COMP5046
Natural Language Processing

Lecture 4: Word Classification and Machine Learning 2

Semester 1, 2019

School of Computer Science

The University of Sydney, Australia

THE UNIVERSITY OF SYDNEY

Caren Han

Caren.Han@sydney.edu.au

0 LECTURE PLAN

Lecture 4: Word Classification and Machine Learning 2

1. Machine Learning and NLP: Finish
2. Seq2Seq Learning
3. Seq2Seq Deep Learning
   1. RNN (Recurrent Neural Network)
   2. LSTM (Long Short-Term Memory)
   3. GRU (Gated Recurrent Unit)
4. Seq2Seq Encoding and Decoding
5. Next Week Preview
   - Natural Language Processing Stack

0 LECTURE PLAN

Lecture 4: Word Classification and Machine Learning 2

1. **Machine Learning and NLP: Finish**
2. Seq2Seq Learning
3. Seq2Seq Deep Learning
   1. RNN (Recurrent Neural Network)
   2. LSTM (Long Short-Term Memory)
   3. GRU (Gated Recurrent Unit)
4. Seq2Seq Encoding and Decoding
5. Next Week Preview
   - Natural Language Processing Stack

1 Machine Learning and NLP

The purpose of Natural Language Processing: Overview

Application

Understanding

Searching

Dialog

Translation

Sentiment Analysis

Topic Classification

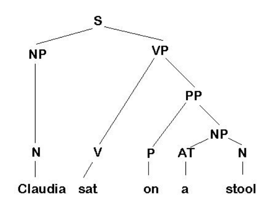
Topic Modelling

....

Search

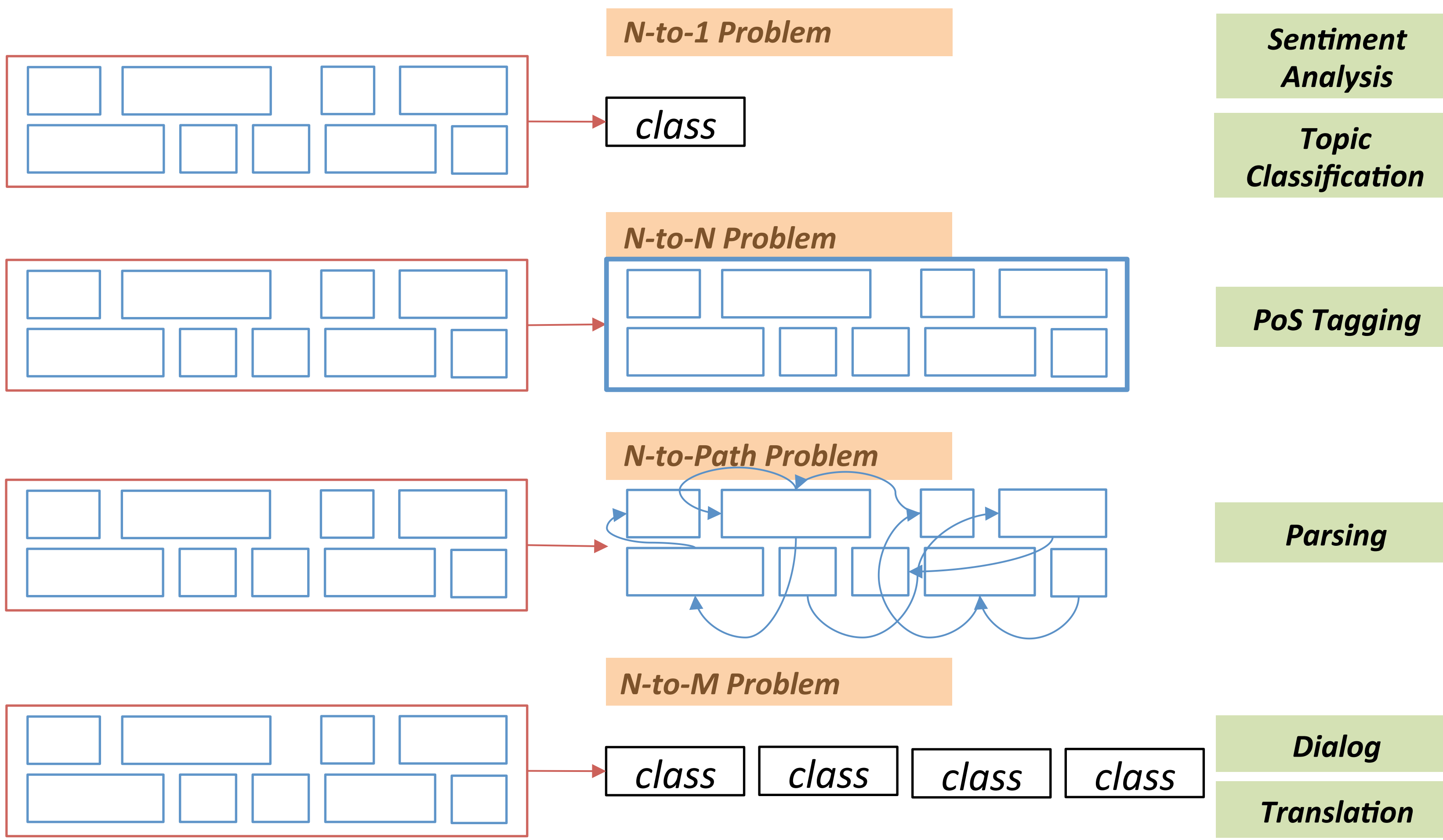
....

NLP Stack

Entity Extraction	When Sebastian Thrun …	When Sebastian Thrun PERSON started at Google ORG in 2007 DATE
Parsing	Claudia sat on a stool	
PoS Tagging	She sells seashells	**[she/PRP] [sells/VBZ] [seashells/NNS]**
Stemming	Drinking, Drank, Drunk	**Drink**
Tokenisation	How is the weather today	**[How] [is] [the] [weather] [today]**

1 Machine Learning and NLP

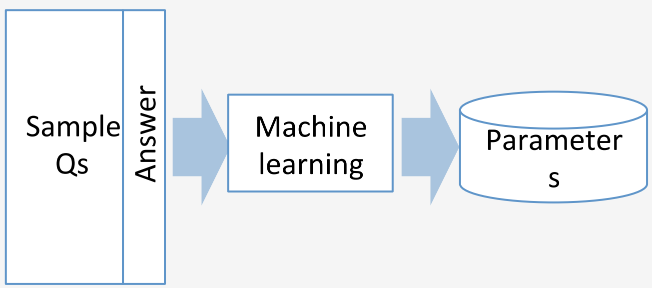
Problem Abstraction



1 Machine Learning and NLP

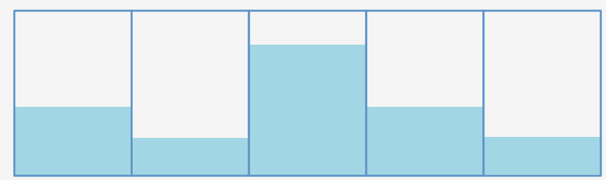
Classification

Supervised Learning



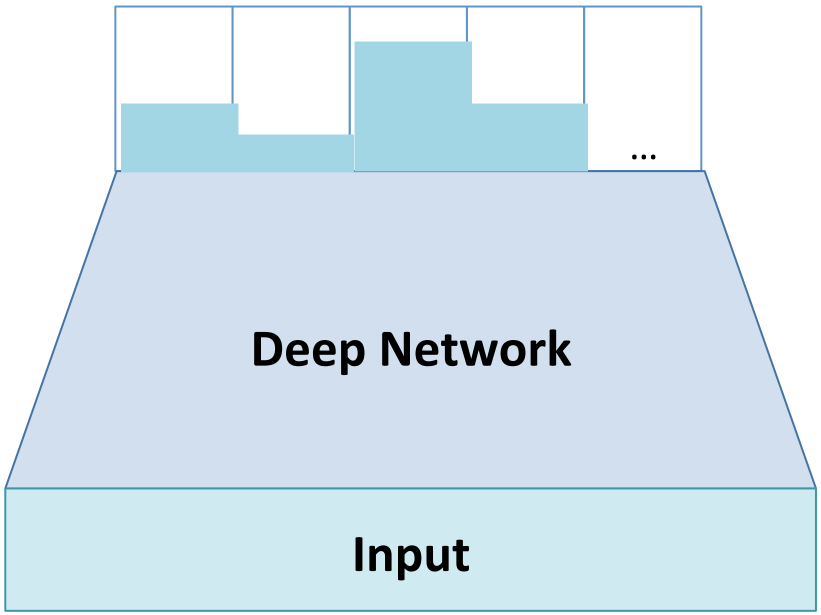
1. Take the sample question
2. Compare with the answer
3. Learn to get the right answer

Probabilistic classification



1. Scoring for given classes
2. Choose the highest scored class

Probabilistic Classification



1. Define N Class

5. Score given classes

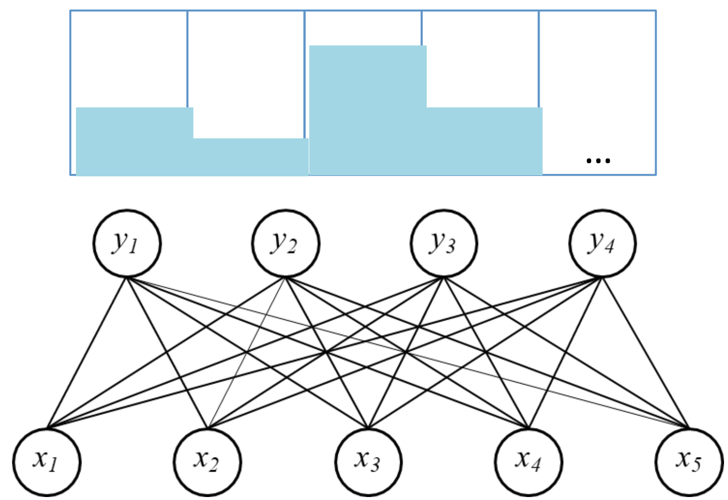
2. Construct Network

4. Classification

3. Input

1 Machine Learning and NLP

Probabilistic Classification



1. Define N Class

5. Score given classes

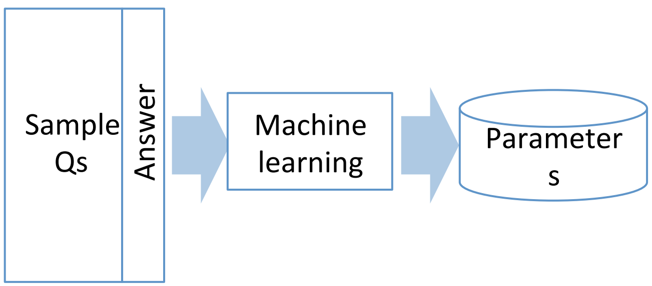
2. Construct Network

4. Classification

3. Input

$$y_1 = f(W_{11}x_1 + W_{12}x_2 + W_{13}x_3 + W_{14}x_4 + W_{14}x_4 + b_1)$$
$$y_2 = f(W_{21}x_1 + W_{22}x_2 + W_{23}x_3 + W_{24}x_4 + W_{25}x_5 + b_2)$$
$$y_3 = f(W_{31}x_1 + W_{32}x_2 + W_{33}x_3 + W_{34}x_4 + W_{35}x_5 + b_3)$$
$$y_4 = f(W_{41}x_1 + W_{42}x_2 + W_{43}x_3 + W_{44}x_4 + W_{45}x_5 + b_4)$$

Probabilistic Classification

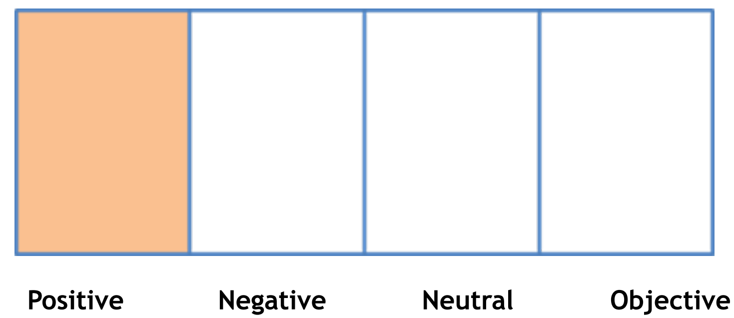


1. *Reference Representation*
2. *Scoring Normalisation*
3. *Cost Function Design*
4. *Parameter Update*

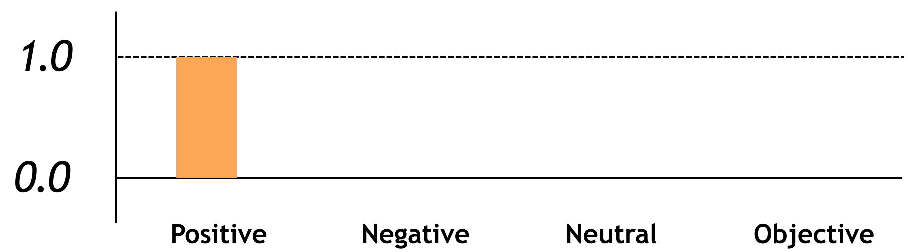
1 Machine Learning and NLP

Reference Representation

"I love this place so much!"



One-Hot Representation

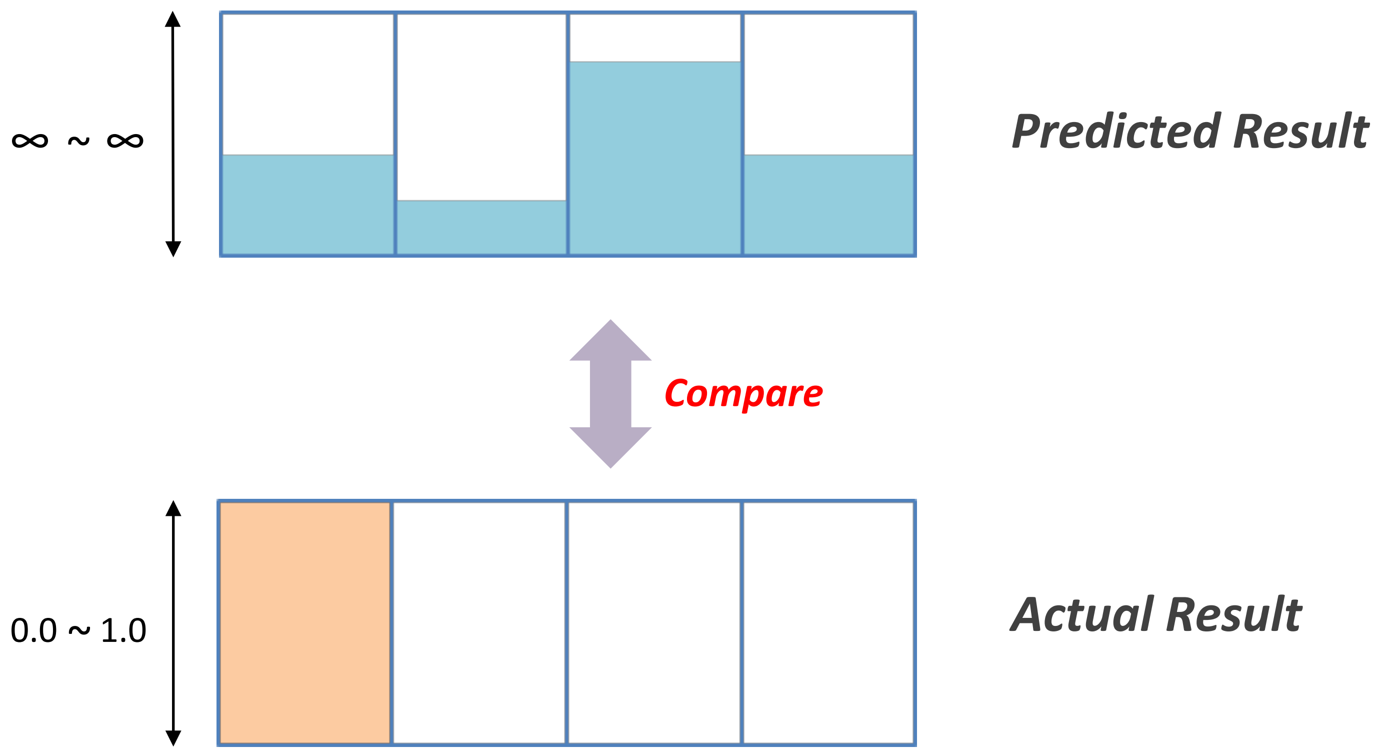


1.0 for reference class

[1, 0, 0, 0]

1 Machine Learning and NLP

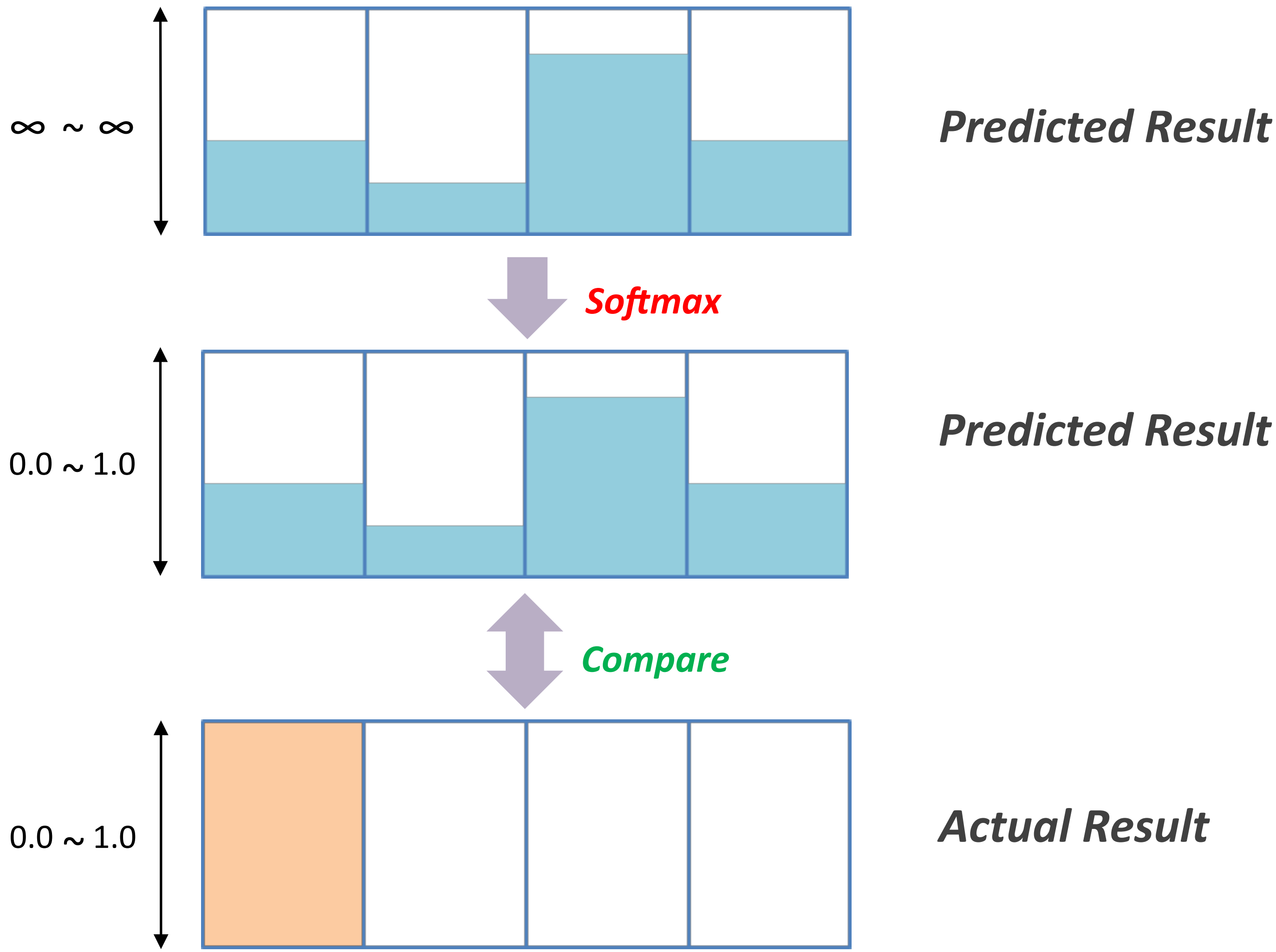
Score Normalisation



In order to compare the predicted and actual result, the range should be in the same scale

