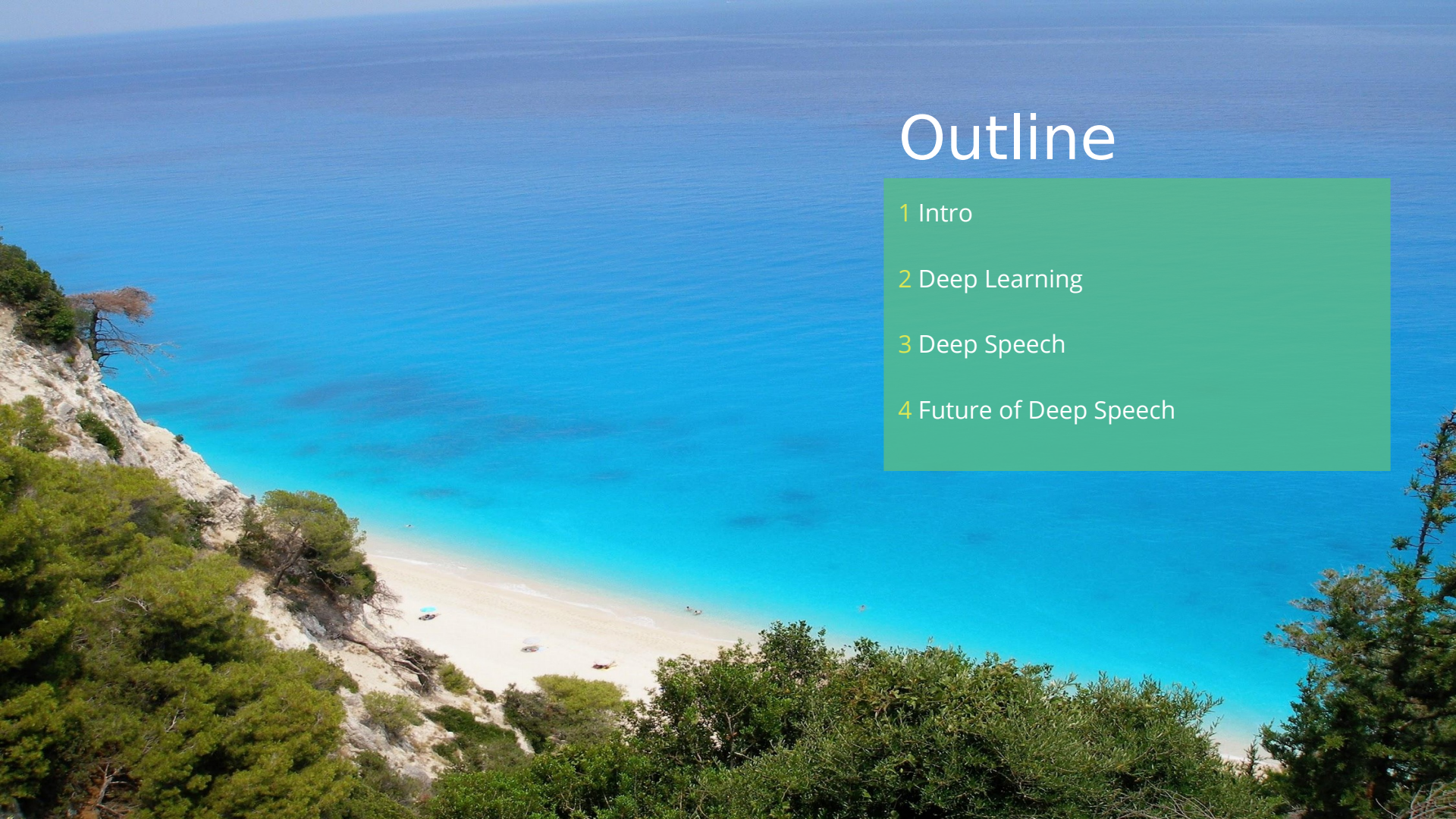
Mozilla’s Deep Speech



Outline

1 Intro

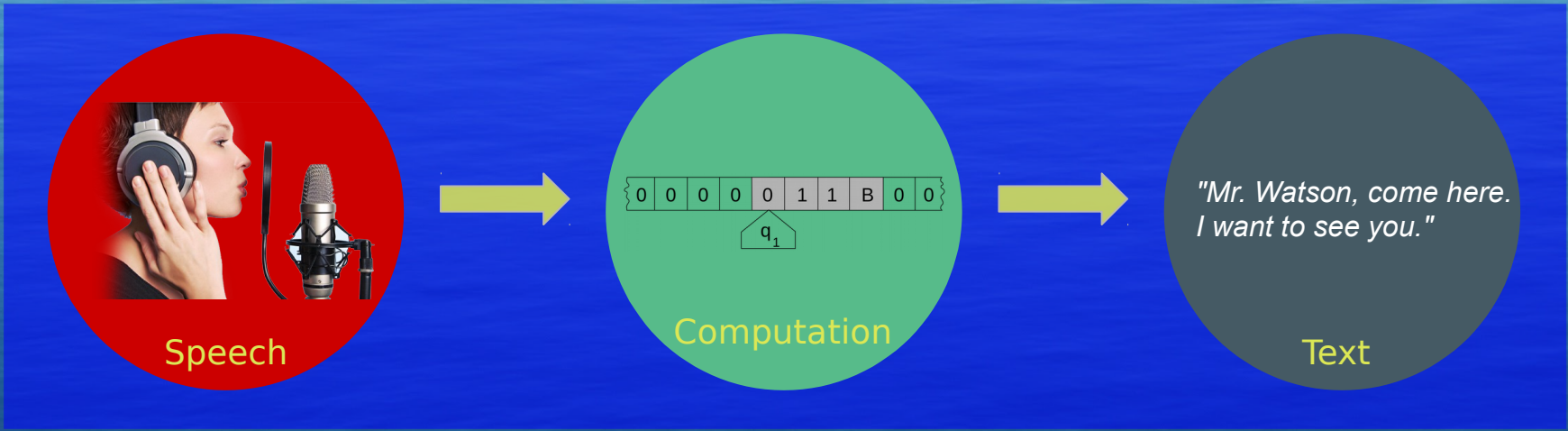
2 Deep Learning

3 Deep Speech

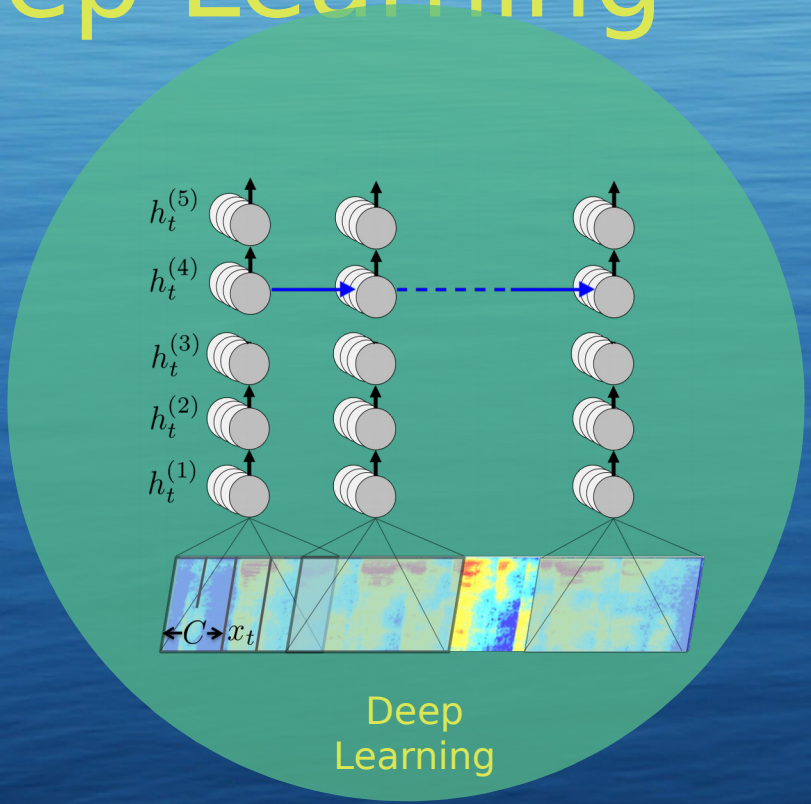
4 Future of Deep Speech

