Speech Recognition

Connectionist Temporal Classification

By: Emily Nguyen

November 2018

Outline

Background: Neural Networks

Speech Recognition and Problems

Recurrent Neural Networks (RNNs)

Long Short Term Memory (LSTM)

Connectionist Temporal Classification (CTC)

Background: Neural Networks

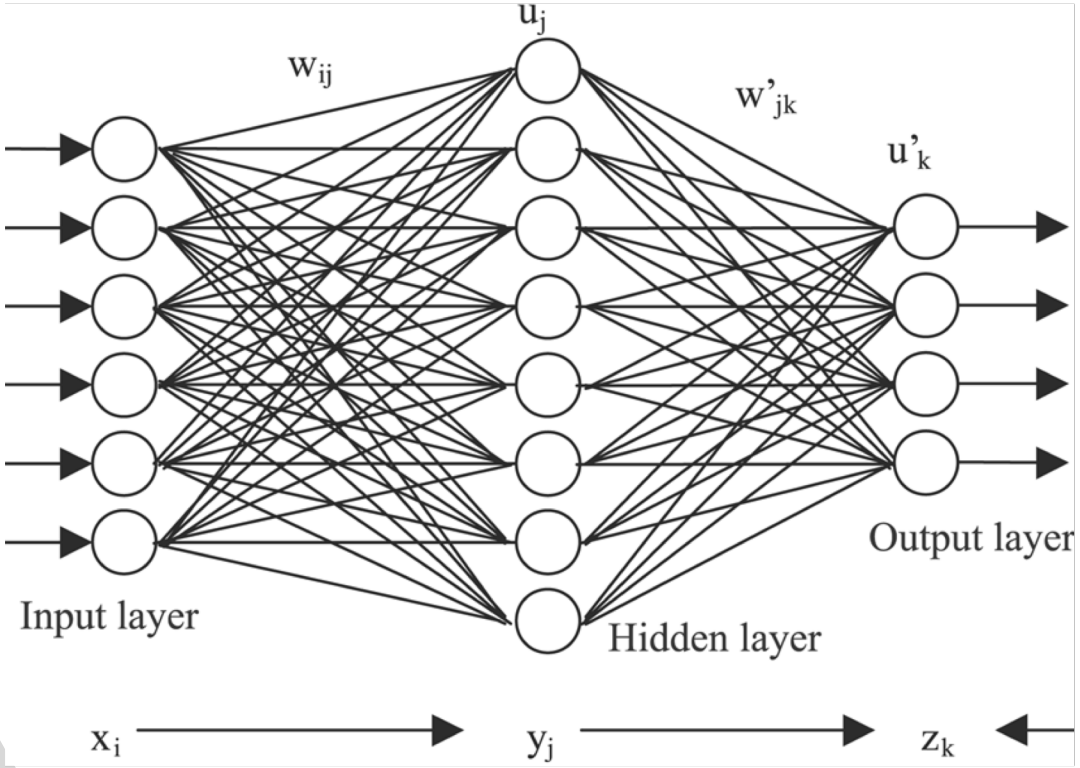
- Output to next layer: $f(Wa + b)$
- Final layer:

$$F(x) = (W^{[L]} f(W^{[L-1]} \dots f(W^{[2]} x + b^{[2]}) + \dots + b^{[L-1]}) + b^{[L]})$$

- Cost function:

$$Cost(W^{[2]}, \dots, W^{[L]}, b^{[2]}, \dots, b^{[L]}) = \frac{1}{2N} \sum_{i=1}^{N} ||y(x^{\{i\}}) - F(x^{\{i\}})||_2^2$$

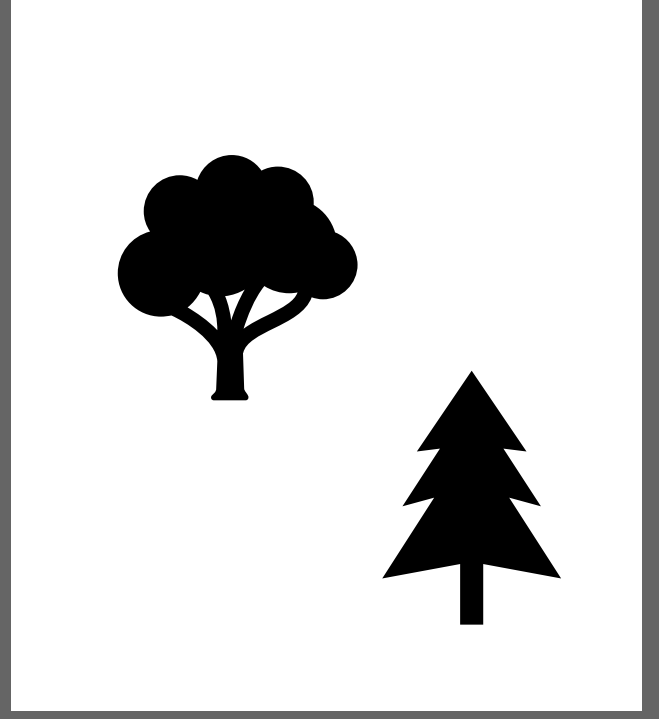
- Training:
  - Optimization algorithm to minimize loss function
  - Back propagation (applied chain rule)
- Good for image classification but not so much for speech recognition

