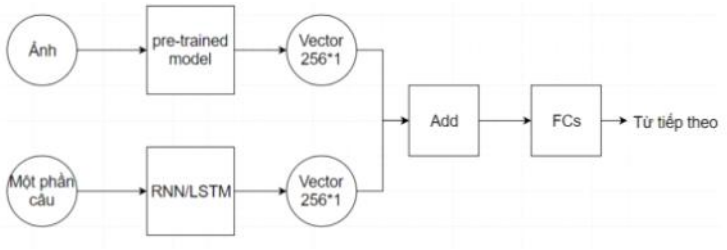
Challenge 2

Monday, January 3, 2022 9:54 PM

Model Architecture



Input: an image
Output: an caption

The input is an image, which is extracted feature though pre-trained model base-on big dataset and Inception v3 model. This process is called image embedding and output is a vector with shape (1,256) .

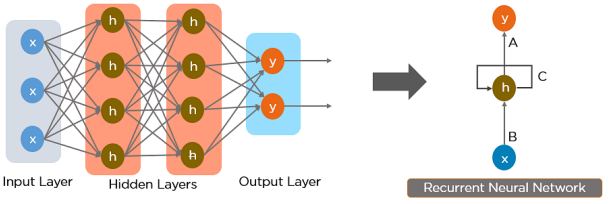
With the captions, we first have to vectorize the words into numbers by using Glove model. Then, we fit the embeded word into LSTM model, this model will predict the probability of which word will appear next after the word we have inputed.

The idea of image captioning is using embedding of the image and using appearing words to predict the next word in caption.

For example:
Embedding vector + A -> girl
- Embedding vector + A girl -> going
- Embedding vector + A girl going -> into
- Embedding vector + A girl going into -> a
- Embedding vector + A girl going into a -> wooden building
- Embedding vector + A girl going into a wooden -> building

To predict the next word, we need to construct a library of words, which is appeared on the training set.

Math - Recurrent Neural Network (RNN)



RNN works on the principle of saving the output of a particular layer and feeding this back to the input in order to predict the output of the layer.

The nodes in different layers of the neural network are compressed to form a single layer of recurrent neural networks. A, B, and C are the parameters of the network.

“x” is the input layer, “h” is the hidden layer, and “y” is the output layer. A, B, and C are the network parameters used to improve the output of the model. At any given time t, the current input is a combination of input at x(t) and x(t-1). The output at any given time is fetched back to the network to improve on the output.