Understanding Long Short-Term Memory (LSTM) networks

Dr. Franziska Boenisch

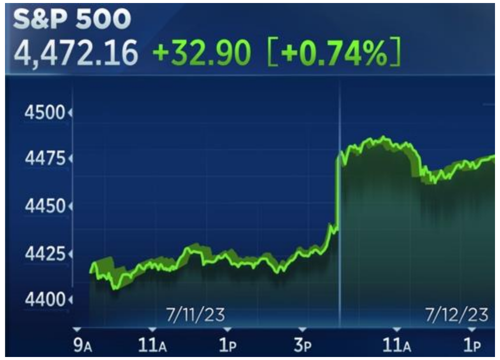
Outline for Today

1. Intuition on why we need LSTMs
2. Short recap of Recurrent Neural Networks (RNNs)
3. Overview on LSTMs
4. LSTM flow and examples
5. Strengths and limitations

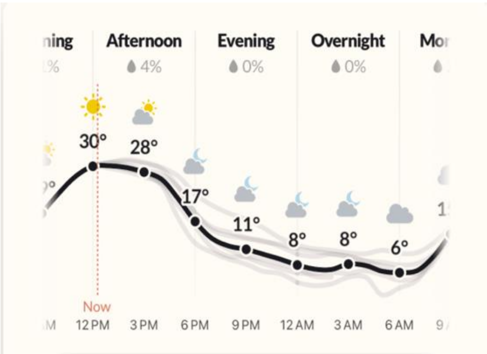
Sequence data is everywhere

Text

Music



Stock prices



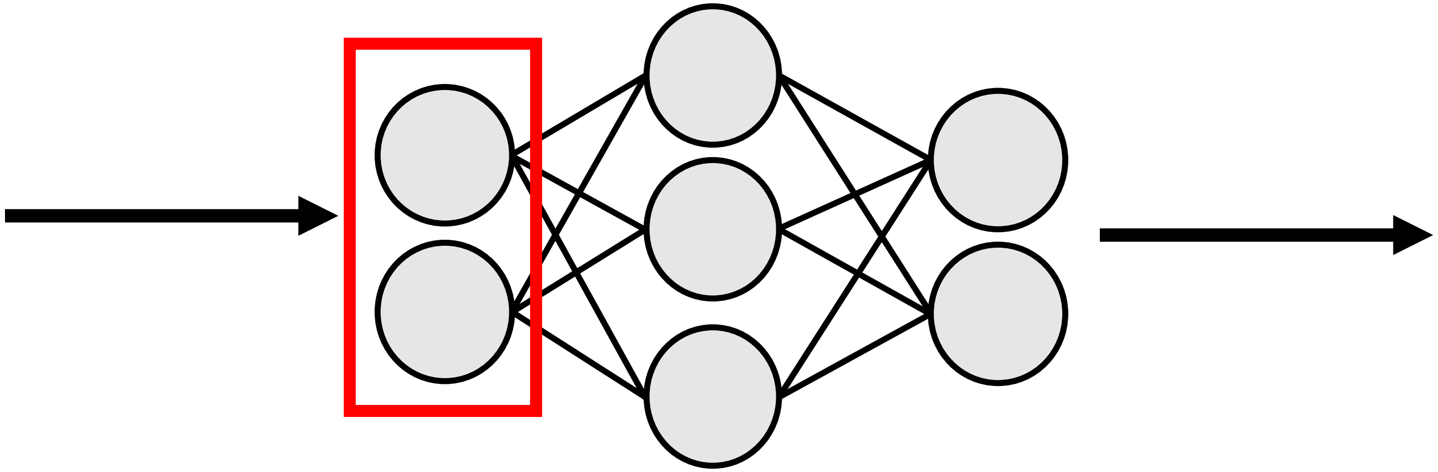
Weather

Needs a “memory” of what came before!

