University of BRISTOL

Deep learning for sequence processing: Section 9.2, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2021).

8.2 Recurrent Neural Networks

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

Sequence Processing with Feedforward Networks

- A sentence can have the same meaning if it comes at the start, middle or end of a document
- But, in a feedforward network…
  - Different weights are applied to each input position, so “good day to you” is processed differently to “a good day to you”.
  - All data must be passed in at once, even if it is a document with thousand of words.
  - This makes learning harder and scalability trickier.

Sequence Labelling and Text Classification

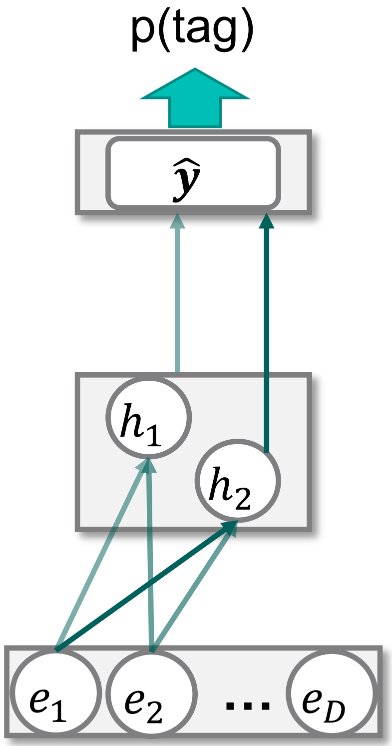
- Sequence labelling:
  - HMMs use information from the previous and next tokens in a sequence.
  - Viterbi algorithm passes messages forward and then backward.
  - Strong assumption that the current tag depends only on the previous tag.
  - How can we perform sequence labelling with a neural network?
- Text classification:
  - We want to process the tokens sequentially, then predict the class label for the whole document.
- Solution: **recurrent neural networks (RNNs)**.

RNN: Recurrent Connections

Task: sequence labelling.

Input to the recurrent layer includes both the current input and the activation of the recurrent layer at the previous time-step.

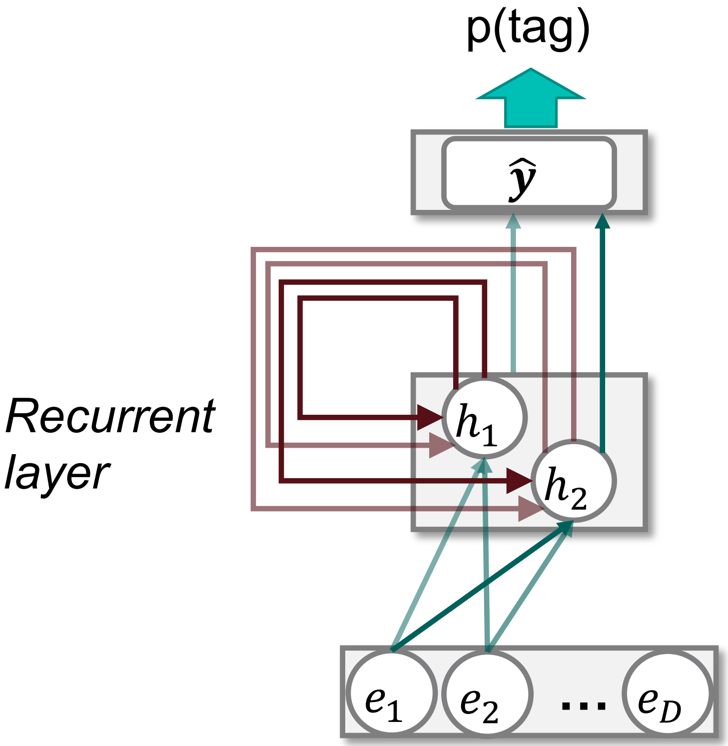
Recurrent layer



RNN: Recurrent Connections

Task: sequence labelling.

Input to the recurrent layer includes both the current input and the activation of the recurrent layer at the previous time-step.

