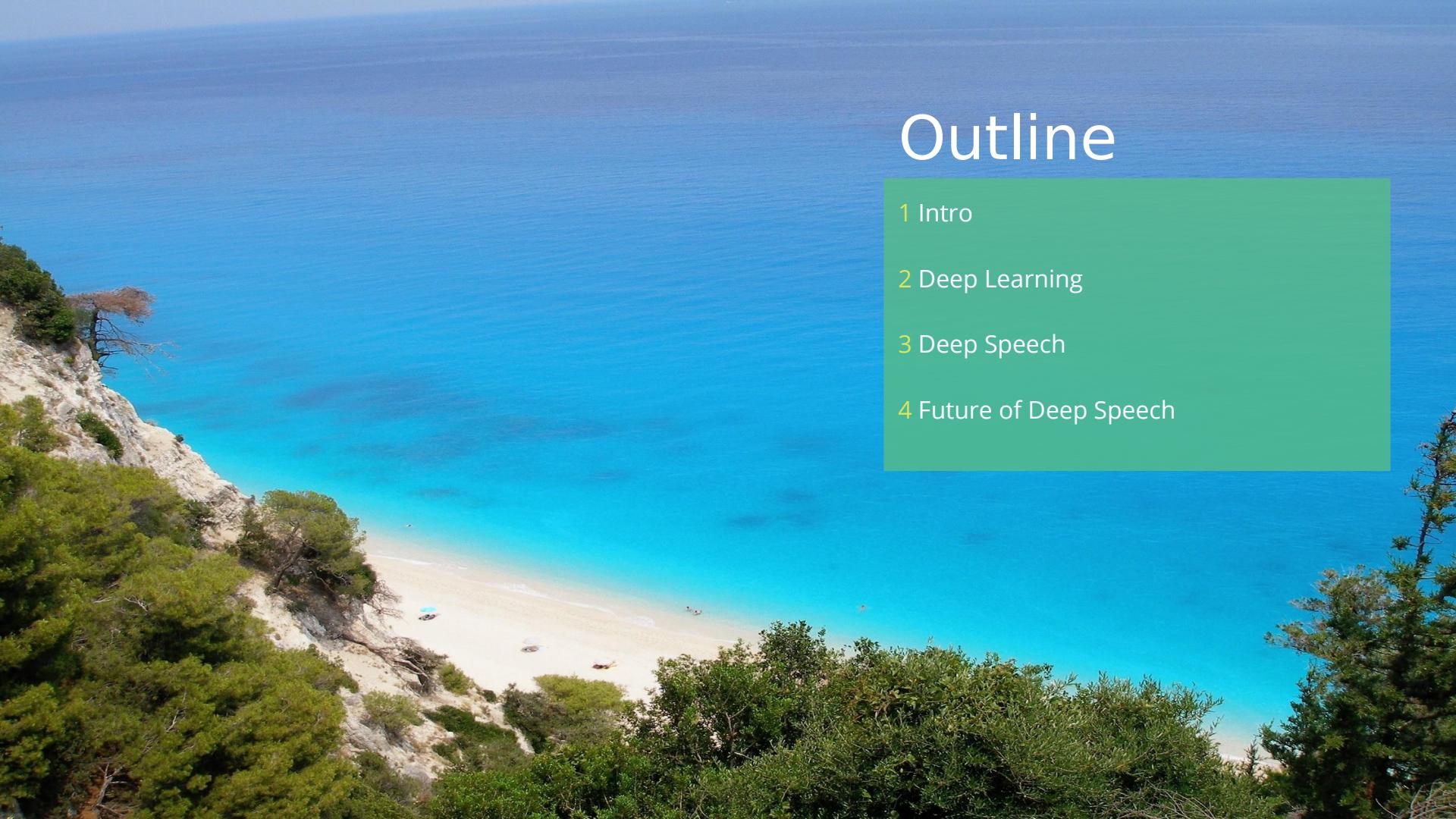


Mozilla's Deep Speech



The background of the slide features a wide-angle photograph of a beautiful coastal scene. On the left, a light-colored cliff face with sparse greenery descends towards a white sandy beach. The water is a vibrant turquoise color, transitioning to a darker blue further out. The sky above the water is clear and bright.

