Connectionist Temporal Classification with Prefix Beam Search Decoding

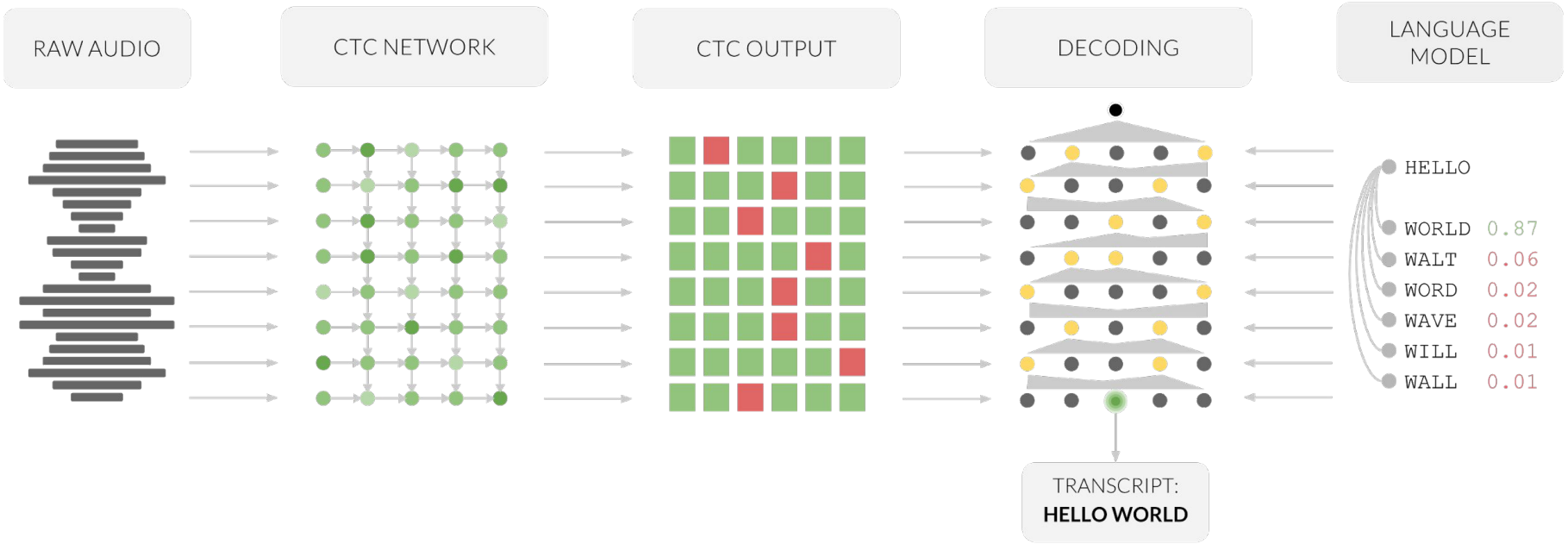
Benjamin Geyer
bgeyer3@masonlive.gmu.edu

The Problem

Automatic Speech Recognition

- Alignment in a continuous sequence
- Connectionist Temporal Classification (CTC)