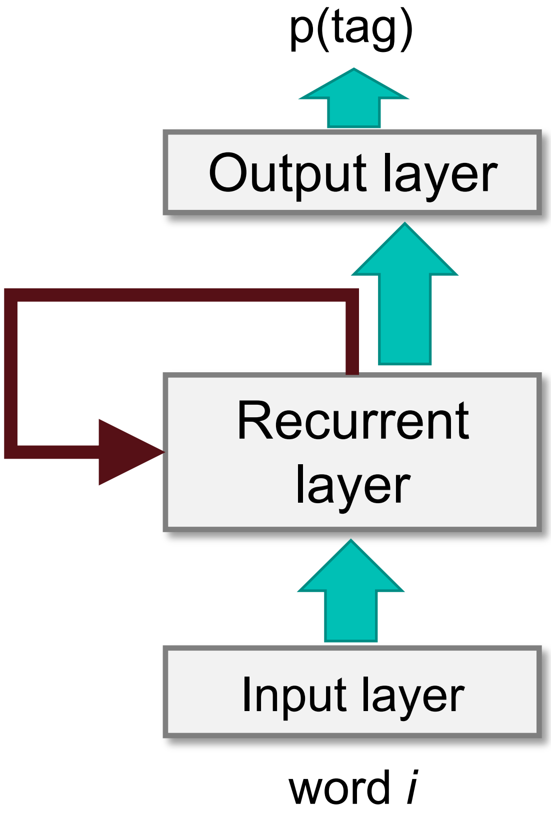


RNN: Recurrent Connections

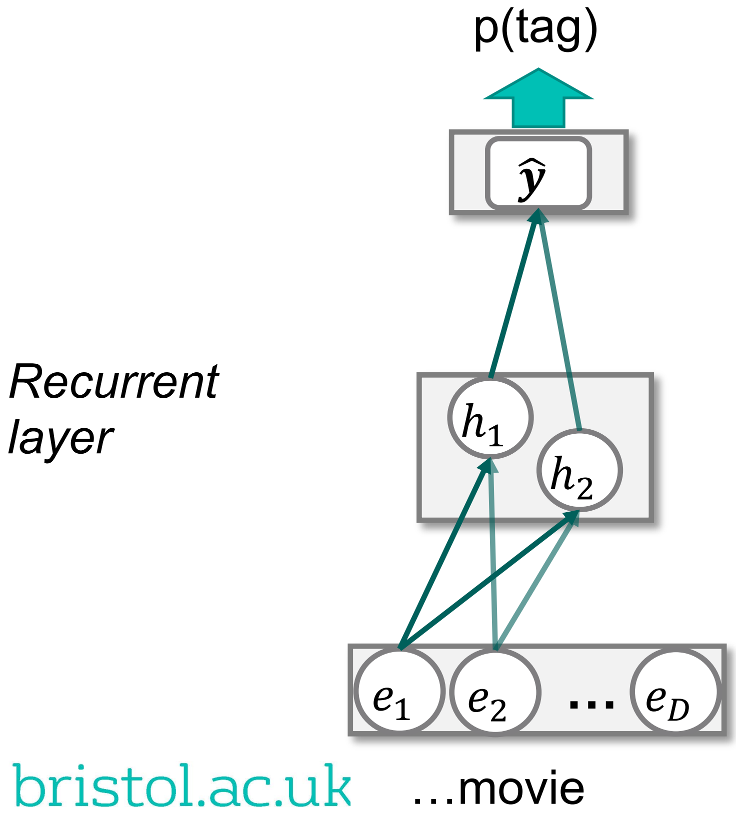
Task: sequence labelling.

Input to the recurrent layer includes both the current input and the activation of the recurrent layer at the previous time-step.