Limitations of neural networks for sequences

Traditional feed-forward networks struggle with sequences because they

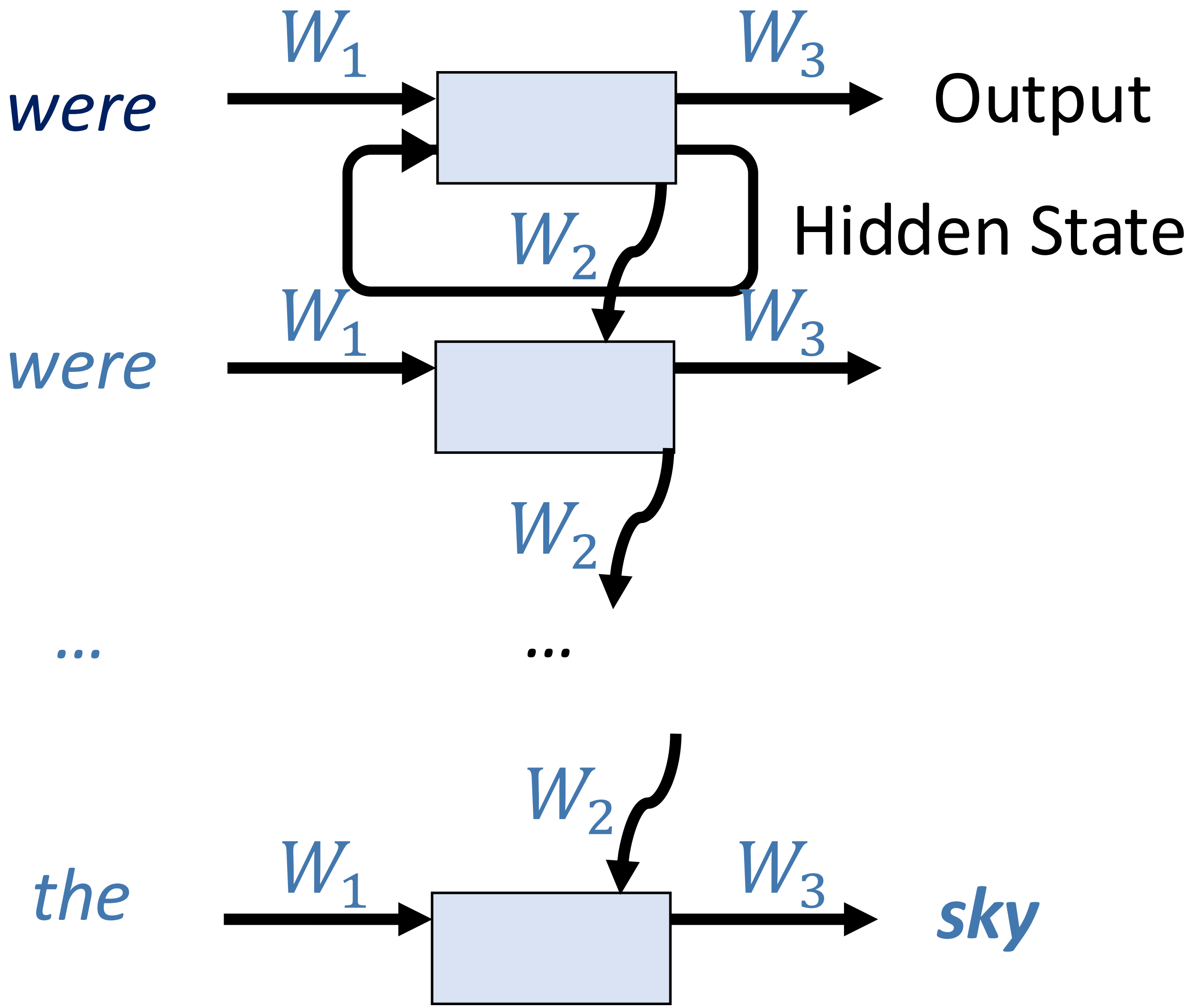
1. have no memory
2. can only process fixed-size inputs



“cat”

