Deep learning for sequence processing: Section 9.6, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2020).

8.3 LSTMs

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

Long-distance Information

- Long-distance information is important: consider these examples.
- Named entity recognition:

The seminar on the actual practice of tax reform was held in Hong Kong.

EVENT

LOC

- Sentiment analysis:

This tool manages to be suitable for both beginners and experienced users, which is very difficult to achieve.

Label=Positive

NER example: Jie, Z., & Lu, W. (2019). Dependency-Guided LSTM-CRF for Named Entity Recognition. EMNLP

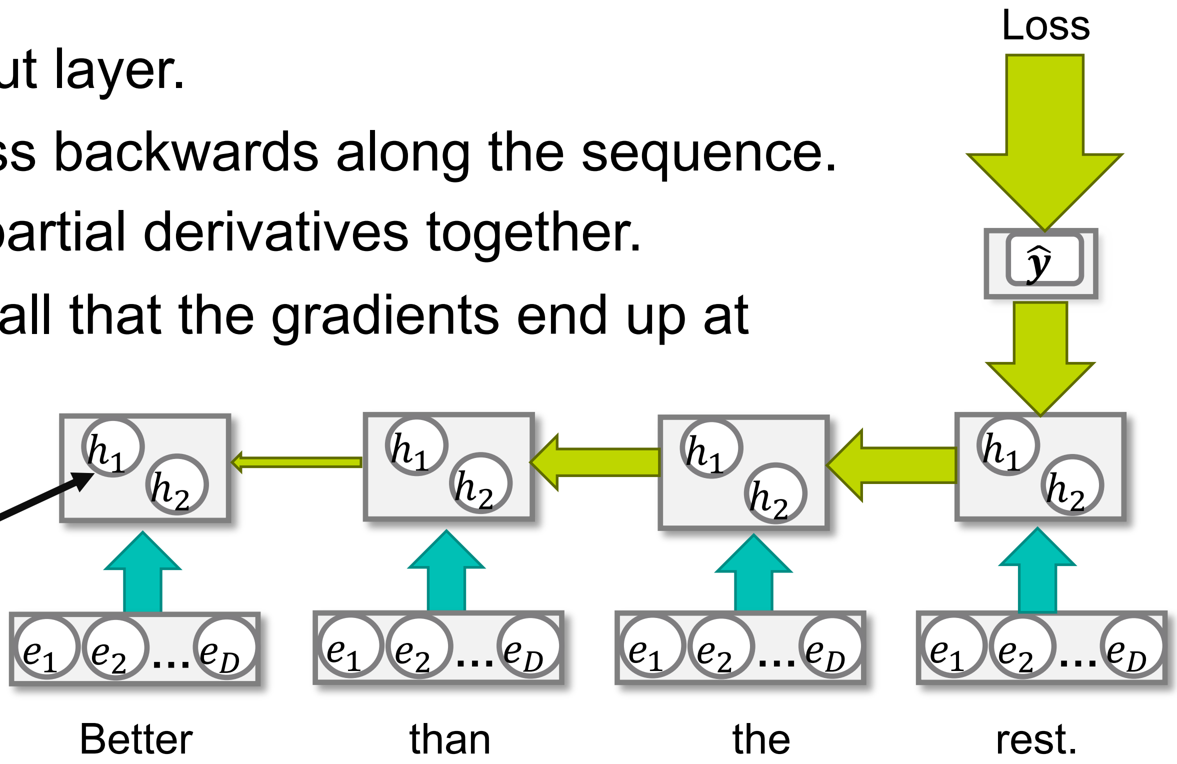
Long-distance Information and RNNs

- Standard RNNs do not perform well on tasks that require distant information.
- The hidden state has to do two things:
  1. Store context information that might be useful for future decisions.
  2. Represent information about the current time-step.
- Information stored in the hidden state can get overwritten or weakened as it passes through several time-steps.

