News Classification Based On DeepSpeech 2 And DeepCNN

Fang Zhu Zixiang Nie Ziyuan Guan

BIG DATA 2018 T2S

2018 Spring Big Data Team InSaiyans Guided By Dr.Andy Lin

Introduction

- Automatic Speech Recognition And Classification
- Significant For Deaf and Introvert People
- DeepSpeech 2 & Classification With Warp CTC
- DeepRNNs For Speech & DeepCNNs for classes
- Evaluation With AN4 and Librispeech Datasets
- Test With Librispeech Model and AGnews get 48% accuracy

Automatic Speech Recognition

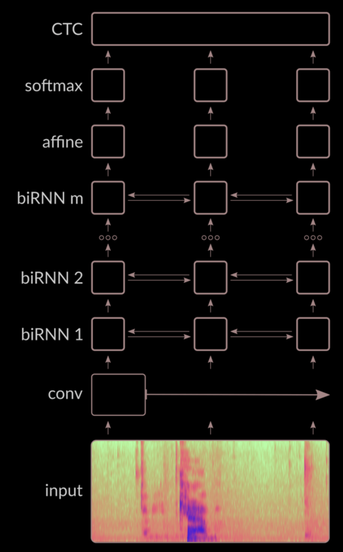
- Based On Deep Speech 2 Model
- Firstly Process the Input Audio With Frequency Convolutions In The first layer
  It can slightly improve the ASR Performance
- Then Put the Sequence Into Bi-Directional RNNs with GRU or LSTM

$$\overrightarrow{h}_t^l = f(\mathcal{B}(W^l h_t^{l-1}) + \overrightarrow{U}^l \overrightarrow{h}_{t-1}^l)$$

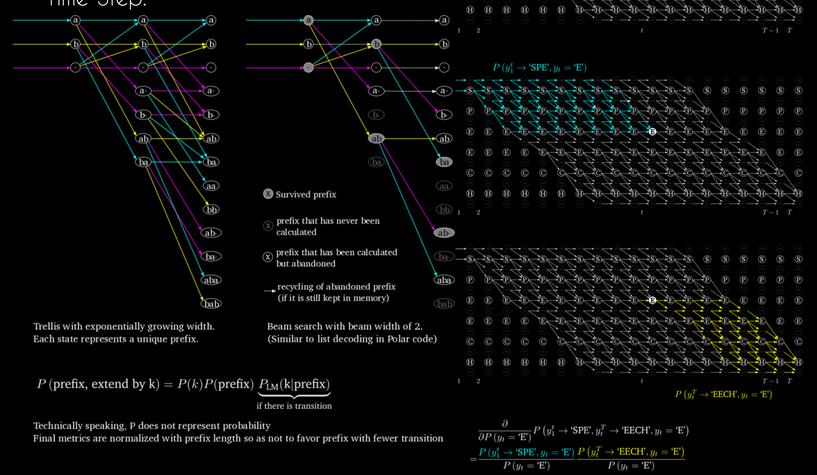
$$\overrightarrow{h}_t^l = f(W^l h_t^{l-1} + \overrightarrow{U}^l \overrightarrow{h}_{t-1}^l + b^l)$$

- We use the Batch-Normalization To Accelerate training For Multi-Layer Network.(Increasing Depth would reduce the performance) To Improve Final Generation Error

$$\overrightarrow{h}_t^l = f(\mathcal{B}(W^l h_t^{l-1}) + \overrightarrow{U}^l \overrightarrow{h}_{t-1}^l)$$



- Finally We used the Row Convolution With Baidu Warp_CTC to speech recognition. It need only little future Information To Accurate Prediction at The Current Time-Step.



- Convert The Voice To Text Scripts And Stored

Natural Language Processing

- Presentation Layer
- Convolutional Layer
- Max-Over-Time Pooling
- Fully Connected Layer

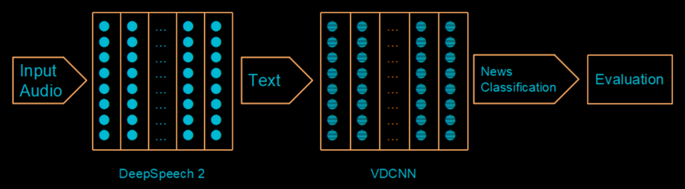
- Weights Initialization & Word Embedding:
  Utilize the pre-trained model from google news to help initialize the weight.
  Using "nn.embedding"from torch.nn api to convert word from the text into word vector as inputs to the network.
- CNNs Convolution Operations:

w: a filter; x: concatenation words; b: bias term; f: non-linear function (ReLU)

