessly identify the environment we are in as well as the objects that surround us. When we see an image or just when we look at the world around us, most of the time we are able to immediately characterize the scene and give each object a label, all without even consciously noticing. These skills of being able to quickly recognize patterns, generalize from prior knowledge, and adapt to different image environments are ones that we do not share with our fellow machines.

When a computer sees an image (takes an image as input), it will see an array of pixel values. Depending on the resolution and size of the image, it will see a 32 x 32 x 3 array of numbers (RGB values). Just to drive home the point, let's say we have a color image in JPG form and its size is 480 480. The representative array will be 480 x 480 x 3. Each of these numbers is given a value from 0 to 255 which describes the pixel intensity at that point. The idea is that we give the computer this array of numbers and it will output numbers that describe the probability of the image being a certain class.

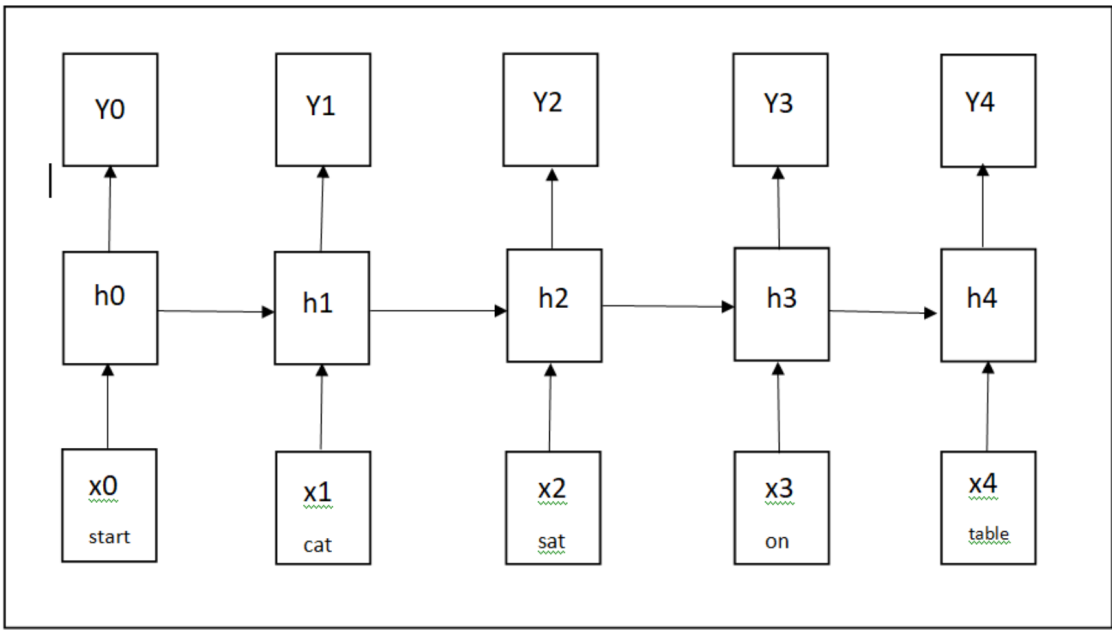


RECURRENT NEURAL NETWORKS (RNN):

Recurrent Neural Network (RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is the Hidden state, which remembers some information about a sequence

Suppose there is a deeper network with one input layer, three hidden layers and one output layer. Then like other neural networks, each hidden layer will have its own set of weights and biases. This means that each of these layers are independent of each other, i.e. they do not memorize the previous outputs.RNN converts the independent activations in dependent activations by providing the same weights and biases to all the layers, thus reducing the complexity of increasing parameters and memorizing each previous outputs by giving each output as input to the next hidden layer. Hence these three layers can be joined together such that the weights and bias of all the hidden layers is the same, into a single recurrent layer.