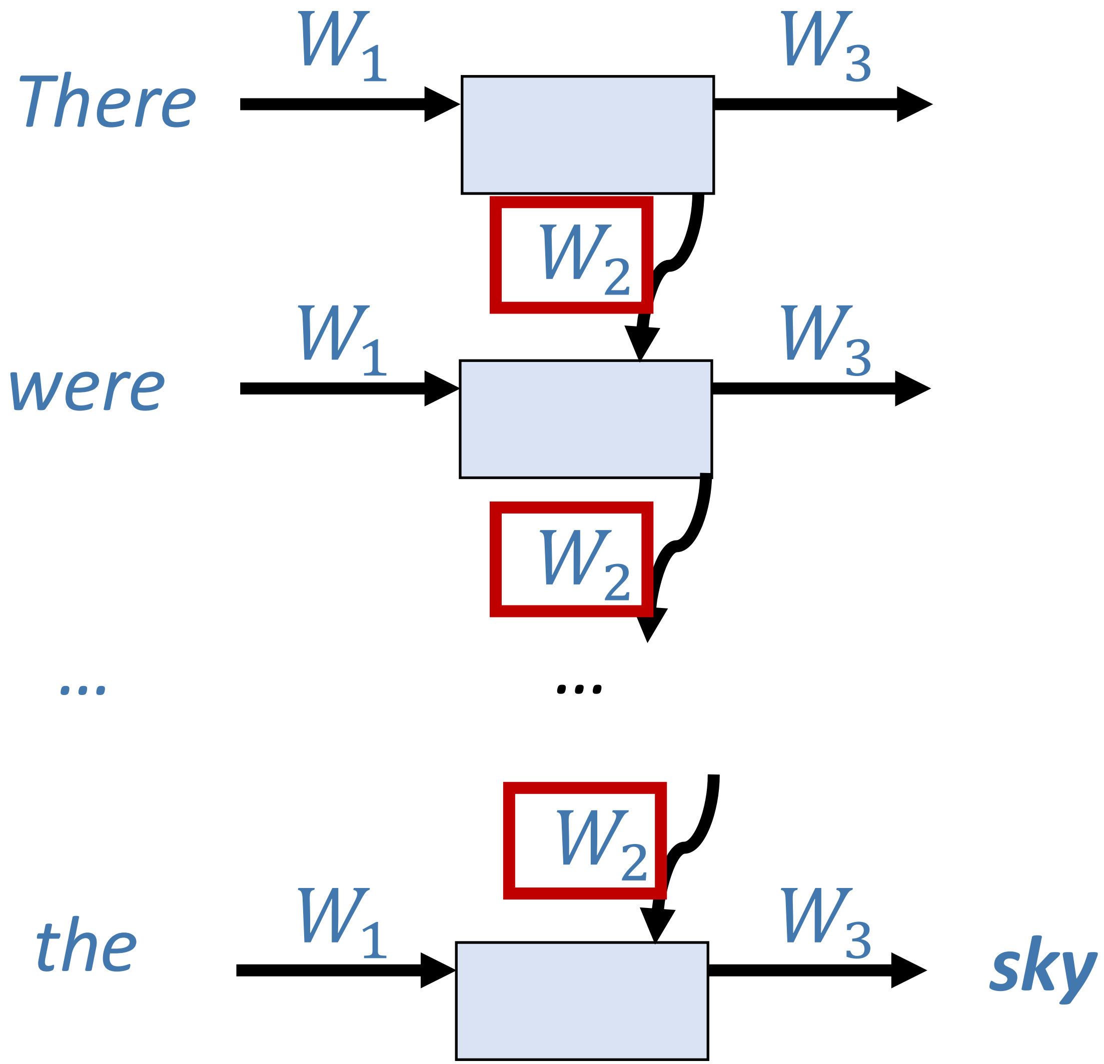
There were dark clouds in the night

Recurrent Neural Networks (RNNs)



Limitations of RNNs

*There were dark clouds in the night **sky***



Standard RNNs struggle with long sequences

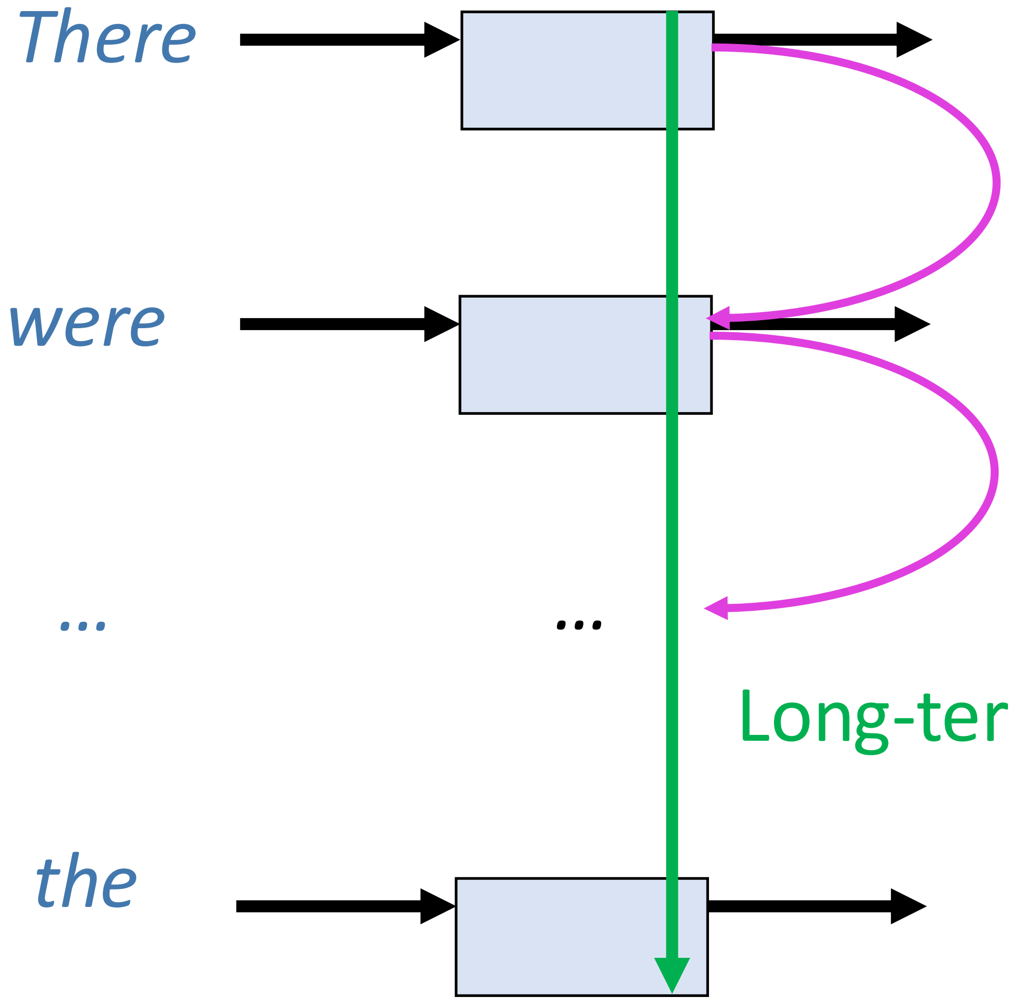
Problem of **exploding** and **vanishing** gradients