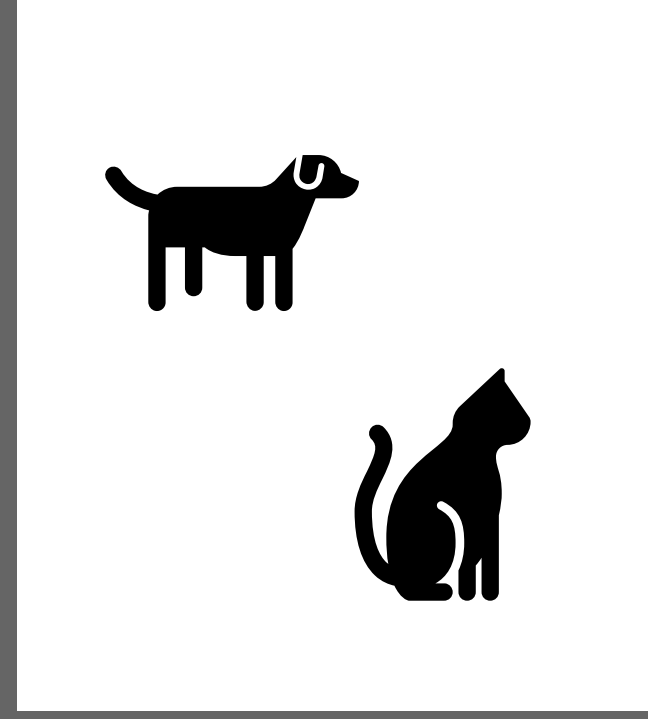


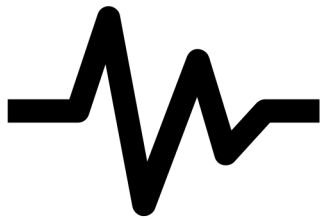
https://www.extremetech.com/extreme/215170-artificial-neural-networks-are-changing-the-world-what-are-they

Speech Recognition

- Want to take in audio files and output what the input is saying
- Difference between preprocessing images and audio:
  - Images can be static
    - Can pick up visual patterns
  - Audio leads to sequenced data
    - Pitch, speed
- Supervised Sequence Labelling with Recurrent Neural Network by Alex Graves[1]