CTC Pipeline



Decoding

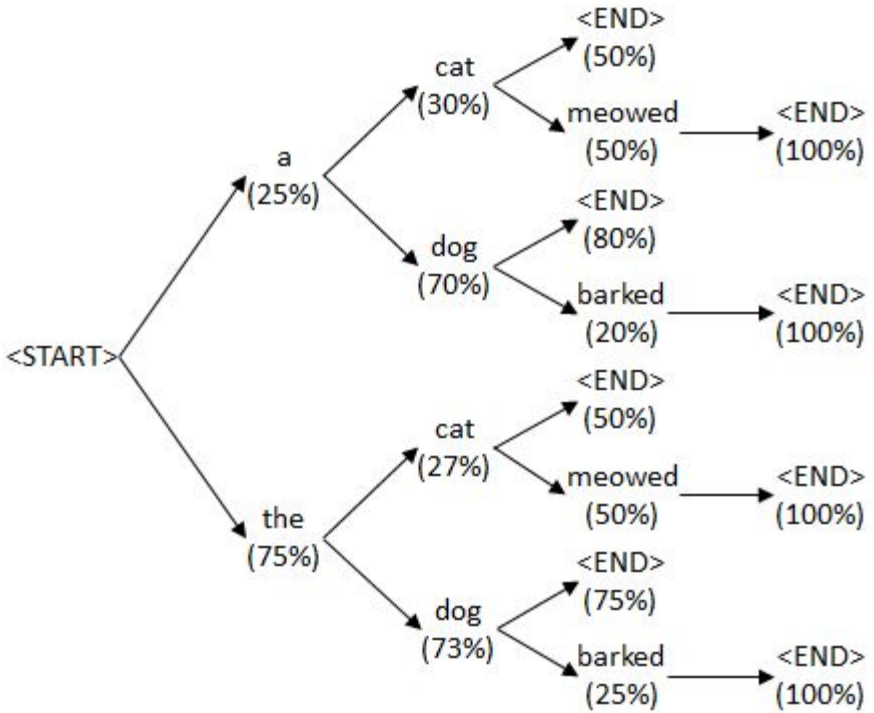
- Alignment in a continuous sequence

The Experiments

- Prefix Beam Search Decoding
  - Language Model
- Large Scale Model (Mozilla Deep Speech 2)

Prefix Beam Search

- Language Model Probability
- Beams
- Pruning



Deep Speech 2

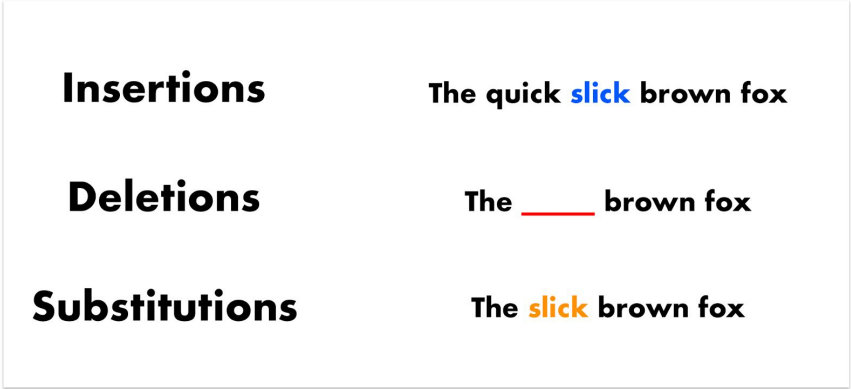
```
Model: "DeepSpeech2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
X (InputLayer)               [(None, None, 160)]       0
_________________________________________________________________
lambda (Lambda)              (None, None, 160, 1)      0
_________________________________________________________________
conv_1 (Conv2D)              (None, None, 80, 32)      14432
_________________________________________________________________
conv_1_bn (BatchNormalizatio (None, None, 80, 32)      128
_________________________________________________________________
conv_1_relu (ReLU)           (None, None, 80, 32)      0
_________________________________________________________________
conv_2 (Conv2D)              (None, None, 40, 32)      236544
_________________________________________________________________
conv_2_bn (BatchNormalizatio (None, None, 40, 32)      128
_________________________________________________________________
conv_2_relu (ReLU)           (None, None, 40, 32)      0
_________________________________________________________________
reshape (Reshape)            (None, None, 1280)        0
_________________________________________________________________
bidirectional_1 (Bidirection (None, None, 1600)        9993600
_________________________________________________________________
dropout (Dropout)            (None, None, 1600)        0
_________________________________________________________________
bidirectional_2 (Bidirection (None, None, 1600)        11529600
_________________________________________________________________
dropout_1 (Dropout)          (None, None, 1600)        0
_________________________________________________________________
bidirectional_3 (Bidirection (None, None, 1600)        11529600
_________________________________________________________________
dropout_2 (Dropout)          (None, None, 1600)        0
_________________________________________________________________
bidirectional_4 (Bidirection (None, None, 1600)        11529600
_________________________________________________________________
dropout_3 (Dropout)          (None, None, 1600)        0
_________________________________________________________________
bidirectional_5 (Bidirection (None, None, 1600)        11529600
_________________________________________________________________
dense_1 (TimeDistributed)    (None, None, 1600)        2561600
_________________________________________________________________
dense_1_relu (ReLU)          (None, None, 1600)        0
_________________________________________________________________
dropout_4 (Dropout)          (None, None, 1600)        0
_________________________________________________________________
dense_2 (TimeDistributed)    (None, None, 29)          46429
=================================================================
Total params: 58,971,261
Trainable params: 58,971,133
Non-trainable params: 128
_________________________________________________________________
```