Outline

1 Intro

2 Deep Learning

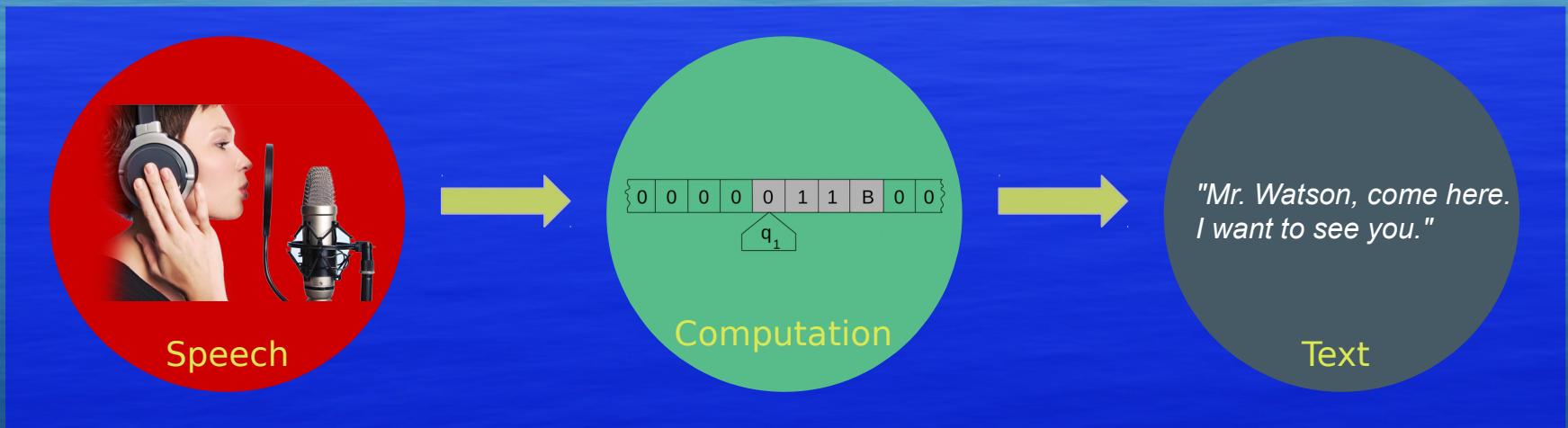
3 Deep Speech

4 Future of Deep Speech

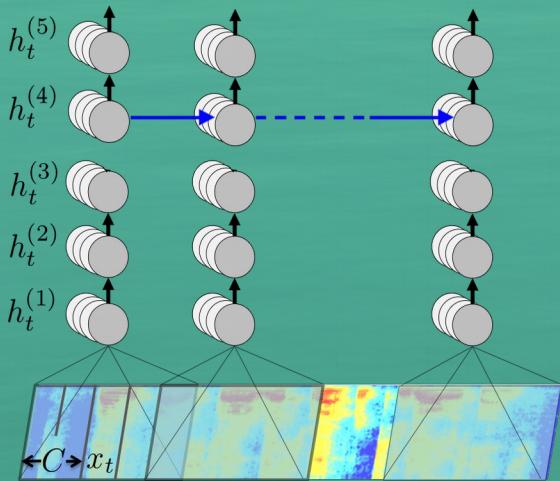
1 Intro



Intro: Speech-to-Text

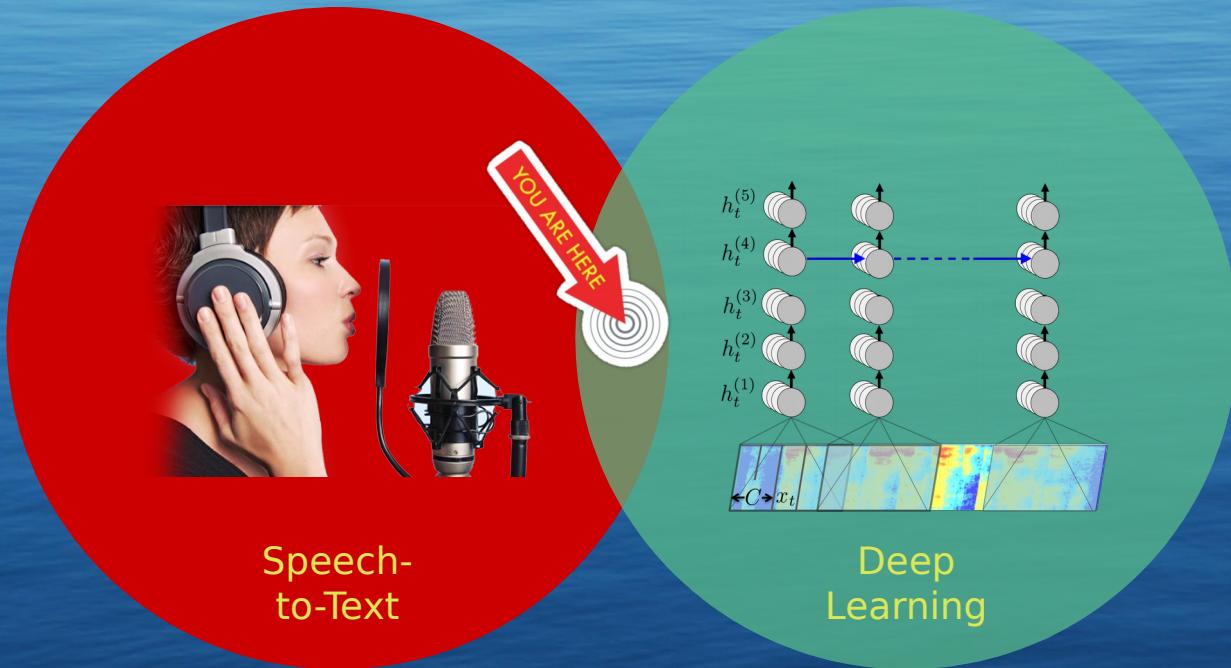