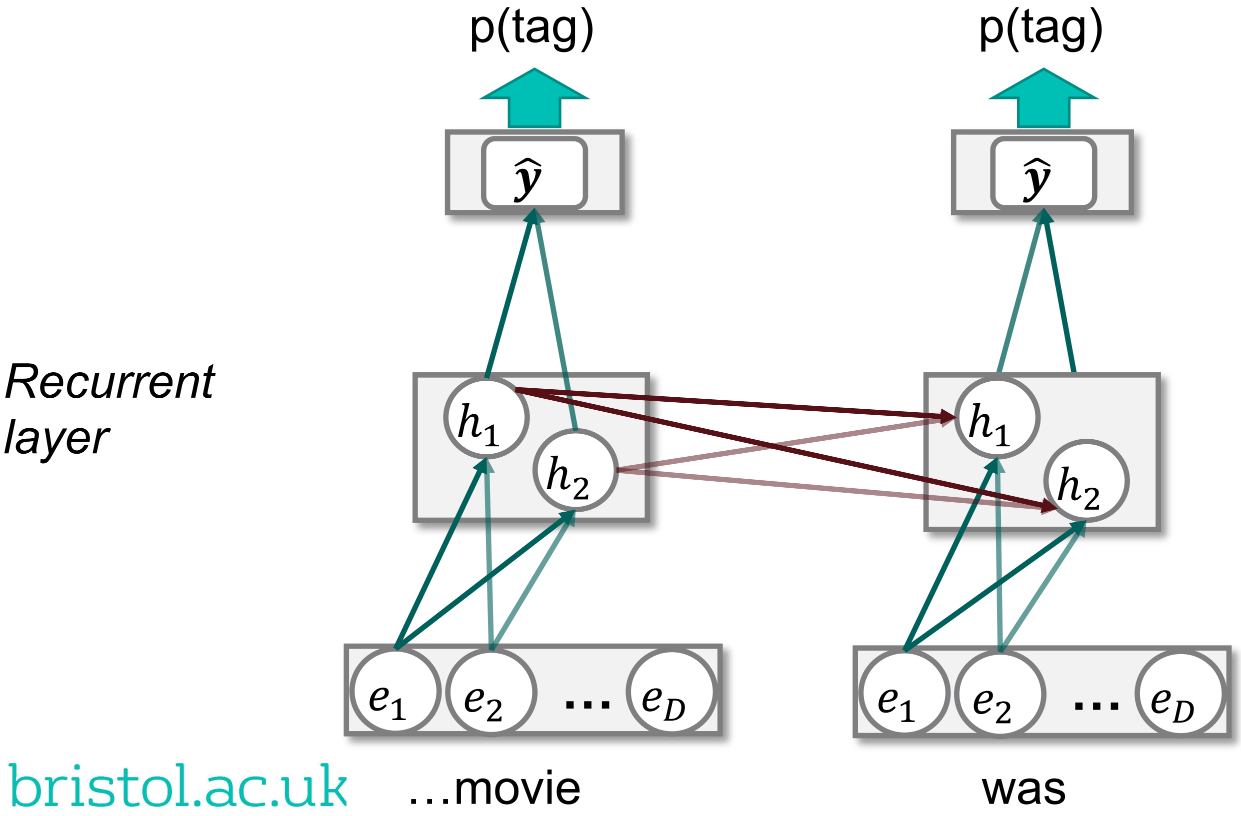
Simplified view:



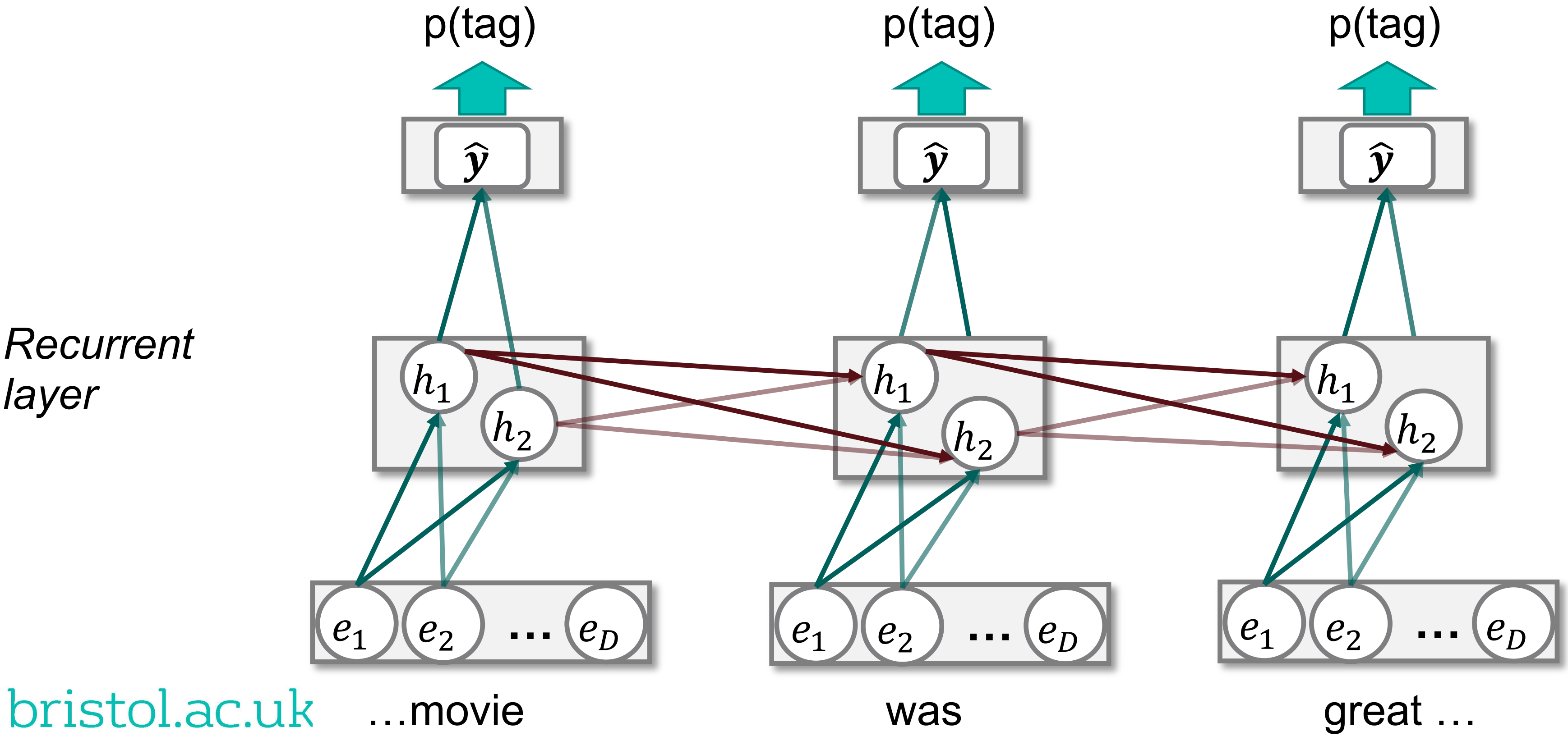
RNN: Unrolled



RNN: Unrolled

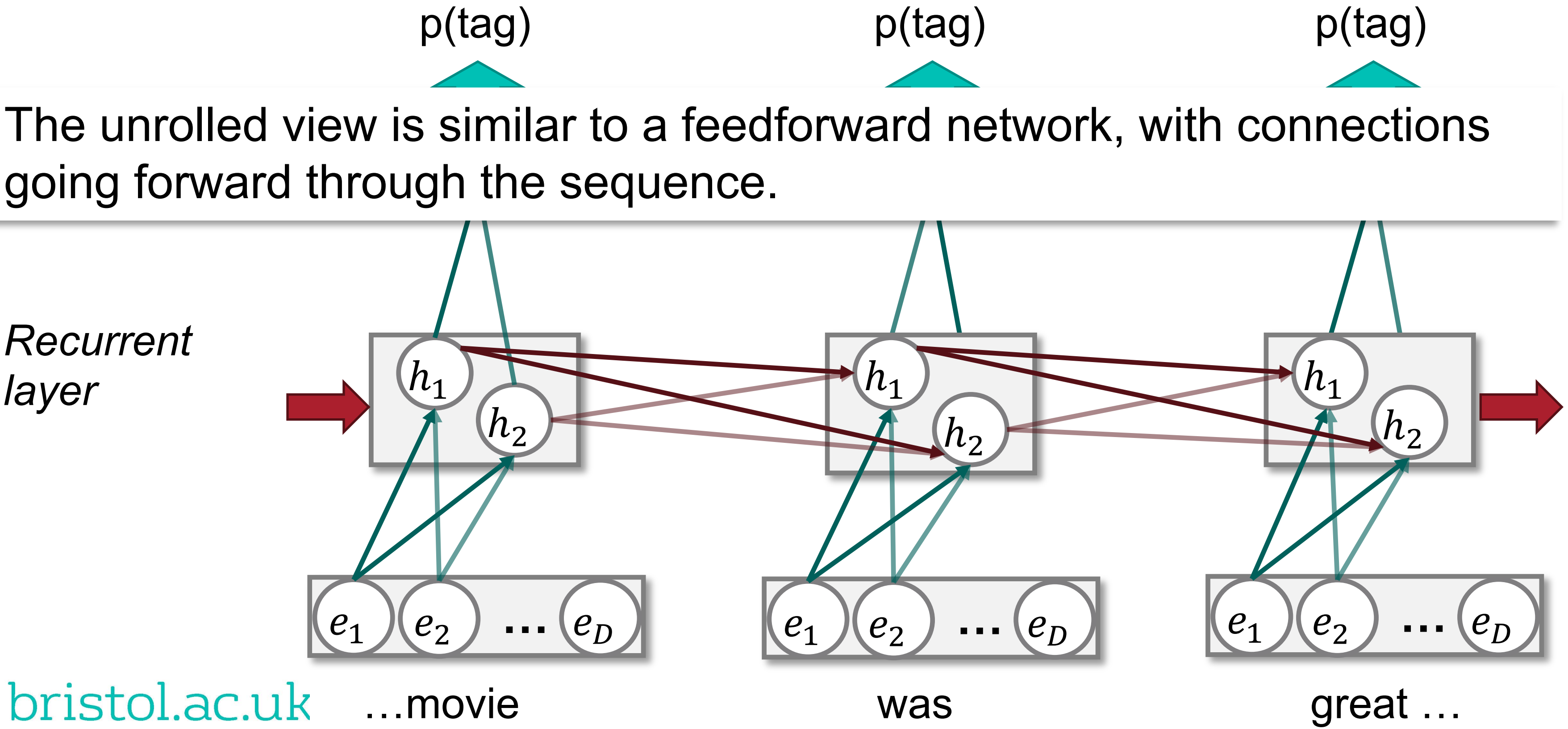


RNN: Unrolled



RNN: Unrolled

The unrolled view is similar to a feedforward network, with connections going forward through the sequence.



RNN Equations

Feedfoward Neural Network:

1. $\boldsymbol{h} = g(\boldsymbol{W}^{(1)}\boldsymbol{x})$
