RNN

Recurrent Neural Network

LSTM

Long Short-term Memory

李叶豪

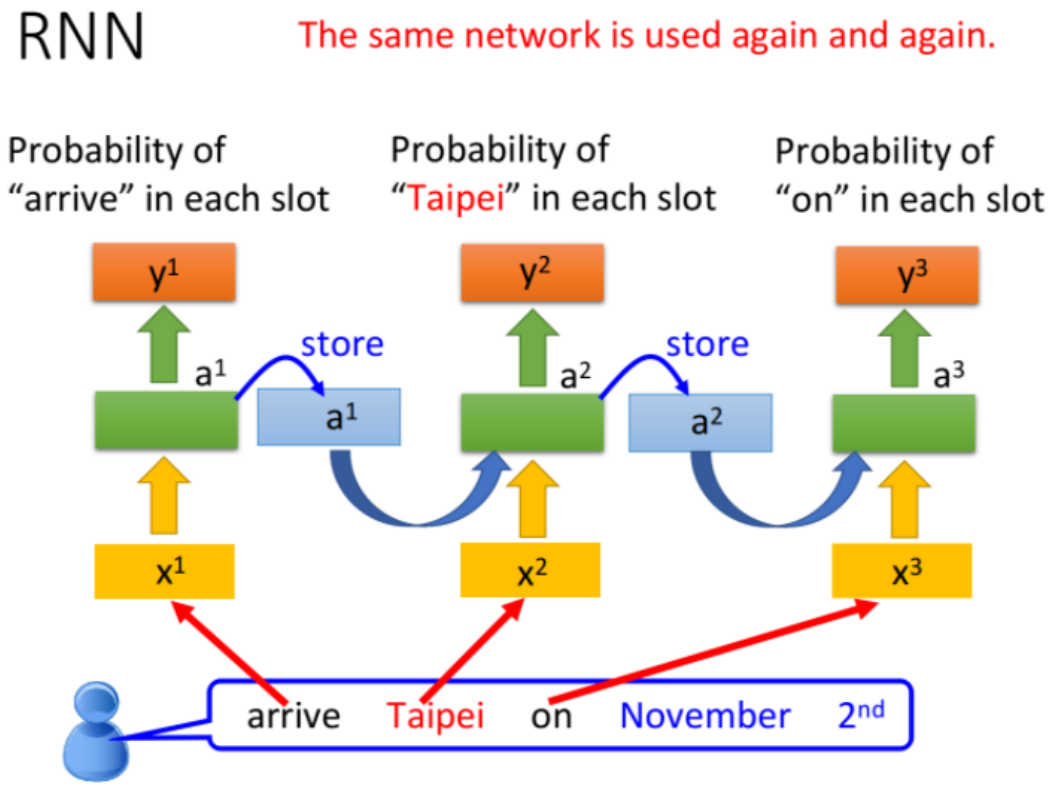
RNN

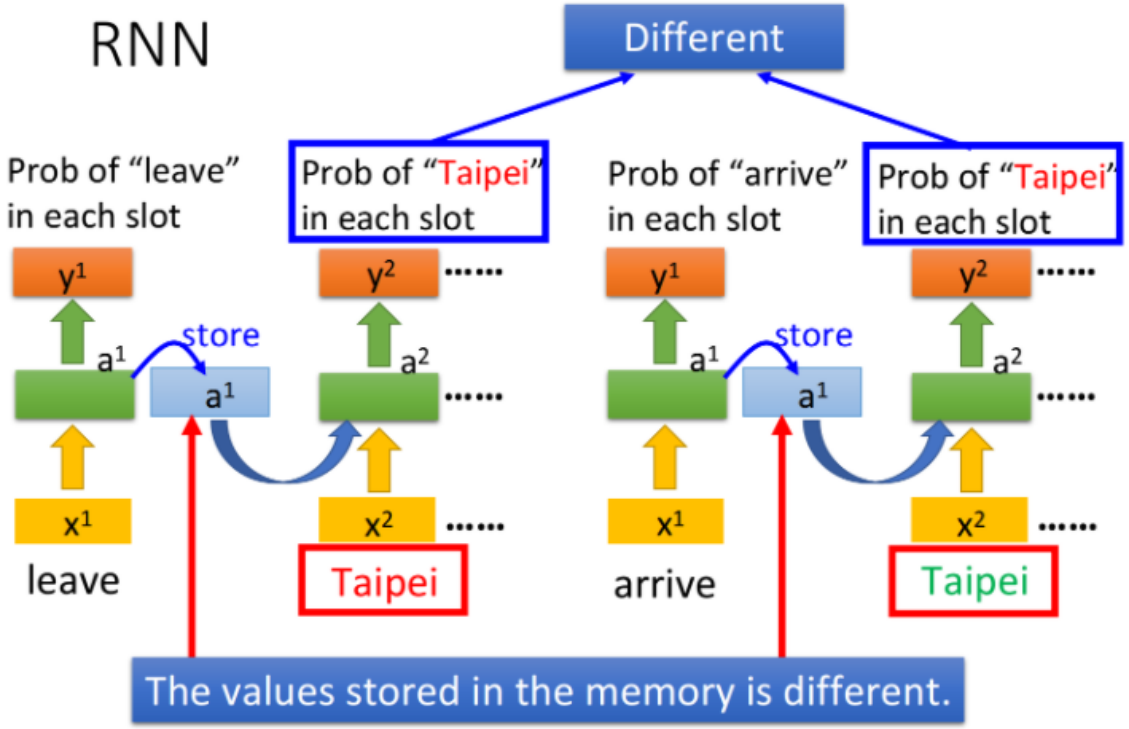
Recurrent Neural Network

RNN

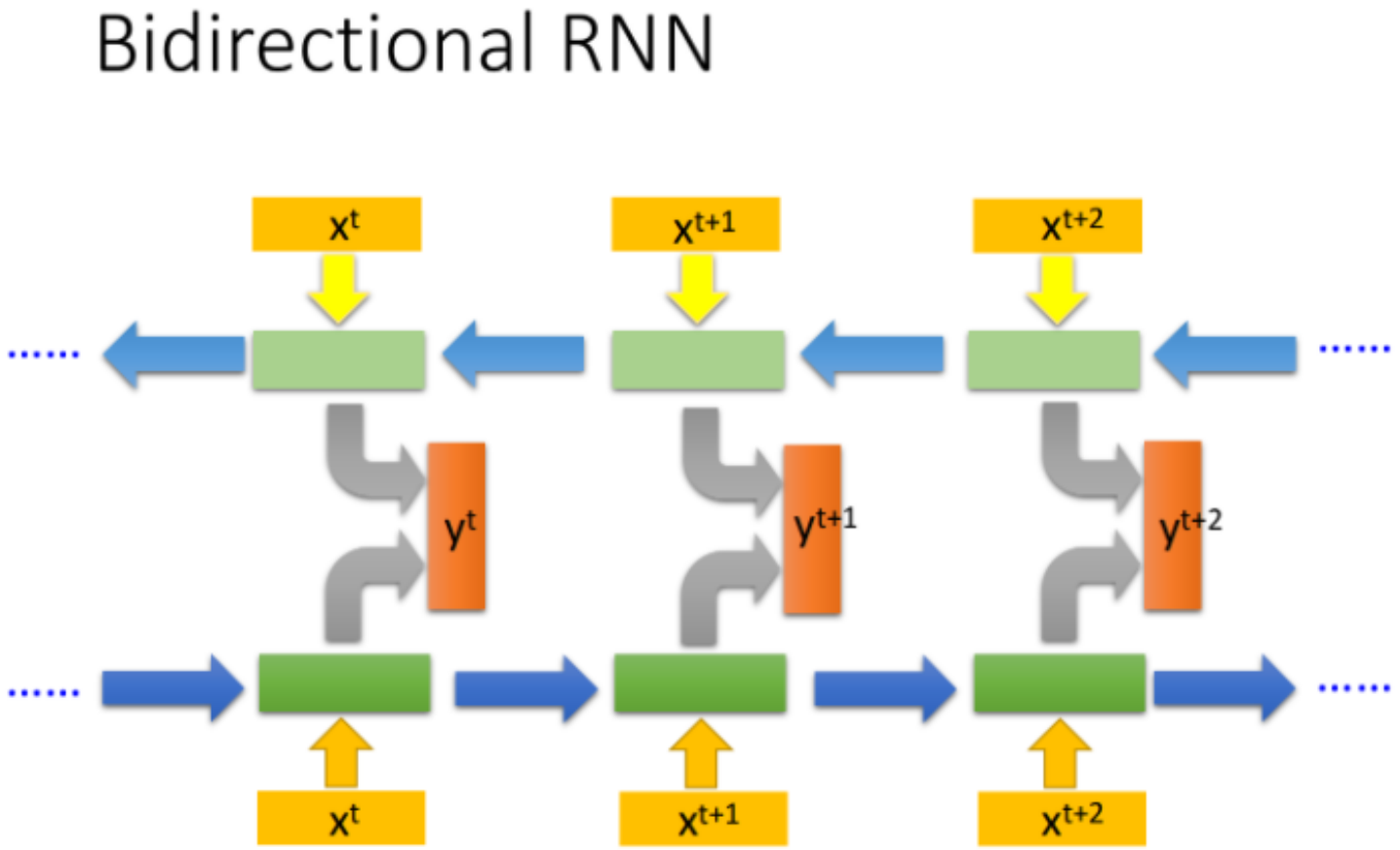
Context Relevance of Language