Intro: Deep Learning

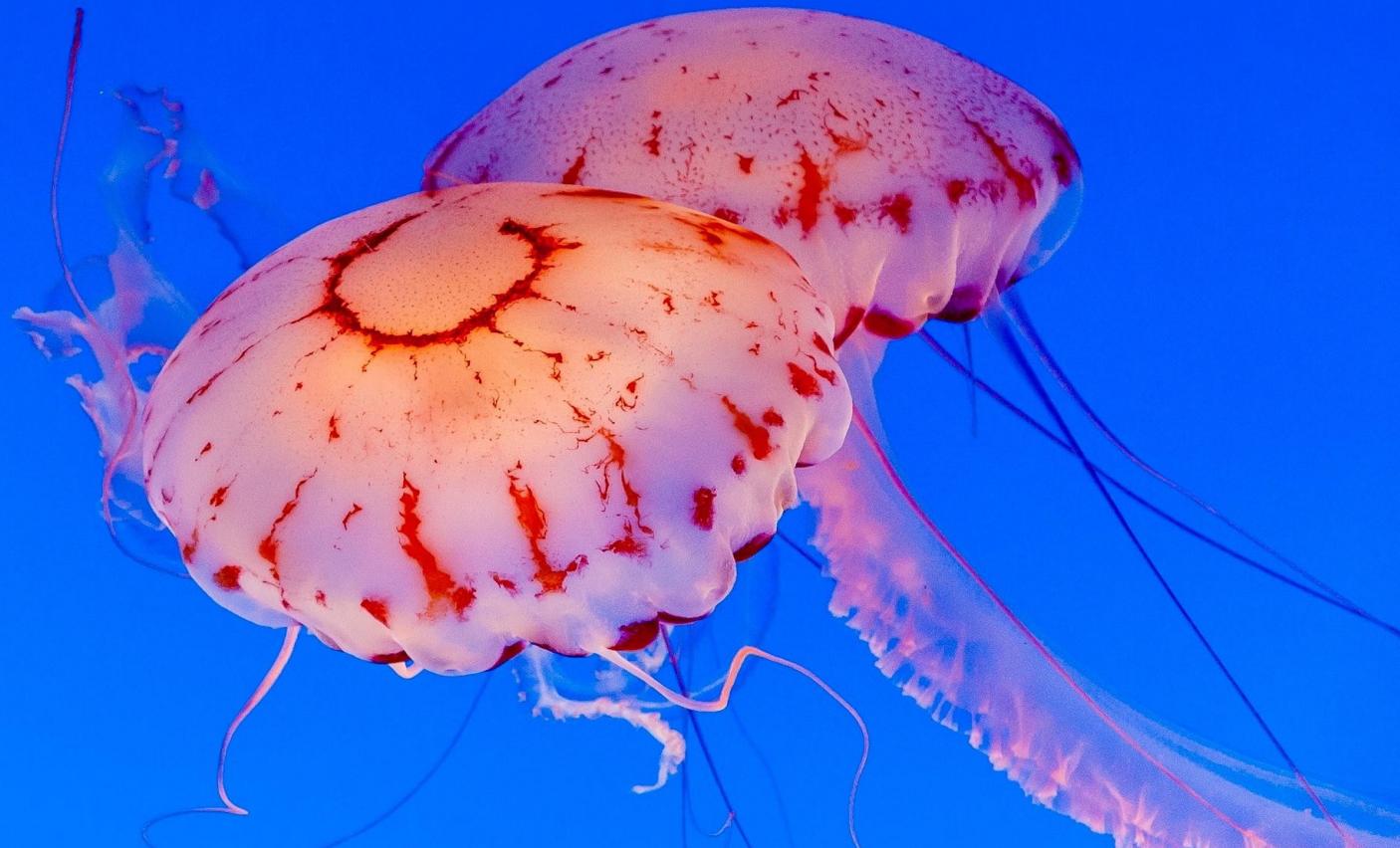


Deep
Learning

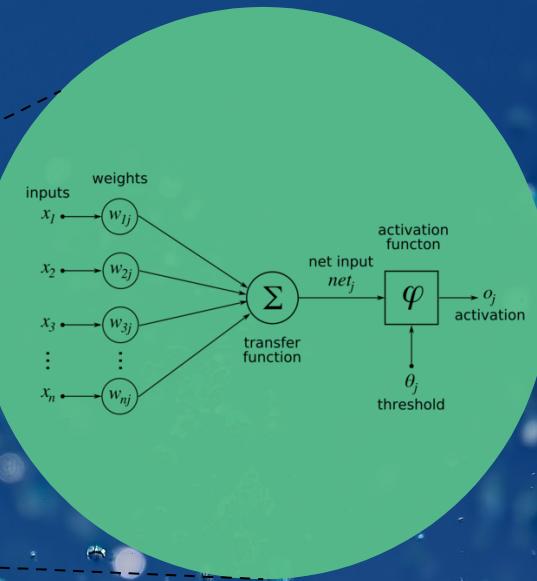
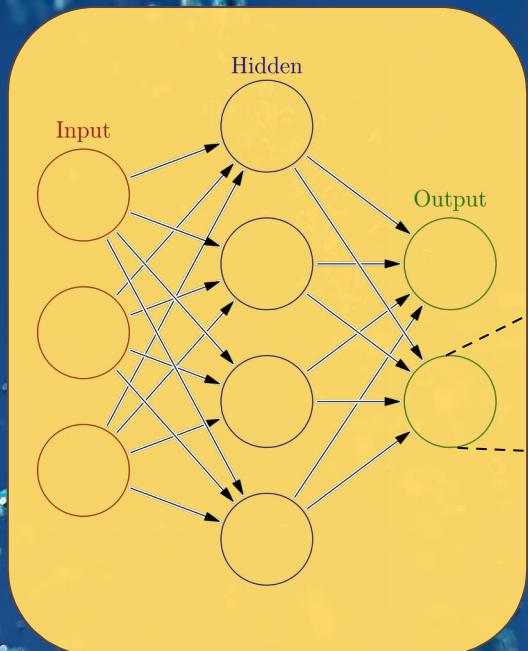
Intro: Speech-to-Text n Deep Learning