Vanishing Gradients

Backpropagation Through Time

- Compute loss at output layer.
- Partial derivatives pass backwards along the sequence.
- Each step multiplies partial derivatives together.
- The values are so small that the gradients end up at zero…

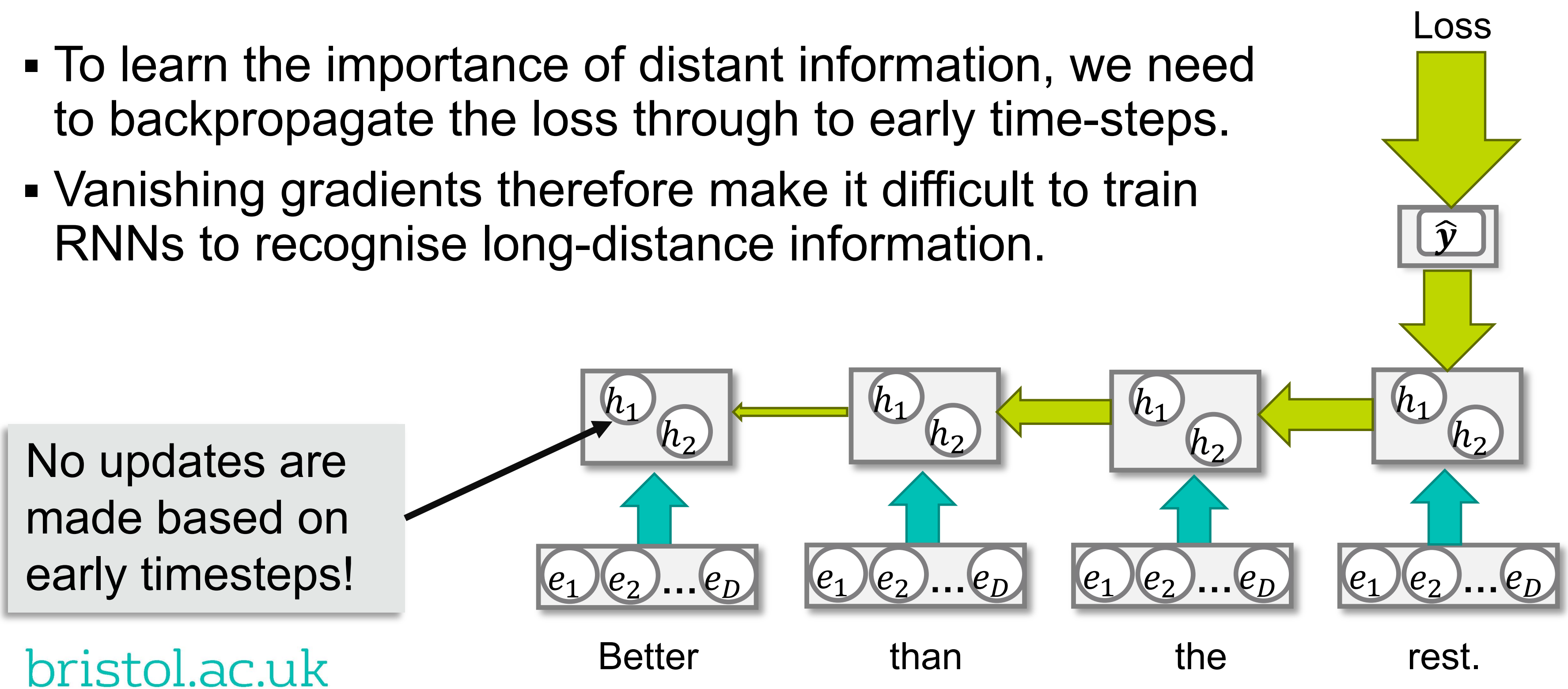


No updates are made based on early timesteps!

Vanishing Gradients

Backpropagation Through Time

- To learn the importance of distant information, we need to backpropagate the loss through to early time-steps.
- Vanishing gradients therefore make it difficult to train RNNs to recognise long-distance information.