RNN

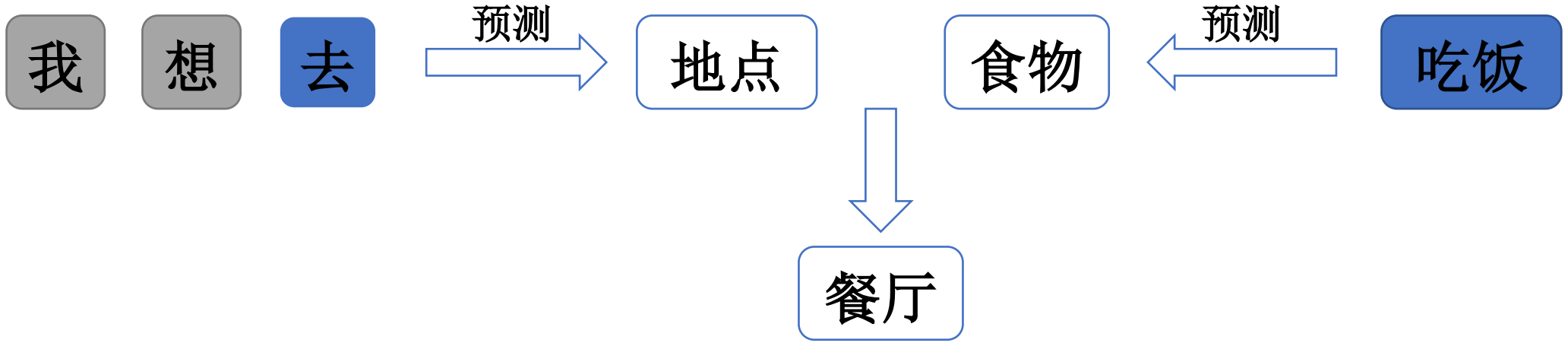

RNN
The same network is used again and again.
Probability of "arrive" in each slot
Probability of "Taipei" in each slot
Probability of "on" in each slot
y1
y2
y3
a1
a2
a3
store
store
a1
a2
x1
x2
x3
arrive
Taipei
on
November
2nd


RNN
Different
Prob of "leave" in each slot
Prob of "Taipei" in each slot
Prob of "arrive" in each slot
Prob of "Taipei" in each slot
y1
y2
a1
a2
store
x1
x2
leave
Taipei
arrive
Taipei
The values stored in the memory is different.