2 Deep Learning

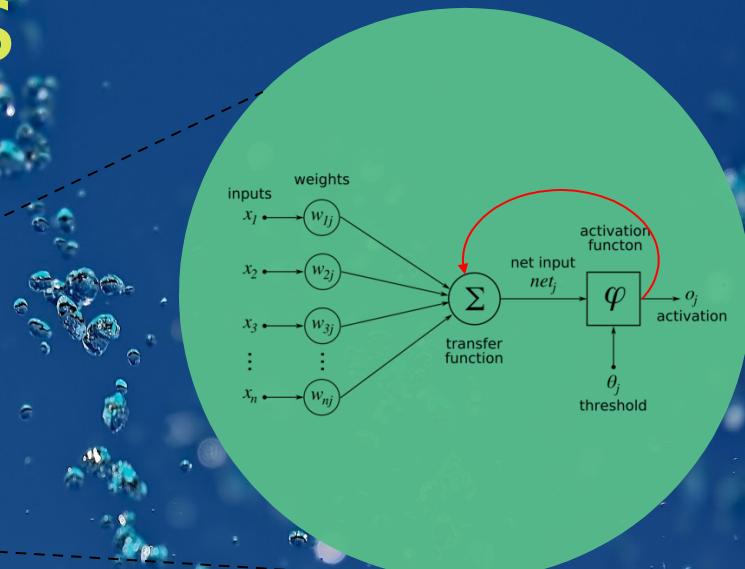
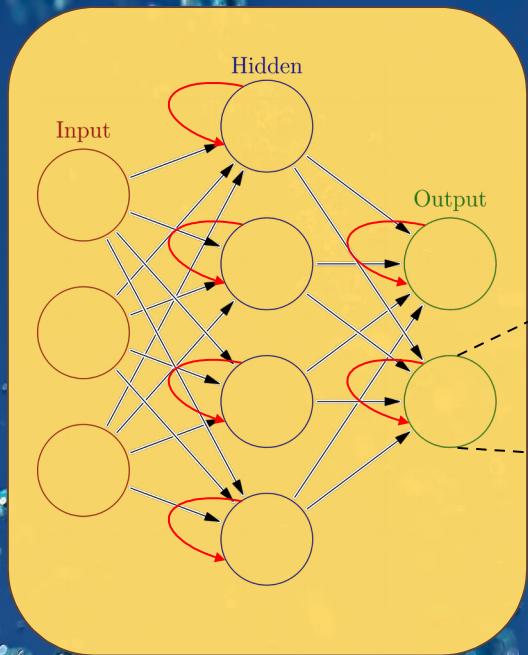


Deep Learning: Feed Forward Networks



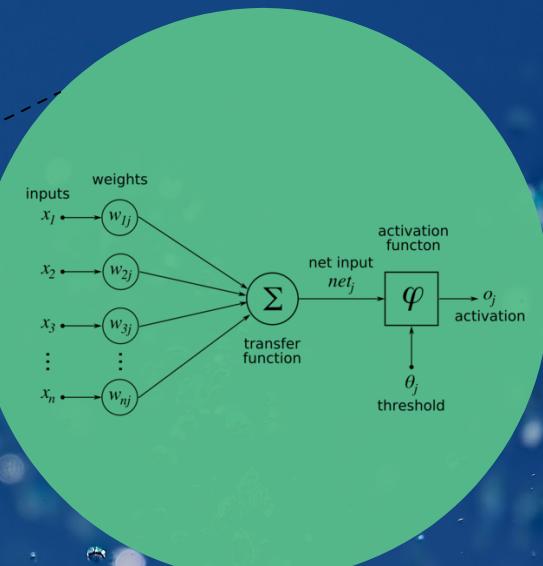
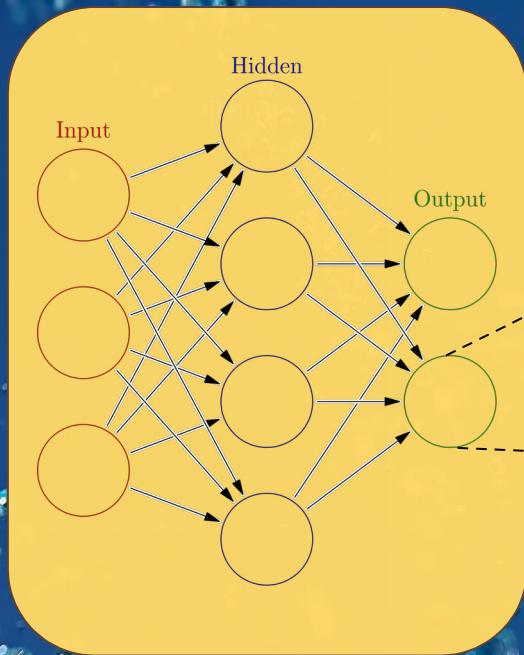
$$o_j = \varphi \left(\theta_j + \sum_{i=1}^n w_{ij} x_i \right)$$

