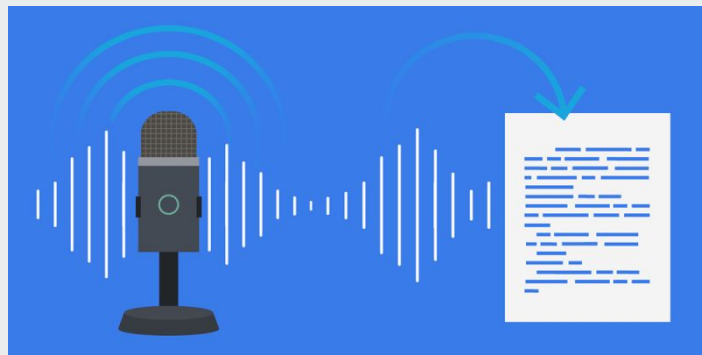


# Connectionist Temporal Classification with Prefix Beam Search Decoding

Benjamin Geyer  
bgeyer3@masonlive.gmu.edu



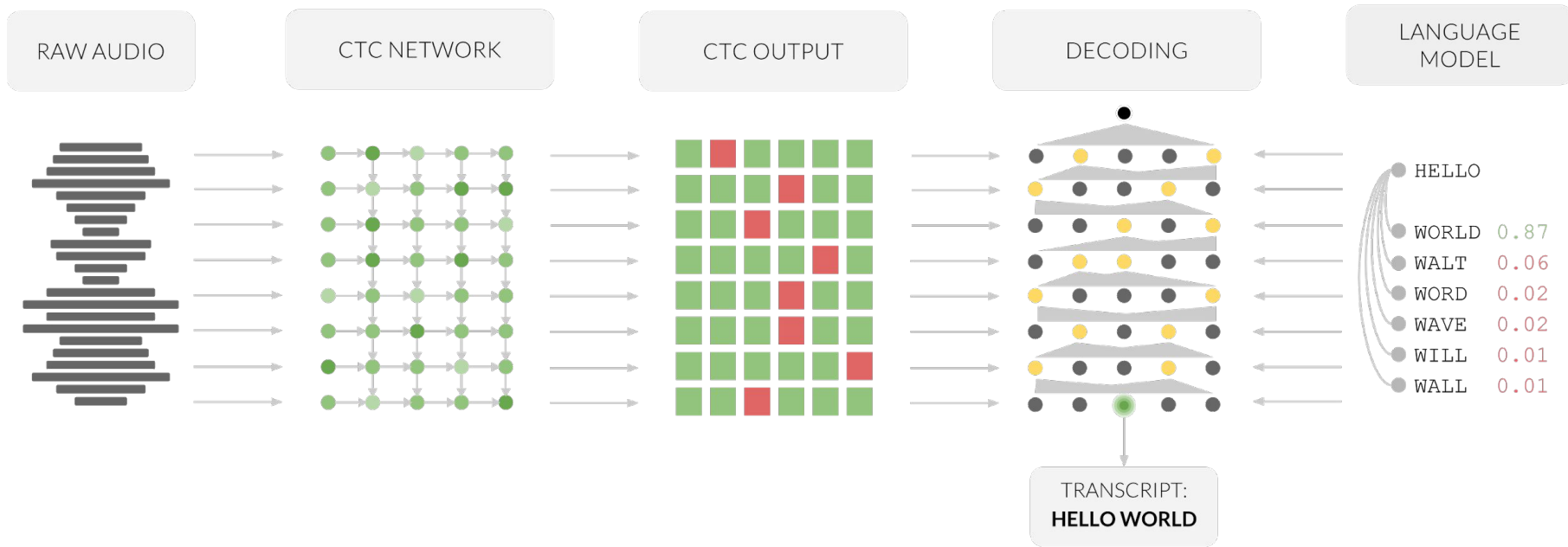


# The Problem

Automatic Speech Recognition

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

# CTC Pipeline



# Decoding

- Alignment in a continuous sequence



We start with an input sequence, like a spectrogram of audio.



The input is fed into an RNN, for example.

h	h	h	h	h	h	h	h	h	h
e	e	e	e	e	e	e	e	e	e
l	l	l	l	l	l	l	l	l	l
o	o	o	o	o	o	o	o	o	o
€	€	€	€	€	€	€	€	€	€

The network gives  $p_t(a | \mathcal{X})$ , a distribution over the outputs  $\{h, e, l, o, \epsilon\}$  for each input step.

h	e	€	l	l	€	l	l	o	o
h	h	e	l	l	€	€	l	€	o
€	e	€	l	l	€	€	l	o	o

With the per time-step output distribution, we compute the probability of different sequences

h	e	l	l	o
e	l	l	o	
h	e	l	o	

By marginalizing over alignments, we get a distribution over outputs.