   - g is the activation function, e.g., ReLU, sigmoid.
   - $\boldsymbol{W}^{(1)}$ is the weight matrix of the first hidden layer
2. $\hat{\boldsymbol{y}} = softmax(\boldsymbol{W}^{(2)}\boldsymbol{h})$
   - $\hat{\boldsymbol{y}}$ is the output probability vector.

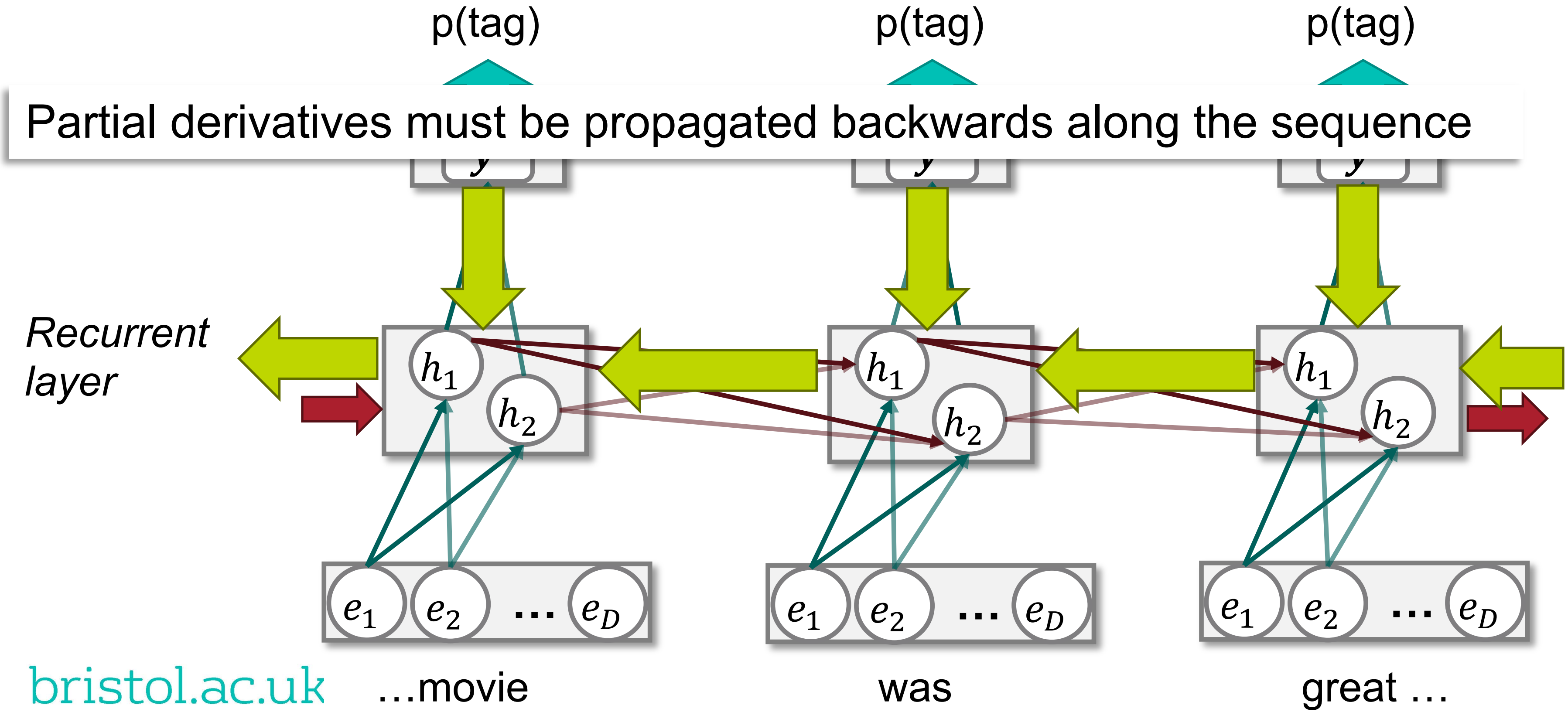
RNN:

1. $\boldsymbol{h}_t = g(\boldsymbol{W}^{(1)}\boldsymbol{x}_t + \boldsymbol{U}\boldsymbol{h}_{t-1})$
   - $\boldsymbol{U}$ is weight matrix for the recurrent connection.
   - $\boldsymbol{h}_t$ is the activation for token t , called the **hidden state**.
2. $\widehat{\boldsymbol{y}_t} = softmax(\boldsymbol{W}^{(2)}\boldsymbol{h}_t)$
   - Predicts the sequence label for time-step t.

[Section 3.1, Introduction to Natural Language Processing, Eisenstein, J. (2019).]

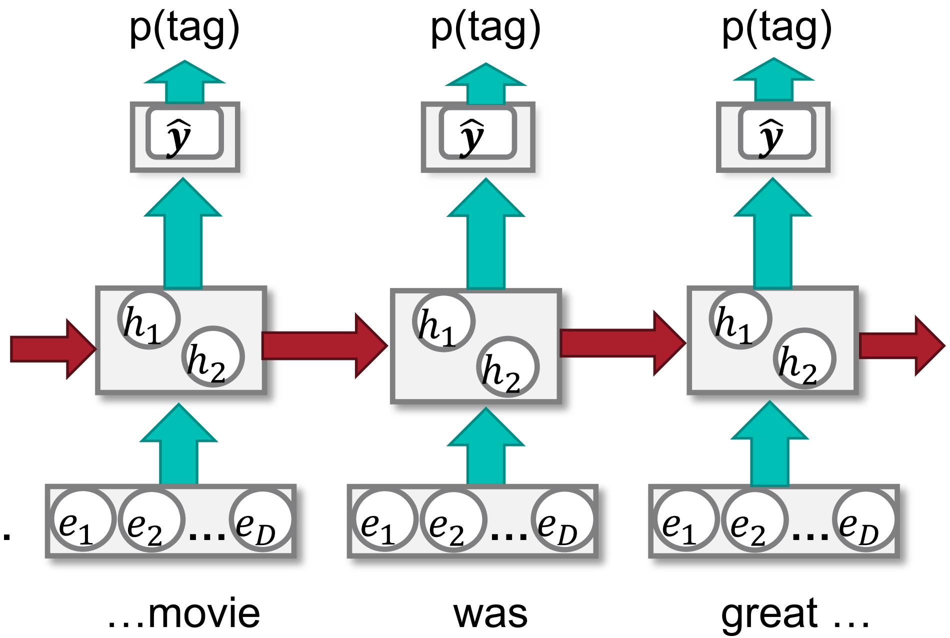
Training: Backpropagation Through Time

Partial derivatives must be propagated backwards along the sequence