Deep Learning: Recurrent Neural Networks



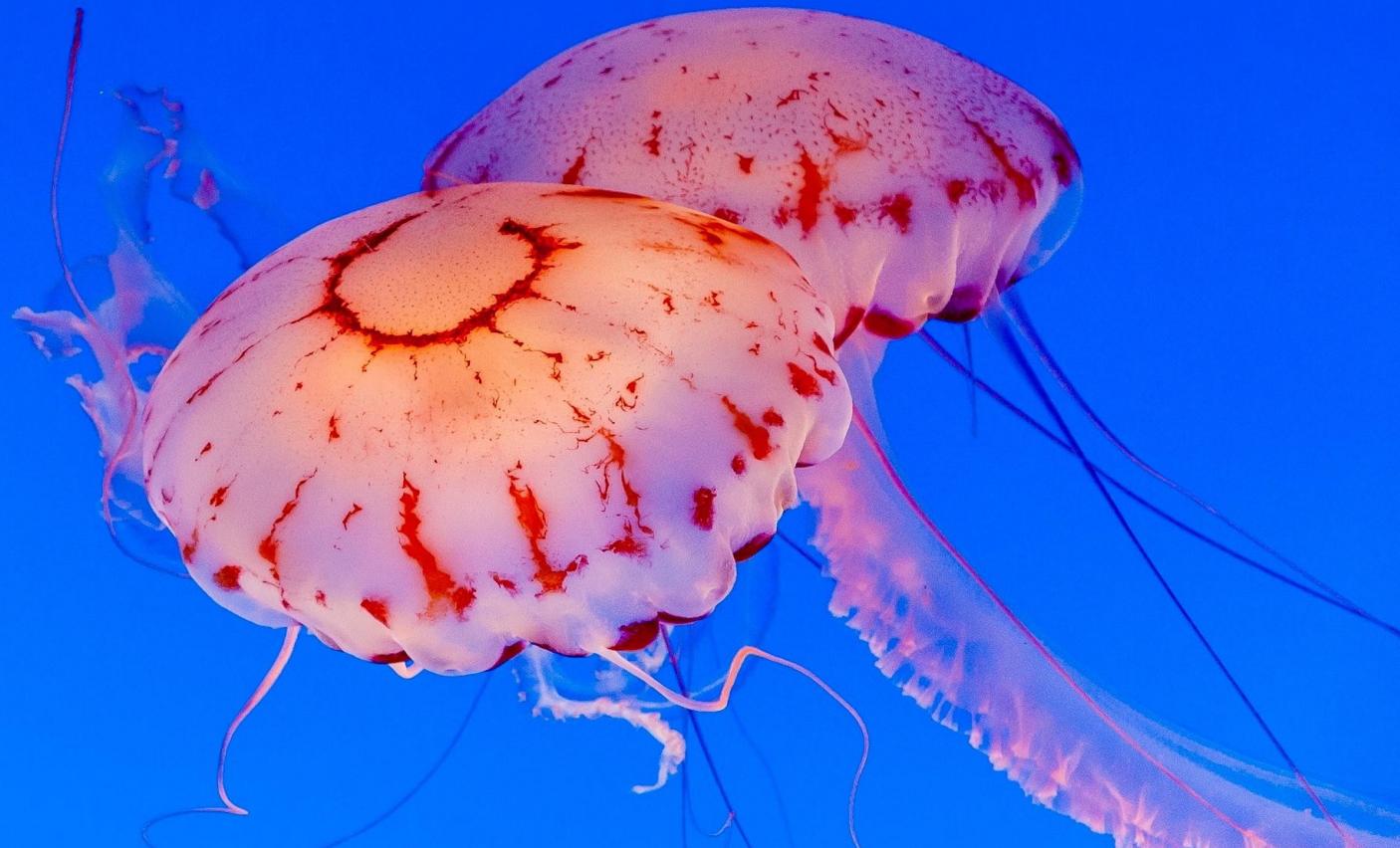
$$t o_j = \varphi \left(\theta_j + \sum_{i=1}^m t^{-1} o_i w_{ij}^f + \sum_{i=1}^n x_i w_{ij} \right)$$

