**Week 6-2: Paper Summaries**

***CE-510 Seminar: Social Media Mining***

Student Name: Jiaqi Guo NetID: JGR9647

* **Deep Speech: Scaling up end-to-end speech recognition**

The author abandons the traditional speech recognition model framework and turns to the end-to-end speech recognition model framework based on deep learning:

1. Discard the original feature engineering, such as MFCC.

2. Discard phoneme dictionaries and the concept of phonemes.

3. Use RNN network

In general, the model structure consists of two parts:

1. AM (Acoustic Model): Bi-directional RNN (loss function CTC, Optimizer NAG, Nesterov Accelerated Gradient)

The model has **five hidden layers**, none of which is a recurrent layer except the fourth, which is bi-Directional layer.

The calculation formula of the first three layers is as follows:

For the 4th layer:

5th layer:

The output is a standard SoftMax:

Where, represents the probability that the model predicts that the letter corresponding to the speech in frame will be the letter in the dictionary.

1. LM (Language Model): N-gram, which is trained with KenLM.

**Possible Improvement Directions:**

1. The assumption that CTC is conditional independent is too strong to be true, so language models are needed to improve conditional dependence to achieve better results

* **Sequence to Sequence Learning with Neural Networks**

Although DNNs performs well on large-scale labeled data sets, it cannot solve the SEQ2SEQ (sequence to sequence) problem. In this paper, the author proposed an end-to-end approach that makes minimal assumptions about sequence structure, using multi-level LSTM to map input sequences to fixed-dimensional vectors.

The LSTM model implemented in this paper is illustrated below:

图表, 箱线图

描述已自动生成

The goal of LSTM is to optimize the conditional probabilities p:

Here, is the input sequence, and is the corresponding output sequence. LTSM calculates this conditional probability by first obtaining a fixed dimensional vector representing of the input sequence (derived from the last hidden state) and then calculating the probability of by a standard LSTM-LM formula.

**Possible Improvement Directions:**

1. The model is insensitive to the active and passive voice of a sentence.