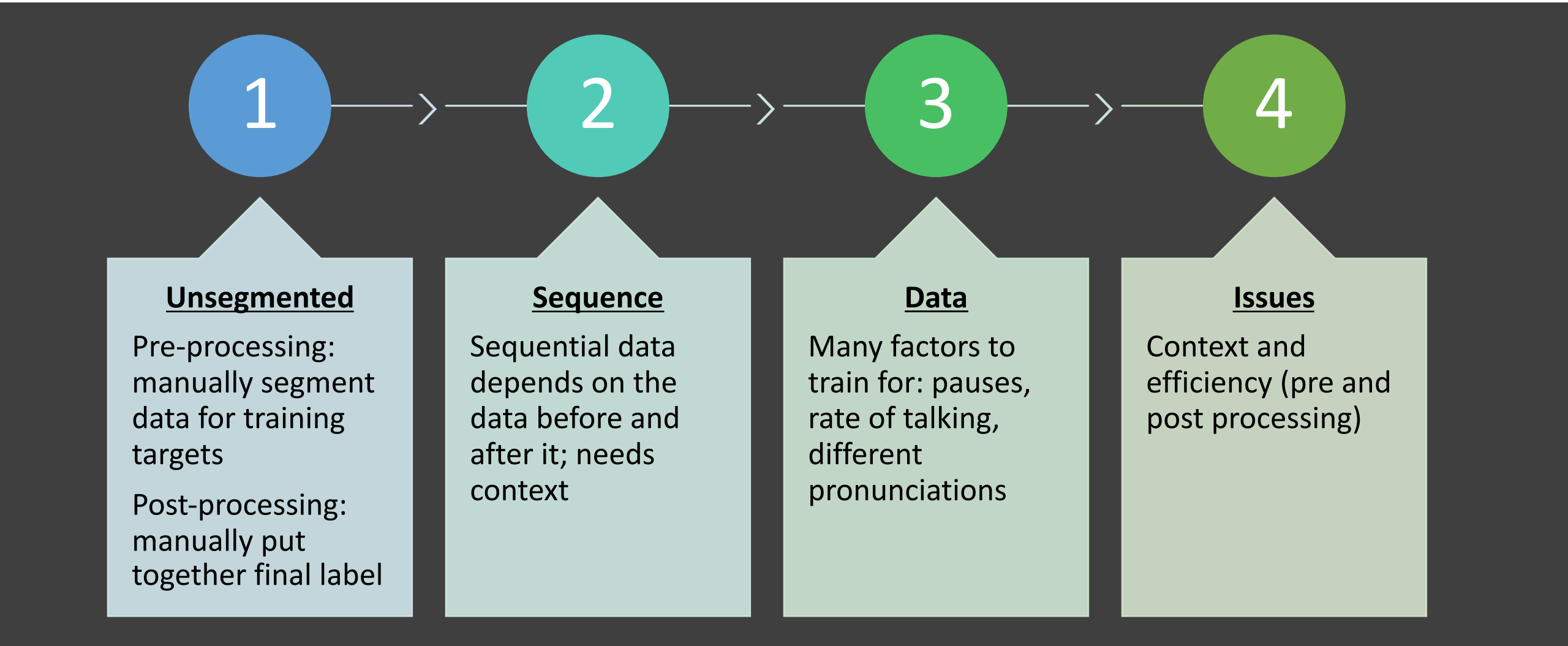






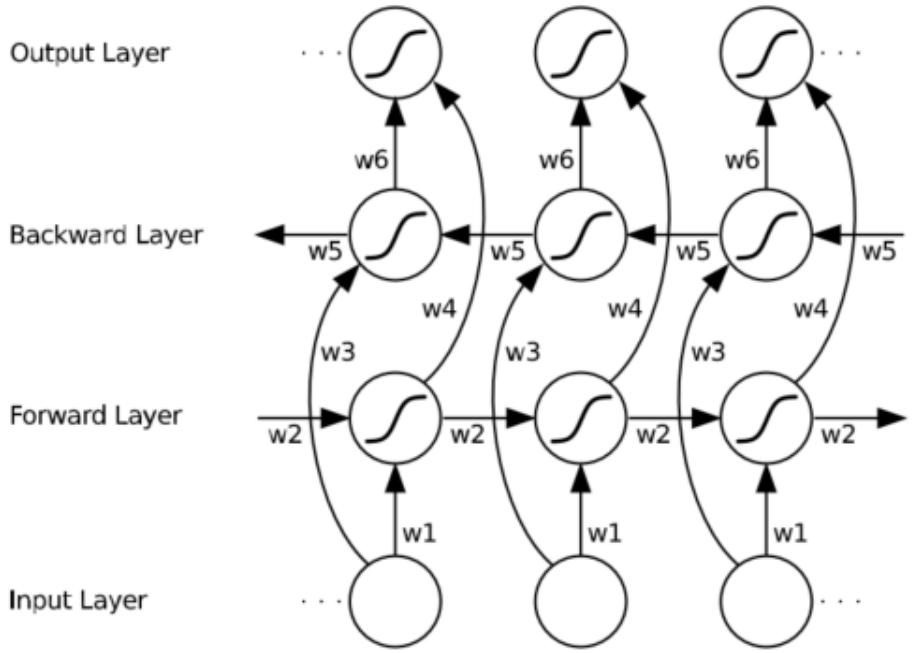
1. Graves, Alex. "Supervised sequence labelling with recurrent neural networks. 2012." *ISBN 9783642212703. URL http://books. google. com/books*.

