Long-Term Recurrent Convolutional Network (LRCN)

The long-term recurrent convolutional network (LRCN) is a unique neural network architecture for generating descriptions of images and videos (which is seen as a sequence of images). These problems can be termed as visual time series modeling. The LRCN architecture combines the ability of the convolutional neural network (CNN) to extract image features together with a recurrent network for learning sequences or long-term dependencies. The LRCN passes visual inputs into a CNN to retrieve image features as outputs. These outputs are then passed into a recurrent LSTM network layer to generate the natural language descriptions. The recurrent layer can contain stacked LSTMs.

One core advantage of LRCN for modeling sequential vision problems such as image captioning and video captioning is that the network is not constrained to fixed lengths of inputs and outputs. Hence, it can be used to model sequential data with different lengths such as textual data and videos.

The following illustrations show how LRCN is applied to a variety of sequence problems:

1. Image captioning: Image captioning can be seen as a one-to-many sequence problem. The input is an image and therefore a static input, and the output is a sequence of text that describes the objects in the image; this is a sequential output. The use of LRCN for image captioning is illustrated in Figure 36-16.

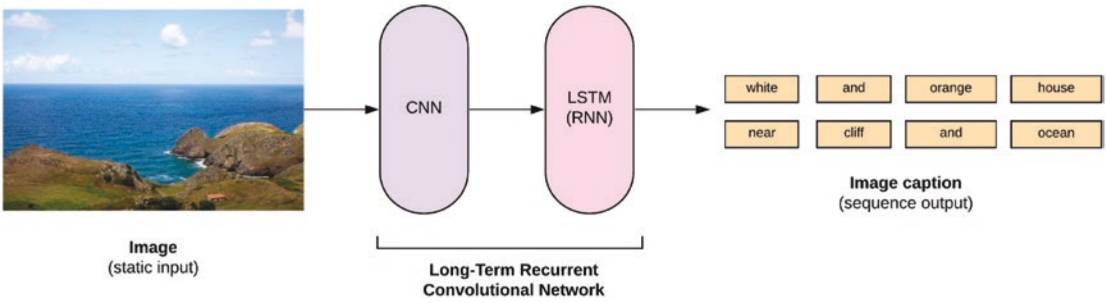

Figure 36-16. *Image captioning (photo by Daniel Llorente on Unsplash)*

2. Video captioning: Video can be seen as a sequence of images. Hence, in a video captioning problem, a sequence of images is passed as input to the LRCN model which in turn returns a sequence of outputs as a textual description for each video frame. Hence, video captioning can be seen as a many-to-many sequence problem. This approach is an example of an Encoder-Decoder LSTM where CNN is used as an image encoder that is initially trained for image classification. The final hidden layer, which is also called a bottleneck, is then passed as input to the RNN decode. It is typical to use an already pre-trained CNN on a large-scale image recognition task. A number of such models exist in the public domain. We will survey Encoder-Decoder LSTMs in more detail shortly. Video captioning is illustrated in Figure 36-17.

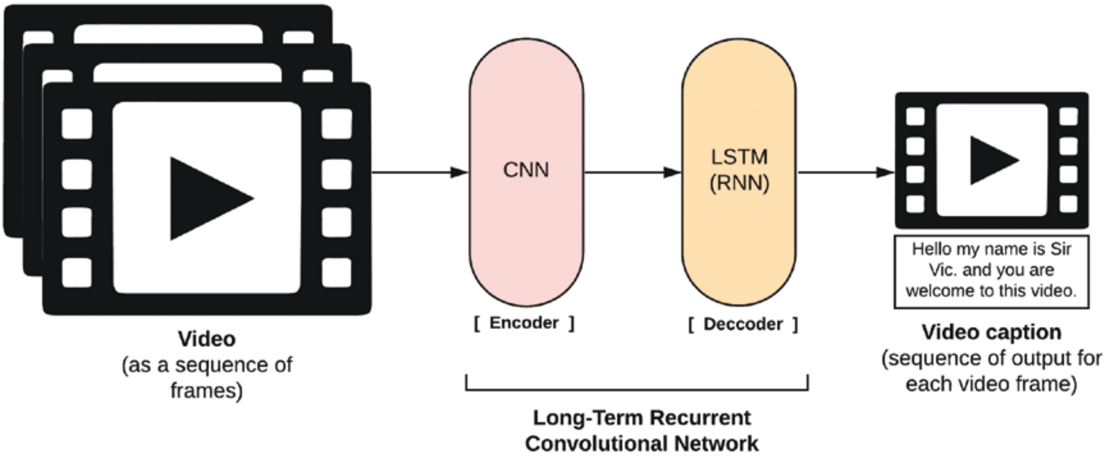
CNN
LSTM
(RNN)
Hello my name is Sir Vic. and you are welcome to this video.
Video
(as a sequence of frames)
[Encoder]
[Deccoder]
Video caption
(sequence of output for each video frame)
Long-Term Recurrent Convolutional Network

Figure 36-17. *Video captioning*

Encoder-Decoder LSTMs

Encoder-Decoder LSTM architecture handles a particular class of sequence problems that takes as input multiple time steps and also returns a multiple time step output. A major challenge of this sort of problems is that both the input and output sequences can have varied lengths.

The first part of the architecture, that is, the Encoder, is responsible for receiving and encoding the input sequence; the second part of the architecture, that is, the Decoder, takes in the output from the Encoder and then predicts the output sequence.