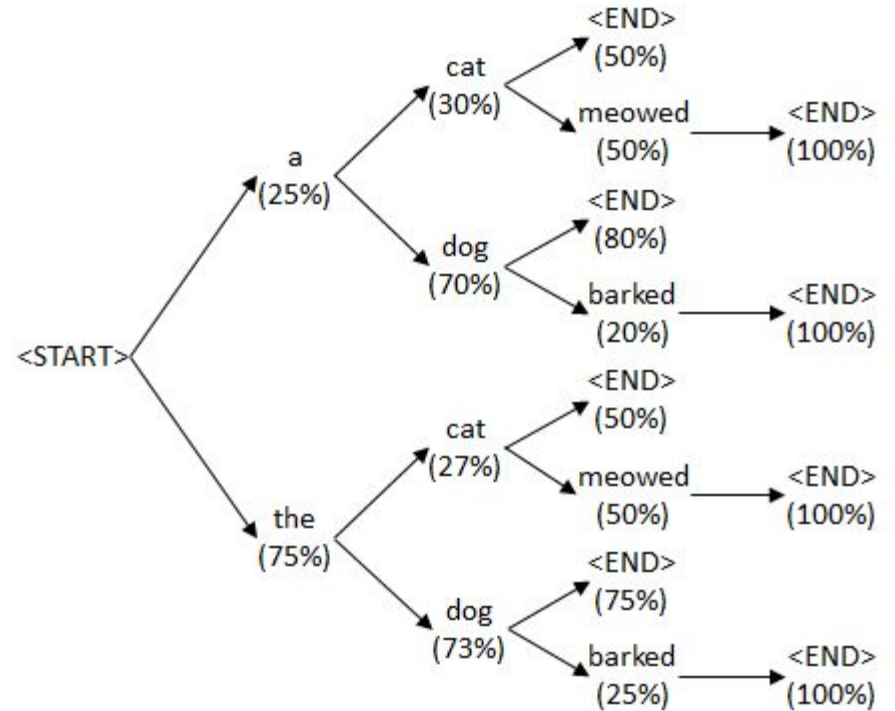


# The Experiments

- 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)

## Insertions

The quick **slick** brown fox

## Deletions

The **\_\_\_\_\_** brown fox

## Substitutions

The **slick** brown fox

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

where...

S = number of substitutions

D = number of deletions

I = number of insertions

N = number of words in the reference

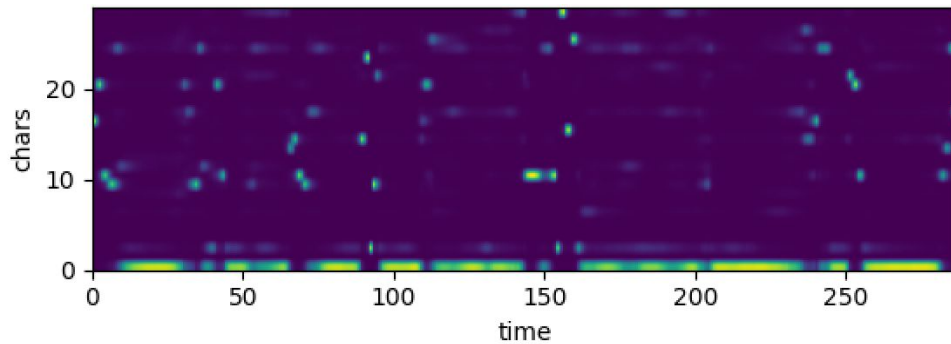


## Decoding Algorithm Results

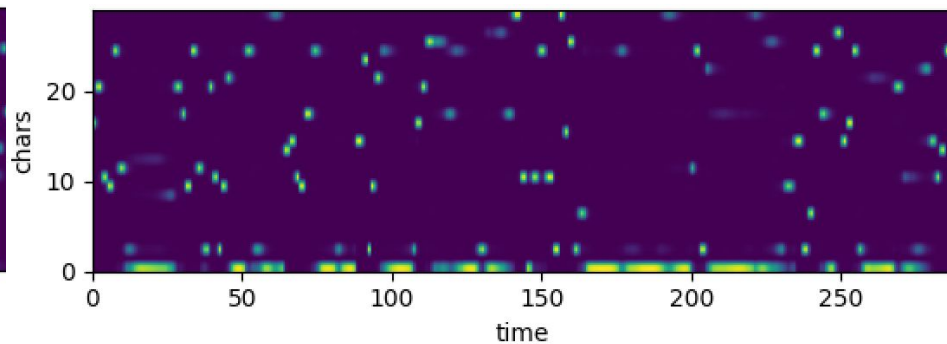
Model	50 Epoch %CER	75 Epoch %CER
Prefix Beam Search	29.2	0.0
Tensorflow Beam Search	37.1	2.2
Harald Scheidl Beam Search	38.2	2.2
Greedy Decoding	66.3	49.6

# 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!**