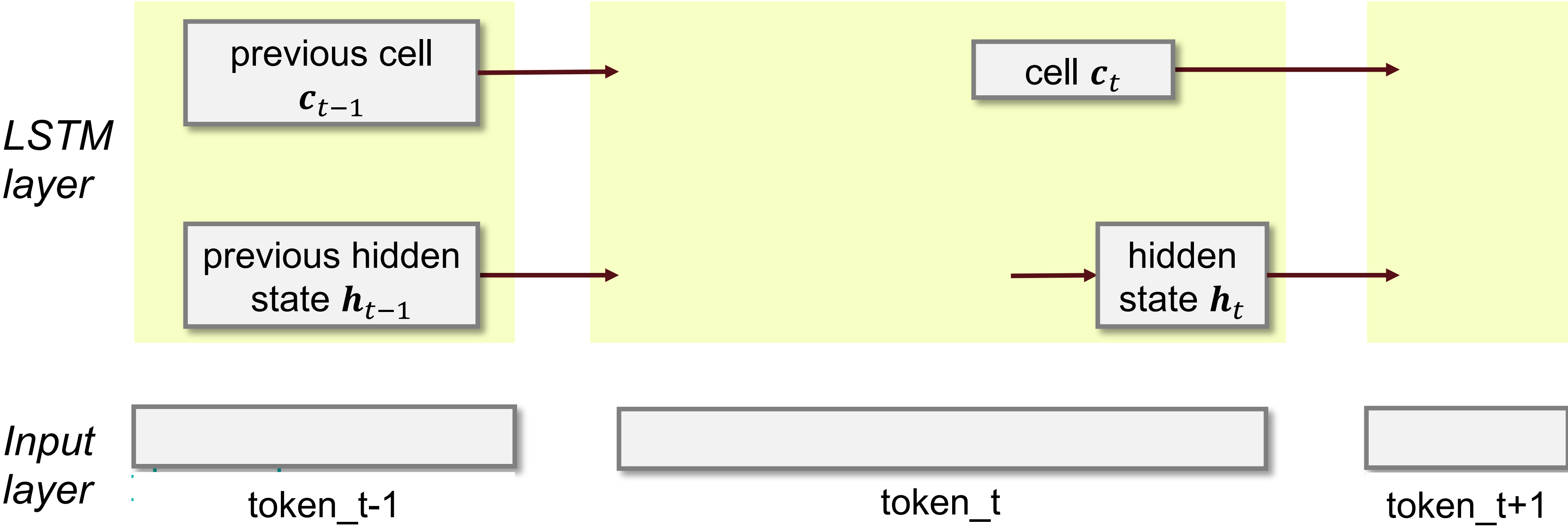
Long Short-Term Memory(LSTM)

- The LSTM redesigns the standard RNN
- It separates the **context** from the **current hidden state**.
  - Introduces a **memory cell** to store context in addition to the hidden layer.
- The model learns three sub-problems for managing context:
  - When to memorise context information,
  - When to forget context information,
  - When to use context information.
  - These tasks are managed by **gates**.

Understanding LSTM Networks:
https://colah.github.io/posts/2015-08-Understanding-LSTMs