Dataset

- LibriSpeech 1000 Hours of Audio Books

Metrics

- Character Error Rate (CER)
- Word Error Rate (WER)



$$WER = \frac{S + D + I}{N}$$

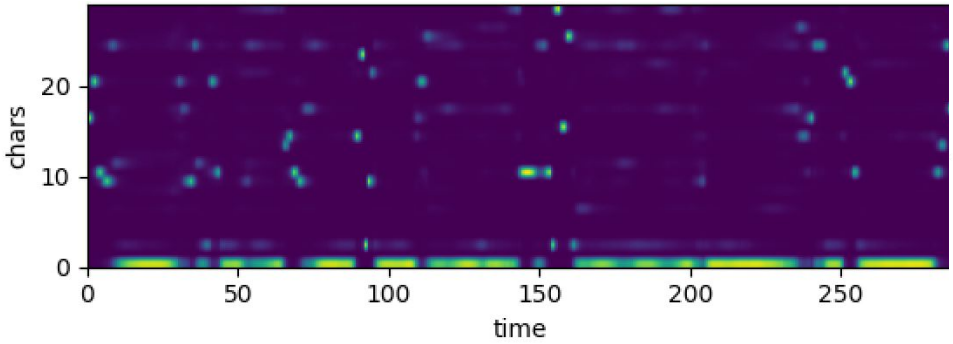
where...
S = number of substitutions
D = number of deletions
I = number of insertions
N = number of words in the reference

Decoding Algorithm Results

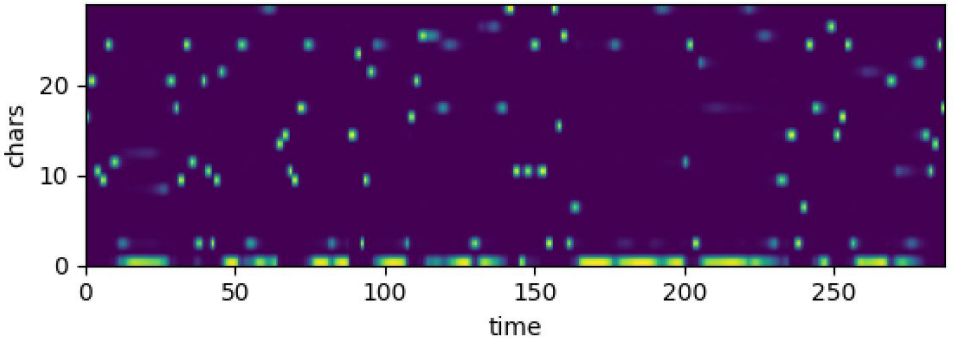
Model	50 Epoch %CER	75 Epoch %CER
Prefix Beam Search	29.2	0.0
Tensorflow Beam Search	37.1	2.2
Harald Scheidl Beam Search	38.2	2.2
Greedy Decoding	66.3	49.6

Decoding Algorithm Results

50 Epochs



75 Epochs



Deep Speech 2 Results

250 Epochs

Model	%CER	%WER
Prefix w/ LM	48.6	67.2
Prefix No LM	54.8	74.0
Greedy	65.2	88.6

Thank You!