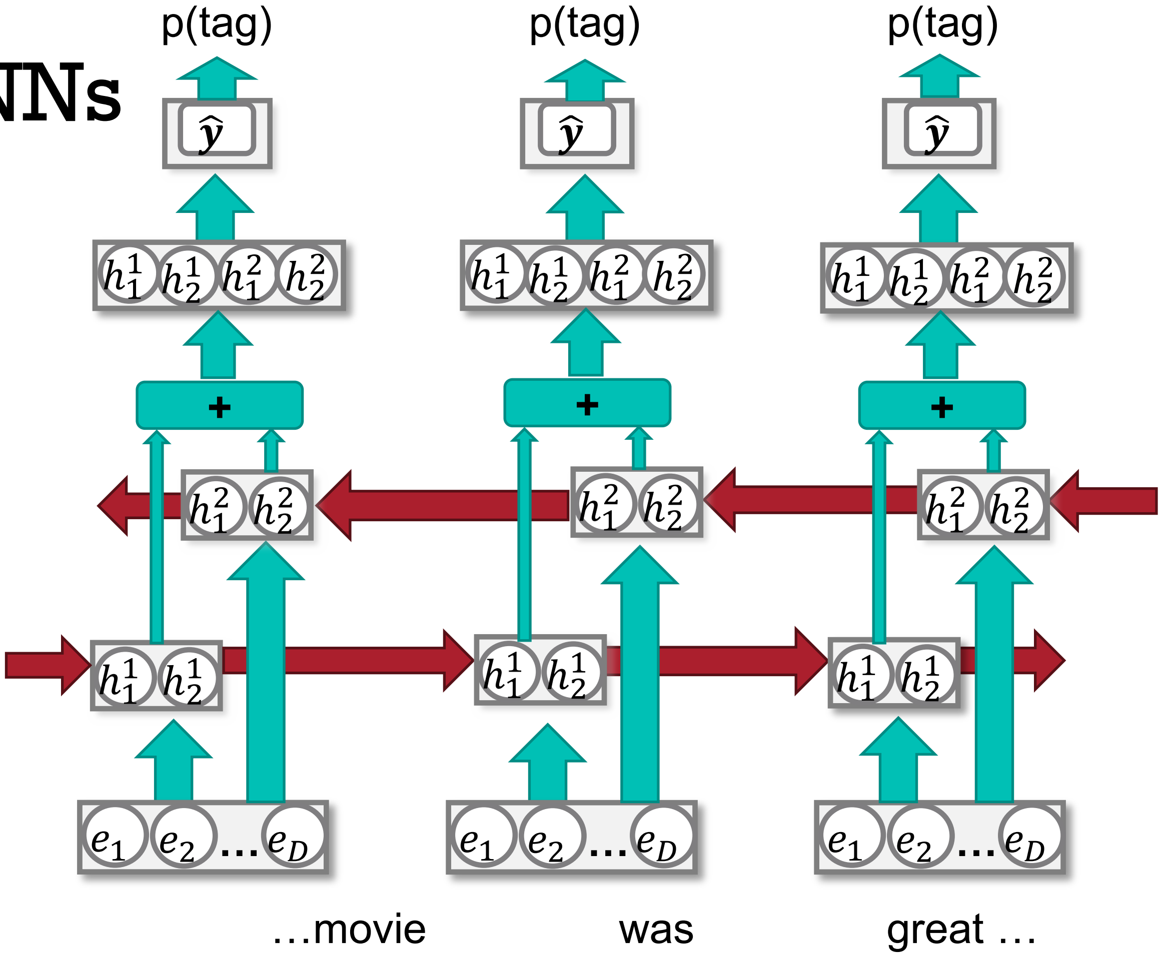
Inference: Uni-directional RNNs

- Forward inference passes information from left to right.
- The tag at time-step *i* depends only on previous tokens.
- But seeing later later tokens can help to choose the label for earlier ones.
- Viterbi algorithm for HMM also passed information backwards to select most likely label sequence.



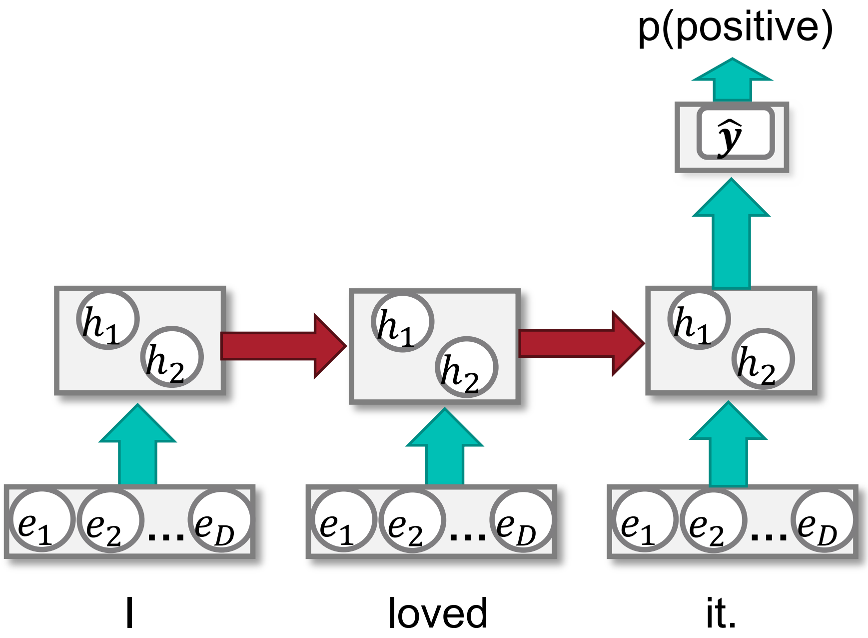
Bi-Directional RNNs

- Introduce a second RNN layer, which runs from right to left.
- Concatenate the hidden states from both layers as input to the next layer.
- Training: backprop for the backward layer passes derivatives from left to right.