- Max Pooling: $\hat{c} = \max\{\mathbf{c}\}$
  Down-sampling
  Won't damage the recognition result
  Keep The Important Features
- Fully Connected (SoftMax,Cross Entropy and Dropout)
  Computing the gap between the sample and the label
  Very computationally convenient Regularization
  (Reduce Overfitting)

$$L_i = -log(\frac{e^{f_{y_i}}}{\sum_j e^j}) \qquad y = \mathbf{w} \cdot (\mathbf{z} \circ \mathbf{r}) + b$$

Experiments



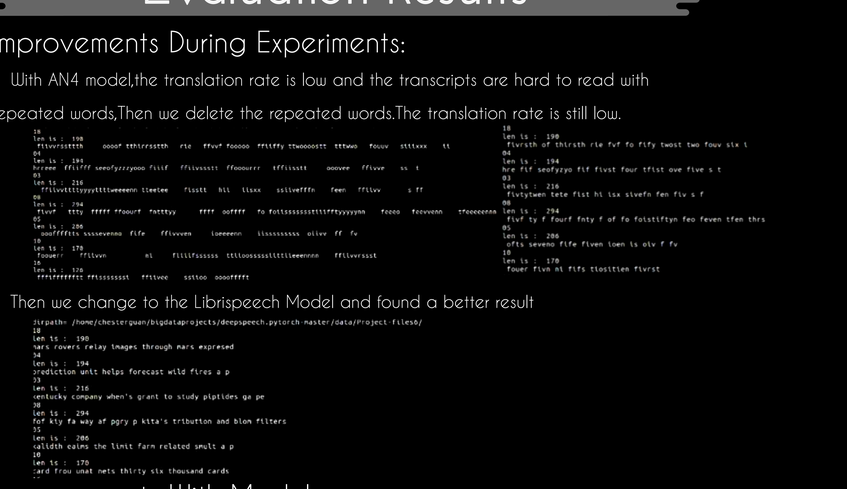
- The Experiments Processed By the Upper Architecture
- The Whole Part Of Experiment Could Be Seen As Two Parts
- The Audio Data-Sets: AN4 ; Librispeech ;Tedlium
  AN4: Focuses On Numbers And Alphabets
  Training Size: 50 minutes Testing Size: 8 minutes
  Librispeech: Focuses On Normal English Speech
  Training Size: 1000 hours Testing Size: 20 hours
  Tedlium: Real TED Speech On Special Topics
  Training Size: 216 hours Testing Size: 4 hours
- The News Data-Sets: AGnews,Sogou News,Yahoo answers
  AGnews: 4 classes
  Sogou News: 5 classes
  Yahoo News: 10 classes
- Since The Experiments Done Separated,The Total Result Will Be Effected Negatively
- Experiment Example:
  Input Voice: "Tomorrow There is a NBA game."
  Voice Recognition: "tomorow there is a nba game"
  Classification: "Sports"

Evaluation Results

- Improvements During Experiments:

With AN4 model,the translation rate is low and the transcripts are hard to read with repeated words,Then we delete the repeated words.The translation rate is still low.

Then we change to the Librispeech Model and found a better result



- Improvements With Models:

Hidden Layers	Hidden Layer Units	WER	CER
5	200	78.54	30.16
5	400	68.43	28.47
5	600	56.936	26.250
5	800	41.32	22.53

Hidden Layers With 600 Units Each	Epoche	WER	CER
5	5	57.324	25.321
5	10	48.304	22.452
5	30	36.532	16.783
5	70	10.689	4.942

- Speech Model Evaluation:

Dataset	WER	CER
AN4 test	10.58	4.88
Librispeech test clean	10.239	2.965
Librispeech test other	28.008	9.791
TED test	31.04	10.00

- Text Model Evaluation:

	AG News	Sogou News	Yahoo anwsers
Training Rate	98.30	97.00	95.16
Validate Rate	85.18	92.45	76.54
Test Rate	67.78	20.99	51.67
Number of class	4	5	10

- Combination Model Evaluation:

Times	Accuracy
1	30.00
2	48.00

Analysis & Conclusion

- Since The Two Models Are Tested Individually,The Total Result Was Effected Negatively
- The Deepspeech 2 Model Applied The Greedy Searching.Beam Search May Get Better Results
- The Experiments Applied The Clean Voice To Test,We also Need To Consider The Noise.
- Since The News Contain Key Words,We May Try With "Attention" To Get A Better Feature Map.
- Basically,The Whole Model Could Classify The Audio News

ziyuan.guan@ufl.edu zhufang@ufl.edu znie@ufl.edu ,Gainesville, Florida, University Of Florida