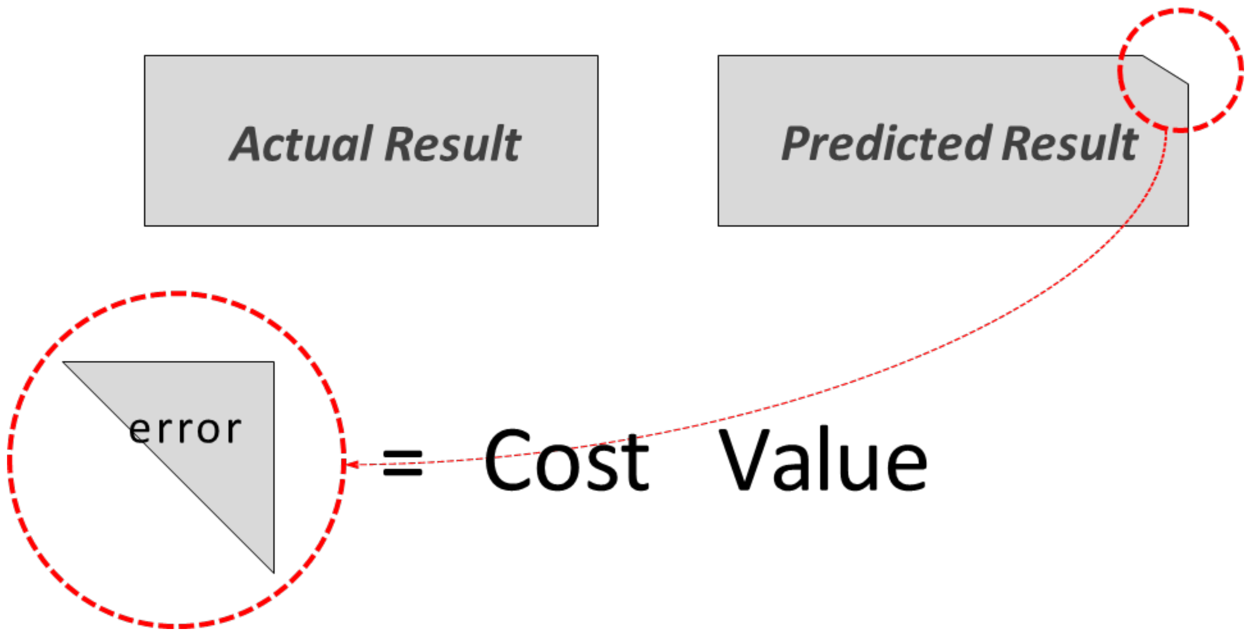
1 Machine Learning and NLP

Score Normalisation and Cost Function



1 Machine Learning and NLP

Score Normalisation and Cost Function



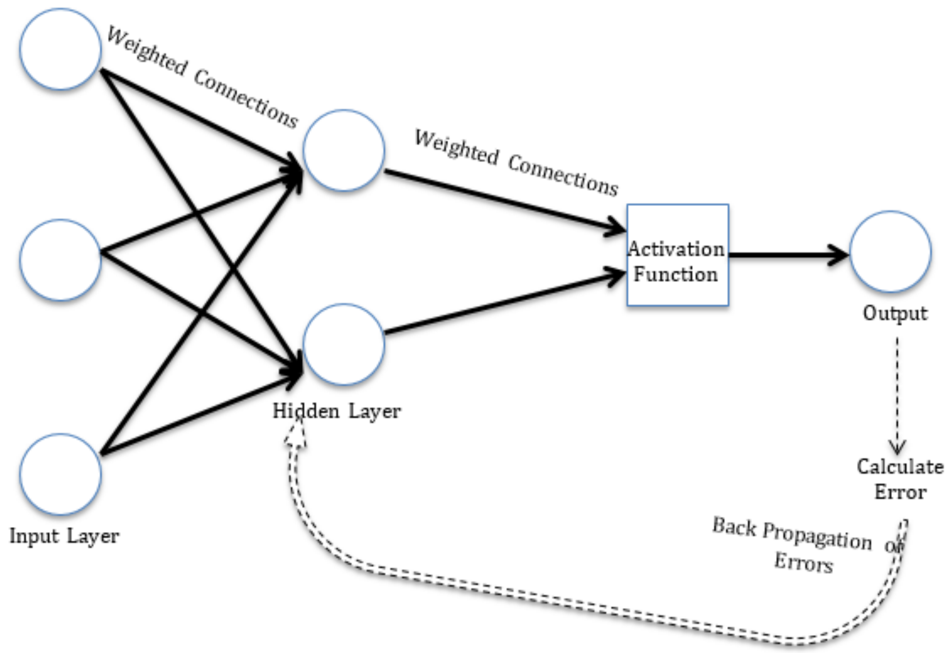
Score Normalisation and Cost Function

Depends on the dataset

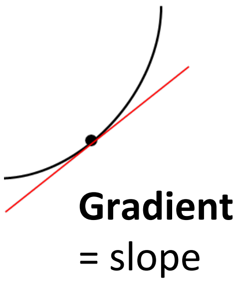
Problem type	Last-layer activation	Loss (Cost) function	Example
Binary classification	sigmoid	Binary Cross Entropy	Sentiment analysis (Positive/ Negative)
Multi-class, single-label classification	softmax	Categorical Cross Entropy	Part-of-Speech tagging Named Entity Recognition
Multi-class, multi-label classification	sigmoid	Binary Cross Entropy	Multi-topic classification, one can have multiple topics
Regression to arbitrary values	None	MSE (Mean Squared Error)	Predict house price
Regression to values between 0 and 1	sigmoid	MSE or Binary Cross Entropy	Engine health assessment where 0 is broken, 1 is new

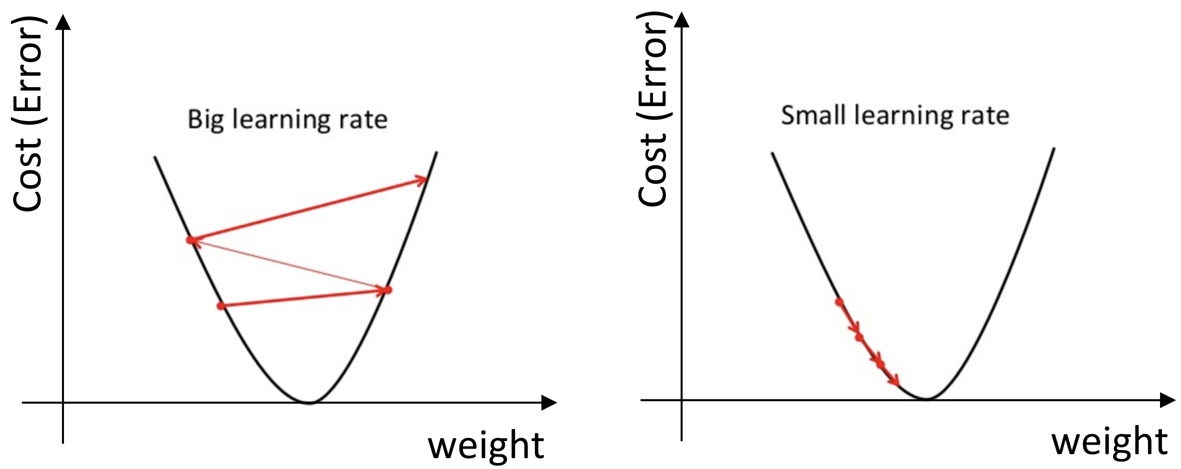
Parameter Update

Backpropagation

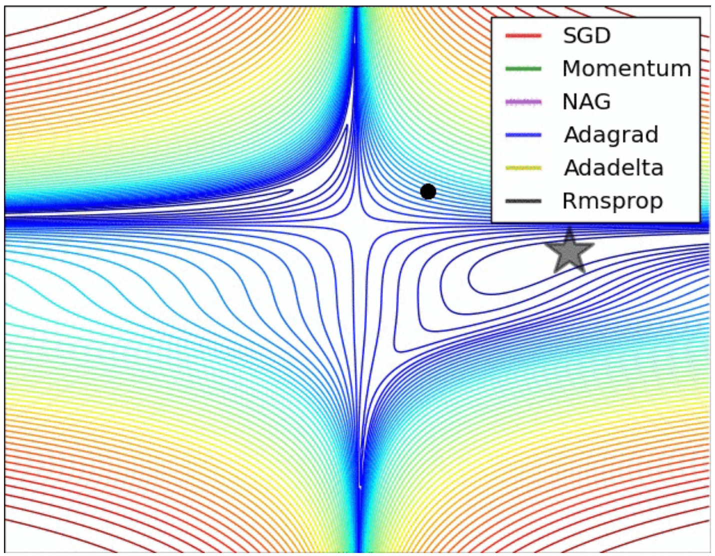


Parameter Update: Gradient

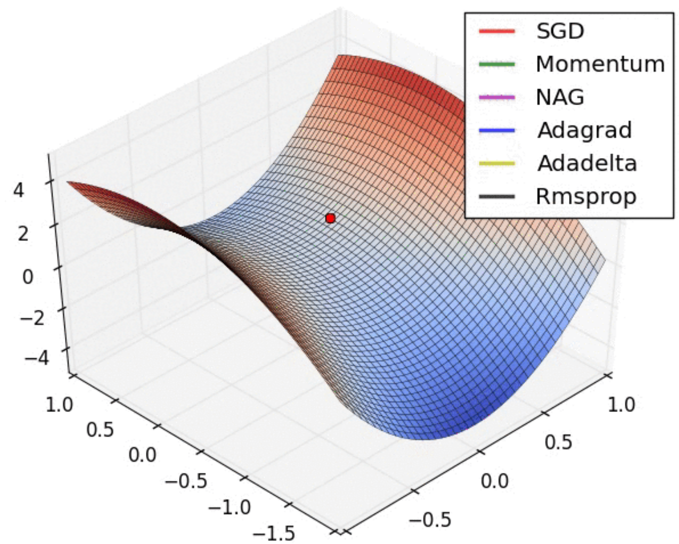




Parameter Update: Gradient descent Optimization



Optimisation on loss surface contours

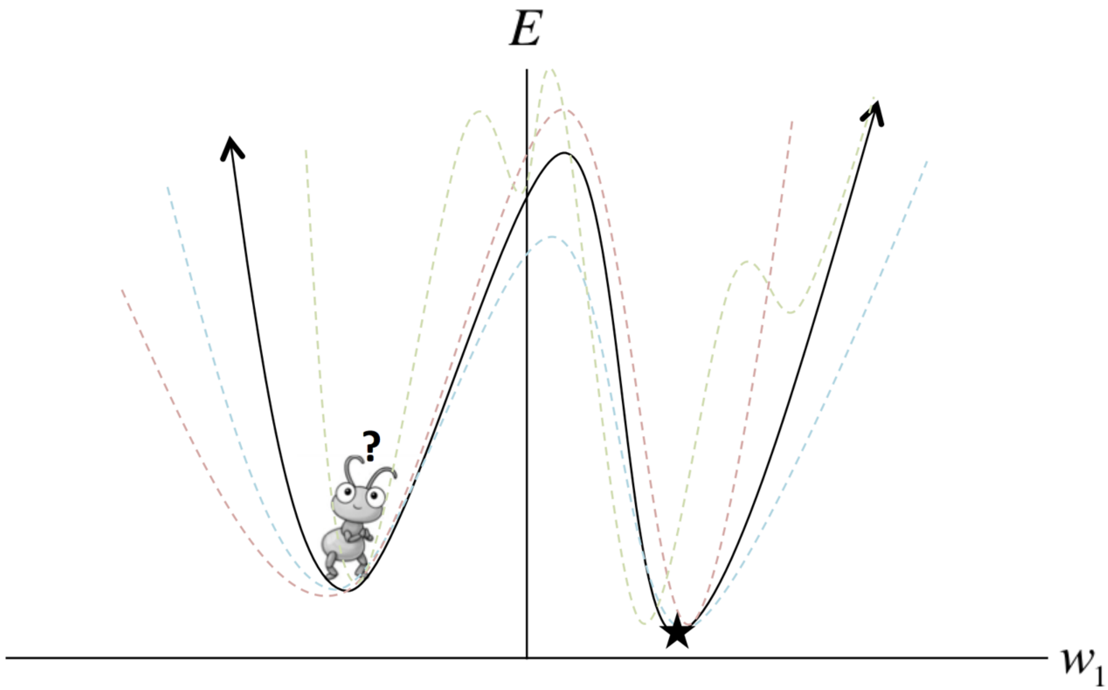


Optimisation on saddle points

There are different types of Gradient descent Optimization algorithm

SGD (Stochastic gradient descent), Momentum, NAG(Nesterov accelerated gradient), Adagrad, Adadelta, Adam,..

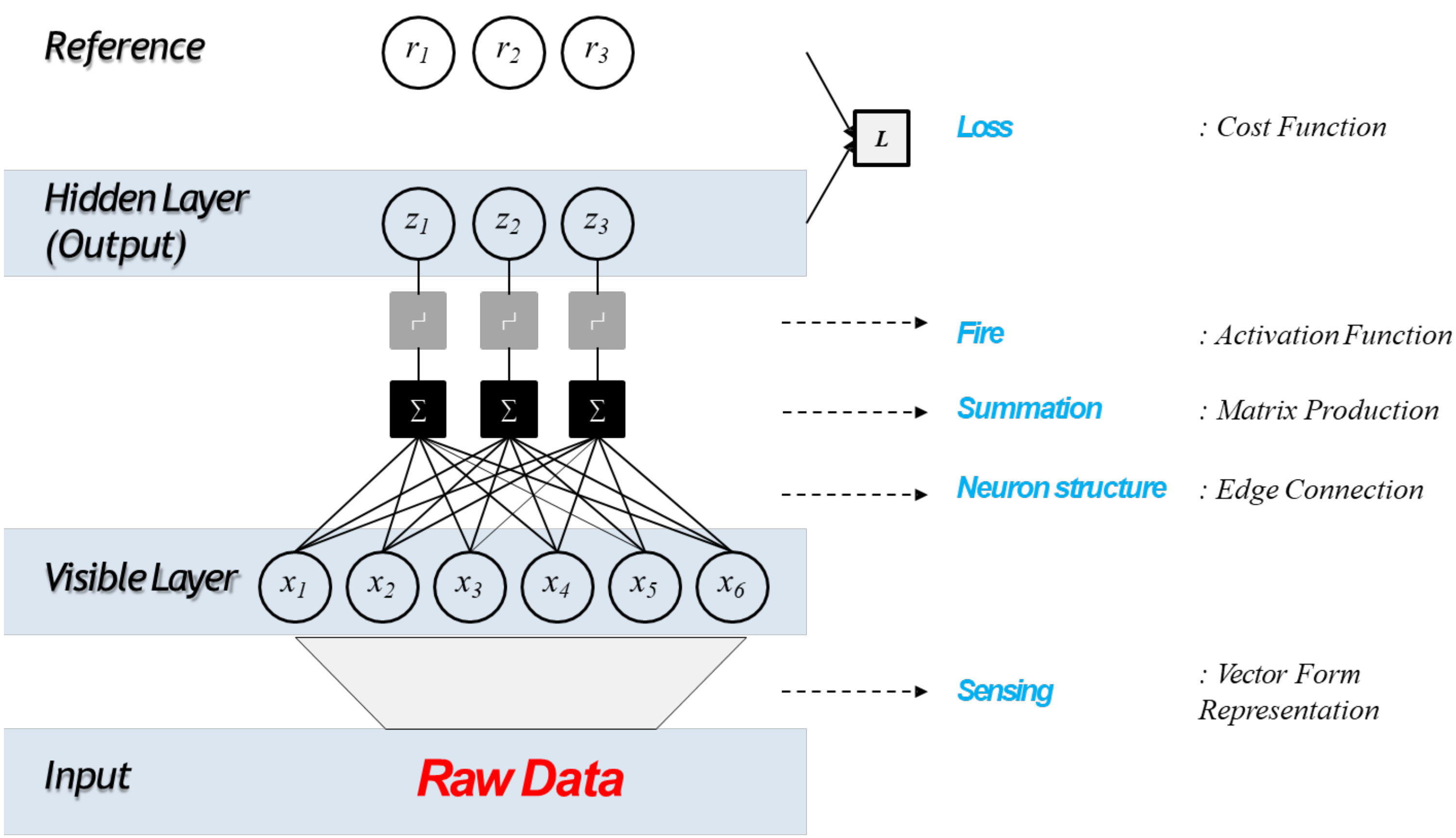
Parameter Update: Gradient descent Optimization



There are different types of Gradient descent Optimization algorithm

SGD (Stochastic gradient descent), Momentum, NAG(Nesterov accelerated gradient), Adagrad, Adadelta, RMSprop,..

Summary

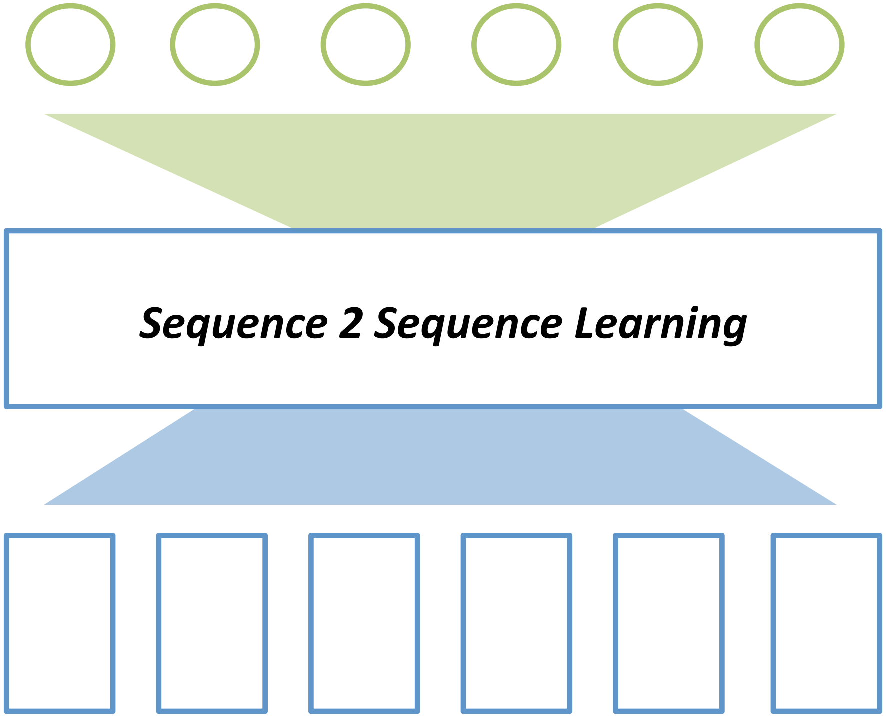

Reference
r_1 r_2 r_3
L
Loss
: Cost Function
Hidden Layer (Output)
z_1 z_2 z_3
Fire
: Activation Function
Σ Σ Σ
Summation
: Matrix Production
Neuron structure
: Edge Connection
Visible Layer
x_1 x_2 x_3 x_4 x_5 x_6
Sensing
: Vector Form Representation
Input
Raw Data