Imagine $W_2 = 5$, and sequence length $l = n$

→ 5^n, e.g. $5^{10} = 9765625$

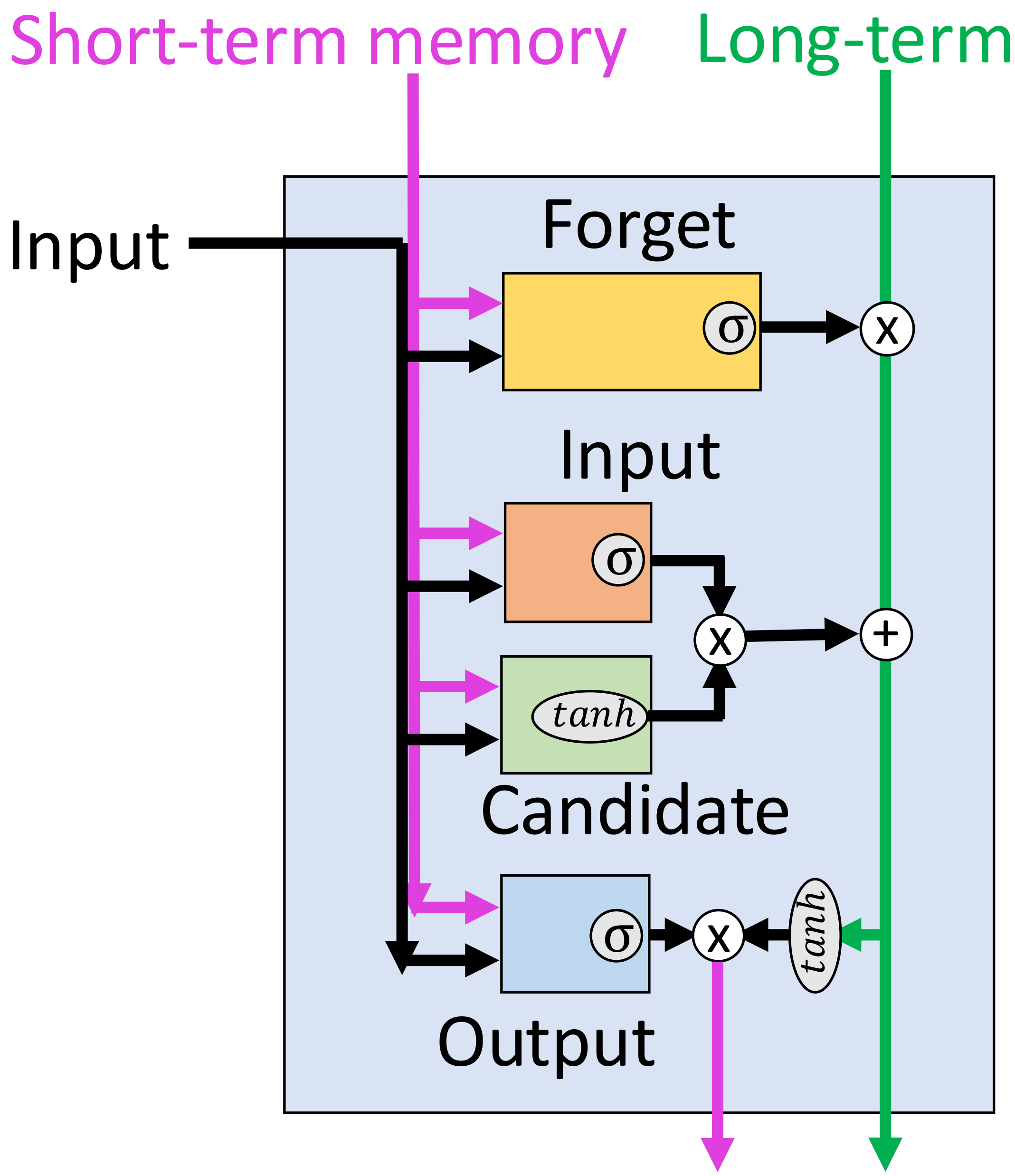
LSTM intuition

LSTMs introduce a separate cell state instead of continuously overwriting the hidden state.



1. What to forget from old memory?
2. What new information to store?
3. What to reveal as output now?