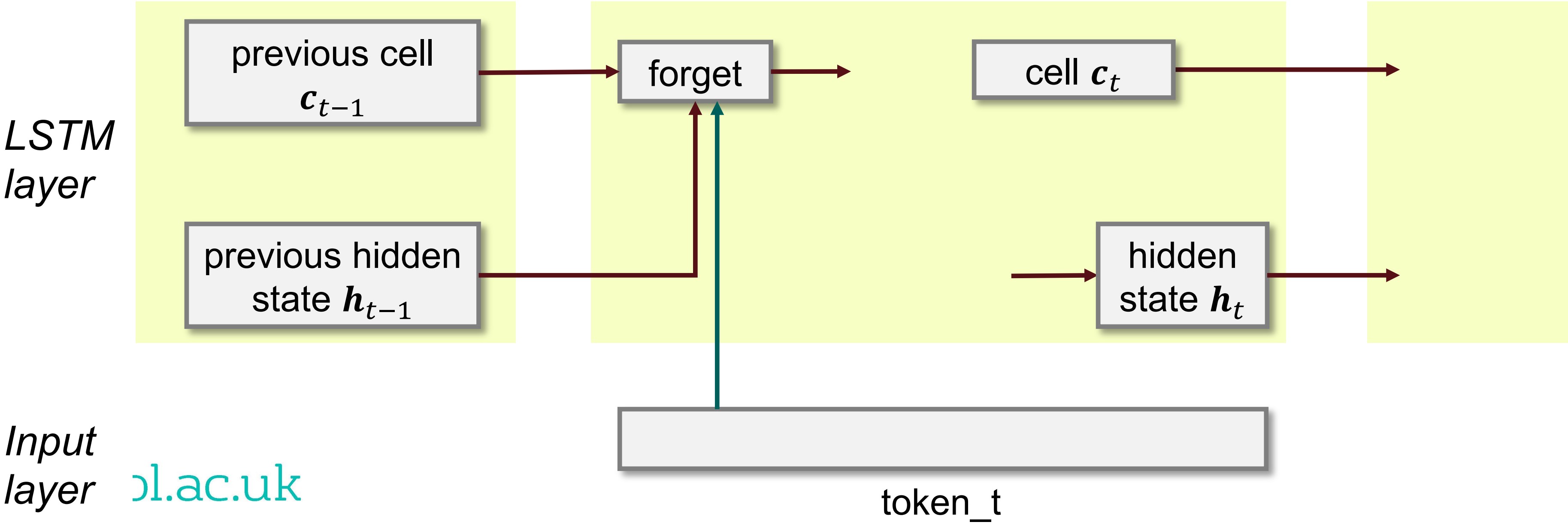
LSTM: Memory Cell

The memory cell is a vector that stores context information from earlier in the sequence.



LSTM: Forget Gate

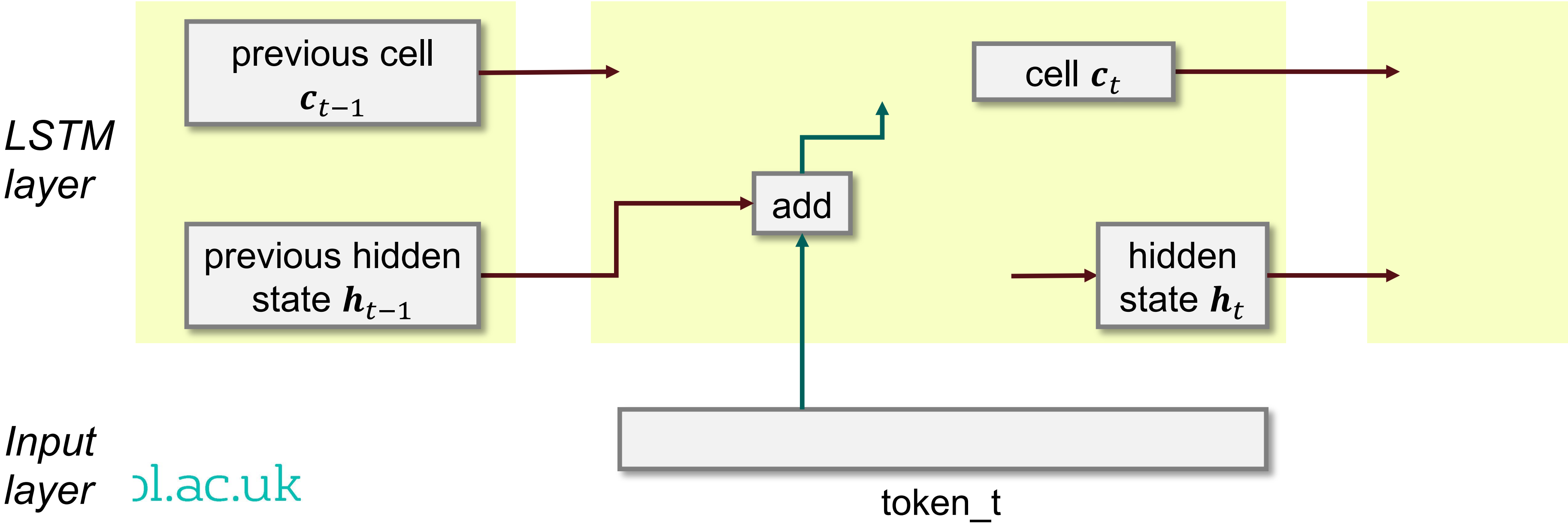
The forget gate erases information from the cell depending on the current input and previous hidden state.