0 LECTURE PLAN

Lecture 4: Word Classification and Machine Learning 2

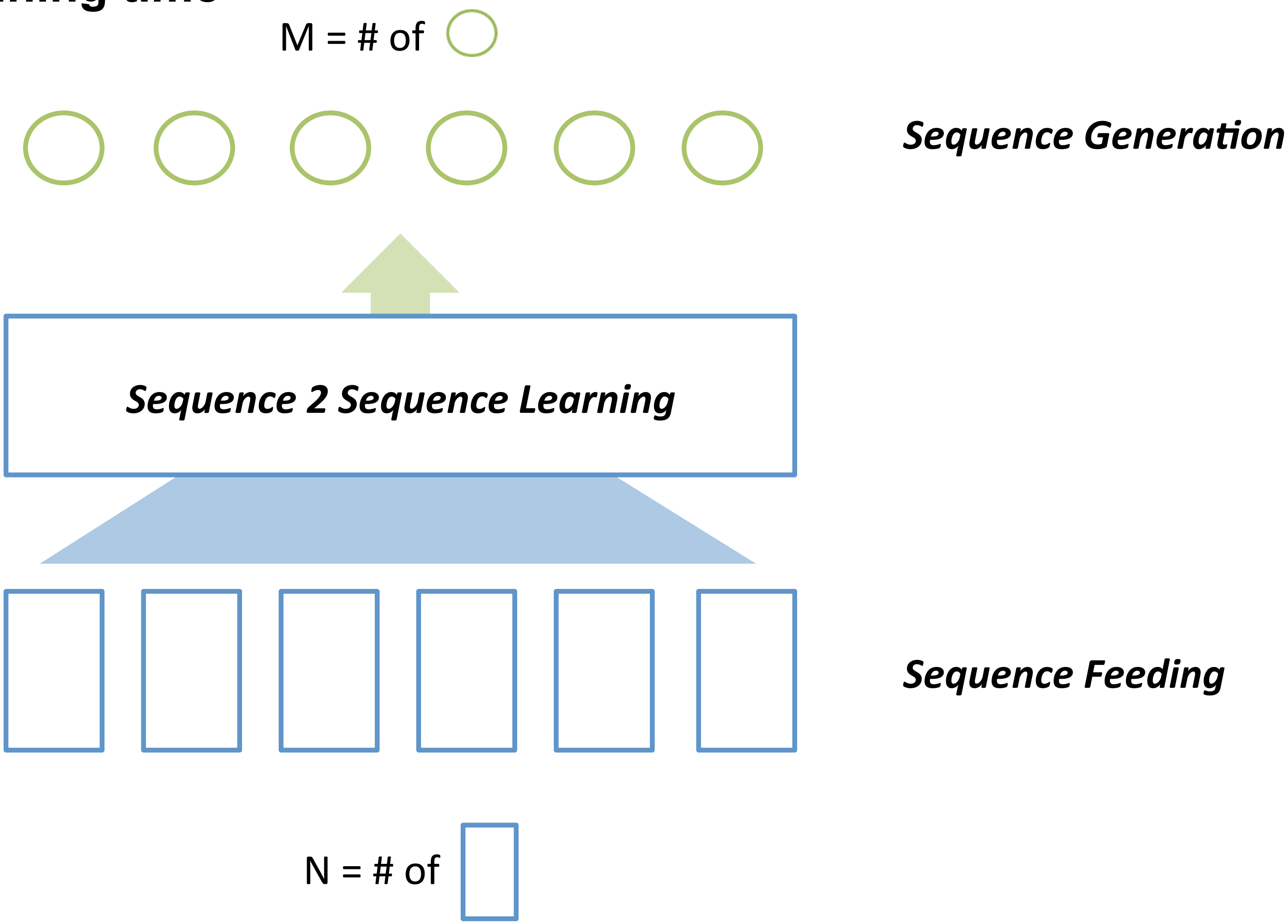
1. Machine Learning and NLP: Finish
2. **Seq2Seq Learning**
3. Seq2Seq Deep Learning
   1. RNN (Recurrent Neural Network)
   2. LSTM (Long Short-Term Memory)
   3. GRU (Gated Recurrent Unit)
4. Seq2Seq Encoding and Decoding
5. Next Week Preview
   - Natural Language Processing Stack

Illustration



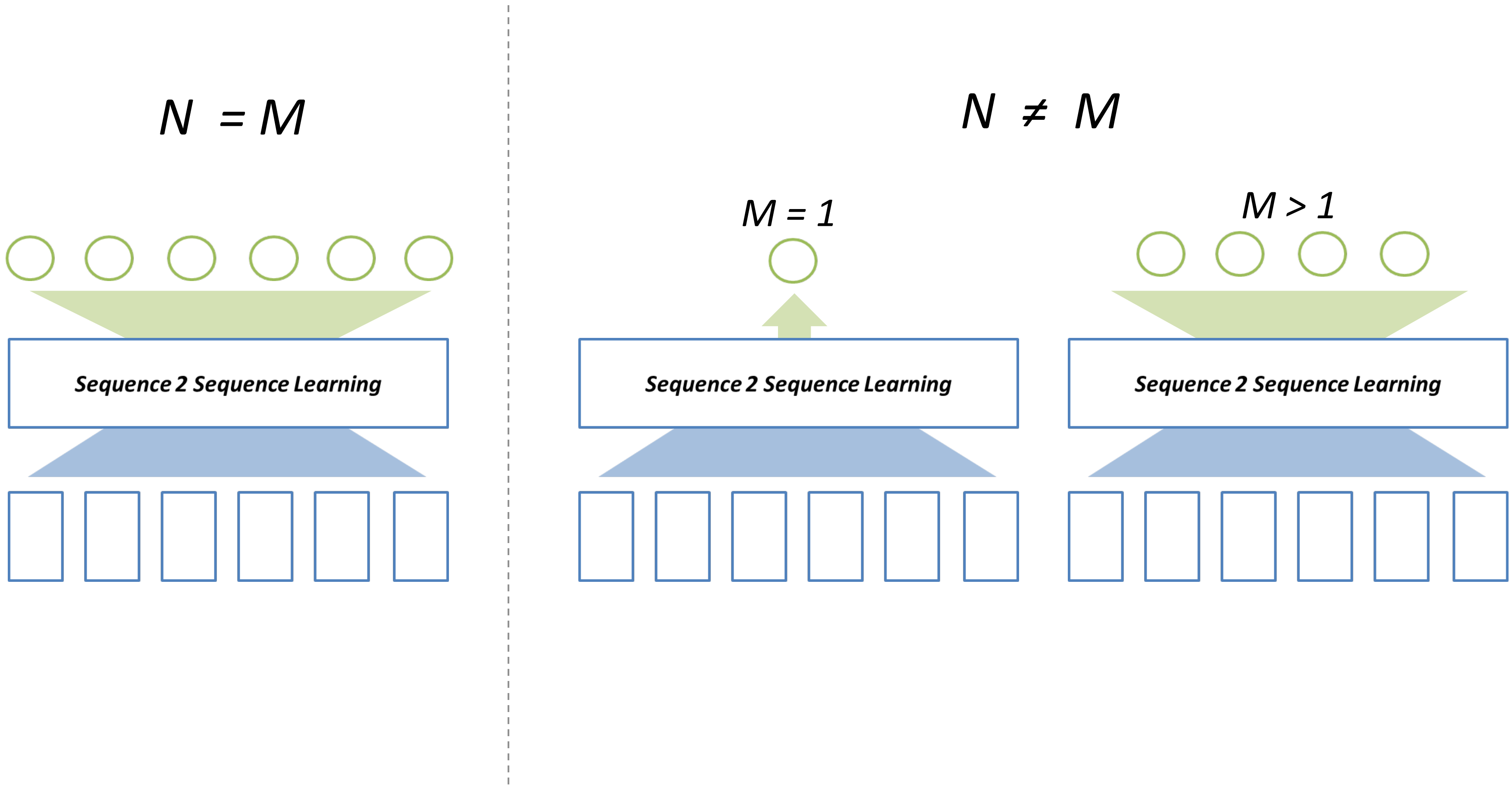
2 Sequence 2 Sequence Learning

Running time



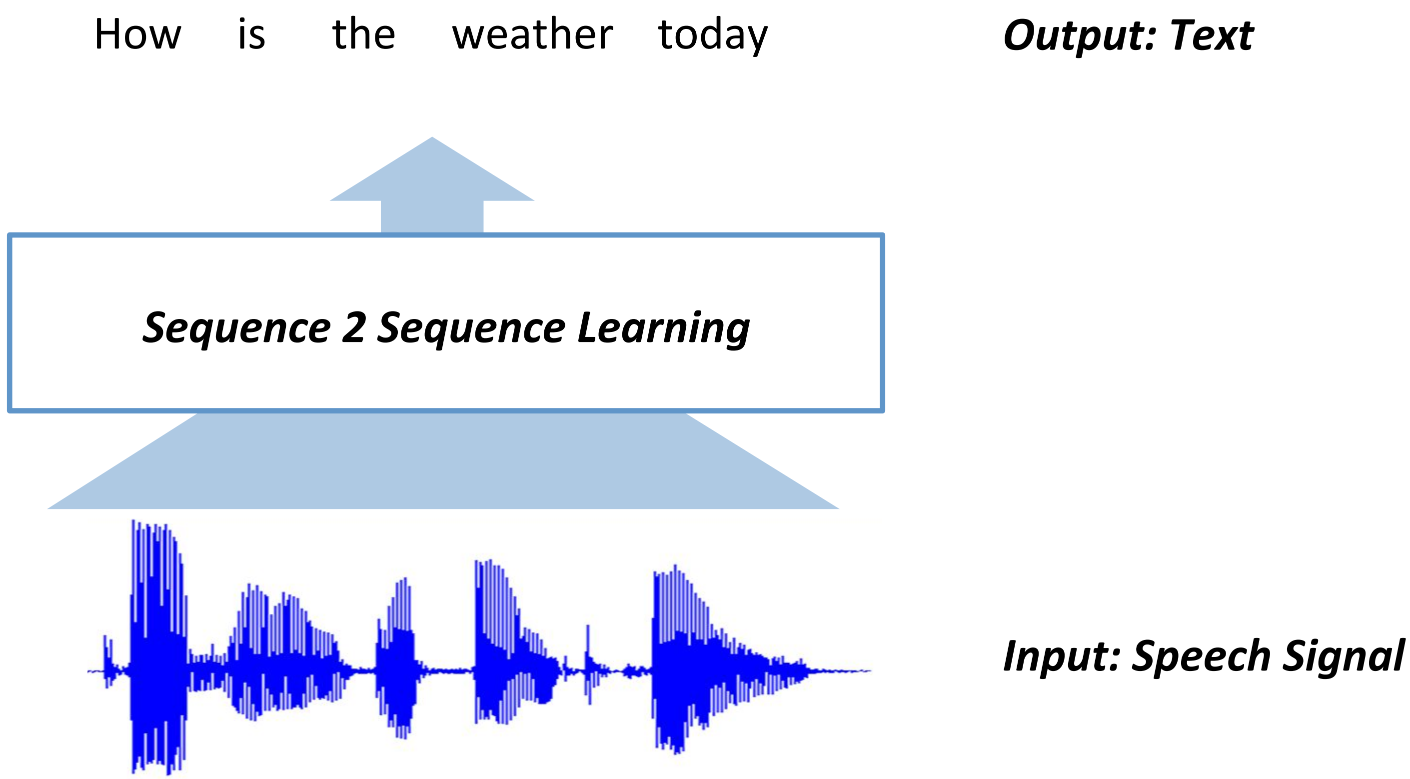
2 Sequence 2 Sequence Learning

Sequence 2 Sequence Learning



Sequence 2 Sequence Learning

Seq2Seq – Speech Recognition



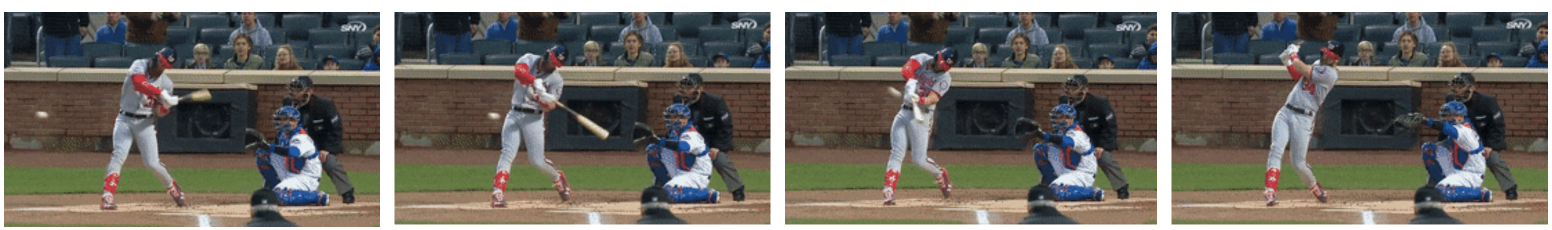
2 Sequence 2 Sequence Learning

Seq2Seq – Movie Frame Labelling

Swing Swing Hit Bat_Broken

Output: Scene Labels

Sequence 2 Sequence Learning

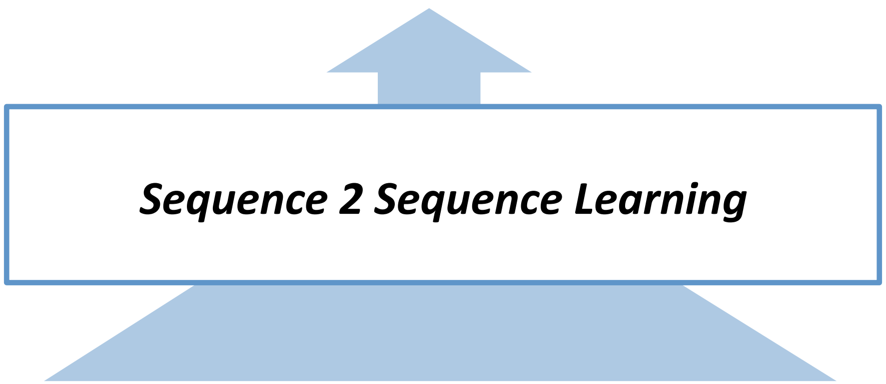


Input: Video Frame

2 Sequence 2 Sequence Learning

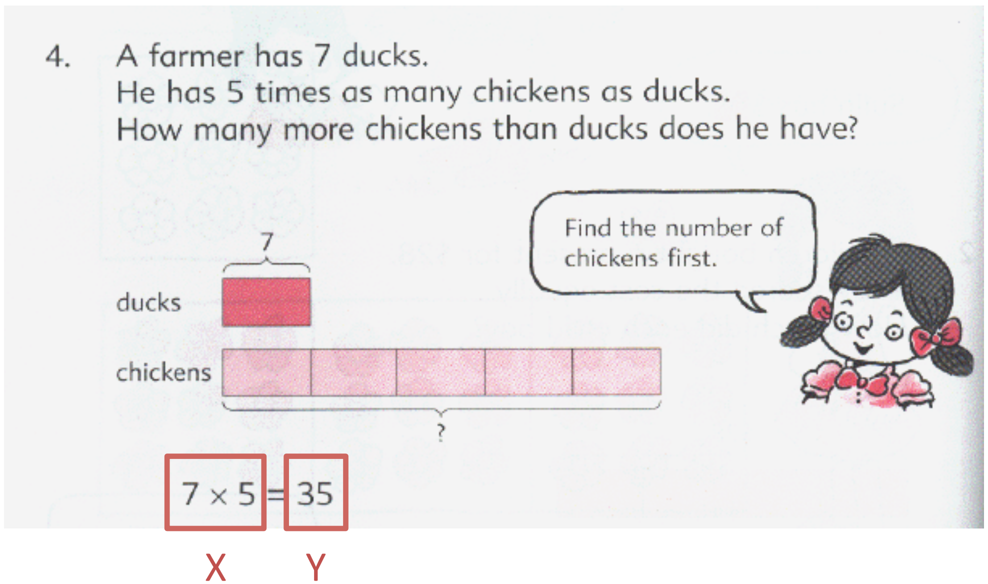
Seq2Seq – PoS Tagging

ADV VERB DET NOUN NOUN ***Output: Part of Speech***

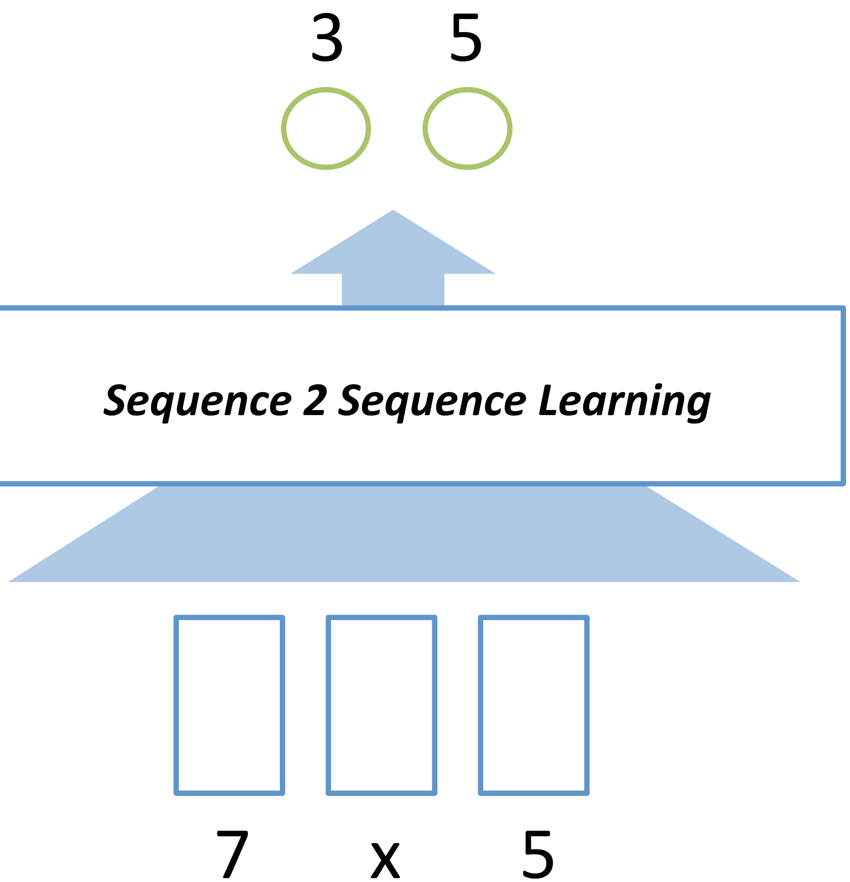


How is the weather today ***Input: Text***

Seq2Seq – Arithmetic Calculation



Seq2Seq – Arithmetic Calculation

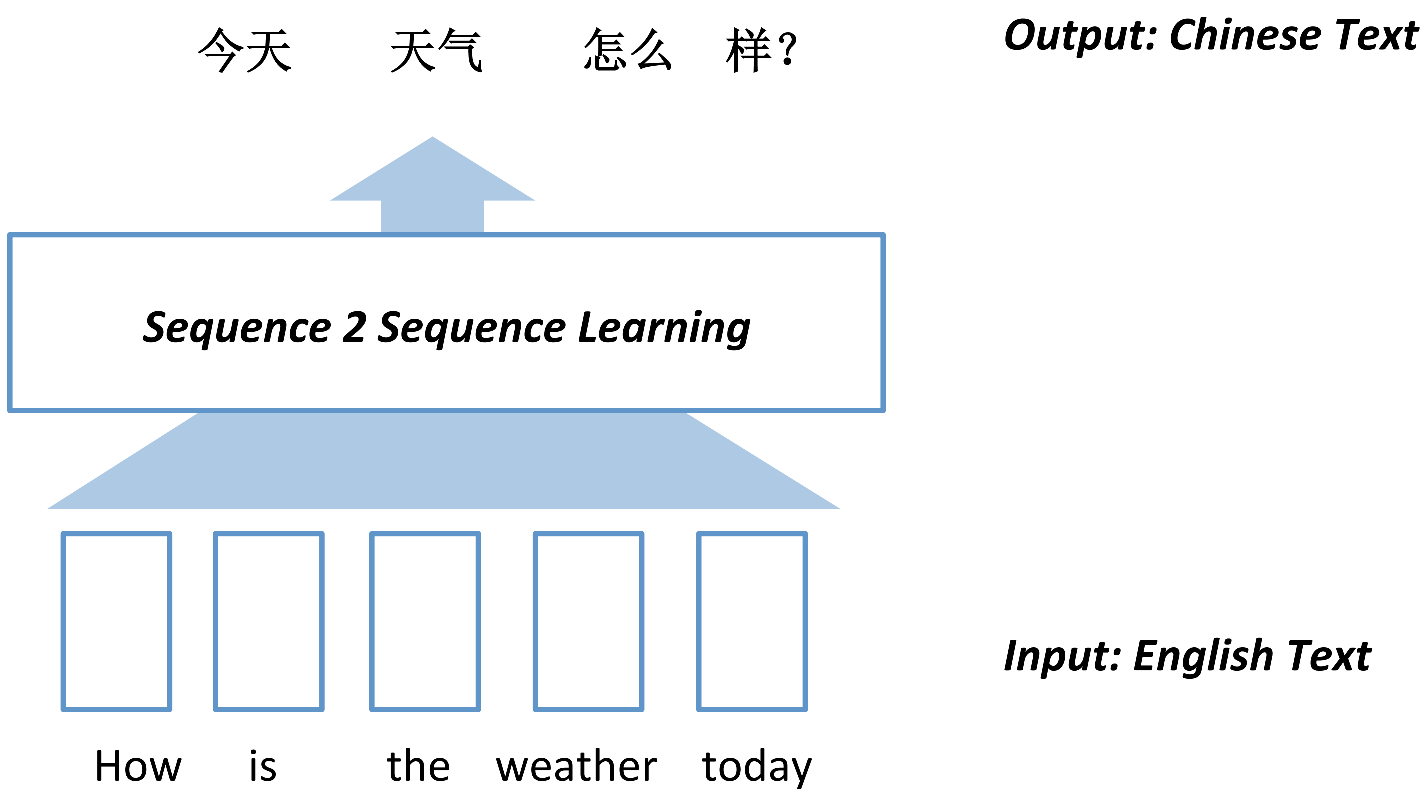


Output: Numbers

Input: Math Expression

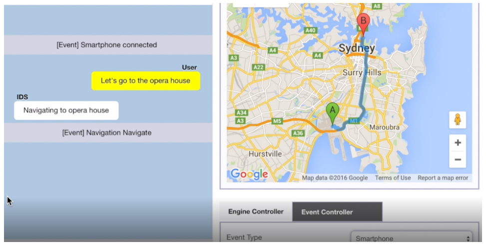
2 Sequence 2 Sequence Learning

Seq2Seq – Machine Translation



2 Sequence 2 Sequence Learning

Seq2Seq – Sentence Completion



How is the weather today?