Anatomy of an LSTM cell



Forget gate: % of long-term memory to remember

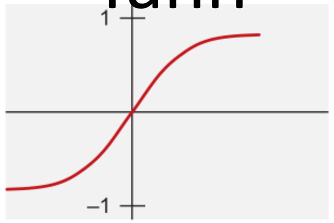
Input gate: % of candidate information from current input to remember

Candidate gate: candidate information from current input

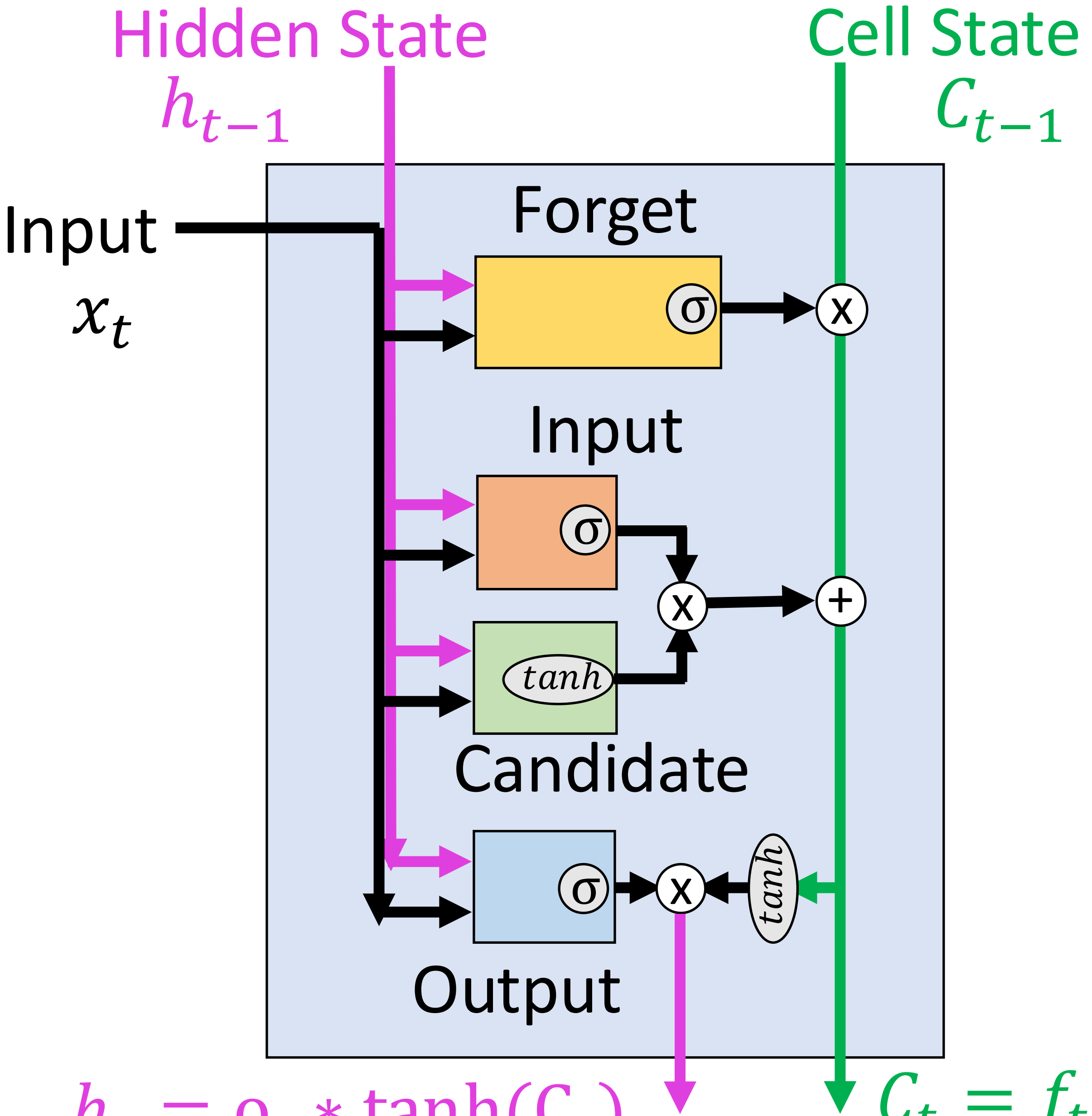
Output gate: % of current memory to affect subsequent layers