1 Intro

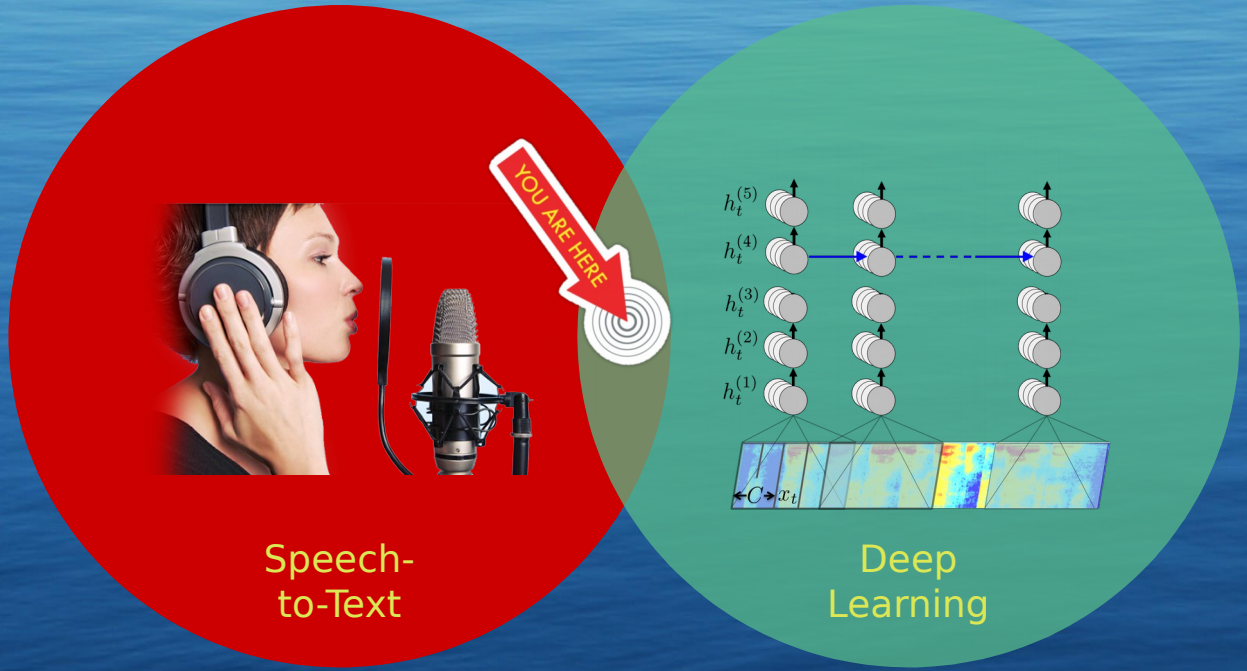
Intro: Speech-to-Text



Intro: Deep Learning

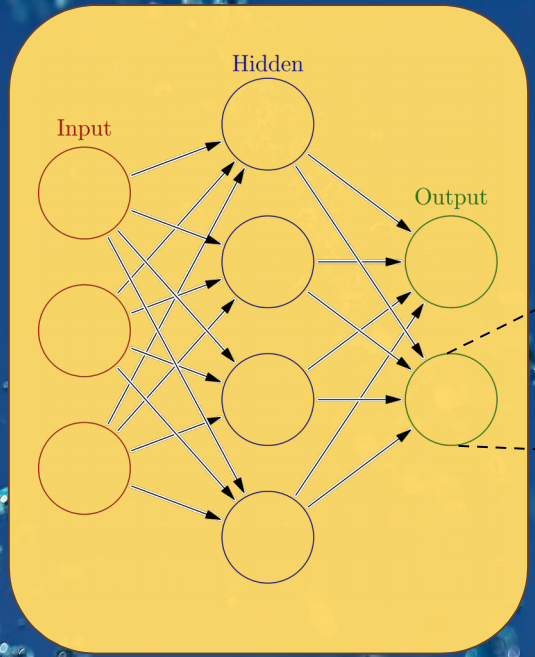


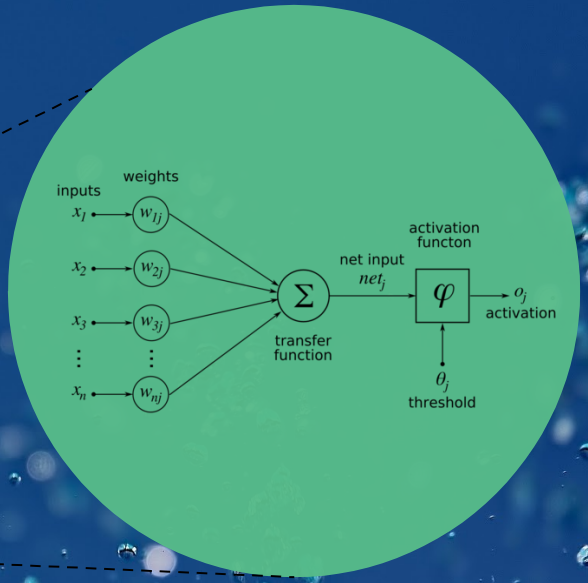
Intro: Speech-to-Text ∩ Deep Learning