How long does it take?

Let's go to the opera house

It is quite hot inside

I may need to stop by Darling Harbour

When is the dinner appointment

Change the schedule

Text him that I cannot meet at 6:30pm

I like learning Natural Language Processing

Sequence 2 Sequence Learning

Seq2Seq – Sentence Completion

How is the weather today?

How long does it take?

Let’s go to the opera house

It is quite hot inside

I may need to stop by Darling Harbour

When is the dinner appointment

Change the schedule

Text him that I cannot meet at 6:30pm

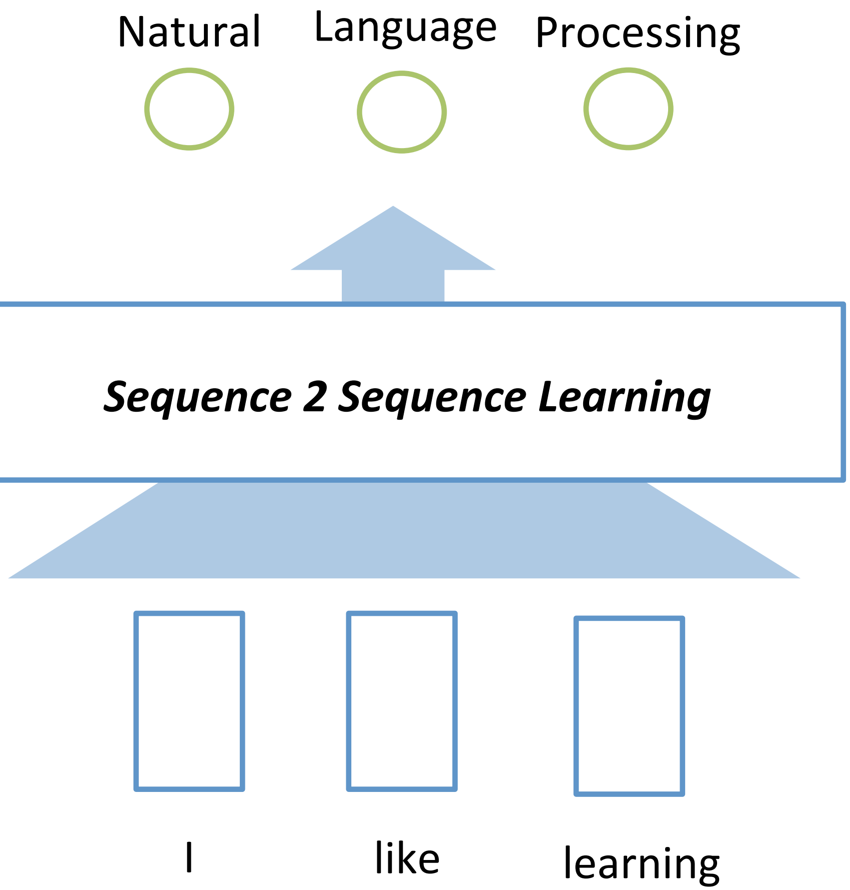
I like learning Natural Language Processing

X *Y*

Sequence 2 Sequence Learning

I like learning Natural Language Processing

Seq2Seq – Sentence Completion

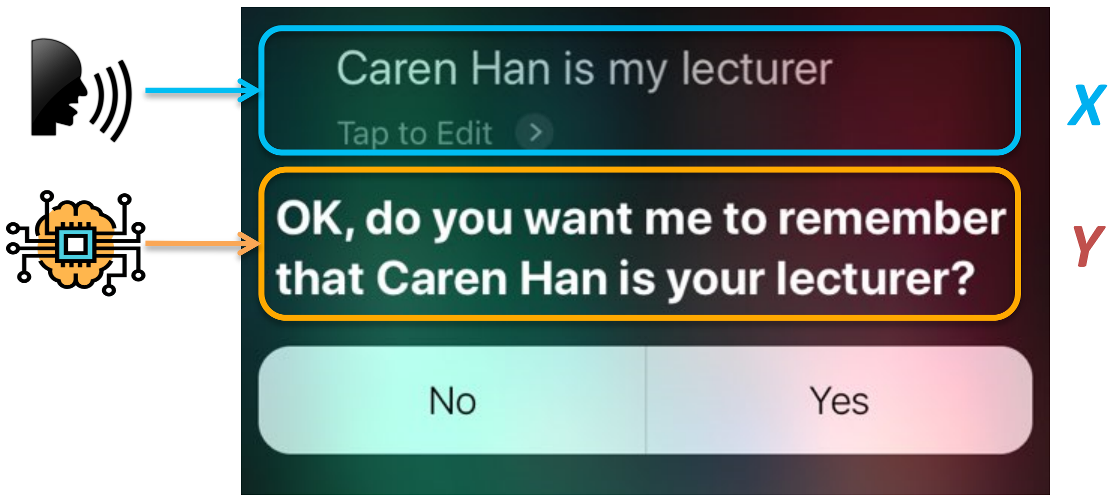


Output: Partial Sentence

Input: Partial Sentence

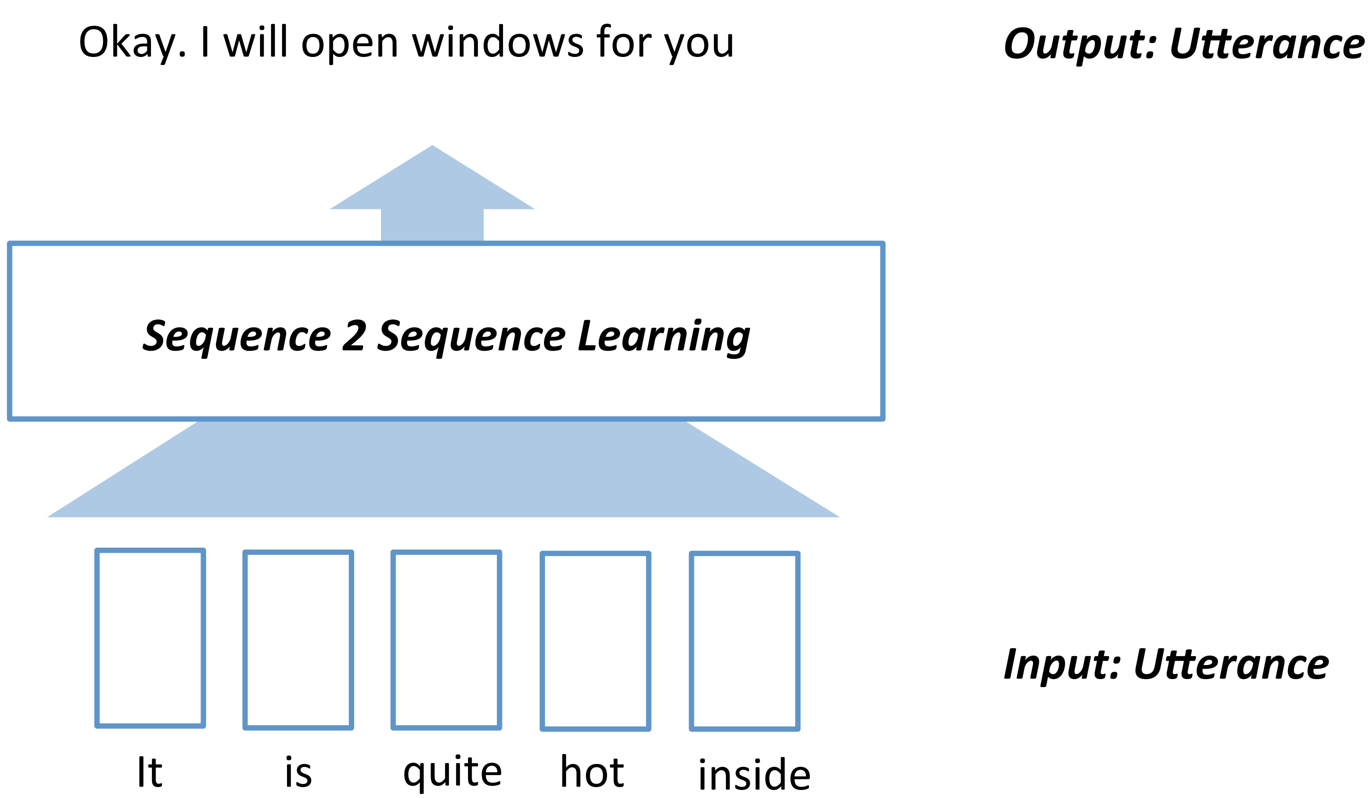
Seq2Seq – Conversation Modelling

Conversation



2 Sequence 2 Sequence Learning

Seq2Seq – Conversation Modelling

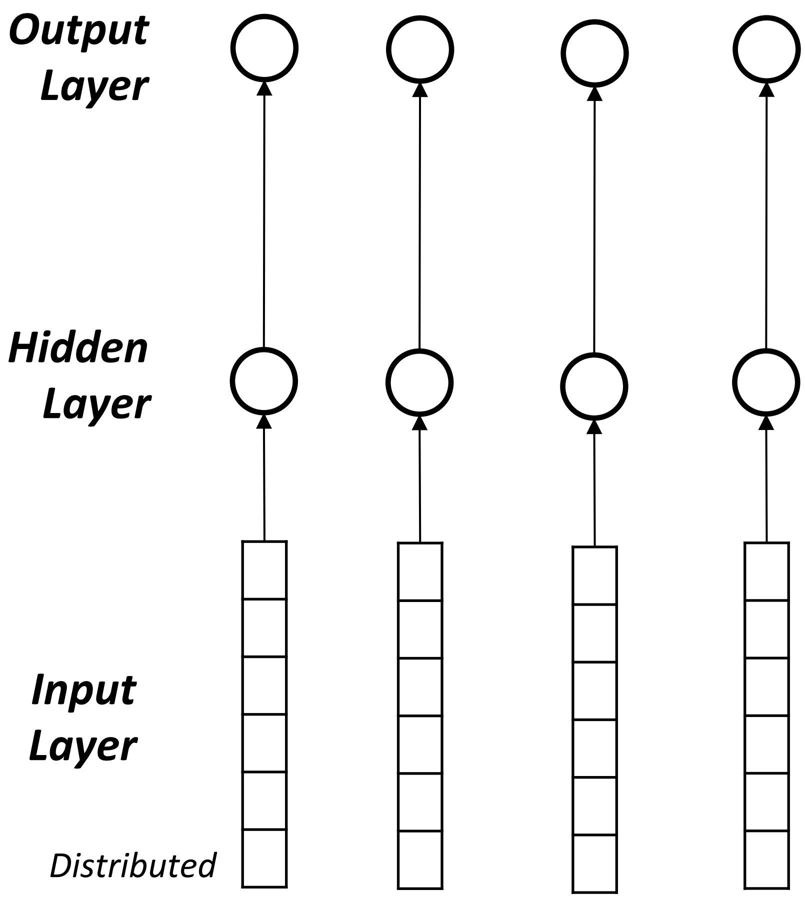


0 LECTURE PLAN

Lecture 4: Word Classification and Machine Learning 2

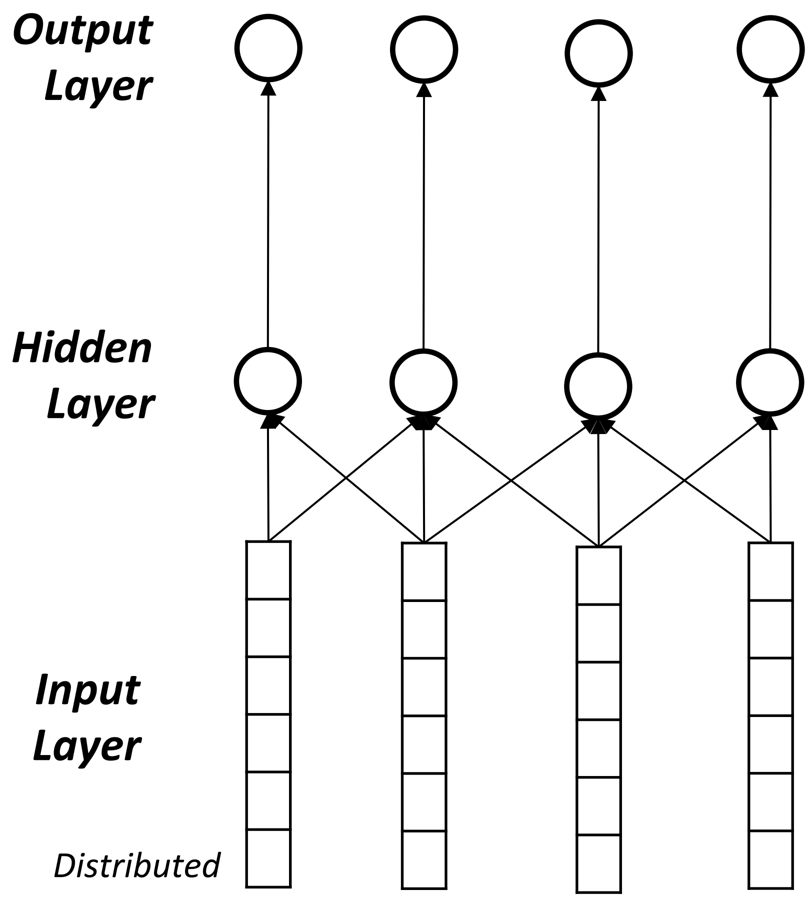
1. Machine Learning and NLP: Finish
2. Seq2Seq Learning
3. **Seq2Seq Deep Learning**
    1. RNN (Recurrent Neural Network)
    2. LSTM (Long Short-Term Memory)
    3. GRU (Gated Recurrent Unit)
4. Seq2Seq Encoding and Decoding
5. Next Week Preview
    - Natural Language Processing Stack