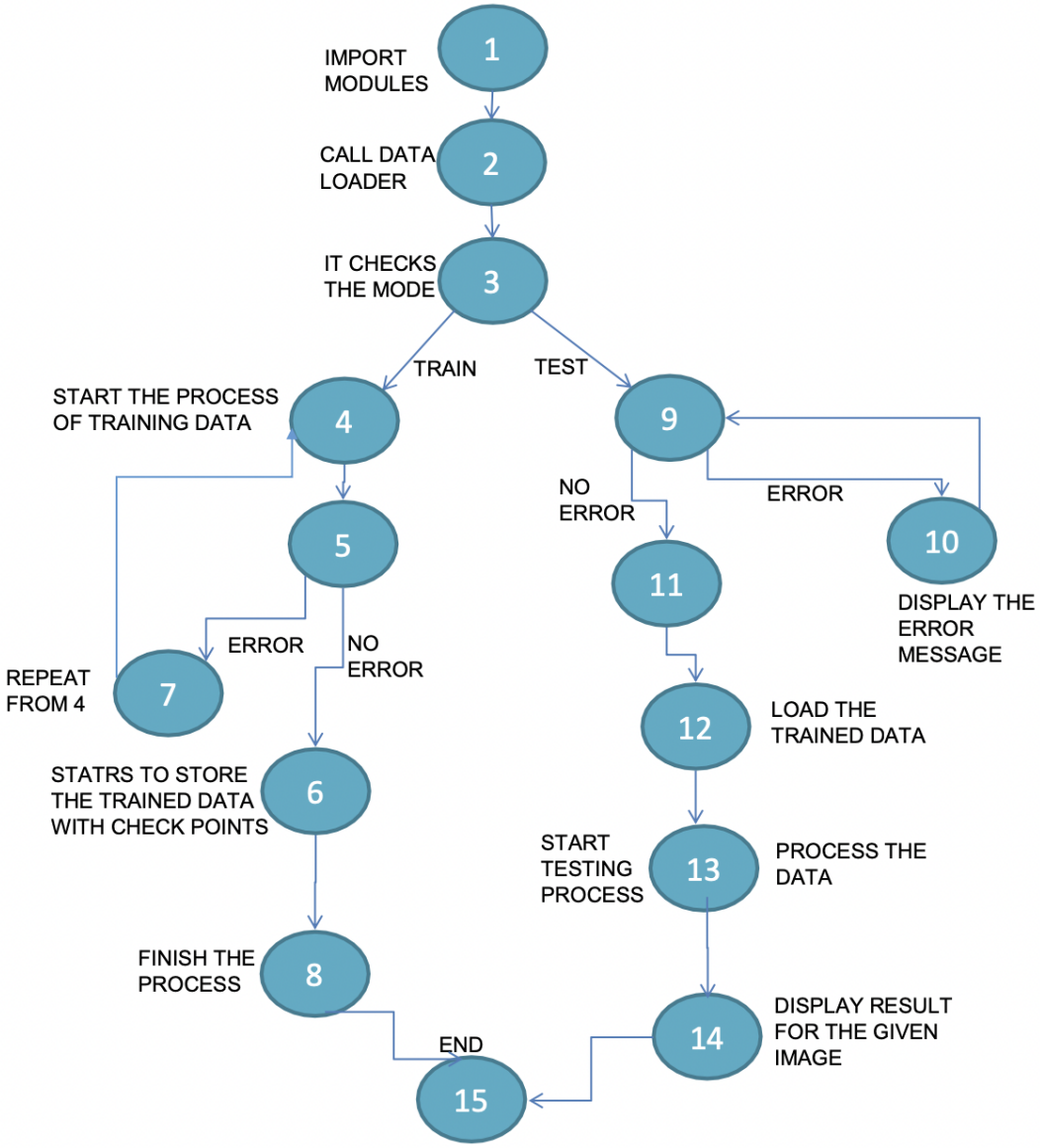
Machine Translation is similar to language modeling in that our input is a sequence of words in our source language (e.g. German). We want to output a sequence of words in our target language (e.g. English). A key difference is that our output only starts after we have seen the complete input, because the first word of our translated sentences may require information captured from the complete input sequence.



xi - number associated with word

hi - hidden representation mediates the contextual information

Flow Diagram:



Cyclomatic Complexity:

Cyclomatic complexity is used to find the source code complexity that is being correlated to a number of coding errors. It is calculated by developing the control flow graph of the code.

$$V(G) = P+1 = 3+1 = 4$$

$$V(G) = R = 4$$

$$V(G) = E - N + 2 = 17 - 15 + 2=4$$

Independent Paths:

Independent paths are the available paths that traverse at least one new edge in the flow graph.

1. 1-2-3-4-5-6-8-15
2. 1-2-3-4-5-7-4-5-6-8-15
3. 1-2-3-9-10-9-11-12-13-14-15
4. 1-2-3-9-11-12-13-14-15

Tools Used:

DJANGO: Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It's free and open source.

PYCHARM: PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django.

TKINTER: Tkinter is Python's de-facto standard GUI (Graphical User Interface) package. It is a thin object-oriented layer on top of Tcl/Tk. Tkinter is not the only GUI Programming toolkit for Python but it's the most popular one.

Conclusion:

We created and deployed a visual application to generate a caption for a given image and audio generation. We also test our model against other recent works in image captioning, by making use of CNN and RNN architecture and simplifying the overall design. In the future, we can optimize our model to perform in the real world with high quality audio.

References:

1. Andrej Karpathy and Li Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, 2015, pp. 3128–3137.
2. Sepp Hochreiter and Jurgen Schmidhuber, “Long short- term memory,” Neural computation, vol. 9, no. 8, pp.1735–1780, 1997.
3. Johnson J Karpathy A and Densecap F, “Fully convolutional localization networks for dense captioning," CoRR, vol. abs/1511.07571, 2015.
4. Vinyals O, Toshev A, Bengio A, and Erhan D,”Lessons learned from the 2015 MSCOCO image captioning challenge," CoRR, vol. abs/1609.06647, 2016.
5. Goodfellow I, Bengio Y, and Courville A, “Book in preparation for MIT Press, 2016.
6. Simonyan K and Zisserman A, “Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014.
7. Kolar M, Hradis M, and Zemc_k, “Image captioning with semantically similar images," CoRR, vol. abs/1506.03995, 2015.
8. Karpathy A and Li F, “Deep visual-semantic alignments for generating image descriptions," CoRR, vol. abs/1412.2306, 2014.
9. Anderson P, Fernando B, Johnson M, and Gould S, “Semantic propositional image caption evaluation," in ECCV, 2016.

.