2 Deep Learning

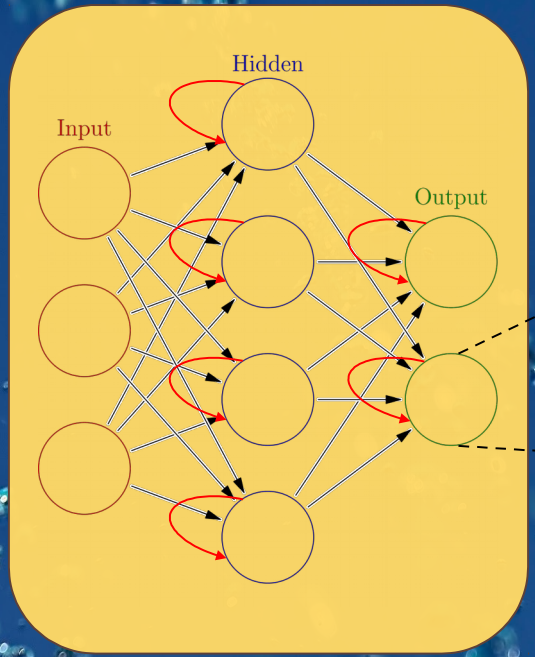
Deep Learning: Feed Forward Networks

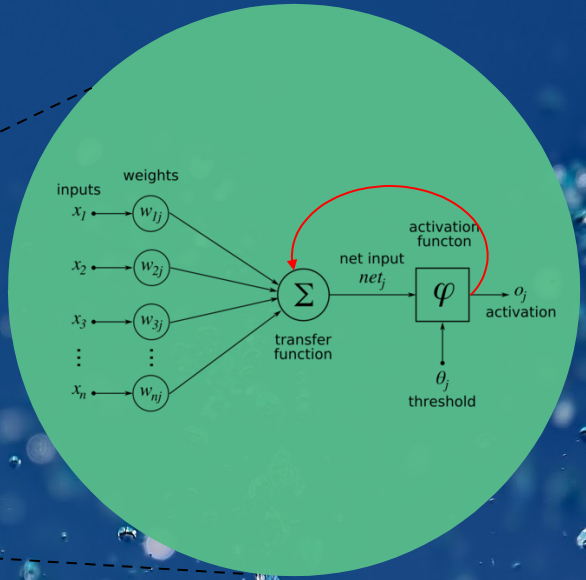




$$o_j = \varphi\left(\theta_j + \sum_{i=1}^{n} w_{ij}x_i\right)$$