Prediction



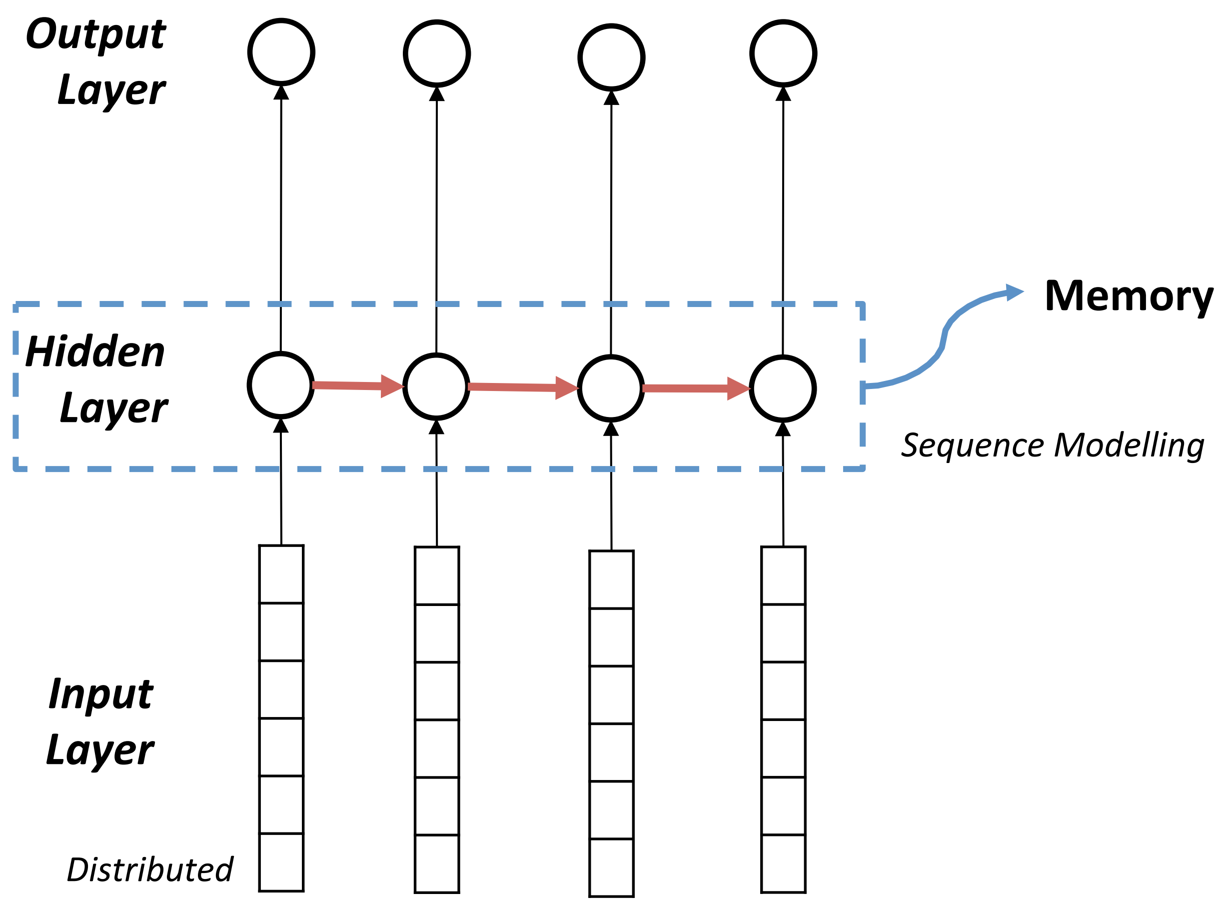
3 Seq2Seq with Deep Learning

Prediction + Convolution Idea



3 Seq2Seq with Deep Learning

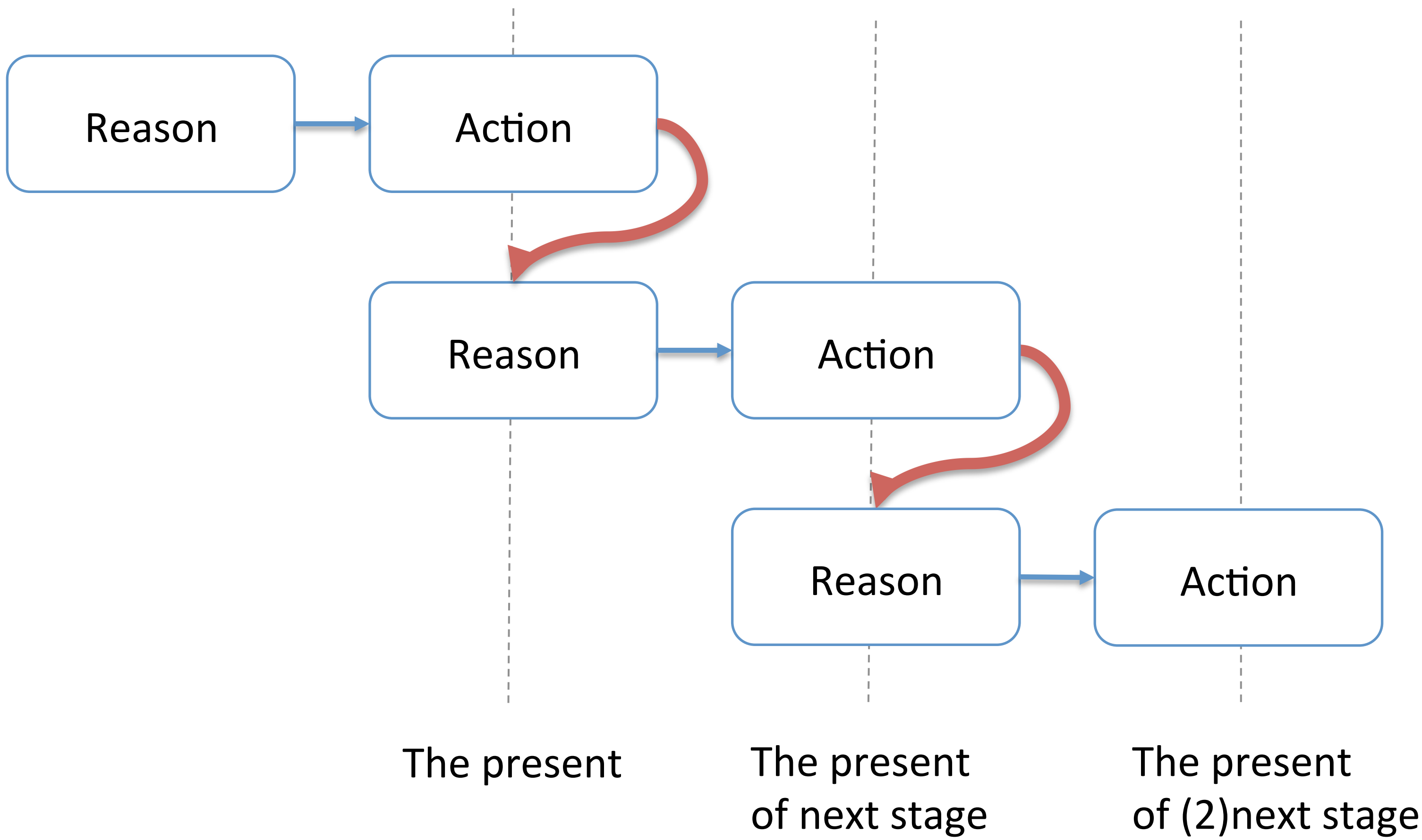
Sequence Modeling



3 Seq2Seq with Deep Learning

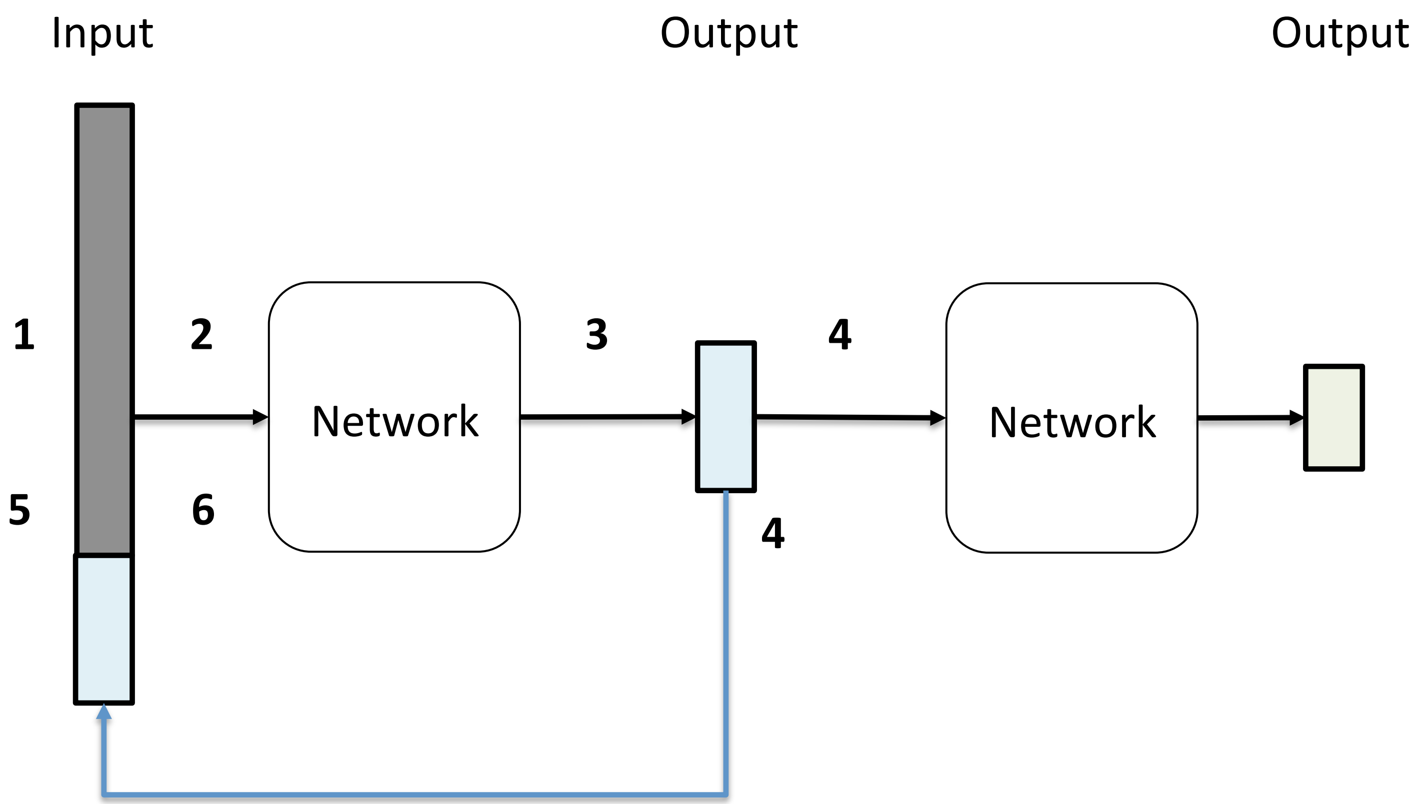
Neural Network + Memory

Memory is vital to experiences, it is the retention of information over time for the purpose of influencing future action



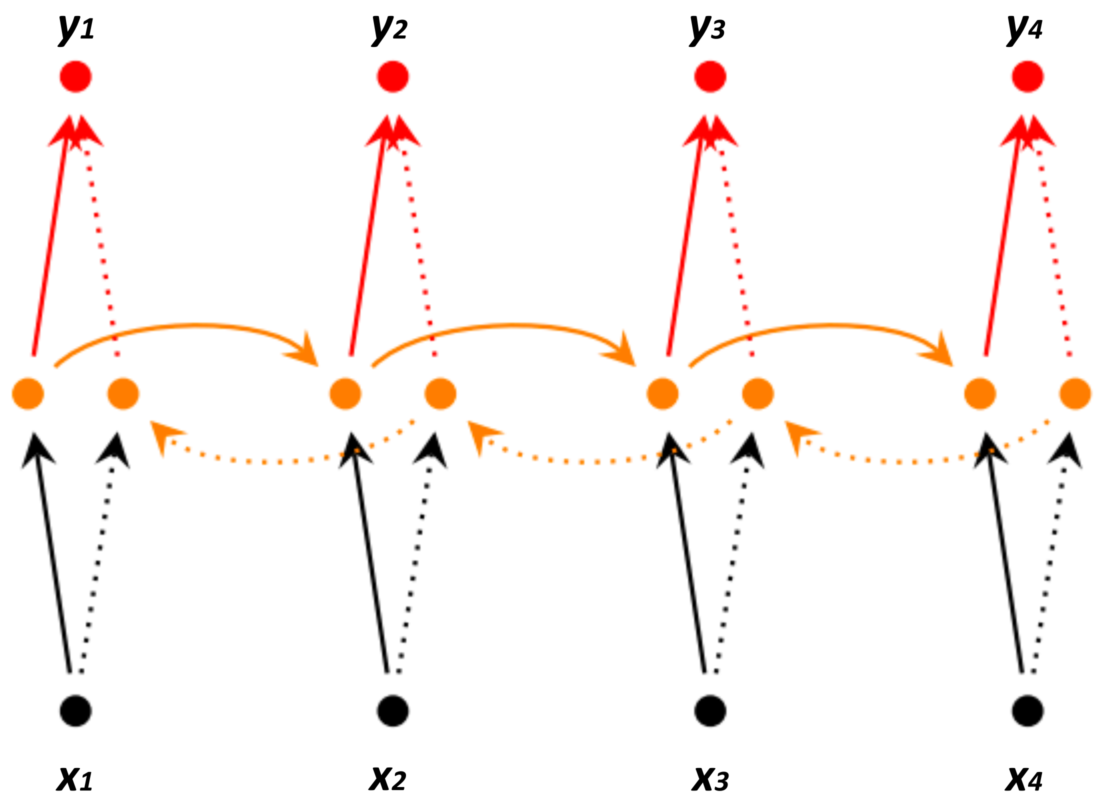
3 **Seq2Seq with Deep Learning**

Neural Network + Memory = Recurrent Neural Network



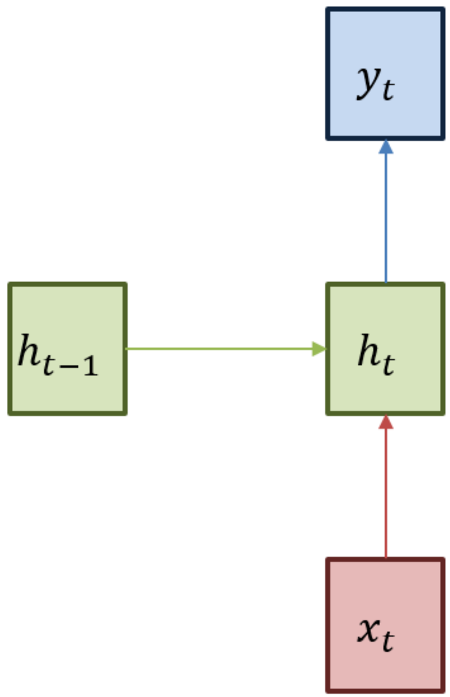
3 Seq2Seq with Deep Learning

Neural Network + Memory = Recurrent Neural Network



3 Seq2Seq with Deep Learning

Neural Network + Memory = Recurrent Neural Network

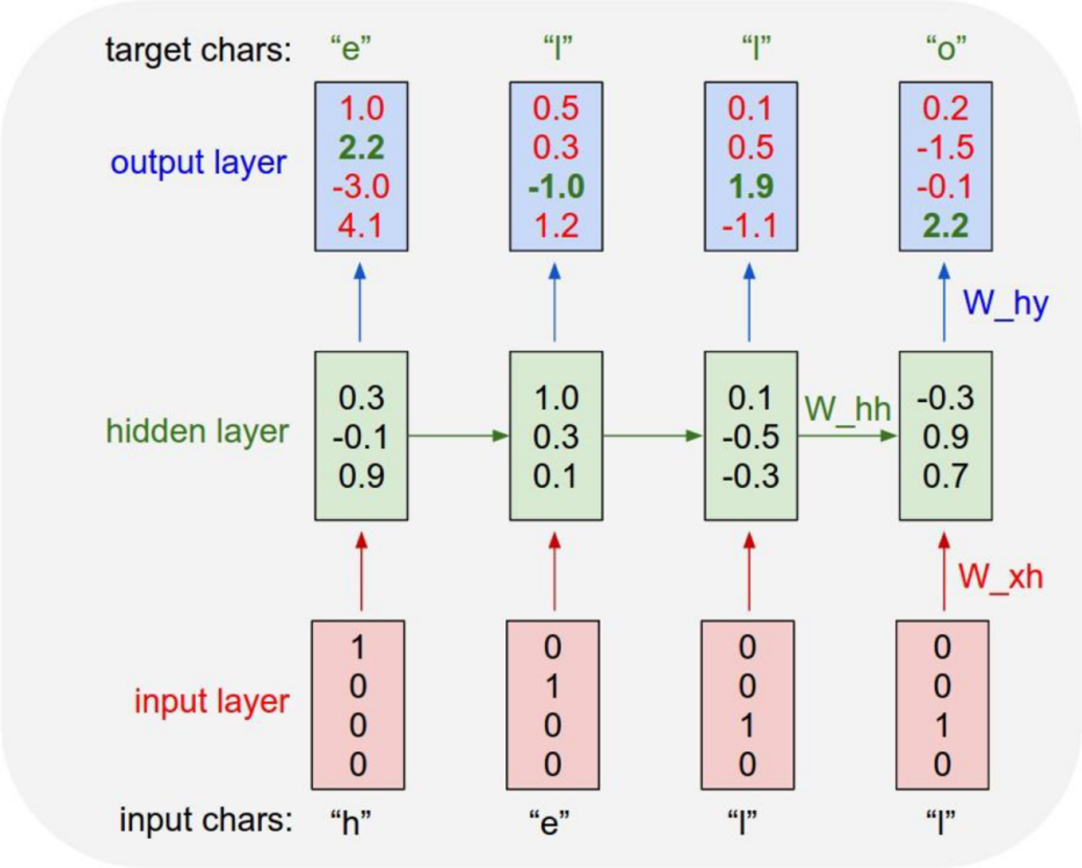


$$y_t = W_{hy} h_t + b_y$$

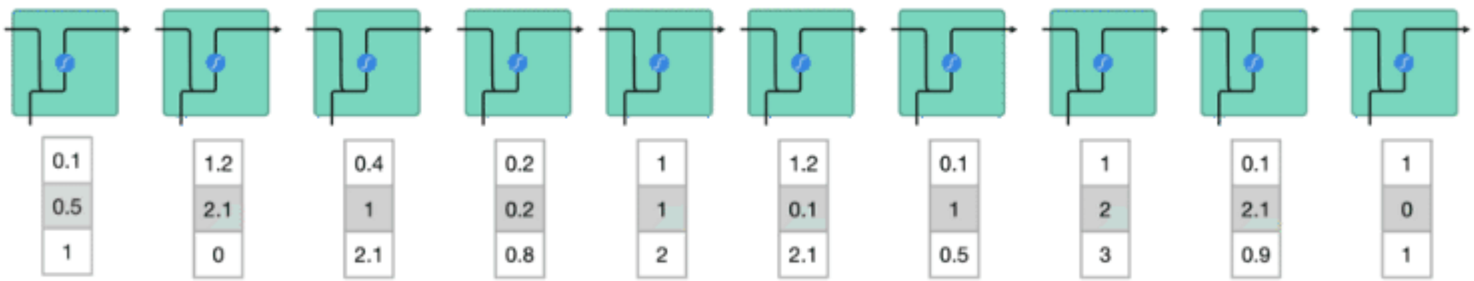
$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t + b_h)$$

Seq2Seq with Deep Learning

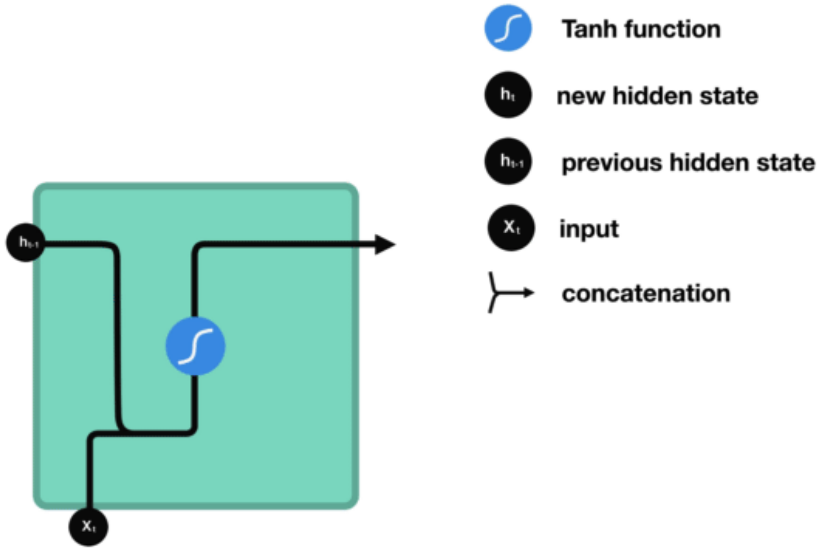
Neural Network + Memory = Recurrent Neural Network



Neural Network + Memory = Recurrent Neural Network



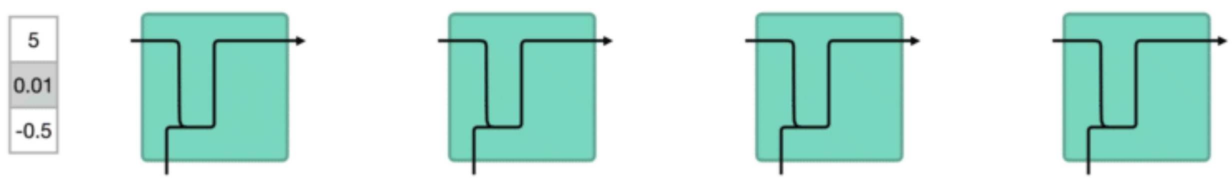
Neural Network + Memory = Recurrent Neural Network