Bi-RNN

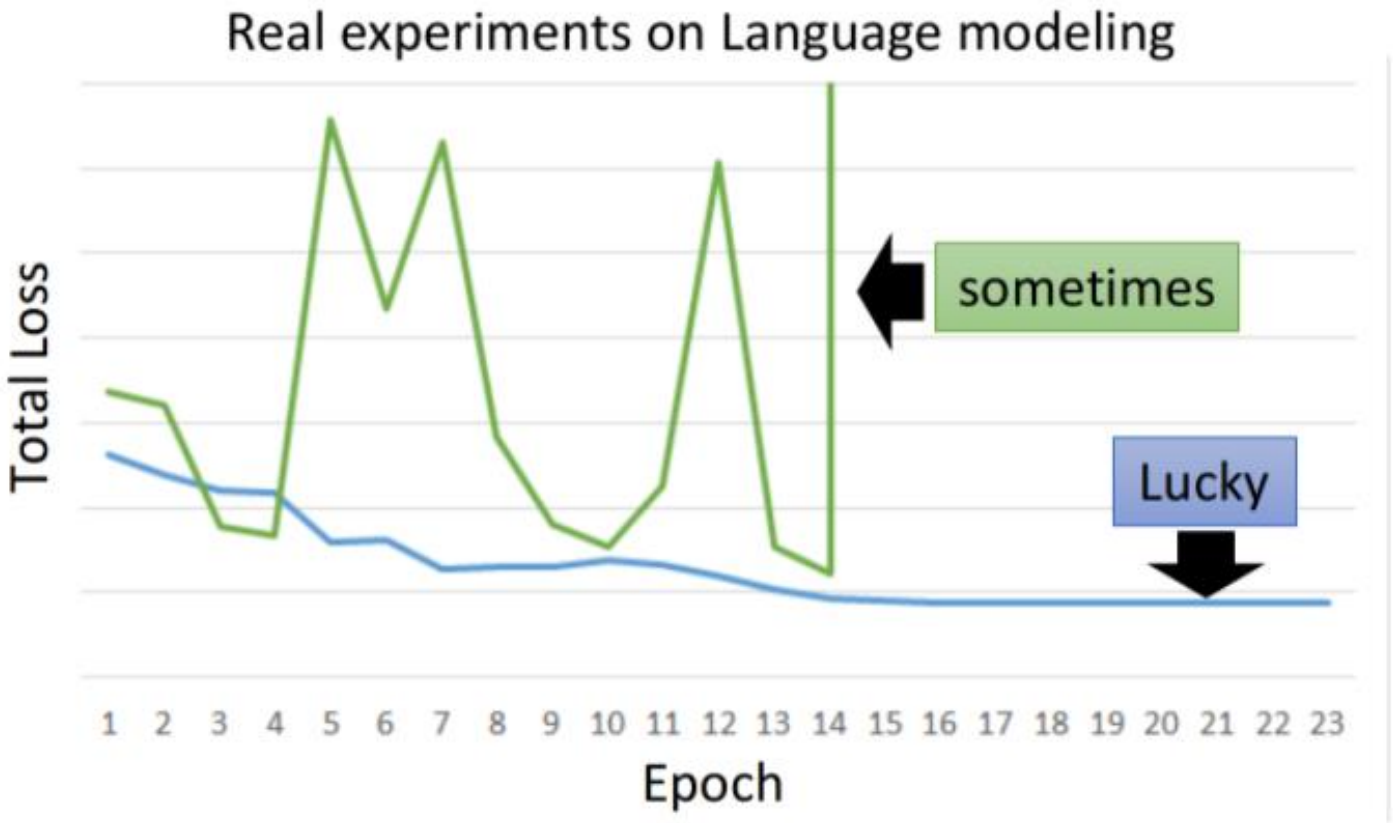


Bi-RNN



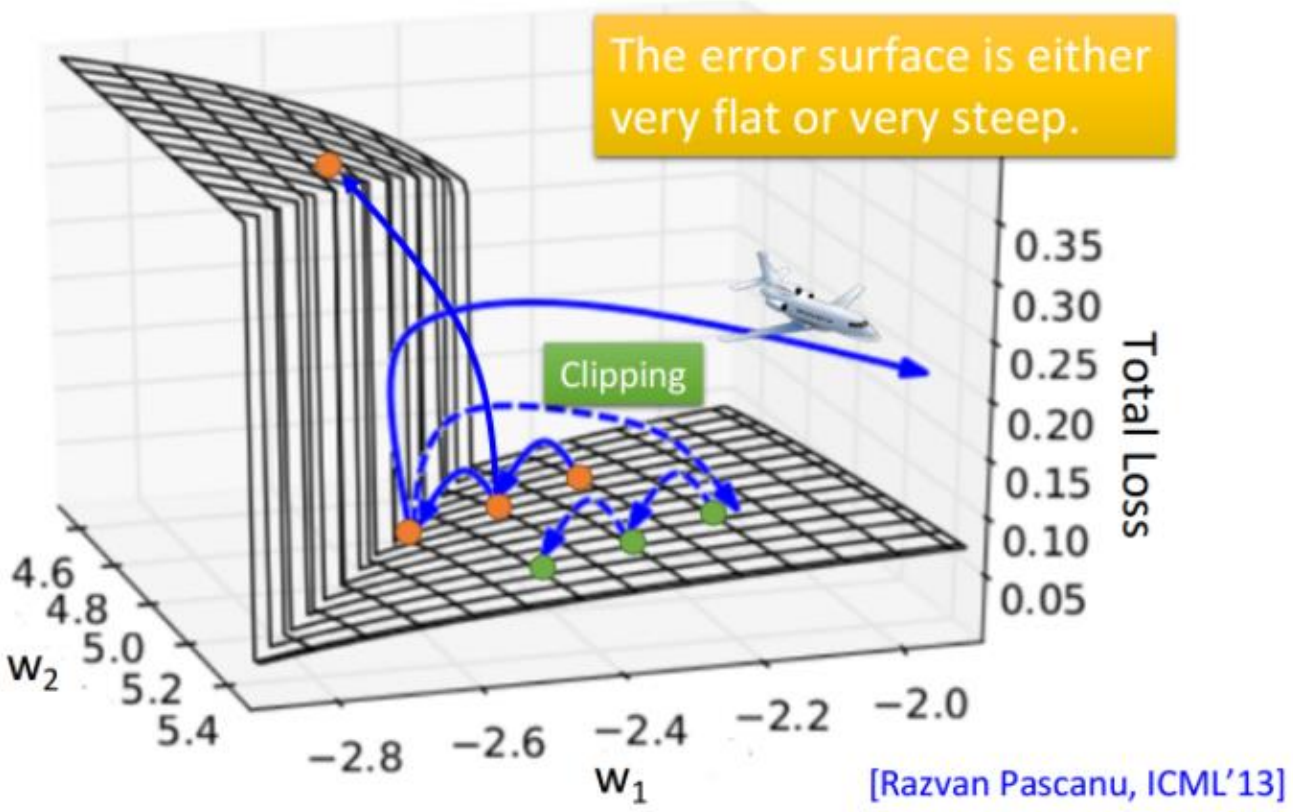
Gradient explosion

- RNN-based network is not always easy to learn



Gradient explosion

The error surface is rough.

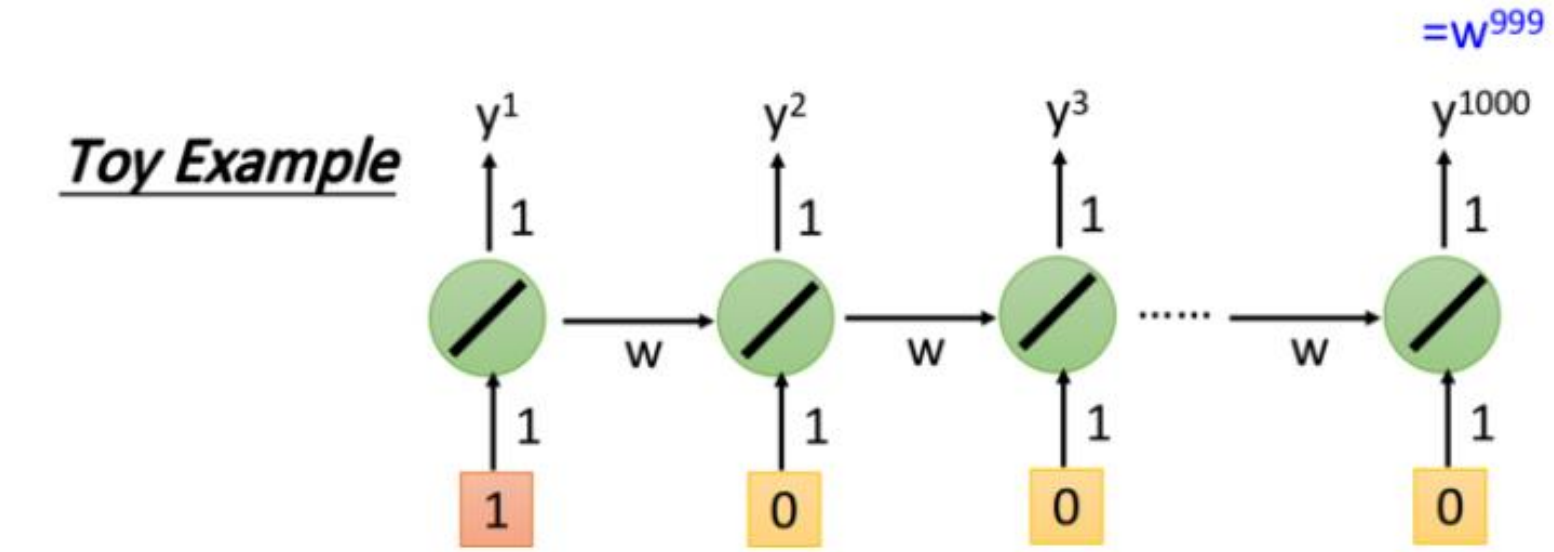


[Razvan Pascanu, ICML'13]

Gradient disappearance

The writer of the books + is/are

Why



$w = 1 \Rightarrow y^{1000} = 1$

$w = 1.01 \Rightarrow y^{1000} \approx 20000$

Large $\partial L/\partial w$ → Small Learning rate?

