Summary/Extended Abstract

Aman Bhansali
B20ME010
bhansali.1@iitj.ac.in

Indian Institute of Technology Jodhpur

Introduction

This paper is based on uncovering some efficient techniques and applying some new combination of algorithms of machine learning so that computers recognize the hand written content accurately. The model made is such that it takes just the raw pixel data and also no language/input based preprocessing, presegmentation or output's post-processing is required.

The applied algorithms are multidimensional recurrent neural networks, multidimensional long short-term memory, connectionist temporal classification output layer and network hierarchy. Each of the above mentioned algorithms are used in an optimized order thus making the model more robust even towards warping data.

The idea behind using recurrent neural networks is that for a sequential data it can easily store information of the previous inputs if important in their memory which may help for further predictions. Since we are dealing with multidimensional data we require multidimensional RNN.
Long short-term memory is being used because a normal RNN while running sequentially cannot store much information of previous important inputs for long textual data. Thus, using this helps storing all the information and can also forget the data not important after a point.
Connectionist Temporal Classification (CTC) is used for sequence problems, helps labelling unsegmented sequence data.
Network Hierarchy is used to obtain more complex features from the smaller and simple local features.

Methodologies

Multidimensional recurrent neural network: It is to replace the single recurrent connection of RNN with connections as per equal to the dimensions of the data. During each forward pass the hidden layer receives a present input and an activation from the previous input for all dimensions. Now it may happen that the order in which this is done does not give the best result. For this we use multi-directionality which means that for n dimensional data we have 2^n possible ways. All the hidden layers made of these combinations are connected to a single output layer, which receives information about all.

Multidimensional LSTM:This contains block which contain internal units called cells. Their activation is controlled by input gate, output gate and forget gate. The forget gate has a main role in either storing or deleting an information as its output decides this. These gates help store information longer. Just adding n connections would make it n-dimensional. This would be jointly used under MDRNN.

Connectionist Temporal Classification (CTC): This is basically a cost function based on probability distribution. A CTC uses soft-max activation function in output layer and the layer contains units = no. of labels(L) + 1. Thus, the activation of the L units give the probabilities of the labels at that time while the extra unit is for detecting no label. These values of probabilities show how much our labels are aligning for all cases with the particular input sequence. The total probability of any one label sequence can be calculated by adding the probabilities of its different alignments. It's important to know that CTC can accept only 1D sequence as its input.

The hierarchical structure of the model created was:

1. The input image is divided into small blocks(say a 4x3 block is reduced to a length 12 vector) and passed through the set of MDLSTM layers. Padding could be done if the image does not divide in blocks as per our dimensions.
2. Since we have a 2D image which means using directionality part we would have 4 layers for 4 different scans.
3. Activation of the layers are collected into blocks.
4. These blocks now serve as input for the forward network. These blocks collects the information of the textual part and also helps reducing dimension.
5. Now repeat the same steps each time until the dimensionality reduces to 1D so that CTC could be applied. Layers play an important role in reducing dimensions too.

Experiment Performed

The model was then participated in ICDAR 2007 Arabic handwriting recognition competition. For the competition we needed to identify the postcodes of Tunisian town and village names which were presented in a word by word format. Each image was supplied with the true value and there were in total 120 different characters. The test data were of two kinds: 'f' and 's'. The 'f' was the data of final evaluation while 's' contained the data which was same just some regional writing variations.

Parameters

Inverted pyramid kind of structures are used were there are more features at the upper layers while the bottom layers have small no. of features. This reduces the no. of weight operations. The activation functions used were: logistic and tanh. The model was trained with online gradient descent, using a learning rate of 104 and a momentum of 0.9. The error rates are calculated on each iterations and once the stage when no further optimization is possible is achieved then the model stops and the output is displayed.

Results

	SET f			SET s		
SYSTEM	top 1	top 5	top 10	top 1	top 5	top 10
CACI-3	14.28	29.88	37.91	10.68	21.74	30.20
CACI-2	15.79	21.34	22.33	14.24	19.39	20.53
CEDAR	59.01	78.76	83.70	41.32	61.98	69.87
MITRE	61.70	81.61	85.69	49.91	70.50	76.48
UOB-ENST-1	79.10	87.69	90.21	64.97	78.39	82.20
PARIS V	80.18	91.09	92.98	64.38	78.12	82.13
ICRA	81.47	90.07	92.15	72.22	82.84	86.27
UOB-ENST-2	81.65	90.81	92.35	69.61	83.79	85.89
UOB-ENST-4	81.81	88.71	90.40	70.57	79.85	83.34
UOB-ENST-3	81.93	91.20	92.76	69.93	84.11	87.03
SIEMENS-1	82.77	92.37	93.92	68.09	81.70	85.19
MIE	83.34	91.67	93.48	68.40	80.93	83.73
SIEMENS-2	87.22	94.05	95.42	73.94	85.44	88.18
Ours	**91.43**	**96.12**	**96.75**	**78.83**	**88.00**	**91.05**

Discussions

As the size of the input image is not fixed, so we can't choose just a particular block height that gives a 1D sequence so as to pass through CTC. To resolve these issues we need to just sum the inputs for each vertical lines thus making it 1D.
Also dimensionality could be changed as per data to make it work for many more systems.

Conclusions

Since, the complete model is framed such that it does not need to be pre-processed for applying on any language. Also the model becomes robust to distorted data.

References

https://arxiv.org/pdf/0705.2011.pdf

https://www.cs.toronto.edu/~graves/icml_2006.pdf