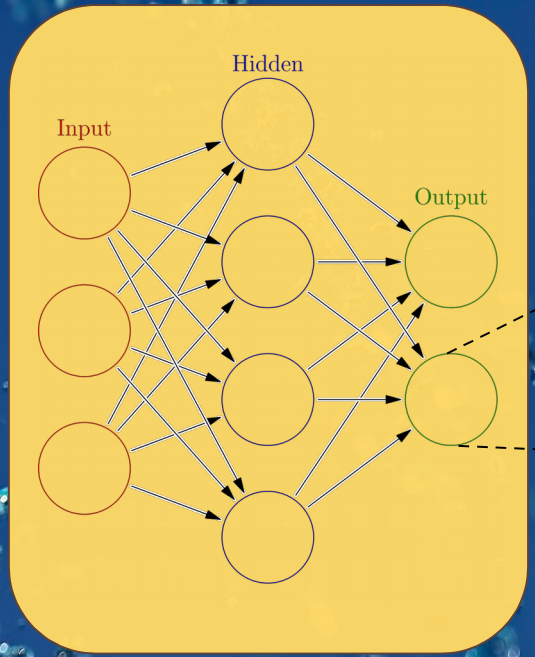
Deep Learning: Recurrent Neural Networks

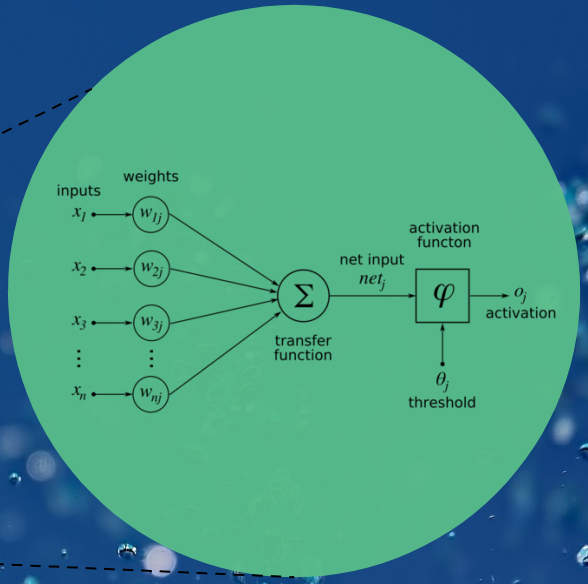




$$ {}^{t}o_j = \varphi\left(\theta_j + \sum_{i=1}^{m} {}^{t-1}o_i w_{ij}^{f} + \sum_{i=1}^{n} x_i w_{ij}\right) $$