Sigmoid

Tanh



LSTM equations and step-by-step flow



$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = tanh(W_C[h_{t-1}, x_t] + b_C)$$

$$o_t = \sigma(W_0[h_{t-1}, x_t] + b_o)$$

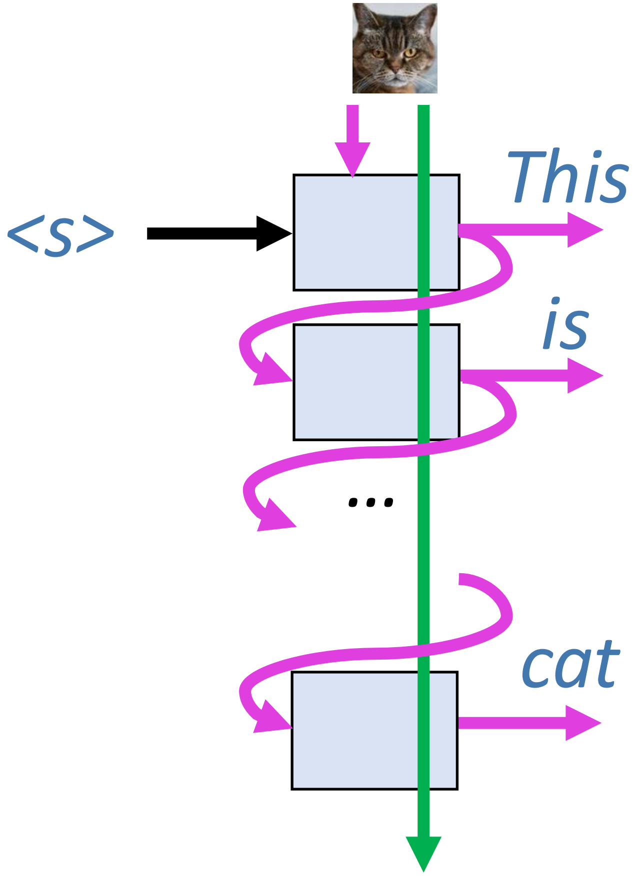
$$h_t = \mathrm{o_t} * \mathrm{tanh(C_t)}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

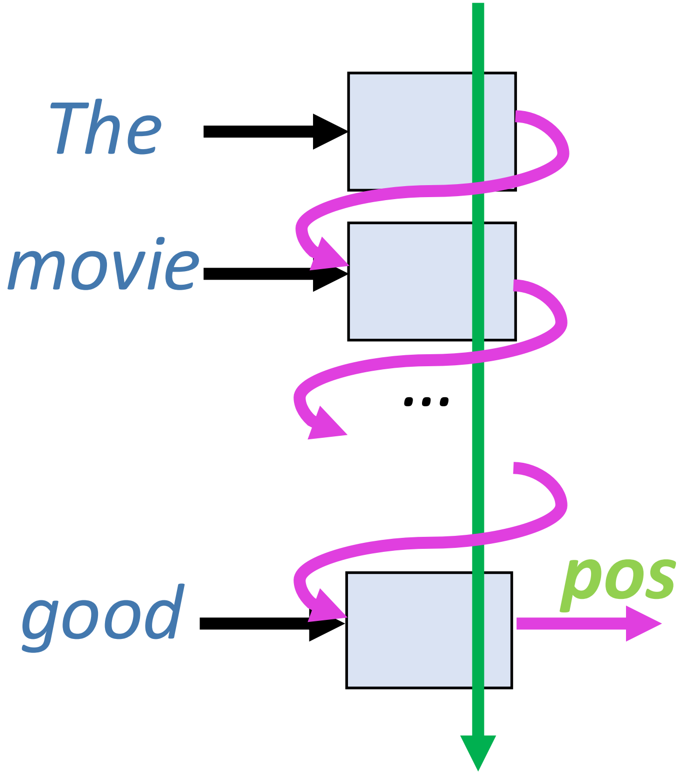
Example applications

Sequence generation	Sequence classification	Sequence translation
One-to-Many	*Many-to-One*	*Many-to-Many*