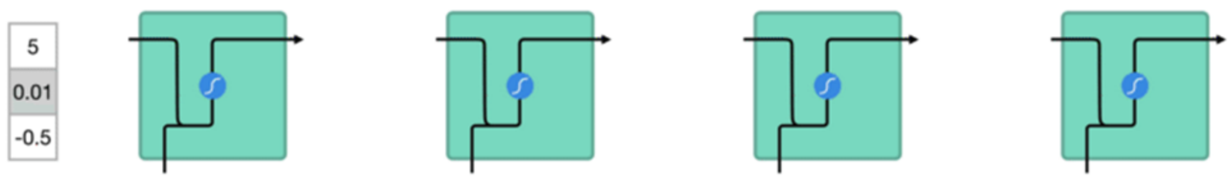
RNN Cell

3 Seq2Seq with Deep Learning

Neural Network + Memory = Recurrent Neural Network



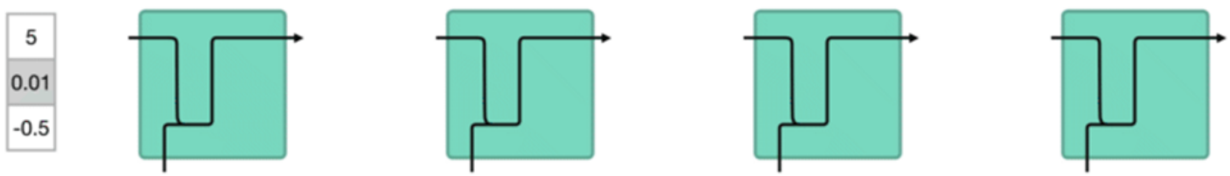
Vector Transformations without tanh



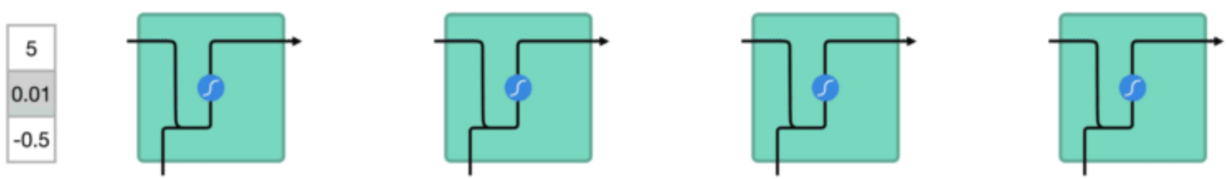
Vector Transformations with tanh

3 Seq2Seq with Deep Learning

Neural Network + Memory = Recurrent Neural Network



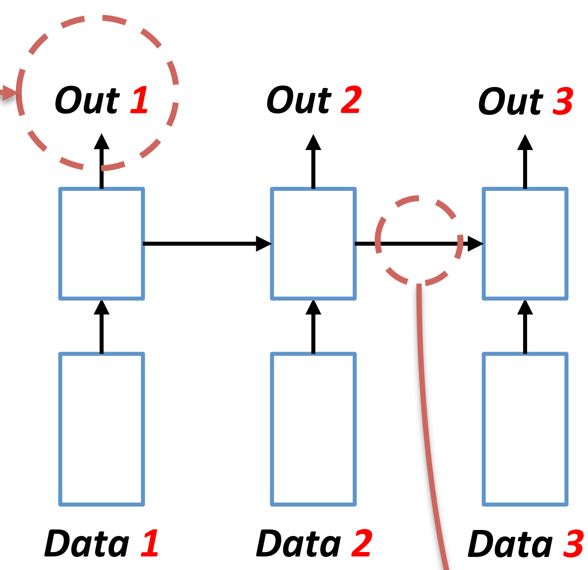
Vector Transformations without tanh



Vector Transformations with tanh

Limitation of Basic RNN

Out1 does not cover the data2 and data3

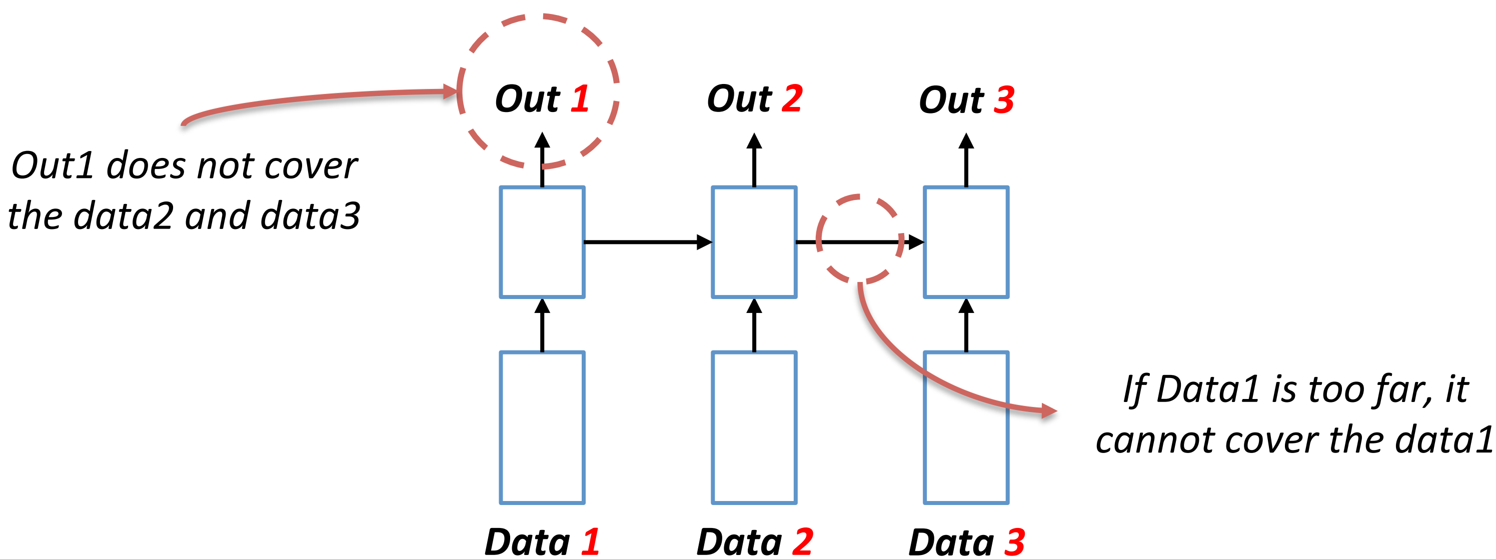


If Data1 is too far, it cannot cover the data1

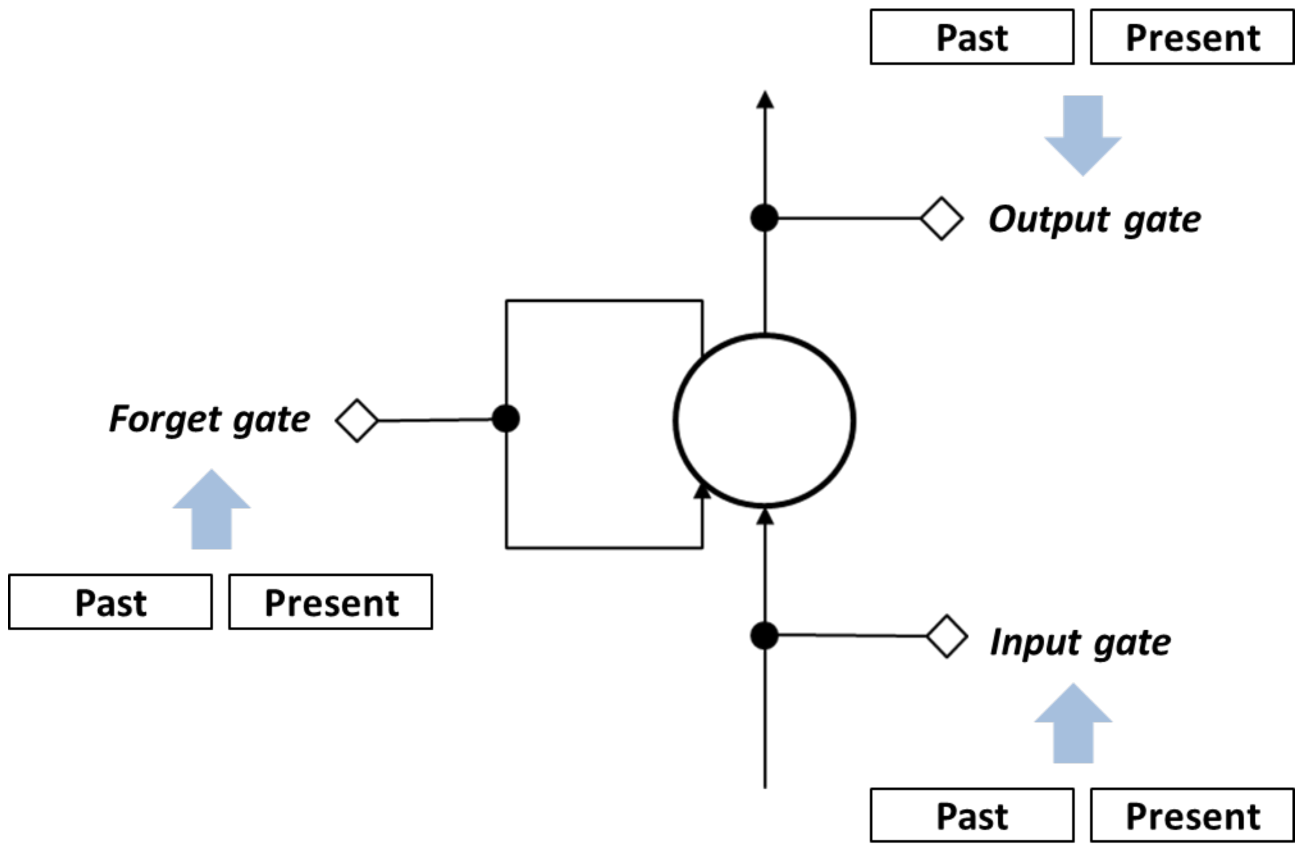
Limitation of Basic RNN

Vanishing Gradient

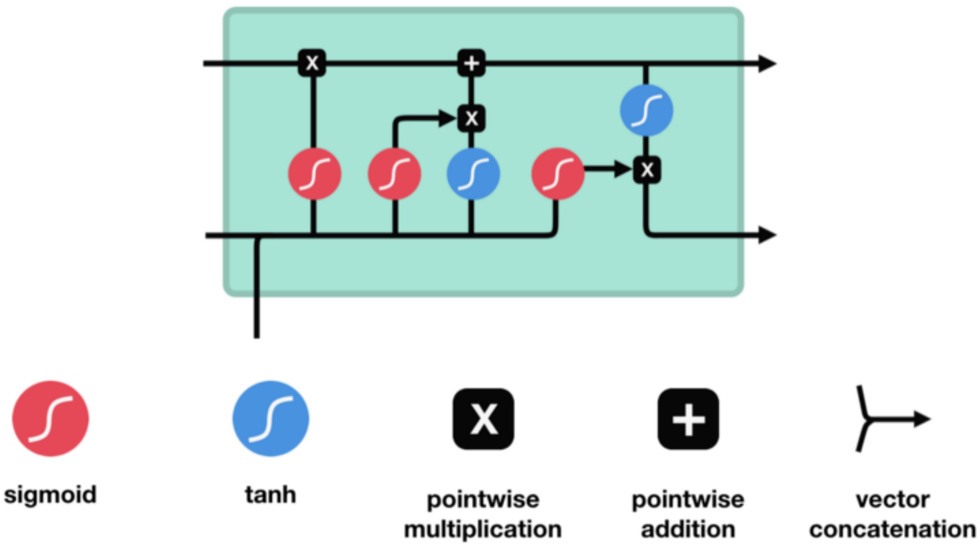
When **the values of a gradient are too small** and the model stops learning or **takes way too long** because of that. This was a major problem in the 1990s and much harder to solve than the exploding gradients



LSTM (Long Short-Term Memory) - Idea

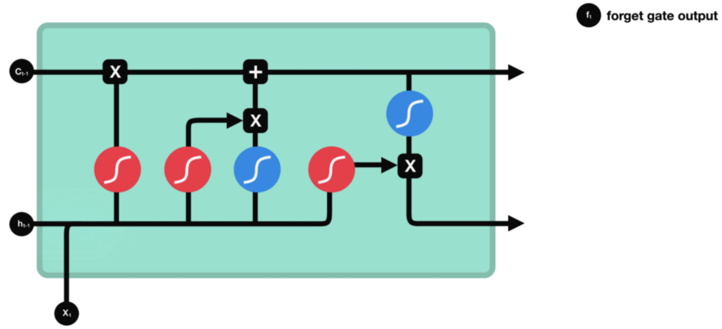


LSTM (Long Short-Term Memory) - Idea

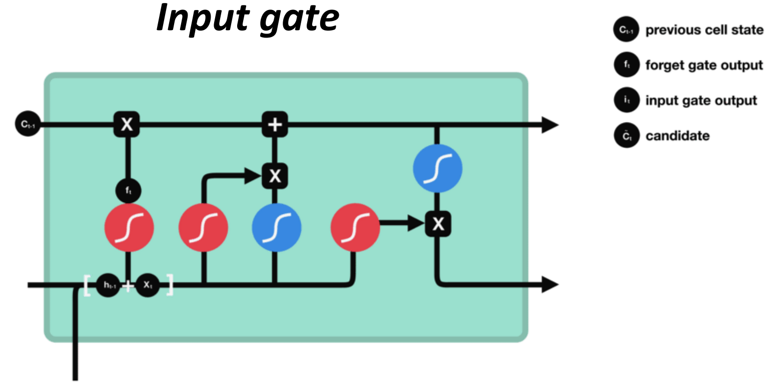


LSTM (Long Short-Term Memory) - Idea

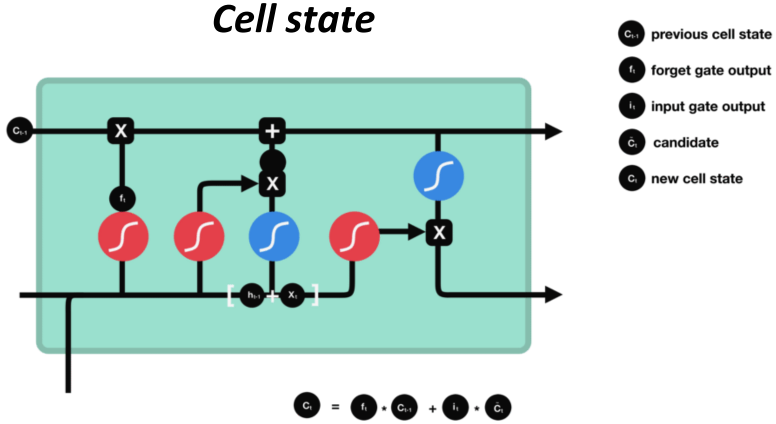
Forget gate



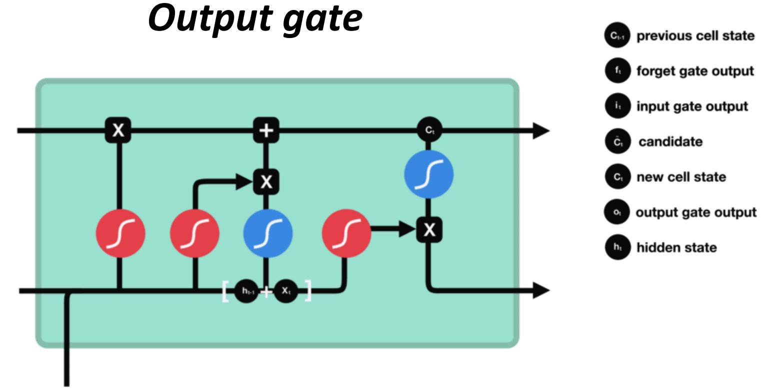
Input gate



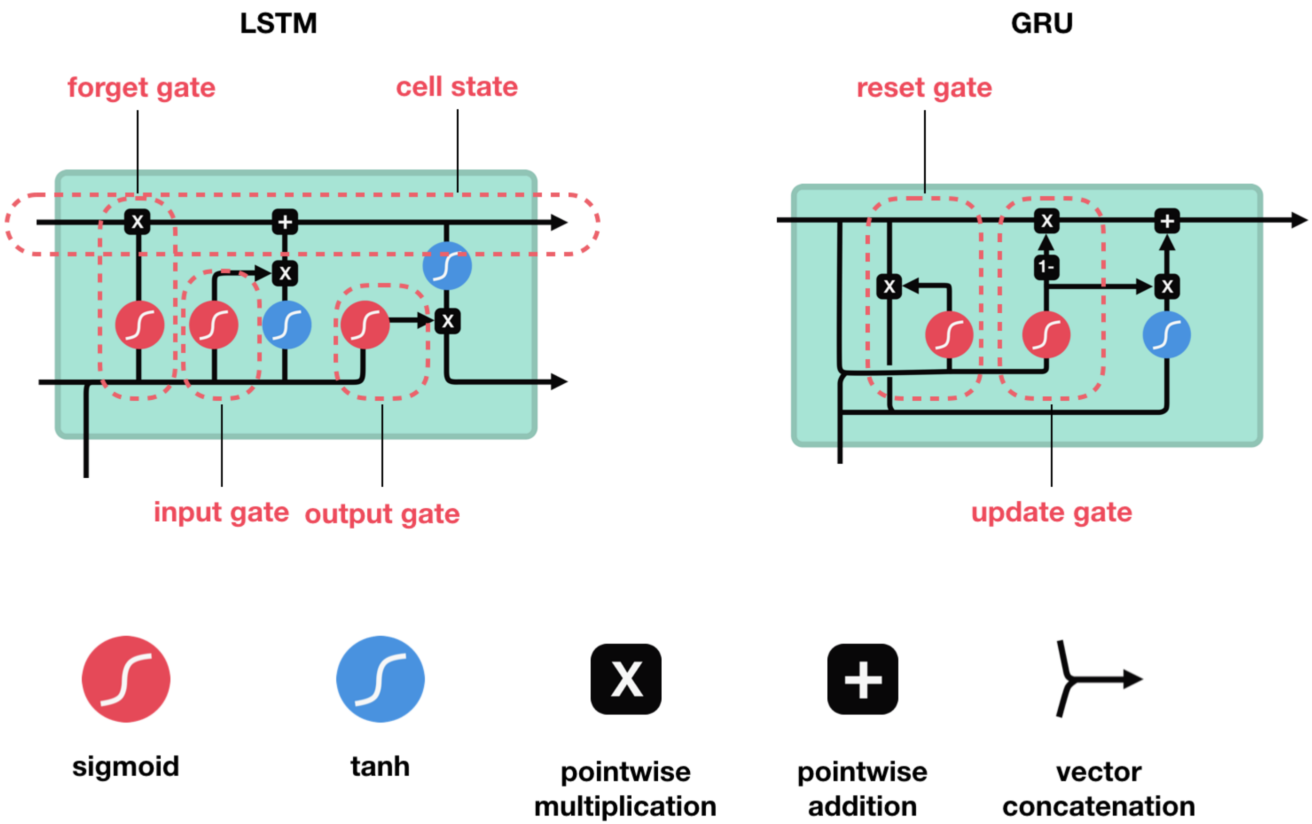
Cell state



Output gate



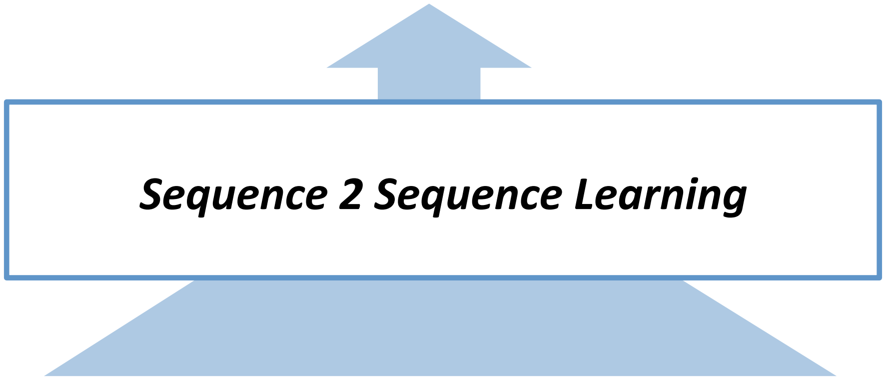
Gated Recurrent Unit



3 Seq2Seq Modelling

Seq2Seq – PoS tagger

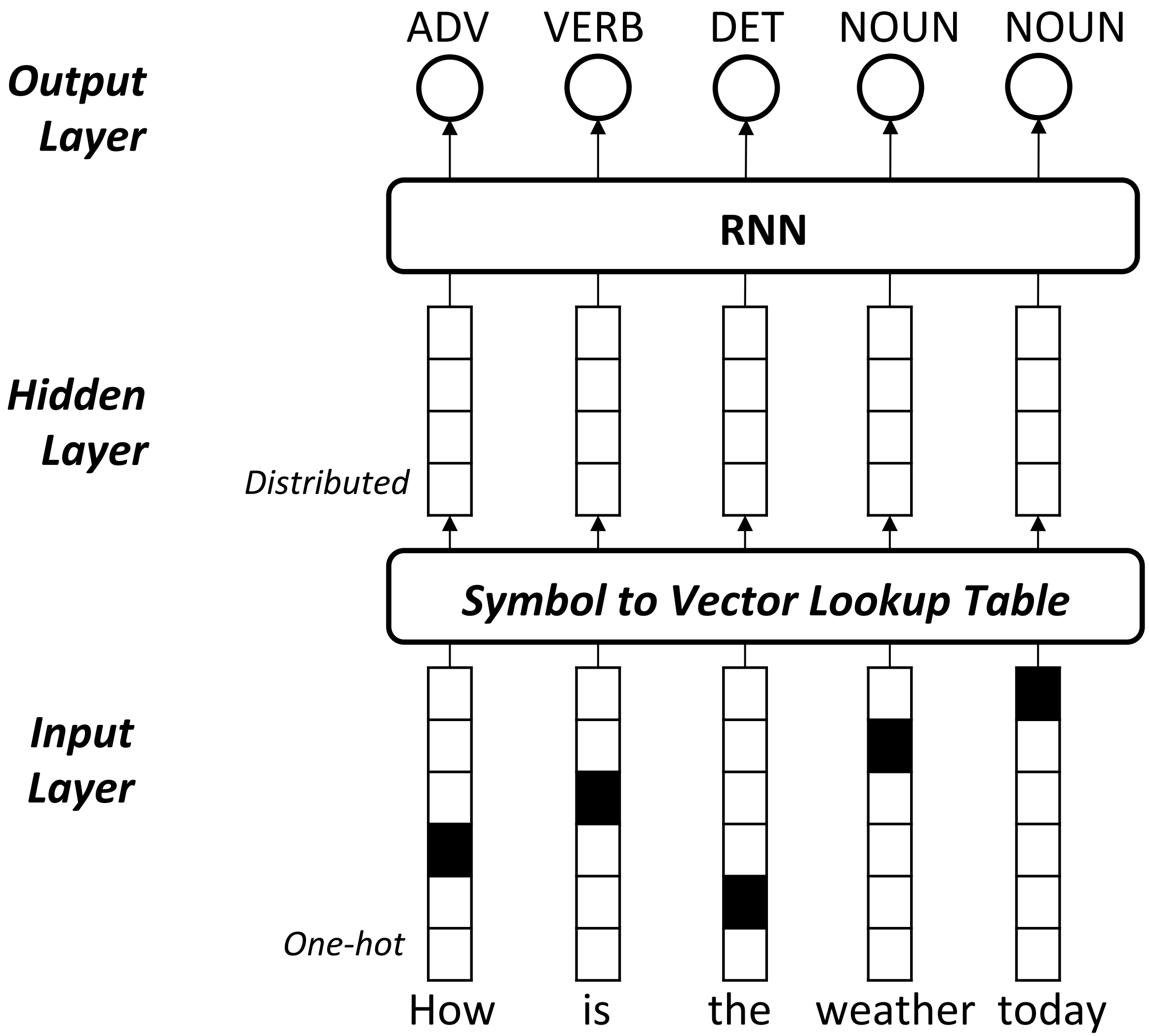
ADV VERB DET NOUN NOUN — ***Output: Part of Speech***



How is the weather today — ***Input: Text***

3 Seq2Seq Modelling

Sequence Modelling for POS Tagging



N to N

0 LECTURE PLAN

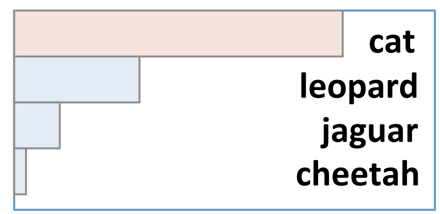
Lecture 4: Word Classification and Machine Learning 2

1. Machine Learning and NLP: Finish
2. Seq2Seq Learning
3. Seq2Seq Deep Learning
   1. RNN (Recurrent Neural Network)
   2. LSTM (Long Short-Term Memory)
   3. GRU (Gated Recurrent Unit)
4. **Seq2Seq Encoding and Decoding**
5. Next Week Preview
   – Natural Language Processing Stack

