Connectionist Temporal Classification for Group Activity Recognition

Bicheng Xu

PROBLEM

Description:

Given a sequence of video frames containing a group of people, recognize the sequence of activities that the group performs.

Input:

A sequence of video frames, length varies from 100 to 300.

Output:

The sequence of activities that the group performs. For example: Left serve, Right pass, Right set, Right spike, Right win point.

Contribution:

1. Construct a volleyball dataset for this task.
2. Use the Connectionist Temporal Classification [1] model to recognize group activities.

Connectionist Temporal Classification [1]

Temporal Classification:

S: a set of training examples

Input space X = (R^m)*: the set of all sequences of m dimensional real valued vectors.

Target space Z = L*: is the set of all sequences over the (finite) alphabet L of labels.

Each example in S consists of a pair of sequences (x, z).

The target sequence z = (z_1, z_2, ..., z_U) is at most as long as the input sequence x = (x_1, x_2, ..., x_T).

The aim is to use S to train a temporal classifier h : X -> Z to classify previously unseen input.

Label Error Rate (LER): (Used when testing.)

The normalized edit distance between its classifications and the targets.

$$LER(h, S') = \frac{1}{Z} \sum_{(\mathbf{x},\mathbf{z})\in S'} ED(h(\mathbf{x})) \tag{1}$$

S': a test set h: the temporal classifier

Z: total number of target labels in S'

ED(p, q): edit distance between the two sequence p and q

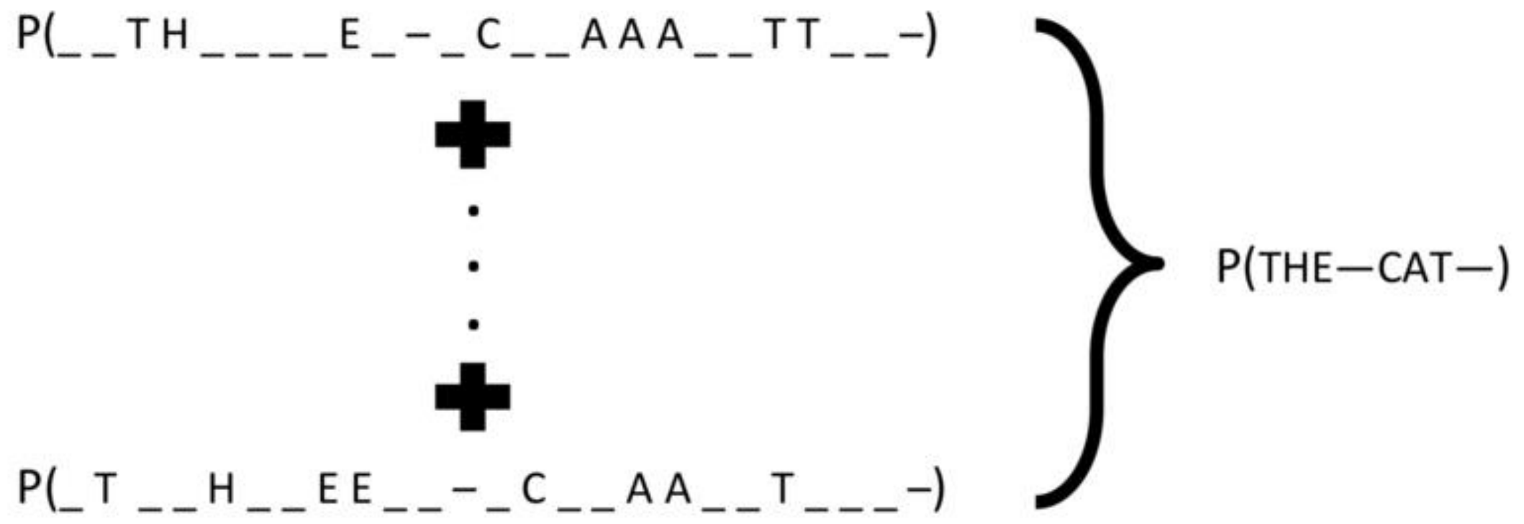
Connectionist Temporal Classification [1]

Connectionist Temporal Classification:

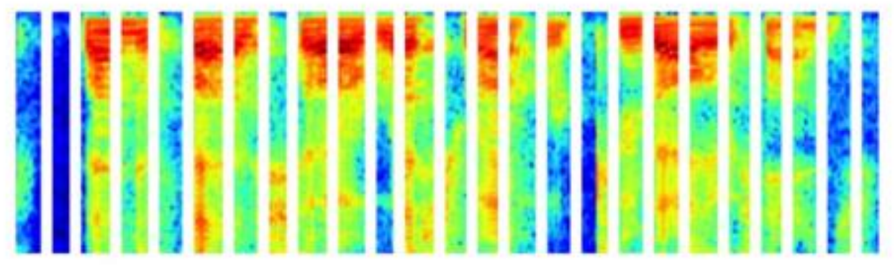
Transform the outputs of a recurrent neural network into a conditional probability distribution over label sequence.

A CTC network has a softmax output layer with one more unit than there are labels in L. The activation of the extra unit is the probability of observing a 'blank', or no label.

The total probability of any one label sequence can then be found by summing the probabilities of its different alignments. (The following picture is from [2].)


P(_ _ T H _ _ _ _ E _ – _ C _ _ A A A _ _ T T _ _ –)
P(_ T _ _ H _ _ E E _ _ – _ C _ _ A A _ _ T _ _ _ –)
P(THE—CAT—)

The output from a RNN:



The way of calculating the conditional probabilities p(l|x):

- The CTC Forward-Backward Algorithm.

Training:

- Maximum Likelihood Training.

Dataset

I manually collect the dataset by myself using the volleyball game videos available on YouTube.

The dataset contains 100 sequences with each sequence having one labeling.

The video sequence length: from 100 to 300.

The labeling length: from 2 to 14.

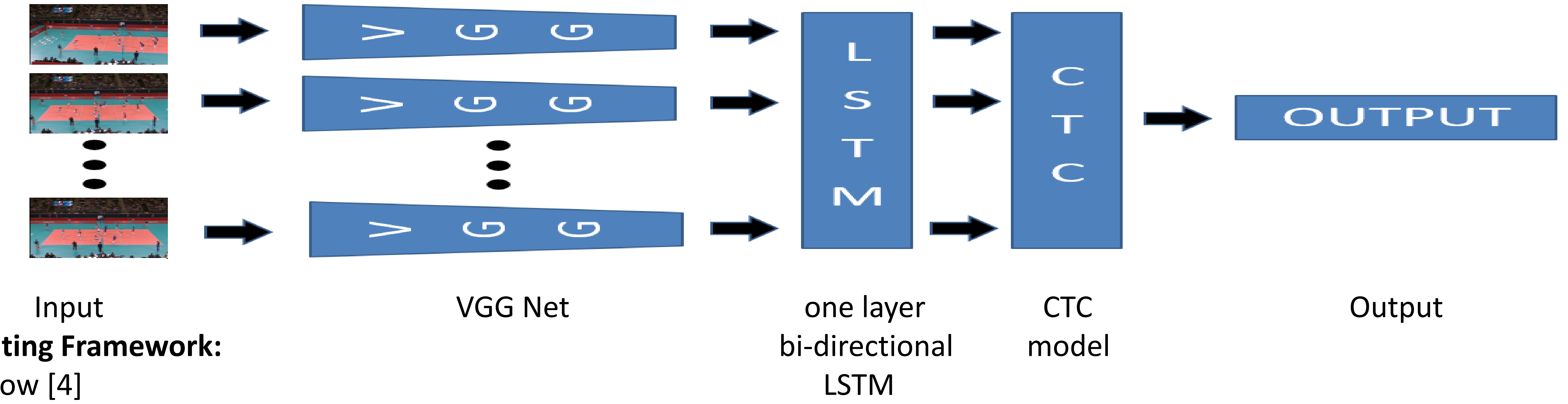
The labels: 10 in total.

Left serve
Left set
Left spike
Left pass
Left winpoint

Right serve
Right set
Right spike
Right pass
Right winpoint

Network Structure

The network consists of VGG [3] net, one layer of bi-directional Long-Short Term Memory (LSTM) recurrent neural network, and the CTC model. A sample output can be [Left serve, Right pass, Right set, Right spike, Right win point].



Implementing Framework:

- TensorFlow [4]

Reference

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *International Conference on Machine Learning (ICML)*, Pittsburgh, USA, 2006.

[2] https://github.com/baidu-research/warp-ctc

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] https://www.tensorflow.org/