# CNN-LSTM Architecture and Image Captioning - Analytics Vidhya - Medium

Deep learning is one of the most rapidly advancing and researched field of study that is making its way into all of our daily lives. It is...

By Shweta Pardeshi

Mar 25, 2023 10:56 AM · 4 min. read ·
View original

This post is co-authored by Kanishkkalra.

Deep learning is one of the most rapidly advancing and researched field of study that is making its way into all of our daily lives. It is simply the application of artificial neural networks using heavy and high-end modern hardware. It allows the development, training, and use of neural networks that are much larger (more layers) than was previously thought possible. There are thousands of types of specific neural networks proposed by researchers as modifications or

tweaks to existing models. Some of the more prominent ones as CNN's and RNN's.

Convolutional Neural Networks were designed to map image data to an output variable. They have proven so effective that they are the go-to method for any type of prediction problem involving image data as an input.

Recurrent Neural Networks, or RNNs, were designed to work with sequence prediction problems. Some of these sequence prediction problems include one-to-many, many-to-one, and many-to-many.

LSTM networks are perhaps the most successful RNN's as they allow us to encapsulate a wider sequence of words or sentences for prediction.

## The CNN-LSTM Model

One of the most interesting and practically useful neural models come from the mixing of the different types of networks together into hybrid models.

### EXAMPLE

Consider the task of **generating captions for images**. In this case, we have an input image and an output sequence that is the caption for the input image.

### Can we model this as a one-to-many sequence prediction task?

Yes, but how would the LSTM or any other sequence prediction model understand the input image. We cannot directly input the RGB image tensor as they are ill-equipped to work with such inputs. Input with spatial structure, like images, cannot be modeled easily with the standard Vanilla LSTM.

### Can we extract some features from the input image?

Yes, this is precisely what we need to do in order to use the LSTM architecture for our purpose. We can use the deep CNN architecture to extract features from the image which are then fed into the LSTM architecture to output the caption.

**This is called the CNN LSTM model,** specifically designed for sequence prediction problems with spatial inputs, like images or videos. This architecture involves using Convolutional Neural Network (CNN) layers for feature extraction on input data combined with LSTMs to perform sequence prediction on the feature vectors. In short, CNN LSTMs are a class of models that are both spatially and temporally deep and sit at the boundary of Computer Vision and Natural Language Processing. These models have enormous potential and are being increasingly

used for many sophisticated tasks such as text classification, video conversion, and so on. Here is a generic architecture of a CNN LSTM Model.

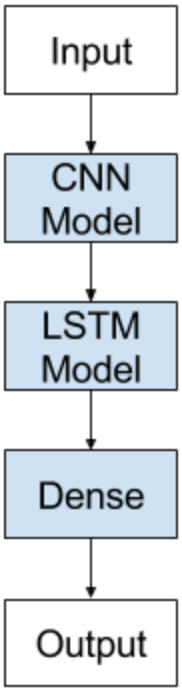


Image Source: https://machinelearningmastery.com/cnn-long-short-term-memory-networks/

## Image Captioning

Describing an image is the problem of generating a human-readable textual description of an image, such as a photograph of an object or scene. It combines both computer vision and natural language processing together.

Neural network models for captioning involve two main elements:

1. Feature Extraction.
2. Language Model.

The rest of the article will elucidate our thoughts and observations while implementing a CNN-LSTM model for image captioning. Note that this post is not a tutorial on image captioning implementation but is aimed at exploring the CNN-LSTM architecture and its practical usage. The code was written in python3 and implemented in Keras. Here are the necessary requirements and pre-requisite for you to be able to understand the implementation completely. If you are interested in the implementation tutorial, you can go to https://bit.ly/2XFCEmN.

The dataset used can be downloaded using the following links.

**Flickr8k_Dataset**(Contains 8092 photographs in JPEG format) — https://bit.ly/35shVWb

**Flickr8k_text**(Contains a number of files containing different sources of descriptions for the photographs.) — https://bit.ly/2DcBAgF

The dataset has a pre-defined training dataset (6,000 images), a development dataset (1,000 images), and test dataset (1,000 images).

The dataset information as well as data preparation for the model can be seen in the same link above.

Here, we will only show the important snippets of the code that has been used to create and run the model. You are encouraged to use a different dataset and prepare your dataset accordingly.

## Feature extraction

The feature extraction model is a neural network that given an image is able to extract the salient features, often in the form of a fixed-length vector. A deep convolutional neural network, or CNN, is used as the feature extraction submodel. This network can be trained directly on the images in your dataset. Alternatively, you can use a pre-trained convolutional model as shown.

## Language Model

For image captioning, we are creating an LSTM based model that is used to predict the sequences of words, called the caption, from the feature vectors obtained from the VGG network.

The language model is trained for 20 epochs. You can play around with other parameters and tune them as much as you want. We are displaying here one result that we obtained after training our network.

Generated Caption: Two girls are playing in the water.

## Conclusion

A CNN-LSTM architecture has wide-ranging applications as it stands at the helm of Computer Vision and Natural Language Processing. It allows us to use state of the art neural models for NLP tasks such as the transformer for sequential image and video data. At the same time, extremely powerful CNN networks can be used for sequential data such as the natural language. Hence, it allows us to leverage the useful aspects of powerful models in tasks they have never been used for

before. This post was just to introduce the concept of hybrid neural models and encourage the people to increasingly use different architectures of the CNN-LSTM models.

## References