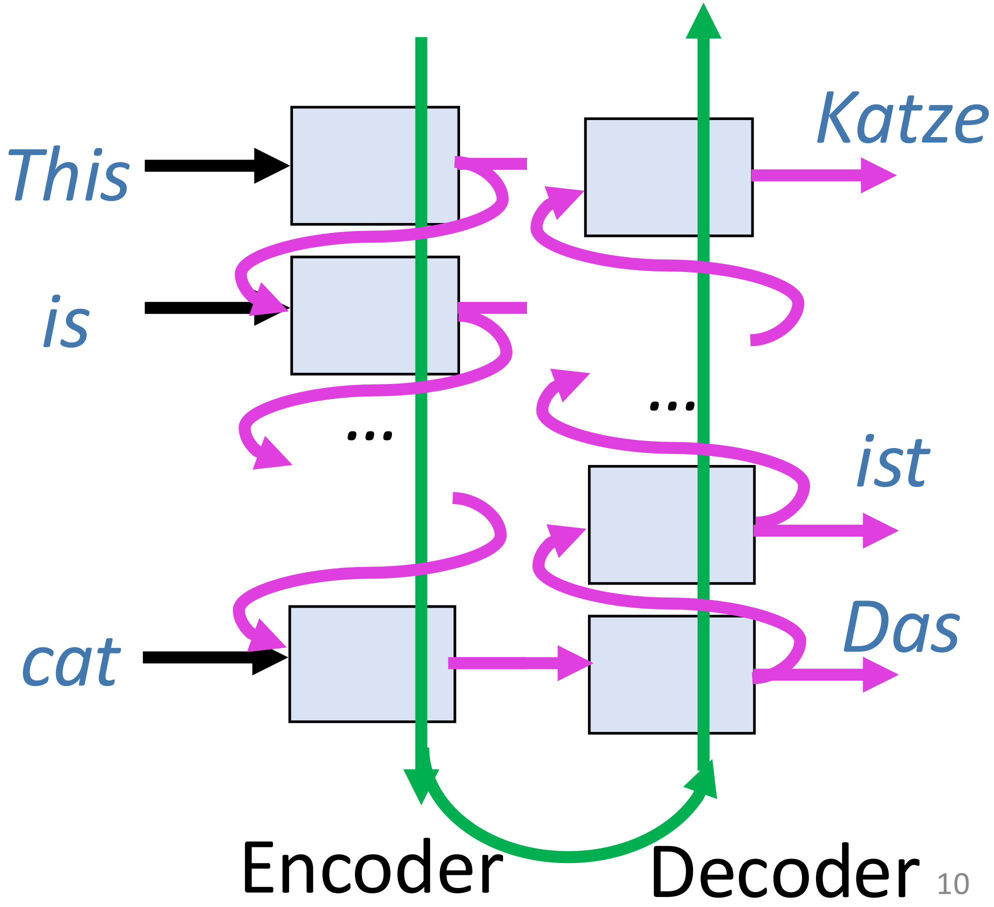
Image captioning



Sentiment analysis



Machine translation



Strengths and limitations

+

Process various-lengths sequences

Better long-t

dependenc

Diverse and practical tasks

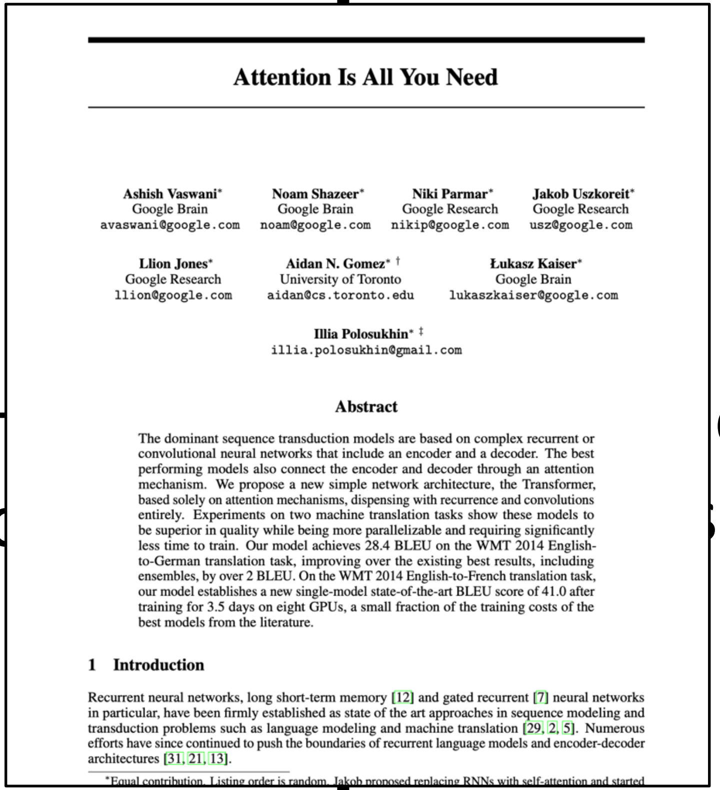
−

Hard to parallelize

e with very

sequences

Complex tuning



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

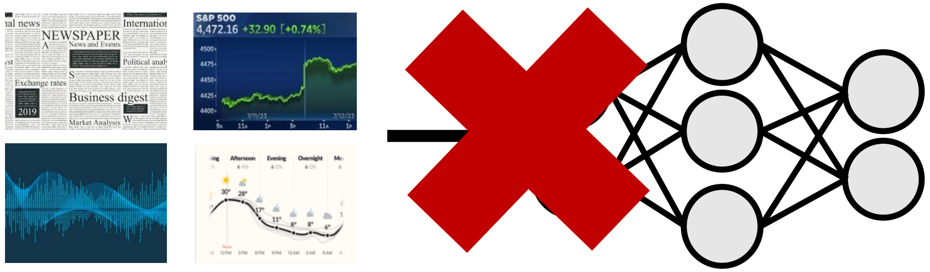