Deep Learning: Softmax



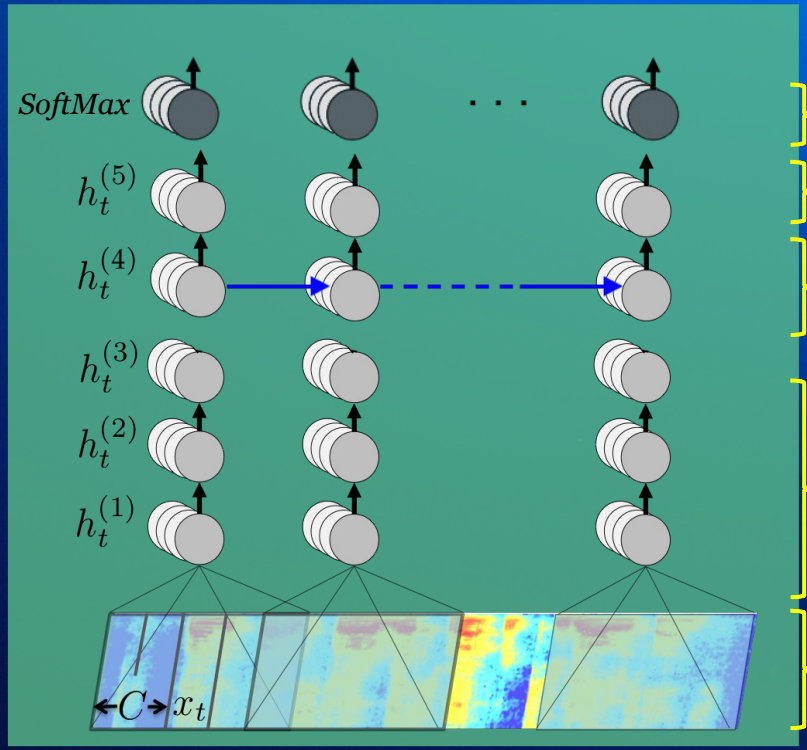


$$softmax(z_i) = \frac{exp(z_i)}{\sum_j exp(z_j)}$$

3 Deep Speech

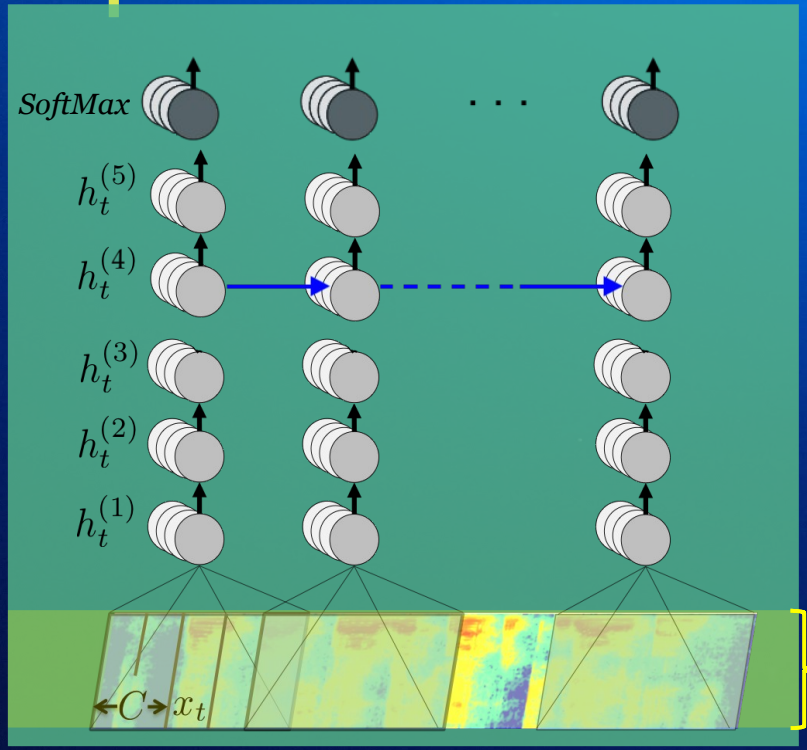
3.1 Deep Speech Architecture

Deep Speech Architecture: Overview



- Softmax Layer
- Feedforward Layer
- Recurrent Layer
- Feedforward Layers
- Input Features

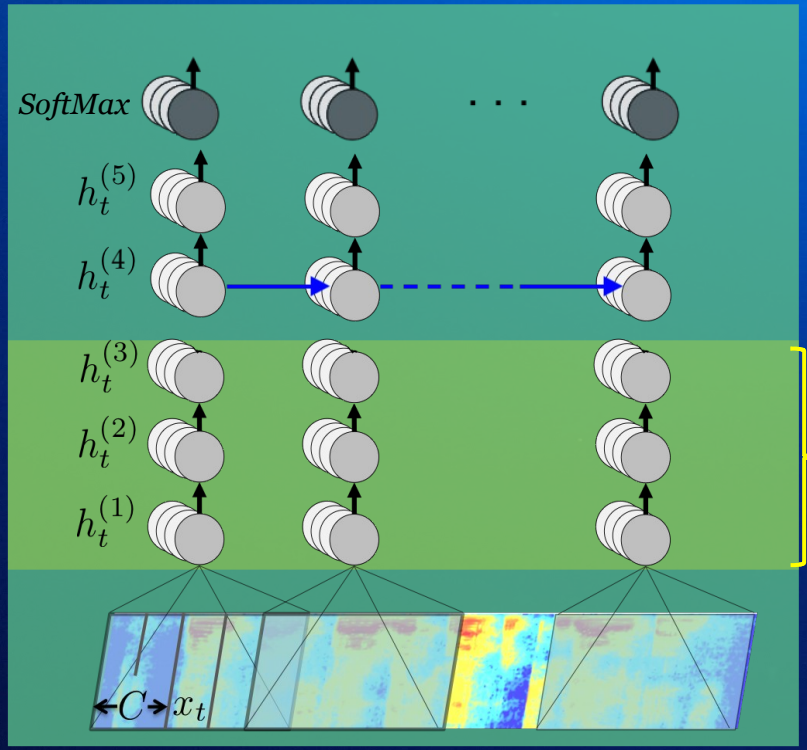
Deep Speech Architecture: Input Features