$w = 0.99 \Rightarrow y^{1000} \approx 0$

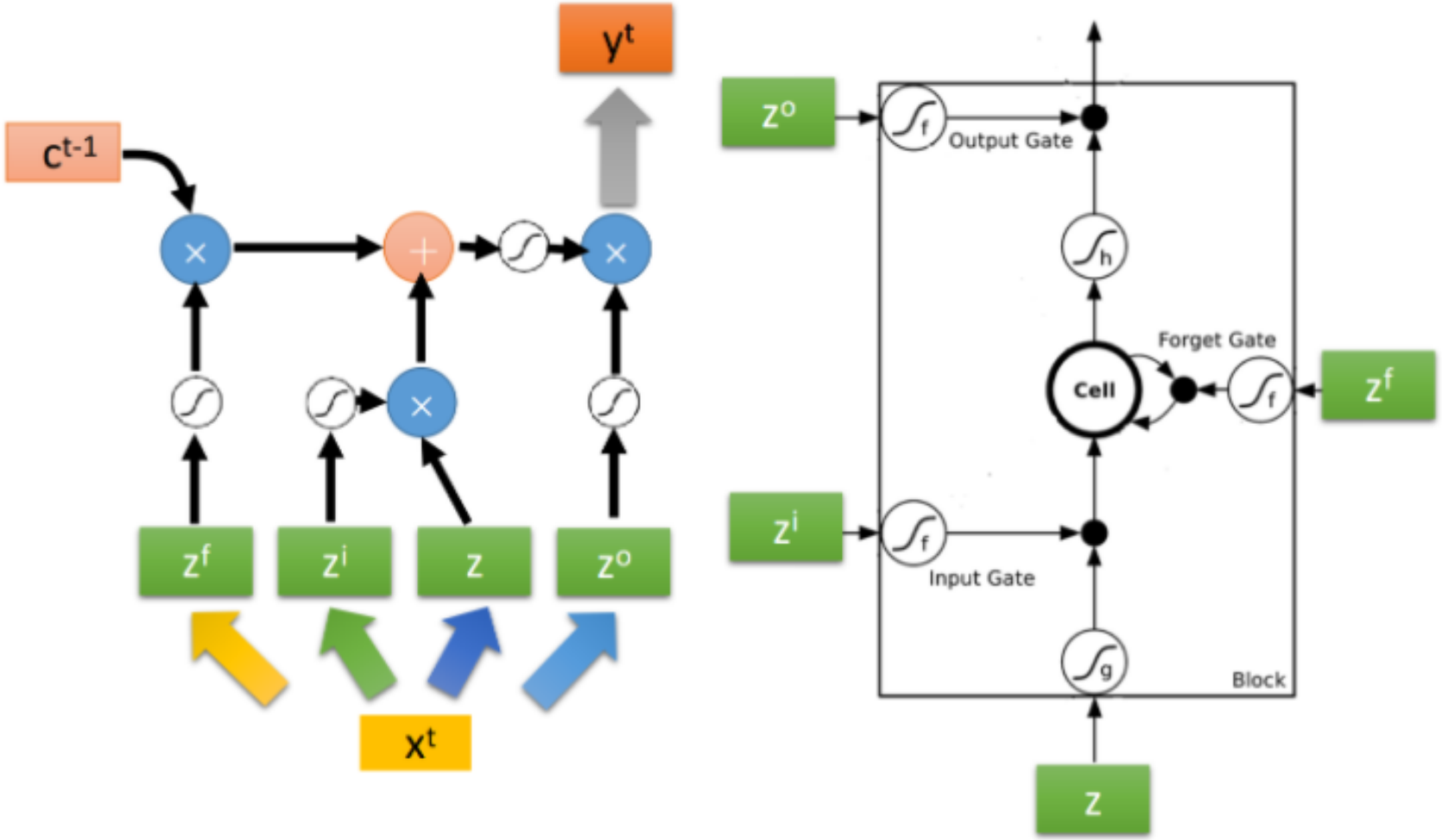
$w = 0.01 \Rightarrow y^{1000} \approx 0$

small $\partial L/\partial w$ → Large Learning rate?