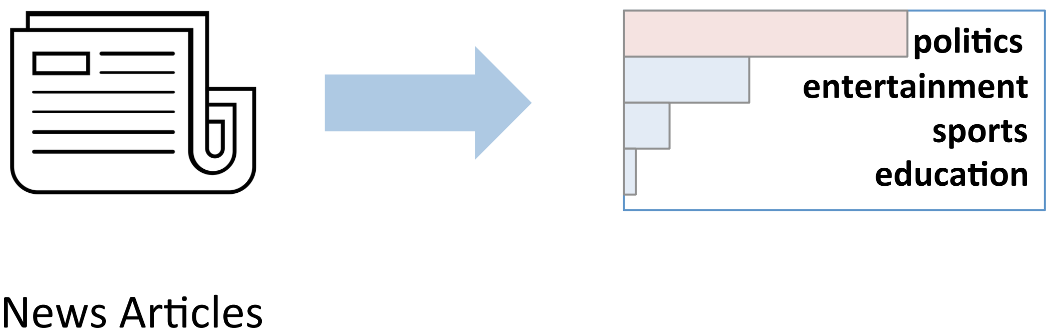
4 Seq2Seq Encoding and Decoding

ImageNet: Image Classification

Image Pixel

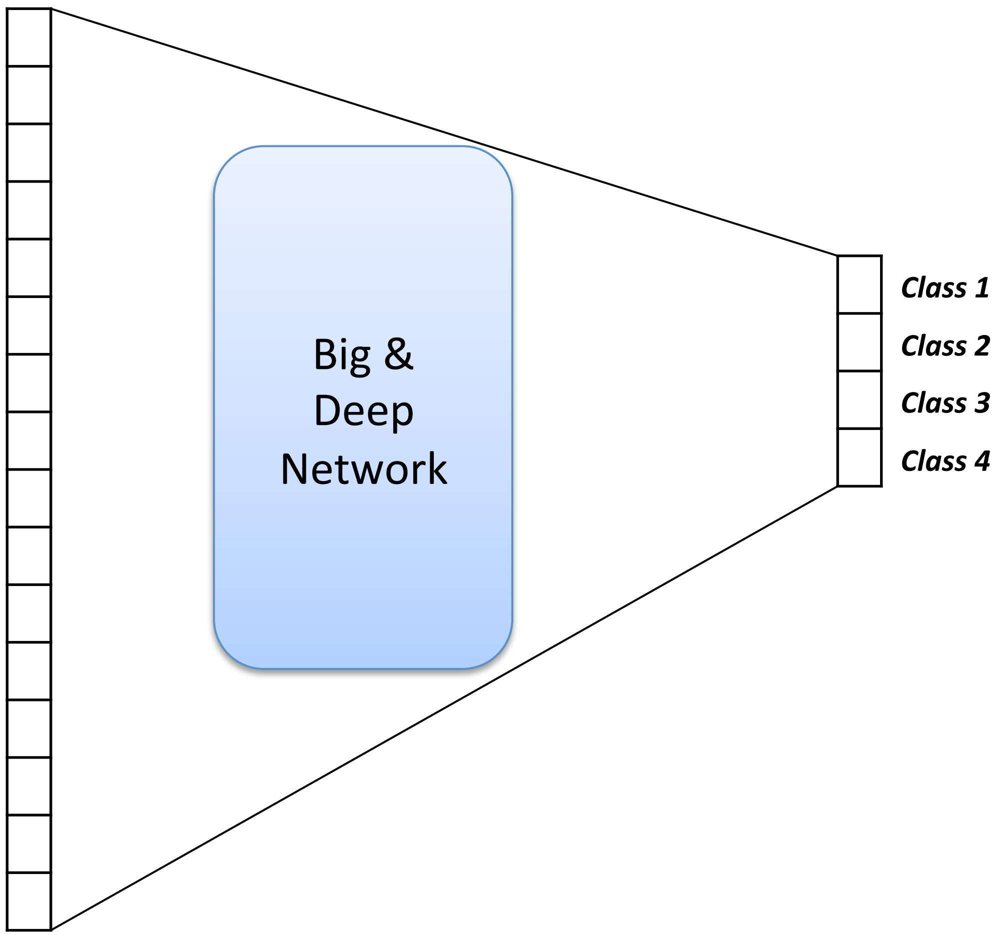


Topic Classification



4 Seq2Seq Encoding and Decoding

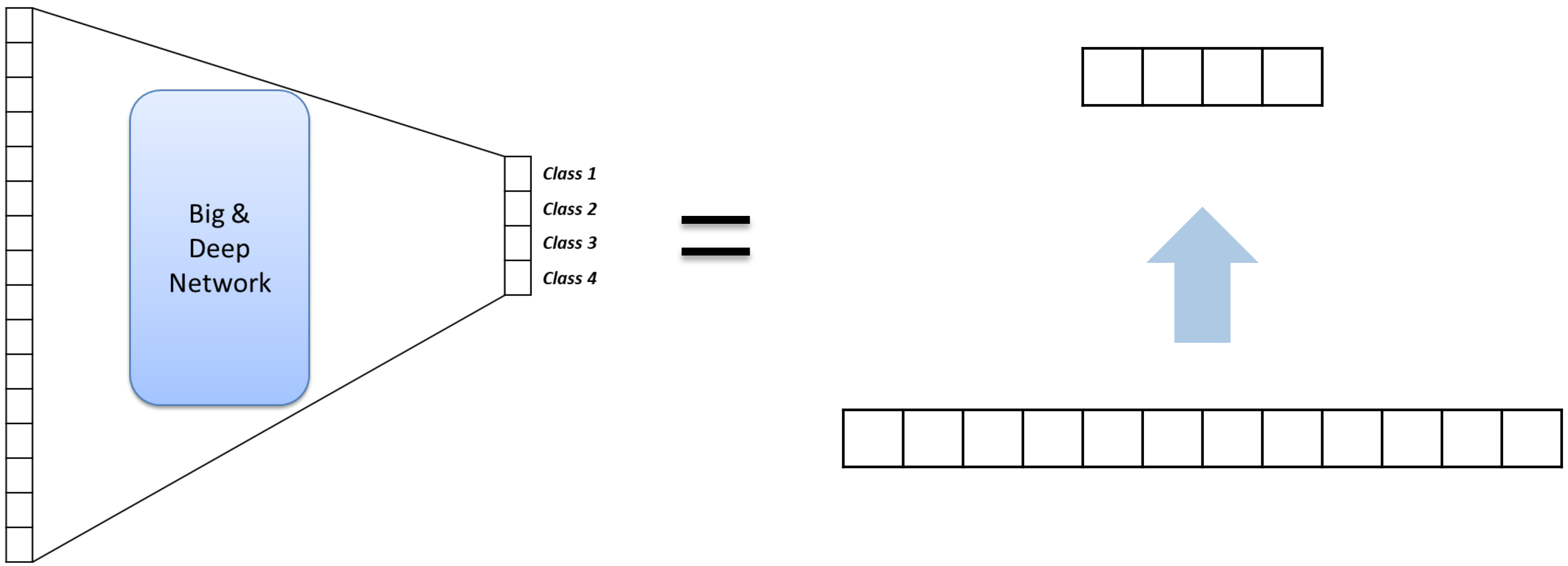
Classification Formulation



0

Seq2Seq Encoding and Decoding

Classification



4 Seq2Seq Encoding and Decoding

Graphical Notation for Data

Data X
10
2
8
2
15
3
5
1
5

=

Data X		
10	2	8
2	15	3
5	1	5

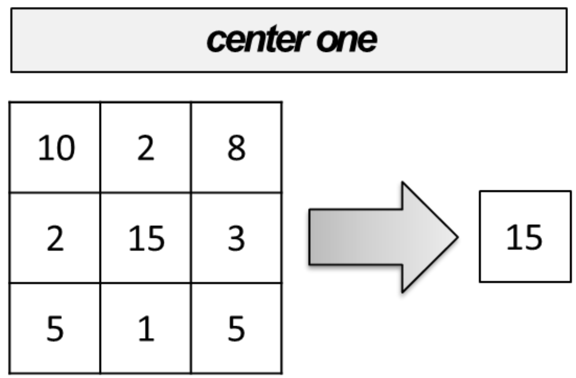
4 Seq2Seq Encoding and Decoding

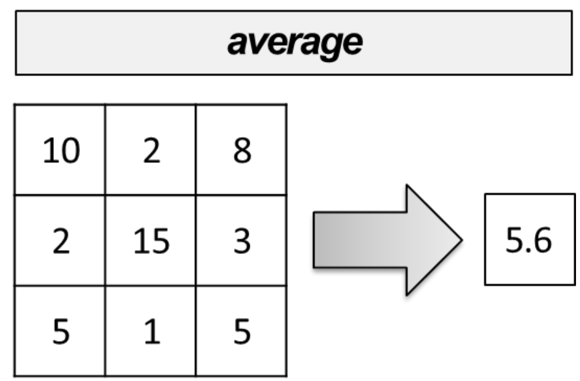
V to 1

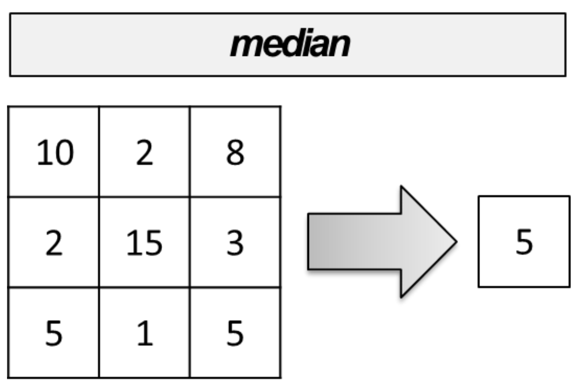
$$\begin{bmatrix} 10 \\ 2 \\ 8 \\ 2 \\ 15 \\ 3 \\ 5 \\ 1 \\ 5 \end{bmatrix} = \begin{bmatrix} 10 & 2 & 8 \\ 2 & 15 & 3 \\ 5 & 1 & 5 \end{bmatrix} \Rightarrow \boxed{?}$$

4 Seq2Seq Encoding and Decoding

V to 1 – Simple Method







V to 1 – Weighted Method

Weighted Sum

Value

10	2	8
2	15	3
5	1	5

Weight

3	1	5
2	6	3
9	3	6

→ 253

Weighted Average

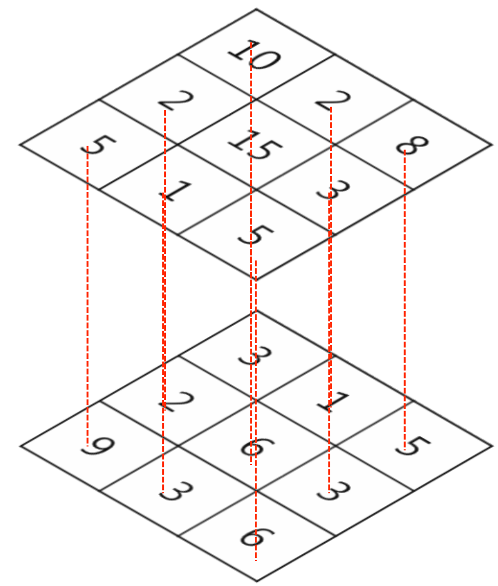
Value

10	2	8
2	15	3
5	1	5

Weight

3/9	1/9	5/9
2/9	6/9	3/9
9/9	3/9	6/9

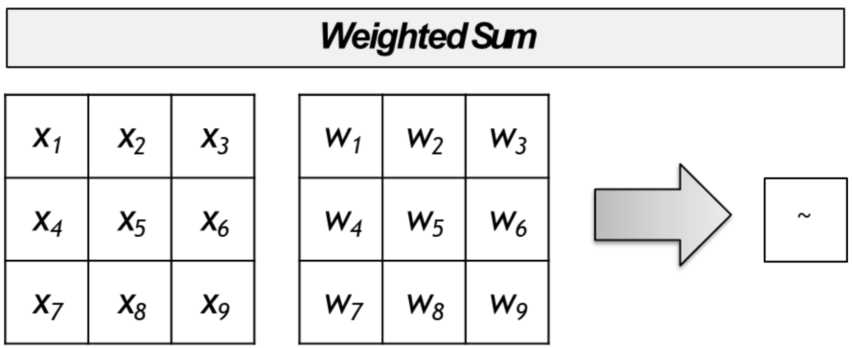
→ 6.65



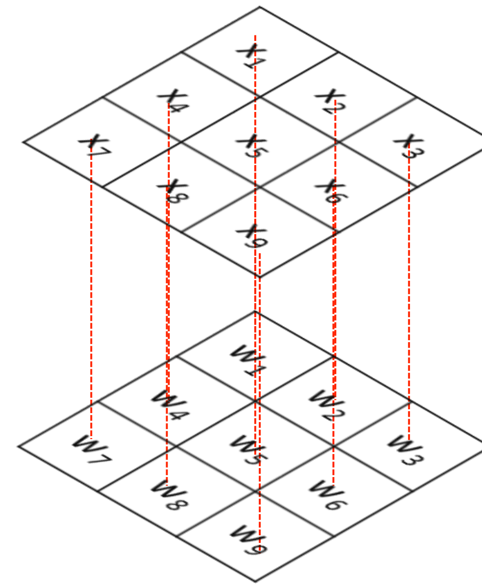
Element-wise multiplication

4 Seq2Seq Encoding and Decoding

V to 1 – General Form



$$v = x_1 * w_1 + x_2 * w_2 + \cdots + x_9 * w_9$$



Element-wise multiplication

4 Seq2Seq Encoding and Decoding

V to 1 – Linear Algebra

[1 x 9] matrix

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \end{bmatrix}$$

X

[9x1] matrix

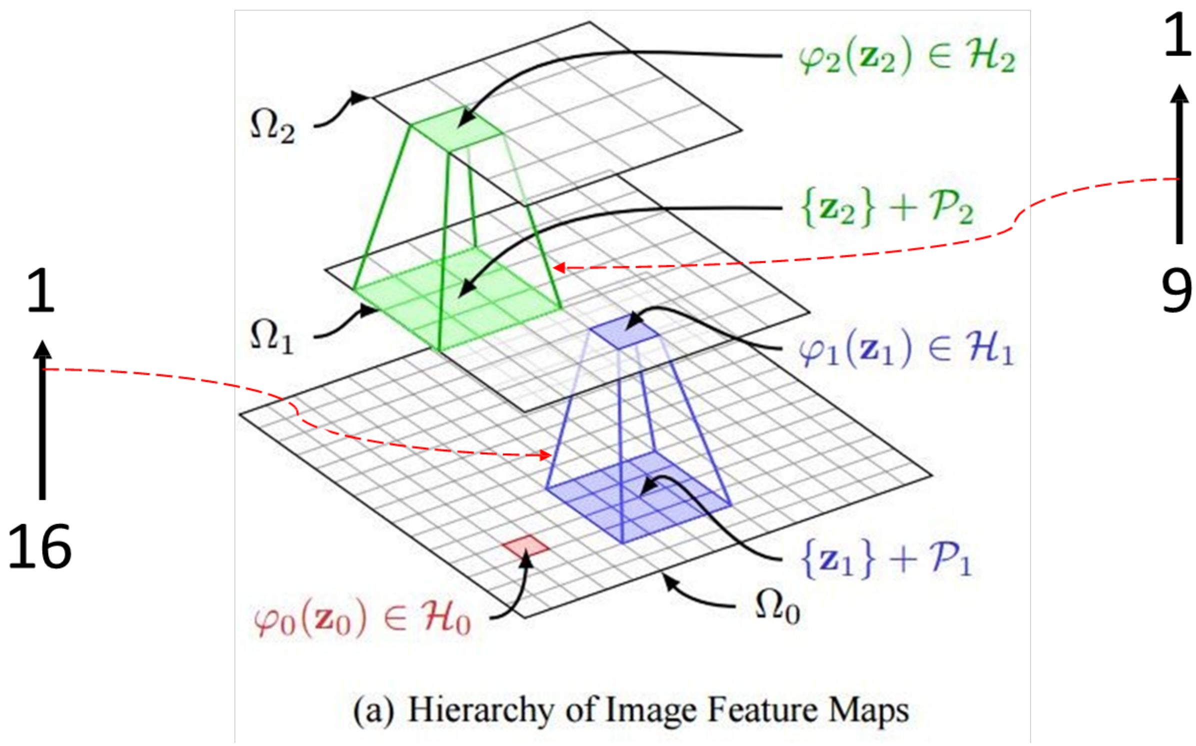
$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \end{bmatrix}$$

=

[1x1] matrix

$$\sum_{i}^{9} x_i * w_i$$

Convolution Neural Network (1)

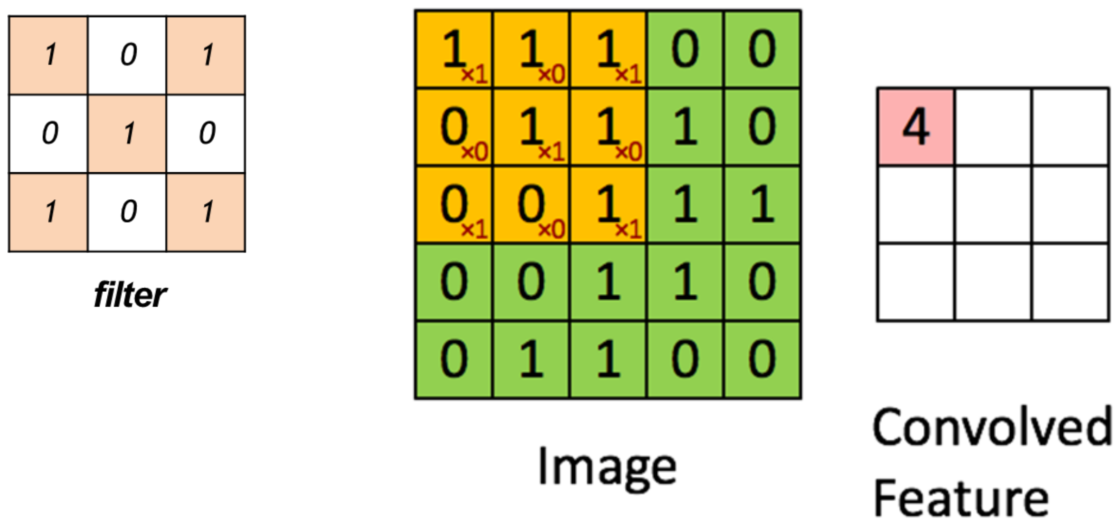


Data Abstraction

Mairal et al., Convolutional Kernel Networks, 2014

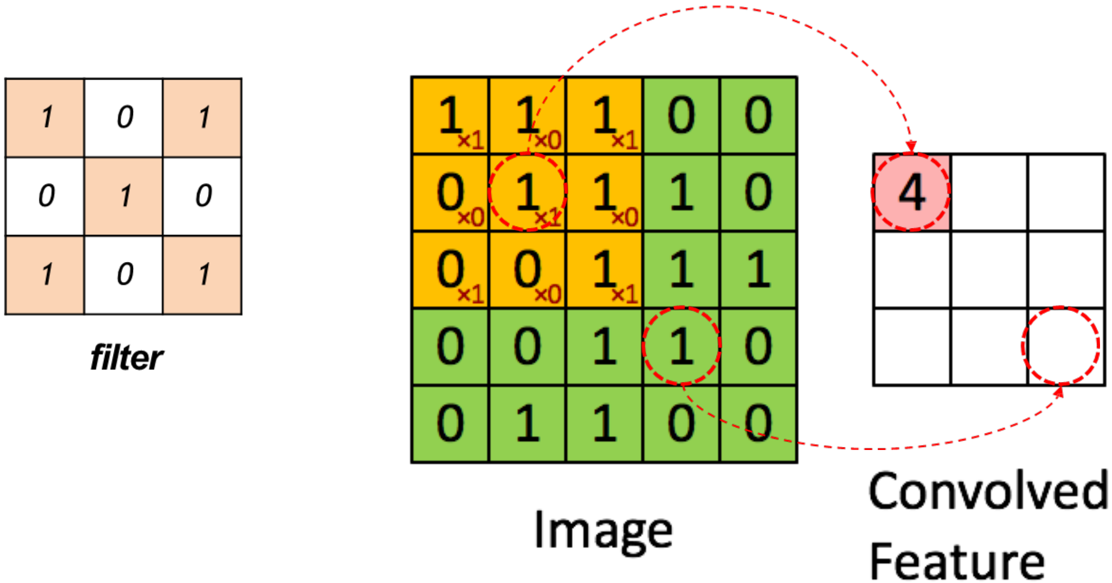
4 Seq2Seq Encoding and Decoding

Convolution Neural Network (2)