Mel-Frequency Cepstrum Coefficients

- 16 bit audio input at 16kHz
- 32ms audio window every 20ms
- 26 Cepstral Coefficients
- Context window width 9

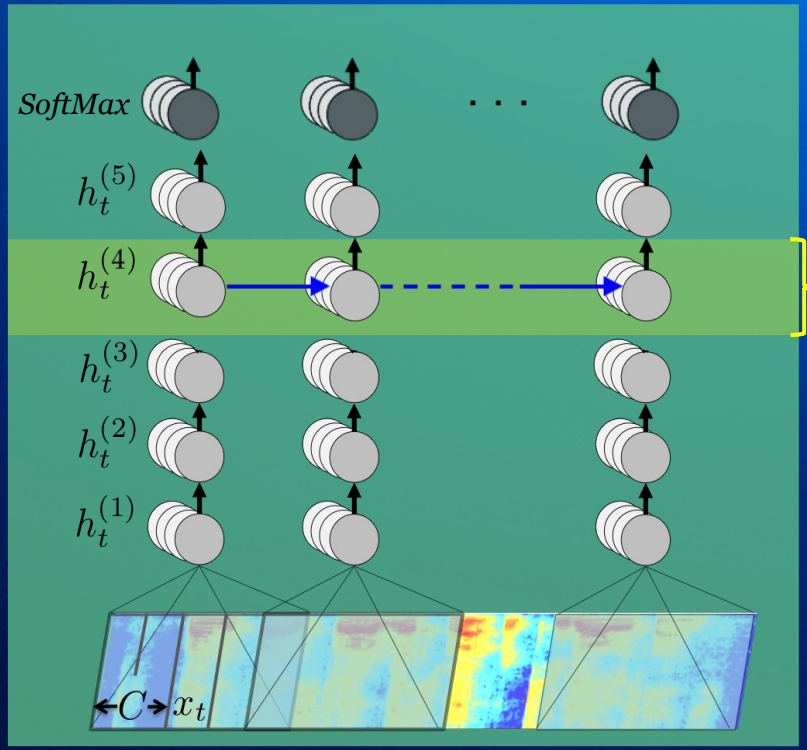
Deep Speech Architecture: Feedforward Layers



Feedforward Layers

- 3 layers
- Layer width 2048
- RELU cells
- RELU clipped at 20
- Dropout 0.20 to 0.30

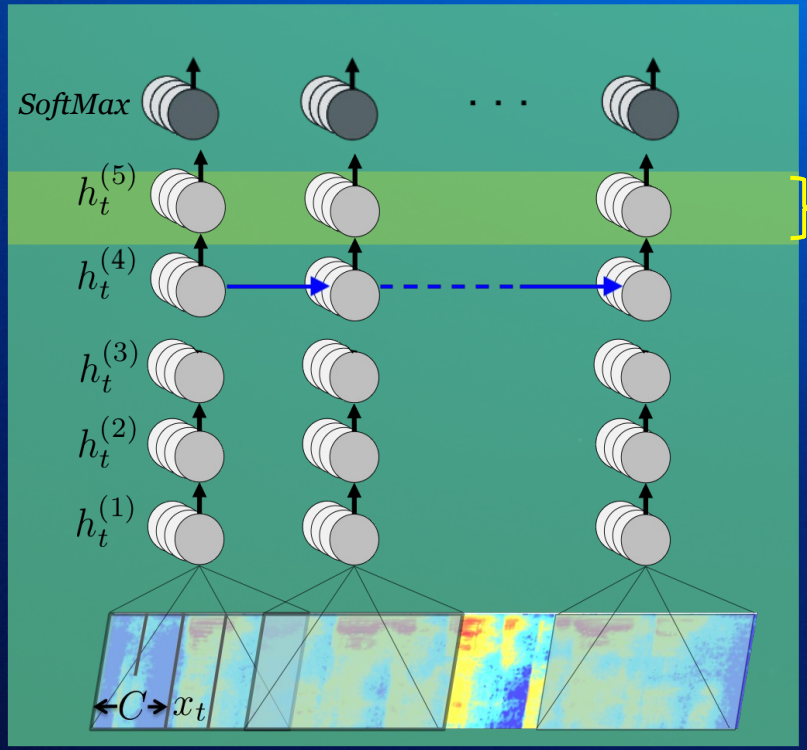
Deep Speech Architecture:
Bidirectional RNN Layer