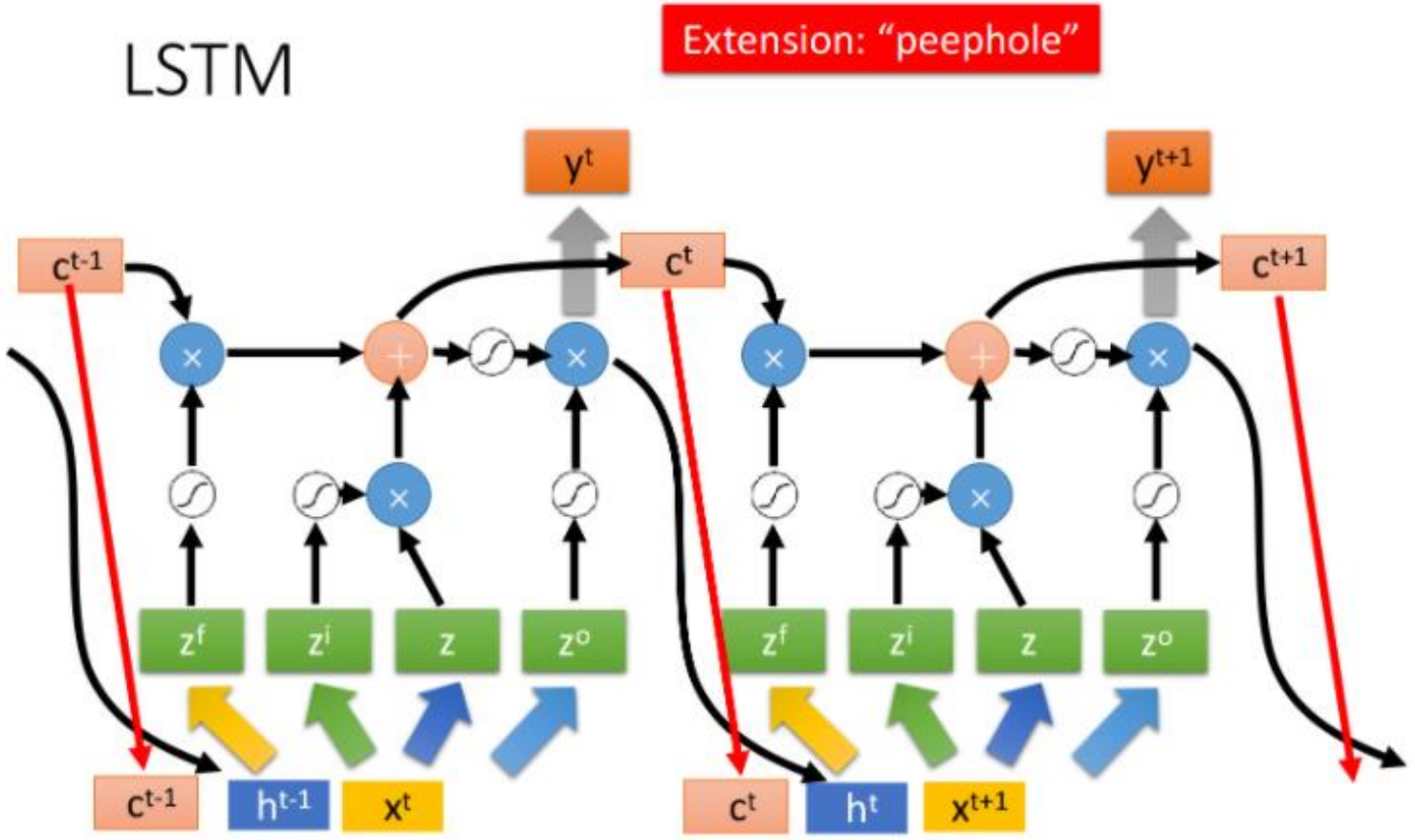
LSTM

Long Short-term Memory

LSTM

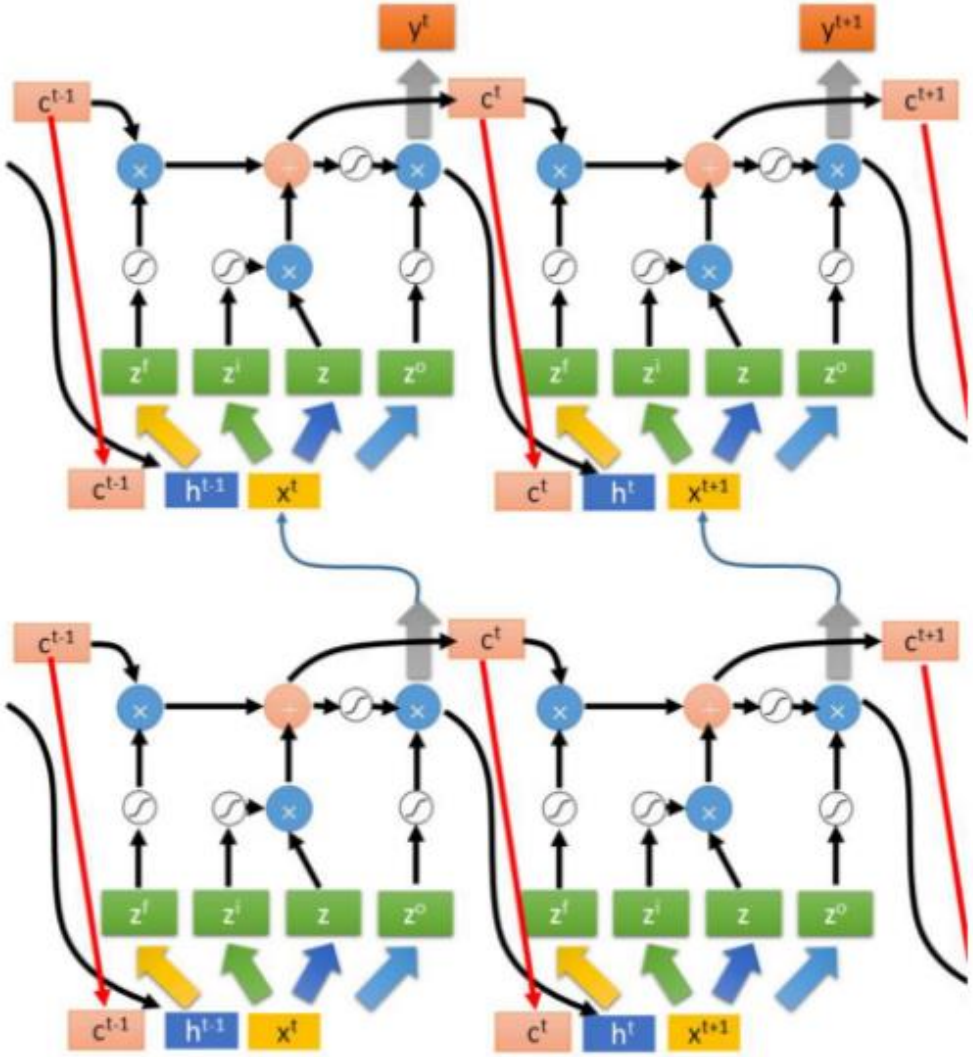

y^t
c^{t-1}
z^f
z^i
z
z^o
x^t
z^o
Output Gate
Forget Gate
Cell
z^f
z^i
Input Gate
Block
z

LSTM


LSTM
Extension: “peephole”
y^t
y^{t+1}
c^{t-1}
c^t
c^{t+1}
z^f
z^i
z
z^o
c^{t-1}
h^{t-1}
x^t
c^t
h^t
x^{t+1}

LSTM