ol.ac.uk

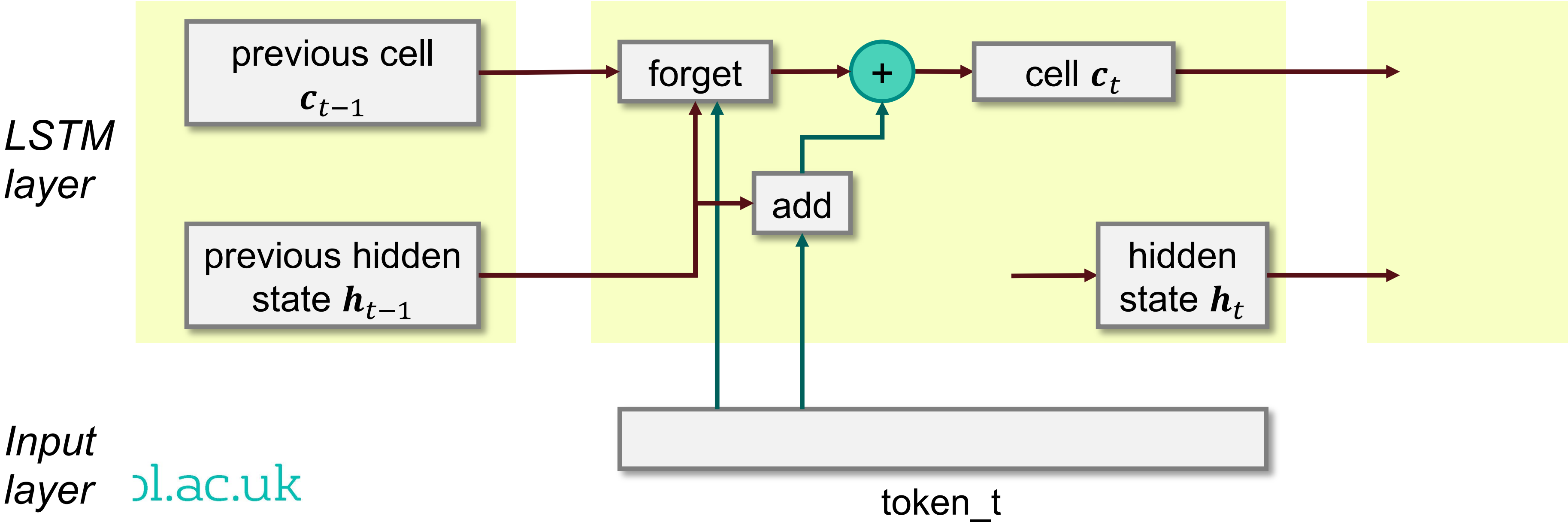
LSTM: Add Gate

The add gate selects new information to add to the cell based on the current input and previous hidden state.