Recurrent Layer

- 1 layer
- Layer width 2048
- LSTM cells
- No clipping
- Dropout 0.20 to 0.30

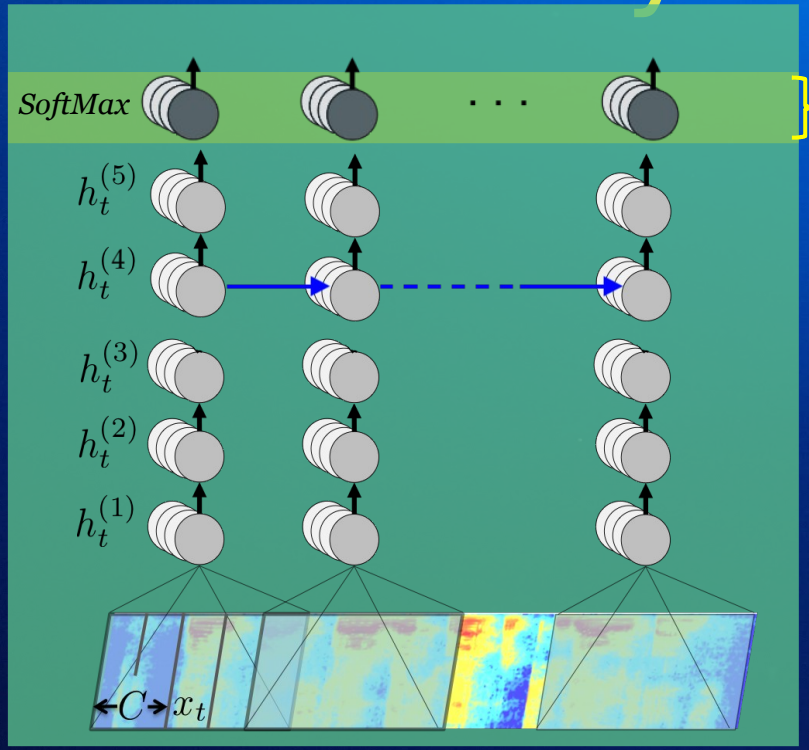
Deep Speech Architecture: Feedforward Layer



Feedforward Layer

- 1 layer
- Layer width 2048
- RELU cells
- RELU clipped at 20
- Dropout 0.20 to 0.30

Deep Speech Architecture:
Softmax Layer



Softmax Layer

- L ≡ Alphabet
- Output width k ≡ |L| + 1
- Extra for a “blank label”

3.2 Performance

Performance: WER

Training Data

- Fisher (2000 hours)
- Switchboard (240 hours)
- Librivox (1000 hours)
- Common Voice (600 hours)

Performance: WER

Training Data

- Fisher (2000 hours)
- Switchboard (240 hours)
- Librivox (1000 hours)
- Common Voice (600 hours)

On Librivox clean test 8.26% WER

Performance: WER

Human
5.8%
WER

Training Data

- Fisher (2000 hours)
- Switchboard (240 hours)
- Librivox (1000 hours)
- Common Voice (600 hours)