Deep Learning: Softmax



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

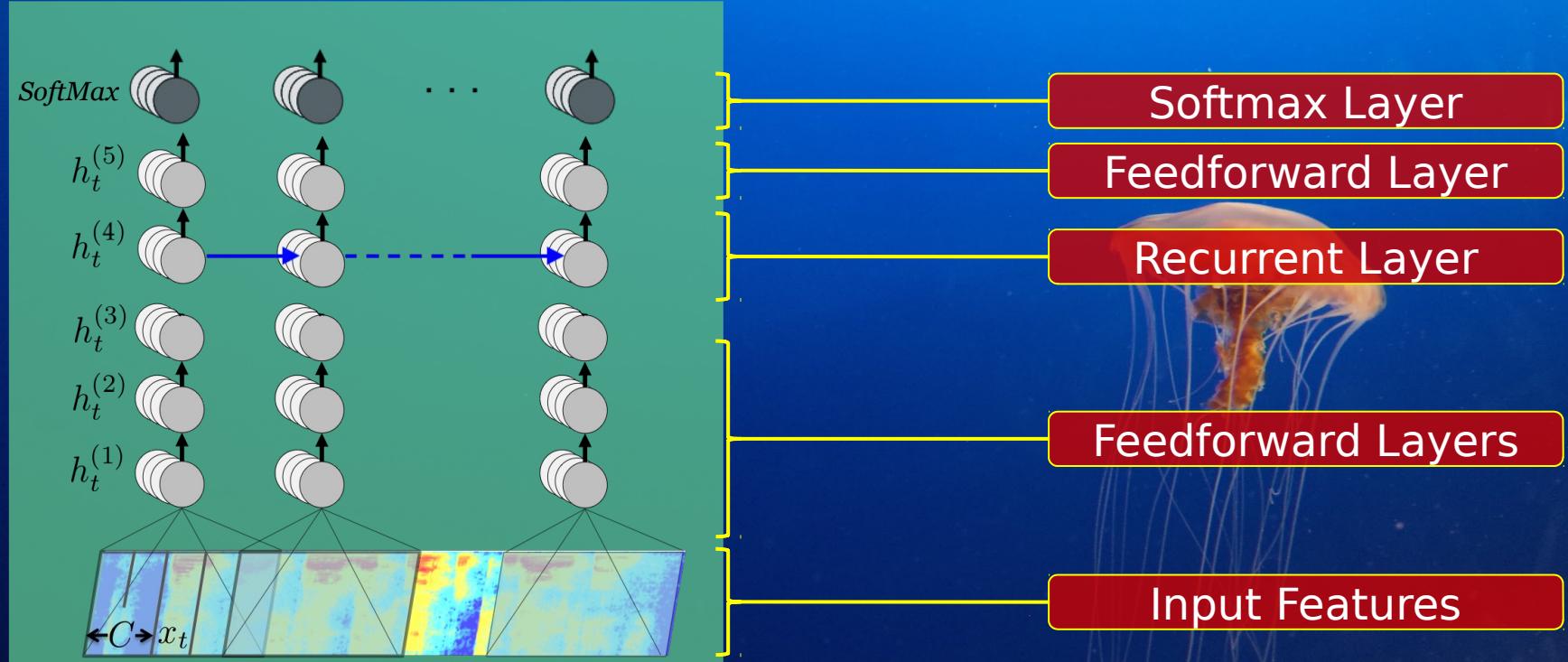
3 Deep Speech



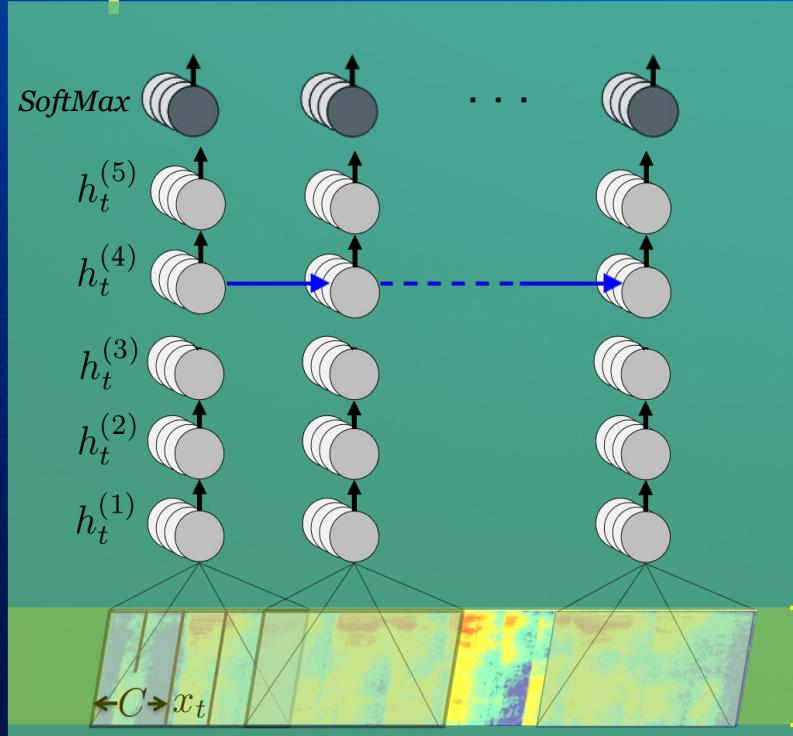
The background image shows a high-angle aerial view of a desert landscape. The terrain is characterized by a series of parallel, reddish-brown ridges and deep, dark blue-grey valleys. The ridges have a distinctively layered or stratified appearance, suggesting geological processes like sedimentary rock formation. Sparse, dry vegetation is scattered across the ridges. The lighting is dramatic, with strong sunlight casting deep shadows in the valleys and highlighting the textures of the rocky surfaces.

3.1 Deep Speech Architecture

Deep Speech Architecture: Overview



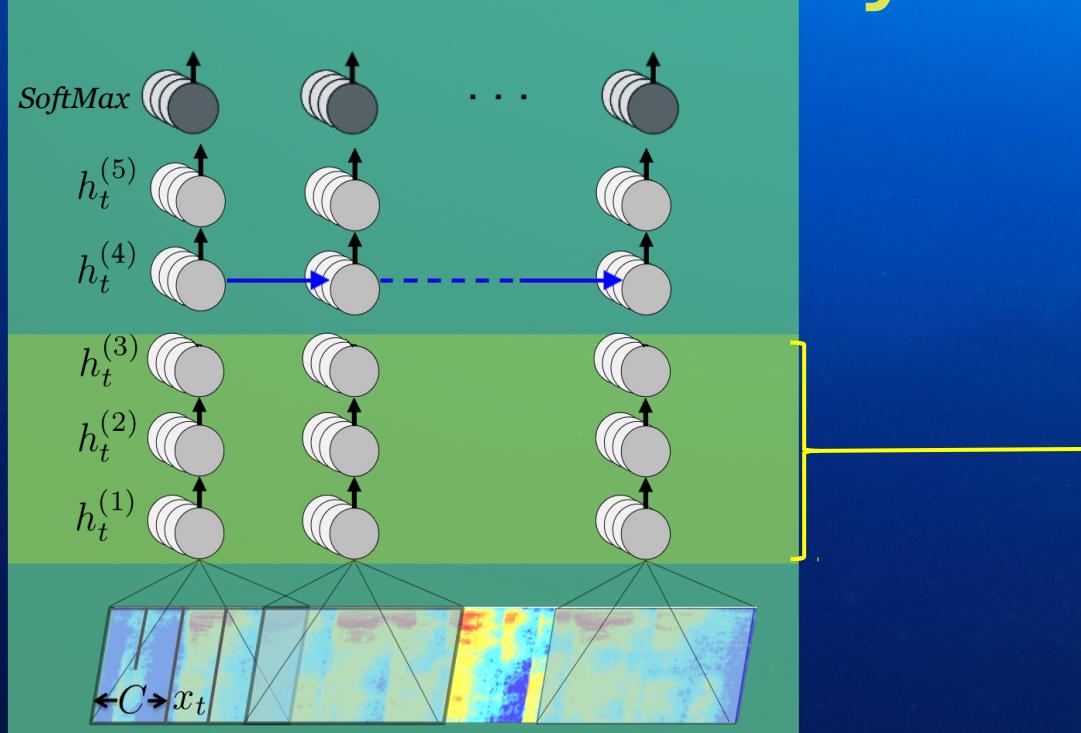
Deep Speech Architecture: Input Features