ol.ac.uk

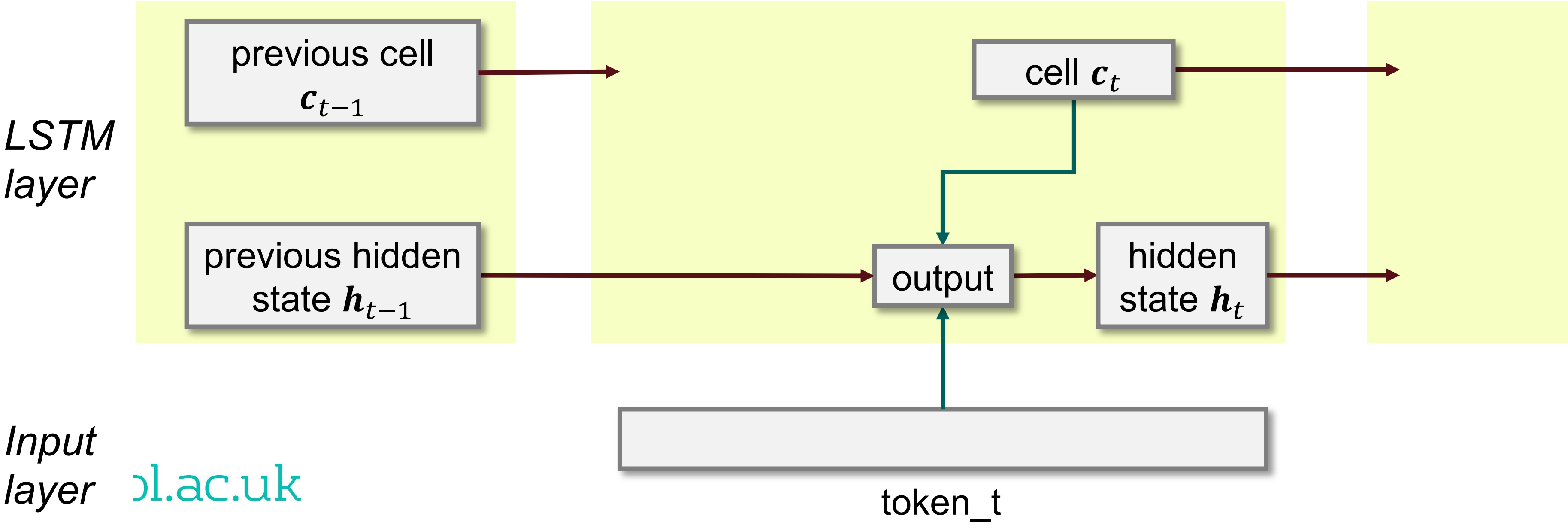
LSTM: Cell Update



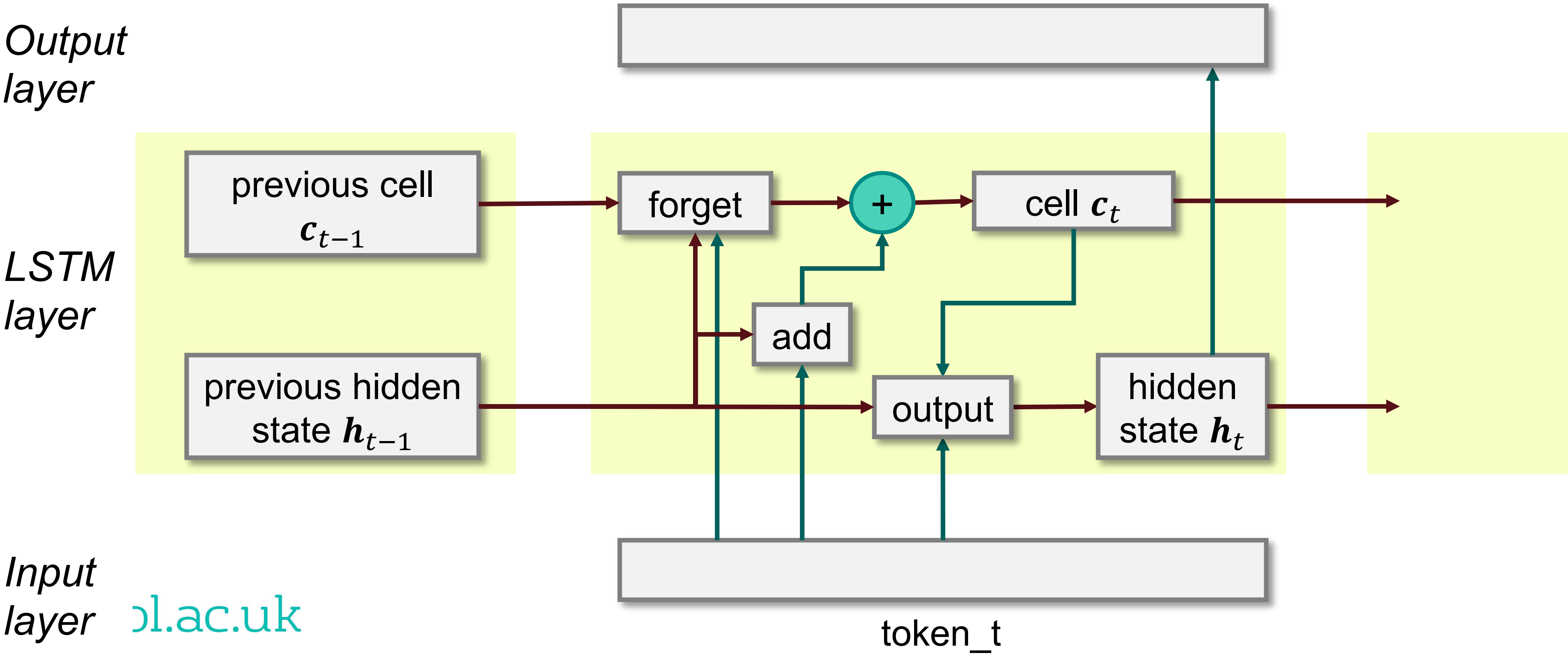
ol.ac.uk

LSTM: Output Gate

The output gate selects information from the cell to combine with the current input and previous hidden state.



LSTM: Complete Picture



ɔl.ac.uk

Gates

- The gates all work in the same way:

$$gateOutput_f = \sigma(\boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{W}_f \boldsymbol{x}_t) \odot gateInput_f$$

Looks like standard recurrent layer with weights $\boldsymbol{U}_f$ and $\boldsymbol{W}_f$.

This is multiplied elementwise with the gate's input to filter the input vector.

- The inputs for each gate are:
  - Forget gate: previous cell vector, $\boldsymbol{c}_{t-1}$.
  - Add gate: the usual computation for a recurrent layer, $g(\boldsymbol{U}\boldsymbol{h}_{t-1} + \boldsymbol{W}\boldsymbol{x}_t)$.
  - Output gate: current cell vector passed through an activation function, $g(\boldsymbol{c}_t)$.

Applications of LSTMs

- Bi-directional LSTMs are the most widely used type of RNN across all kinds of NLP tasks that needs to process text sequentially.
- For example:
  - Sequence labelling tasks like named entity recognition and PoS tagging.
  - Relation extraction where the context of entity mentions provides important information.
  - Sentiment analysis on long documents.
- LSTMs are widely used have practical limitations:
  - Sequential processing makes parallelisation hard.
  - The model of context is still very simplistic.

Summary

- Standard RNNs do not retain long-distance information, which is important for many tasks.
- LSTMs introduce a memory cell to store context across time-steps.
- A series of gates control how information is added, forgotten and output from the memory cell.
- BiLSTMs perform well on many tasks, especially sequence labelling, but do have practical limitations.
- The final lecture will explore some recent methods that can address some of these limitations.