On Librivox clean test 8.26% WER

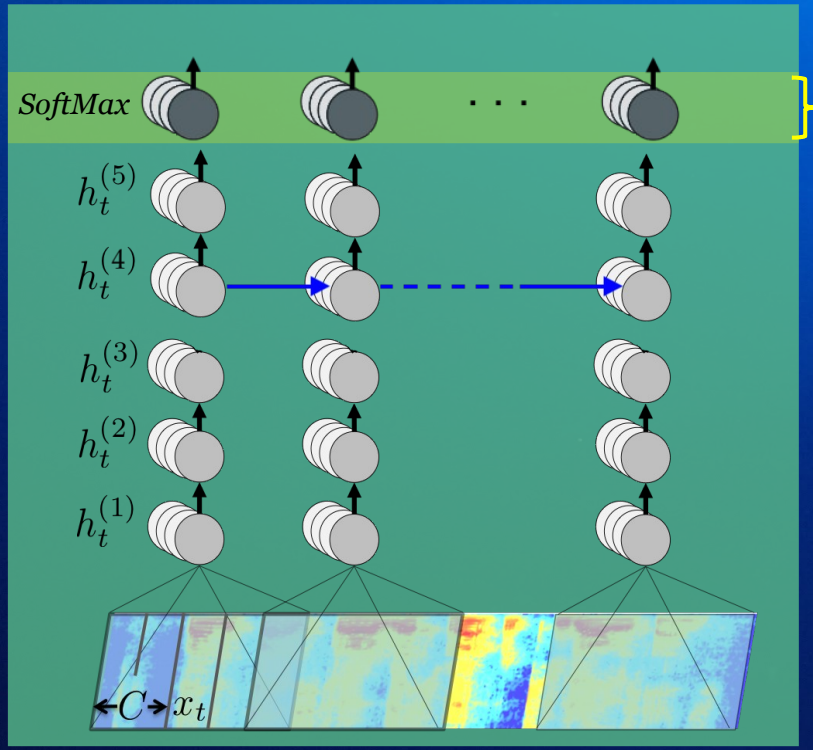
Google
12.1%
WER

[1]Dernoncourt et. al,
A Framework for Speech Recognition Benchmarking

4 Future of Deep Speech

4.1 UTF-8

UTF-8: Softmax Layer

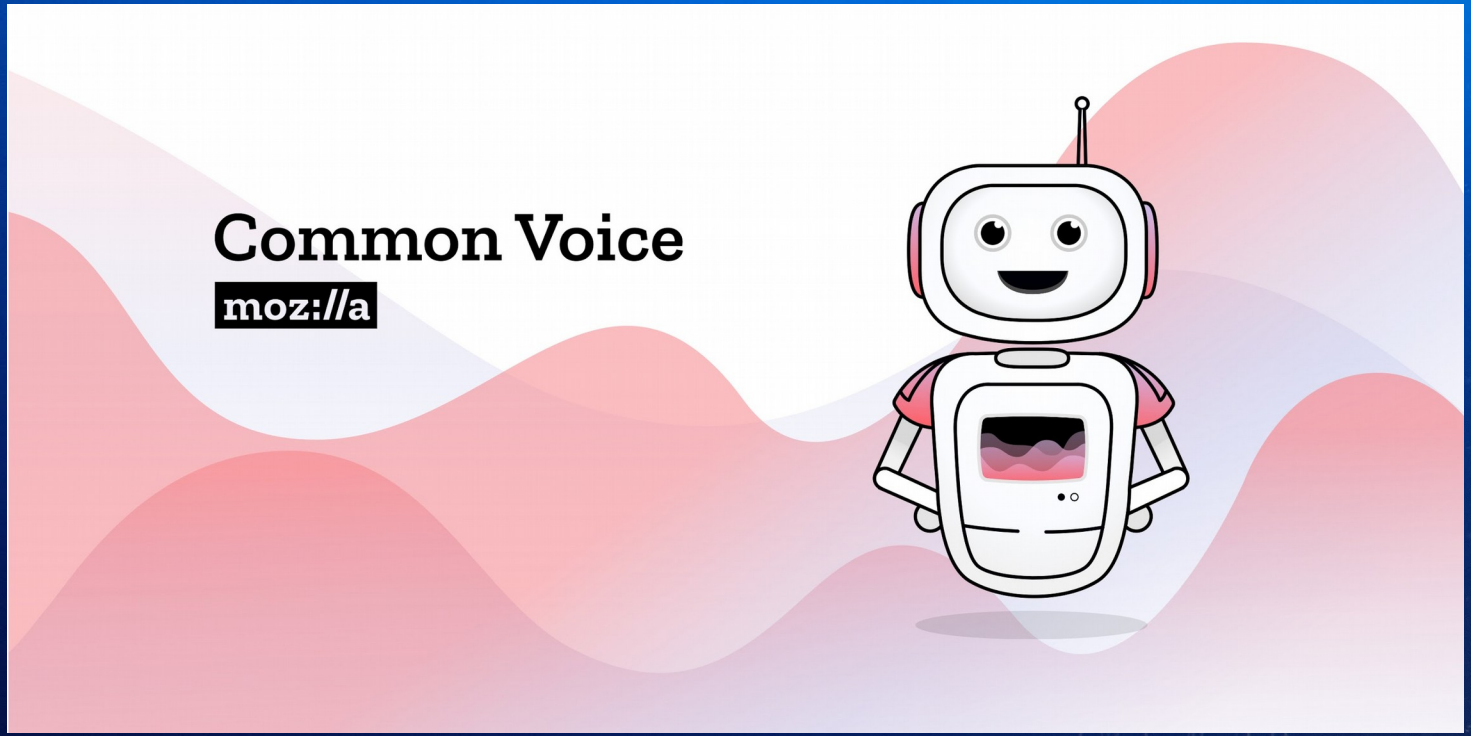


Softmax Layer

- L ≡ 256
- Output width k ≡ |L| + 1
- Extra for a “blank label”

4.2 Non-English Languages

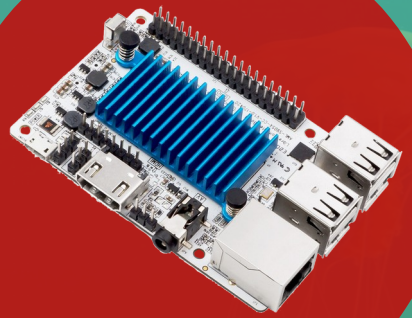
Non-English Languages: Common Voice



4.3 Small Platform Devices

Small Platform Devices

Snapdragon 835



Le Potato

Рахмат!