Mel-Frequency Cepstrum Coefficients

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

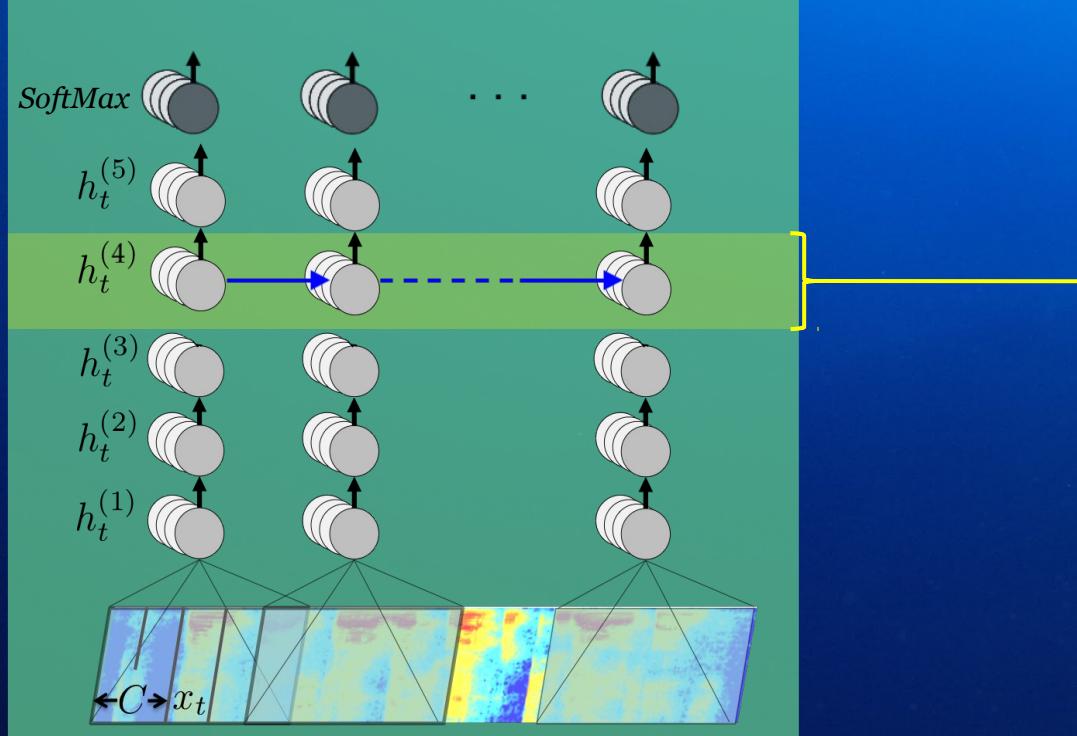
Deep Speech Architecture: Feedforward Layers



Feedforward Layers

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

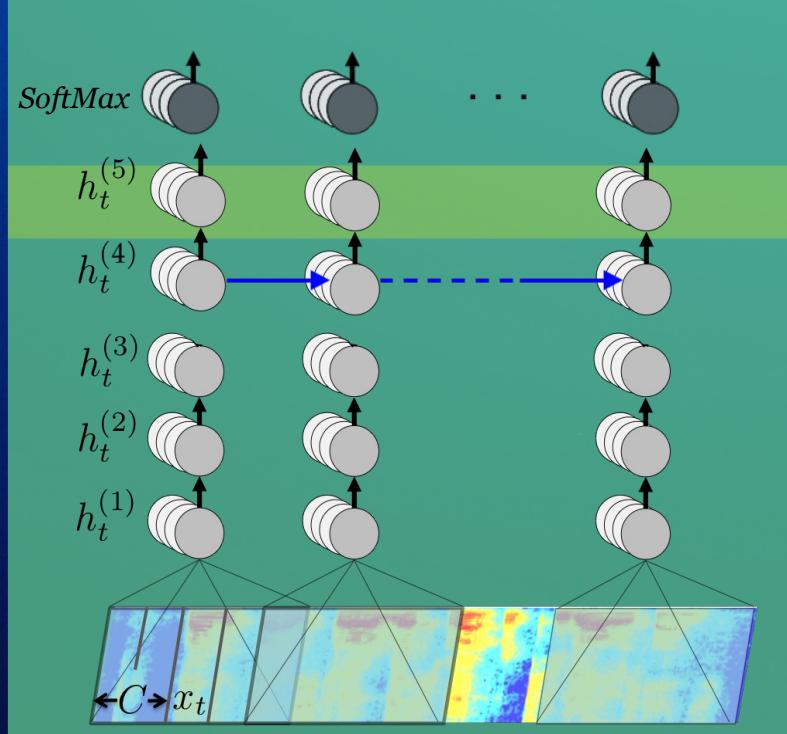
Deep Speech Architecture: Bidirectional RNN Layer