Problem: Labelling Unsegmented Sequence Data

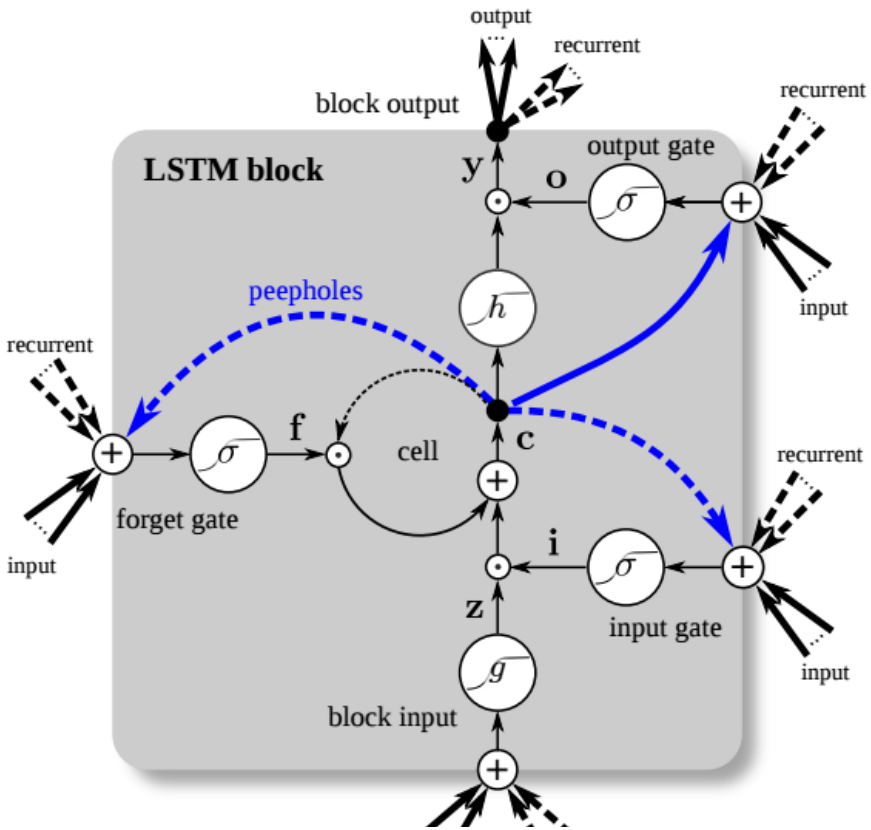


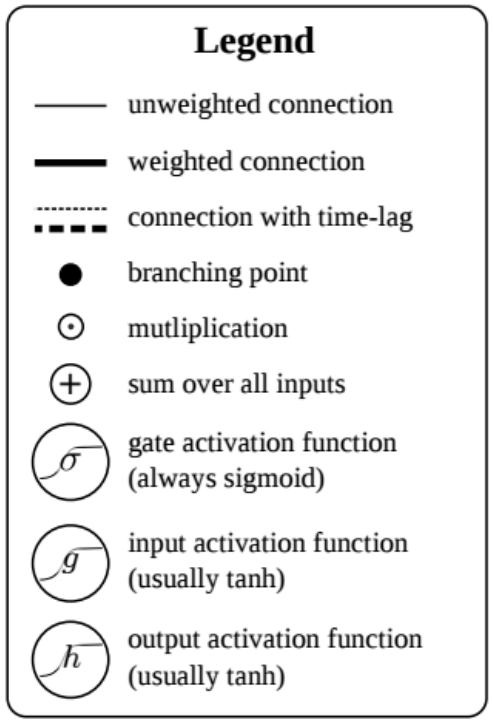
Recurrent Neural Networks

- Can be used to deal with sequenced data
  - Uses cyclical connections
  - Can learn what to store and what to ignore
- Cons:
  - Standard RNNs don’t use future info
  - Vanishing gradient problem
    - Chain rule; small gradients -> not learning as much
- Bidirectional RNNs
  - Input sequence forwards and backwards to two separate RNNs but connected to same output



https://www.researchgate.net/figure/An-unfolded-Bidirectional-Recurrent-Neural-Network-taken-from-Gra12-p-22_fig4_309549275





https://developer.nvidia.com/discover/lstm

Long Short-Term Memory

- LSTM: a set of recurrently connected subnets (memory blocks)
  - obtain one or more self-connected memory cells and three units
  - input, output, and forget gates
- Pro
  - remembering context
- Con
  - network forgets first inputs as new inputs overwrite activations of the hidden layer
- Bidirectional LSTM: access to long range context in both input directions

Connectionist Temporal Classification

- Is an output layer
- y_k^t : activation of output k at time t
- π : paths
- F : many-to-one function mapping set of paths onto set of possible labellings
- Conditional probability of paths

$$p(\pi|x) = \prod_{t=1}^{T} y_{\pi_t}$$

- Probability of a labelling

$$p(l|x) = \sum_{\pi \in F^{-1}(l)} p(\pi|x)$$

- Forward backward algorithm
- CTC loss function

$$L(S) = -\sum_{(x,z)\in S} lnp(z|x)$$

- Training can be done with backpropagation through time and any gradient-based non-linear optimization algorithm

1

Unsegmented

CTC can directly output probabilities of complete label sequences

2

Sequence

CTC can be paired with RNNs for context

3

Data

CTC can model all aspects of sequence with **one** neural network

CTC Continued

Summary and Conclusions

CTC is good for speech recognition, handwriting recognition, and other problems that have sequenced data

Efficient with one network and removes need for pre and post processing

The authors of the book actually ran tests, and the BLSTM with CTC actually outperforms other architectures

Questions, Comments, Concerns?

Emily.pham.nguyen@utexas.edu