Multiple-layer LSTM



This is quite standard now.

Talk is cheap,
show me the code!

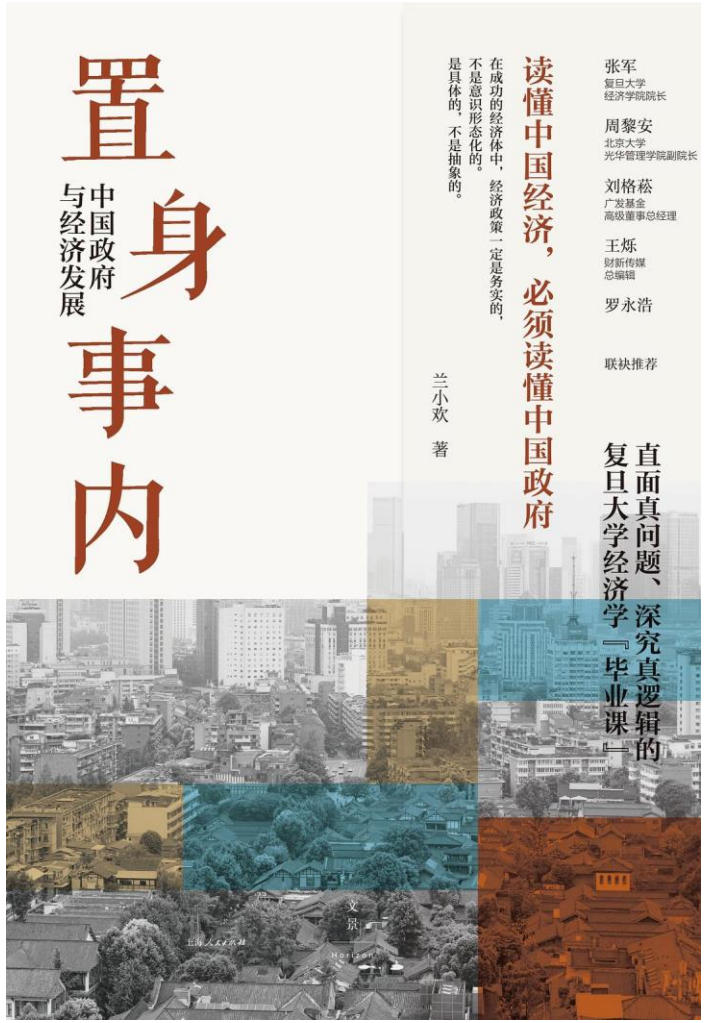
Takeaways

- The art of sequence prediction
- The 20th century code, but programmed in the 21st century way

See Also

- GPT-3 | openAI
- GitHub copilot | CodeX
- Transformer
- BERT

Recommendation



Thanks