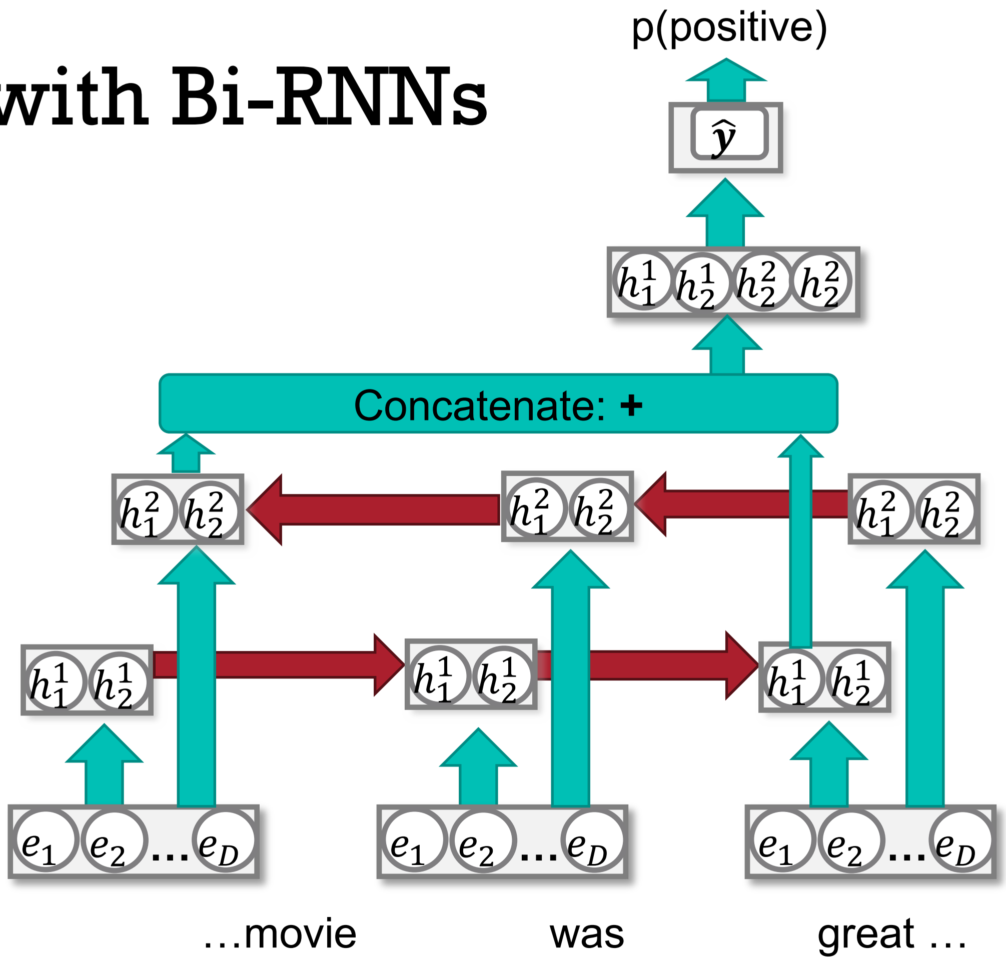
Text Classification with RNNs

- How can we use an RNN to classify an entire document or sentence?
- The last hidden state is taken as a representation of the whole sequence.
- Only the final hidden state is passed to the output layer.



Text Classification with Bi-RNNs

- How can we use an RNN to classify an entire document or sentence?
- The last hidden state is taken as a representation of the whole sequence.
- Only the final hidden state is passed to the output layer.



Summary

- Sequential processing must rely on syntactic structure rather than the position of features within a document.
- Recurrent neural networks (RNNs) have an additional input connection, which is the activation of the previous time-step.
- This allows them to pass information in one direction during inference.
- Bi-directional RNNs concatenate the hidden states (RNN activations) of two RNNs running in opposite directions.
- The final hidden states can represent the whole sequence.