LSTM for Image Annotation with Relative Visual Importance

Geng Yan[1]

Yang Wang[2]

Zicheng Liao[1]

[1] College of Computer Science
Zhejiang University

[2] Department of Computer Science
University of Manitoba

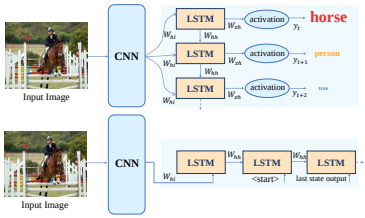

Figure 1: Illustration of the LSTM model. (Top) In our model, the image feature is used as an input to the LSTM at each time step. (Bottom) In the LSTM model used for image captioning, the image feature is only used to start the initial state in the LSTM model.

We consider the problem of image annotations that takes into account of the relative visual importance of tags. Humans have the remarkable ability to selectively process very narrow regions of the scene that are important to us. So when asked to annotate an image, we only mention a subset of the objects appearing in the image, and we mention the important objects first. In this paper, we propose a method for producing such ranked tag list for a given image. Such a ranked tag list can be useful for various applications including image retrieval, image parsing and image caption generation.

Our proposed approach combines the convolutional neural network (CNN) for images and the LSTM for sequential data. Fig. 1 illustrates our model and compare it with the RNN model for image captioning.

Image representation: Following prior work (e.g. [1]), we represent an image as a 4096-dimensional CNN feature vector using pre-trained VGGNet. We then use a fully connected layer to reduce the dimension to d. In other words, given an input image I_m, we represent it as a d-dimensional feature vector as:

$$I = W_I \cdot CNN(I_m) + b_I \tag{1}$$

where $W_I \in \mathbb{R}^{d \times 4096}$ and $b_I \in \mathbb{R}^d$ are the parameters to be learned. $CNN(I_m)$ is the 4096-dimensional CNN feature extracted on the image I.

LSTM for tag list prediction: We modify the standard LSTM, so that the hidden state at each time step considers the image feature $v(I)$ as one of the inputs. In other words, our LSTM model is defined as follows:

$$i_t = \sigma(W^{(i)} I + U^{(i)} \mathbf{h}_{t-1}) \tag{2}$$
$$f_t = \sigma(W^{(f)} I + U^{(f)} \mathbf{h}_{t-1}) \tag{3}$$
$$o_t = \sigma(W^{(o)} I + U^{(o)} \mathbf{h}_{t-1}) \tag{4}$$
$$\tilde{\mathbf{c}}_t = \tanh(W^{(c)} I + U^{(c)} \mathbf{h}_{t-1}) \tag{5}$$
$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \circ \tilde{\mathbf{c}}_t \tag{6}$$
$$\mathbf{h}_t = o_t \odot \tanh(\mathbf{c}_t) \tag{7}$$

At each time step t, we need to predict a tag from a vocabulary of size V. We use another linear layer to project the hidden state h_t into a vector of dimension V, followed by a softmax operator. This will give us the probability of choosing each of the V possible tags as the predicted tag at time t:

$$\mathbf{z}_t = W^{(z)} \mathbf{h}_{(t)} + \mathbf{b}^{(z)} \tag{8}$$
$$p_{t,v} = \frac{\exp(z_{t,v})}{\sum_{k=1}^{V} \exp(z_{t,k})} \tag{9}$$

where $\mathbf{z}_t \in \mathbb{R}^V$, and $p_{t,v}$ denotes the probability of picking the v-th tag in the vocabulary as the predicted tag at time t.

We demonstrate the effectiveness on the PASCAL2007 dataset and the LabelMe dataset.

[1] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.