Recurrent Layer

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

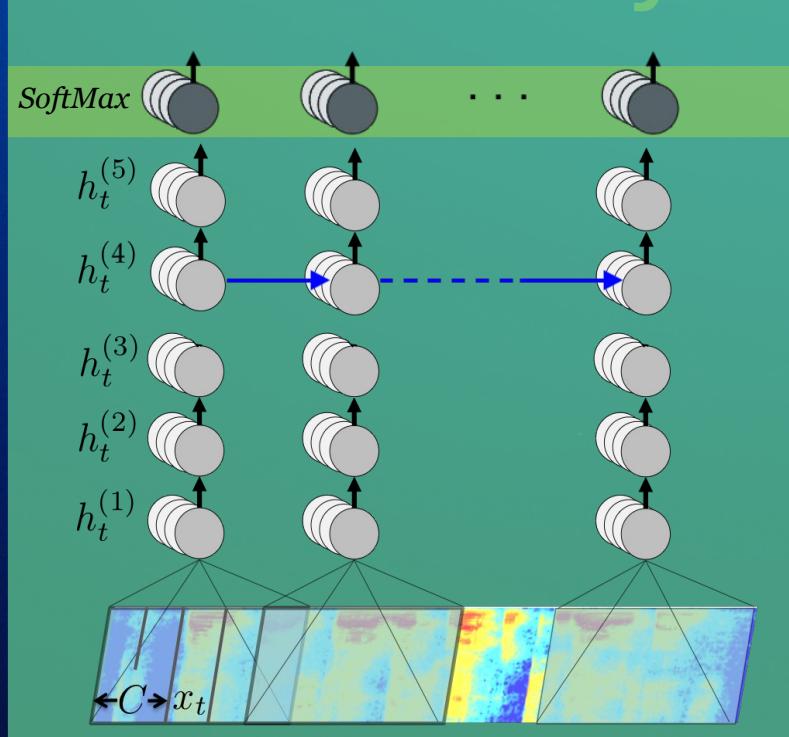
Deep Speech Architecture: Feedforward Layer



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

- $L \equiv \text{Alphabet}$
- Output width $k \equiv |L| + 1$
- 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

Training Data

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



Human
5.8%
WER

On LibriVox clean
test 8.26% WER

Google
12.1%
WER

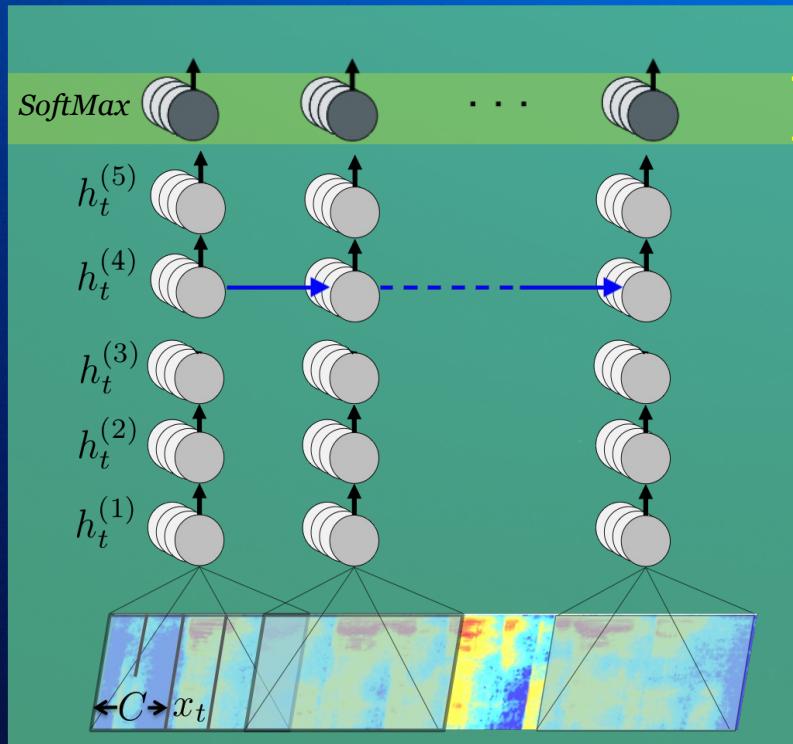
4 Future of Deep Speech



A high-angle aerial photograph of a desert landscape. The terrain is a mix of light brown and tan colors, with darker, more rugged areas representing rocky outcrops and sand dunes. The lighting creates strong shadows, emphasizing the three-dimensional nature of the terrain.

4.1 UTF-8

UTF-8: Softmax Layer



Softmax Layer

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

The background image shows a vast desert landscape from an aerial perspective. The terrain is dominated by large, undulating sand dunes with a reddish-brown hue. Interspersed among the dunes are several rocky, light-colored outcrops and ridges, which appear to be made of limestone or similar sedimentary rock. The lighting suggests either early morning or late afternoon, casting long shadows and highlighting the textures of the rock formations.

4.2 Non-English Languages

Non-English Languages: Common Voice



The background image shows a high-angle aerial view of a rugged, reddish-brown landscape, possibly a dry riverbed or a series of gullies on a planetary surface like Mars. The terrain is textured with various shades of brown, tan, and orange, indicating different geological materials and lighting conditions. The perspective is from above, looking down the length of the gullies.

4.3 Small Platform Devices

Small Platform Devices



Snapdragon 835



Le Potato

Рахмат!