1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

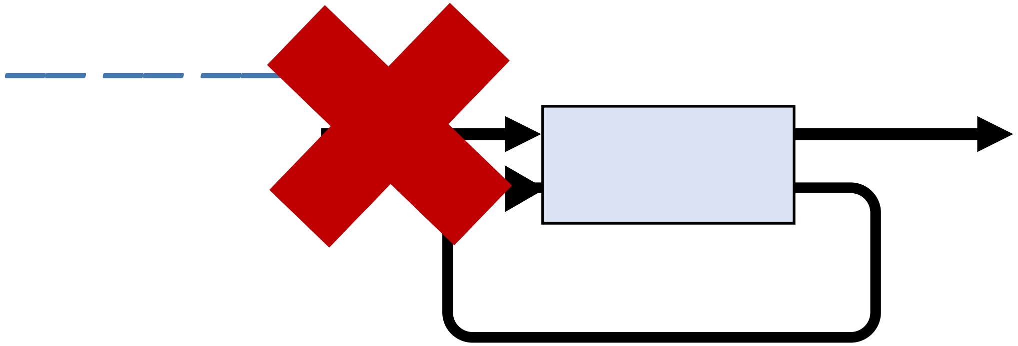
*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started

TRANSFORMERS!

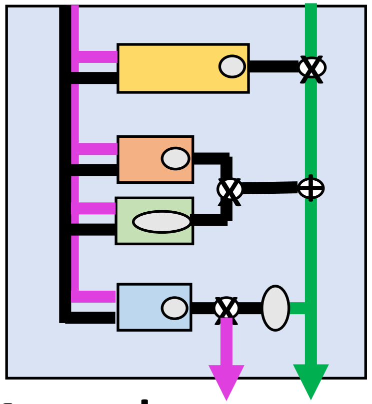
Summary



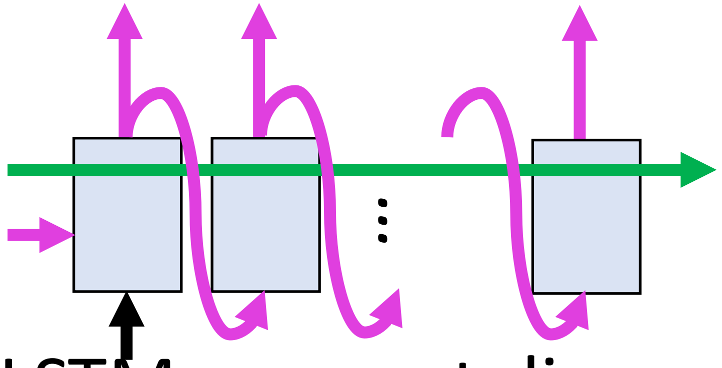
Standard networks unsuited for sequences



RNNs struggle with long-term dependencies



LSTMs rely on cell state and gates to improve



LSTMs support diverse practical tasks