Mairal et al., Convolutional Kernel Networks, 2014

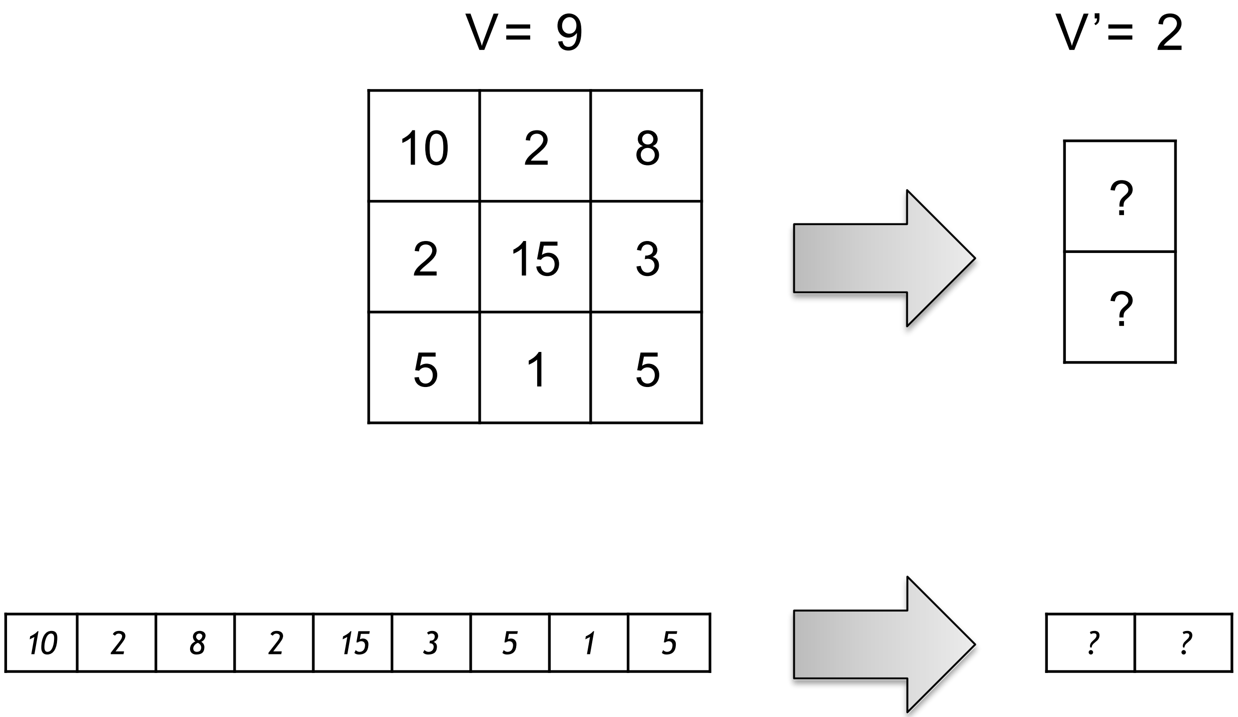
Convolution Neural Network (2)



Mairal et al., Convolutional Kernel Networks, 2014

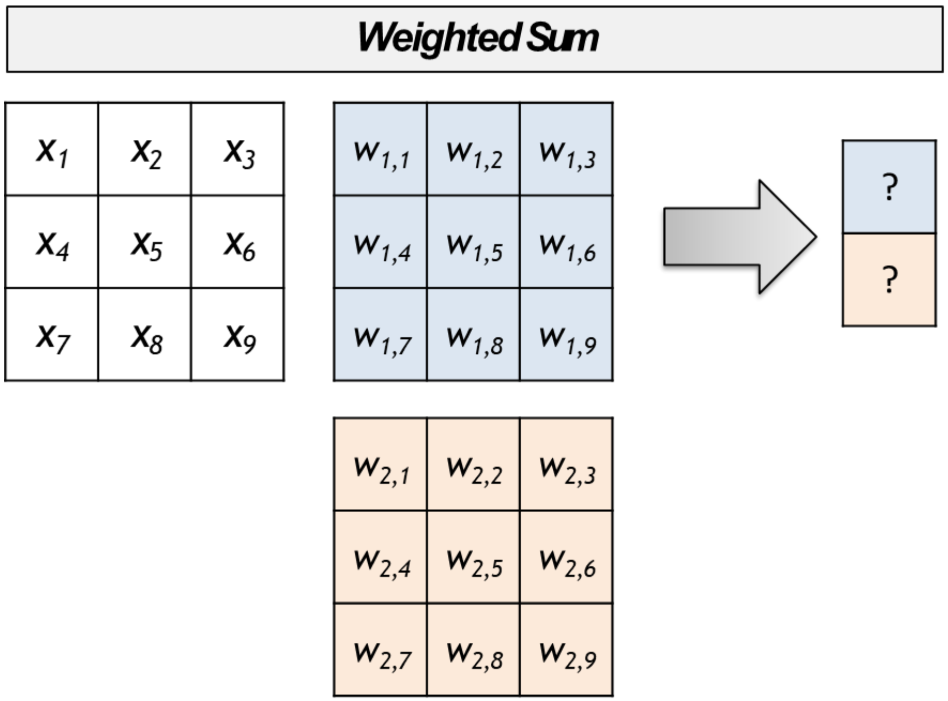
4 Seq2Seq Encoding and Decoding

V to V'



4 Seq2Seq Encoding and Decoding

V to V' – generalized method



$$\begin{bmatrix} v_1 = x_1 * w_{1,1} + x_2 * w_{1,1} + \cdots + x_9 * w_{1,9} \\ v_2 = x_1 * w_{2,1} + x_2 * w_{2,1} + \cdots + x_9 * w_{2,9} \end{bmatrix}$$

4 Seq2Seq Encoding and Decoding

V to V' – generalized method

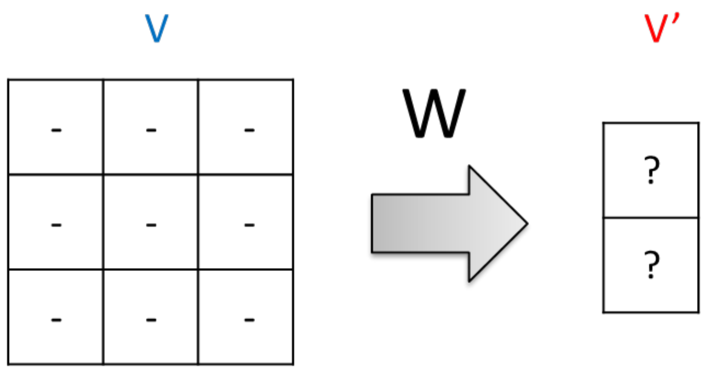
Weighted Sum

[1 x 9] matrix × [9x2] matrix = [1x2] matrix

$$
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \end{bmatrix}
\times
\begin{bmatrix}
w_{1,1} & w_{2,1} \\
w_{1,2} & w_{2,2} \\
w_{1,3} & w_{2,3} \\
w_{1,4} & w_{2,4} \\
w_{1,5} & w_{2,5} \\
w_{1,6} & w_{2,6} \\
w_{1,7} & w_{2,7} \\
w_{1,8} & w_{2,8} \\
w_{1,9} & w_{2,9}
\end{bmatrix}
=
\left[\sum_{i}^{9} x_i * w_{1,i} \; , \; \sum_{i}^{9} x_i * w_{2,i} \right]
$$

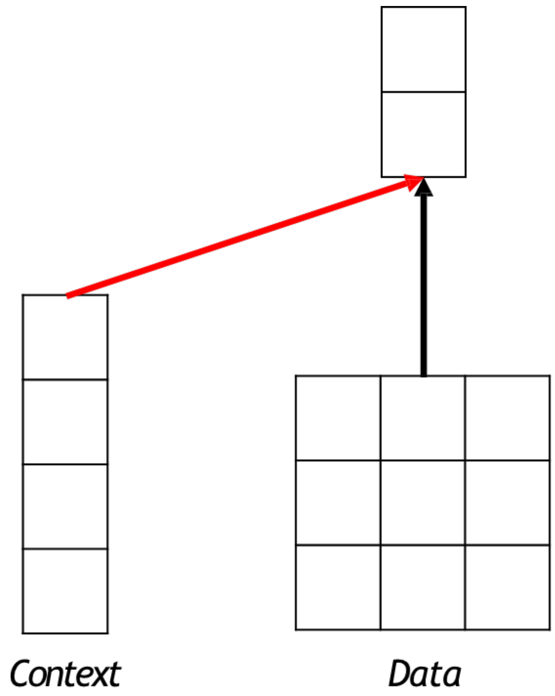
Fully Connected Network

V to V’ – Projection Notation

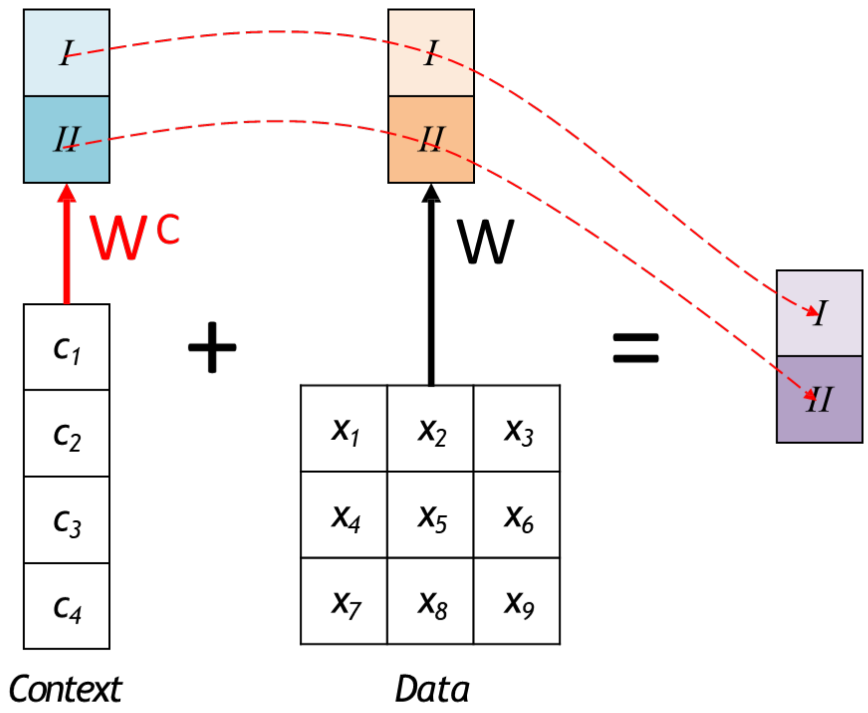


$$W = \begin{array}{cc} & V' \\ V & \begin{bmatrix} \quad & \quad \\ \quad & \quad \\ \quad & \quad \end{bmatrix} \end{array}$$

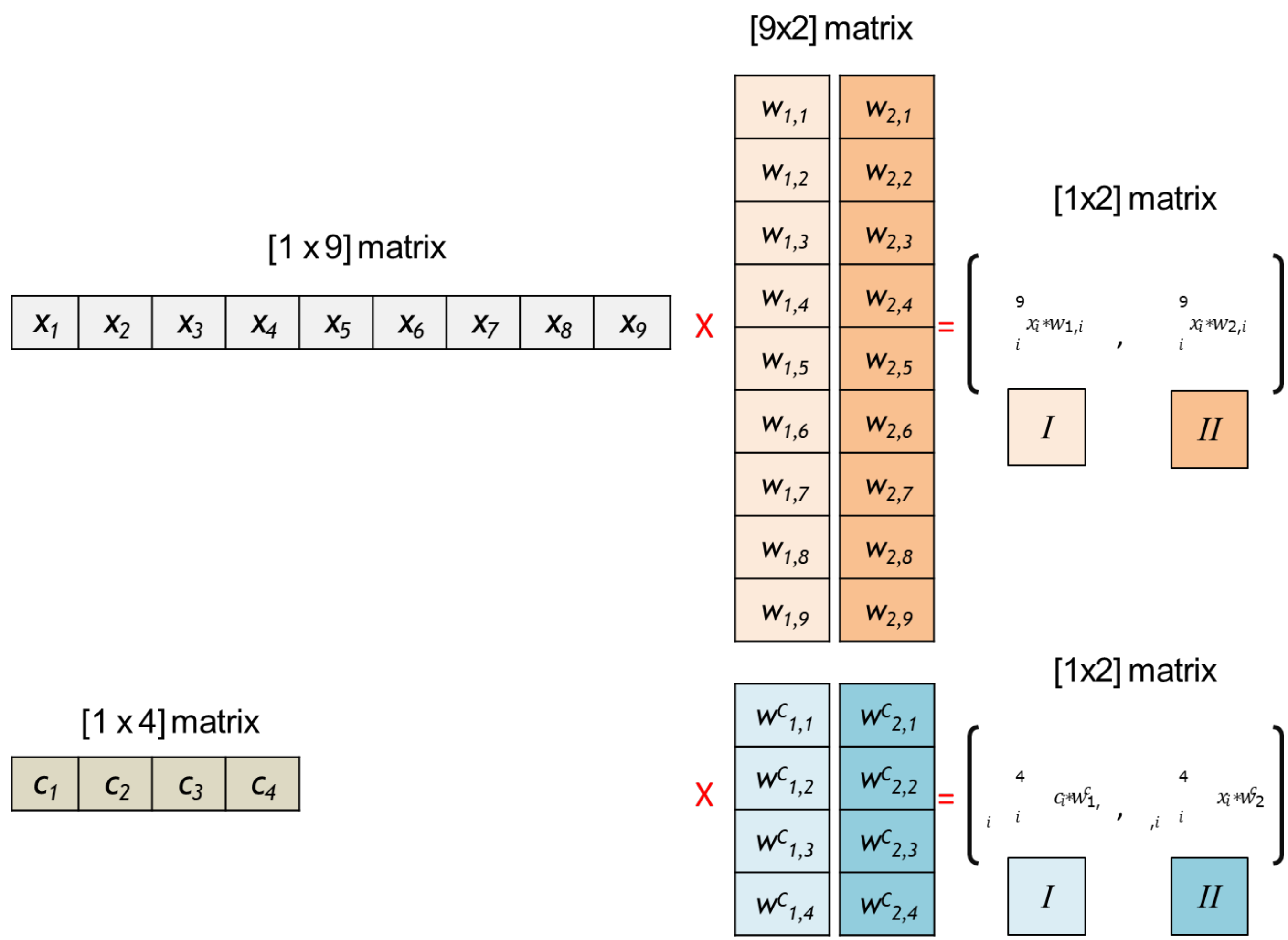
V to V’ – Projection with Context (1)



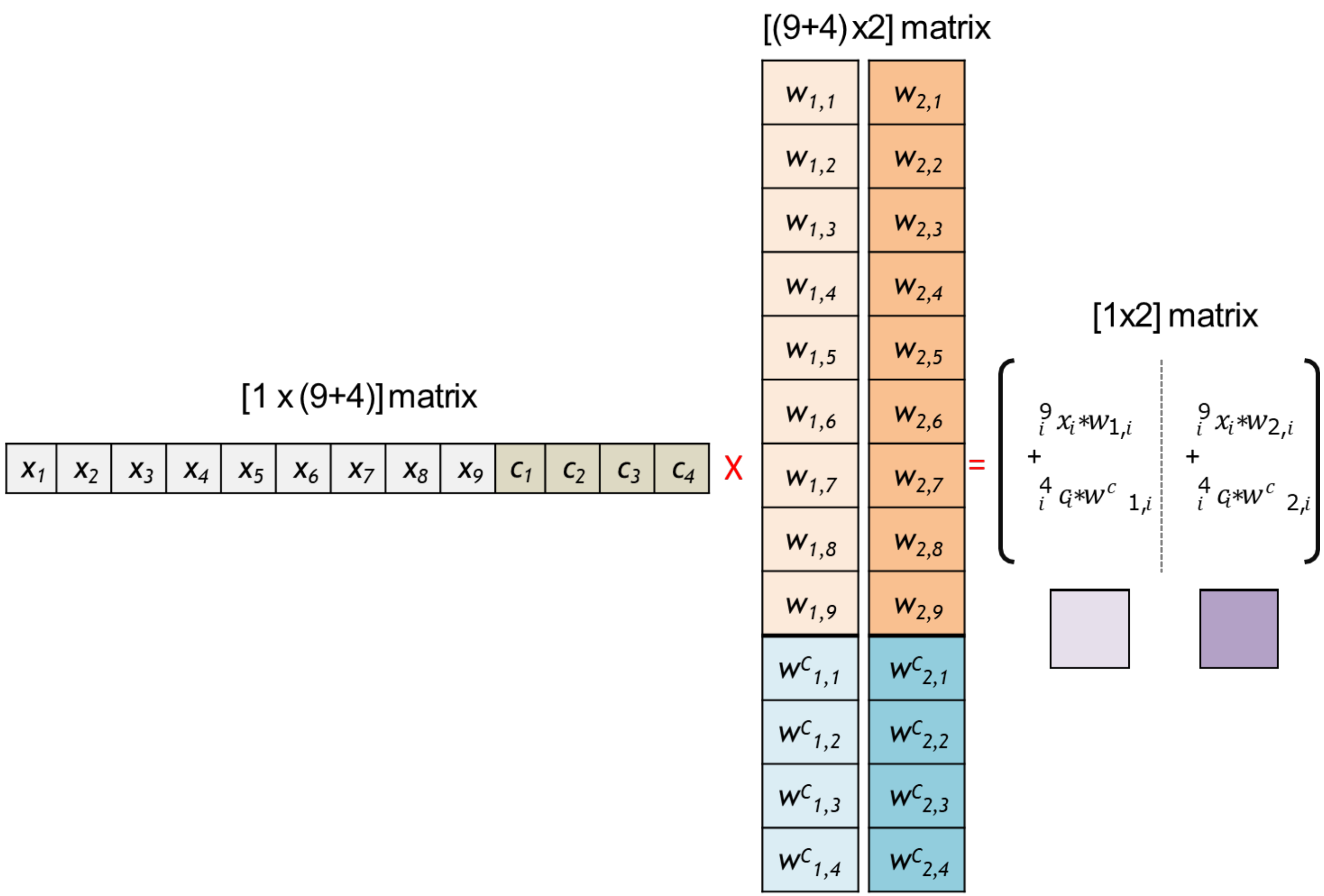
V to V' – Projection with Context (2)



V to V' with Context - Linear Algebra



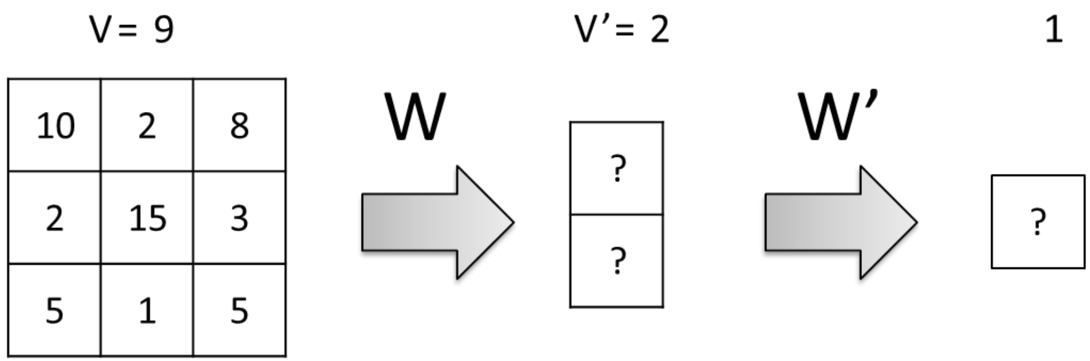
V to V' with Context - Linear Algebra (Simplified)

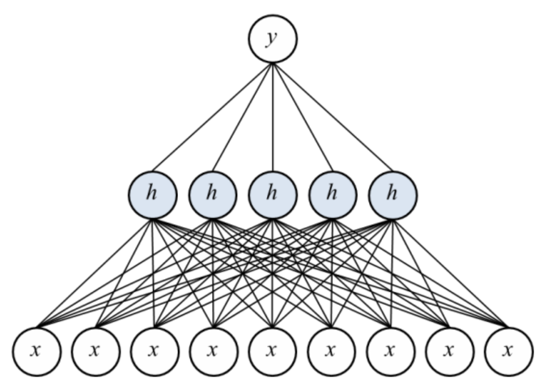


THE UNIVERSITY OF SYDNEY

4 Seq2Seq Encoding and Decoding

V → V' → 1

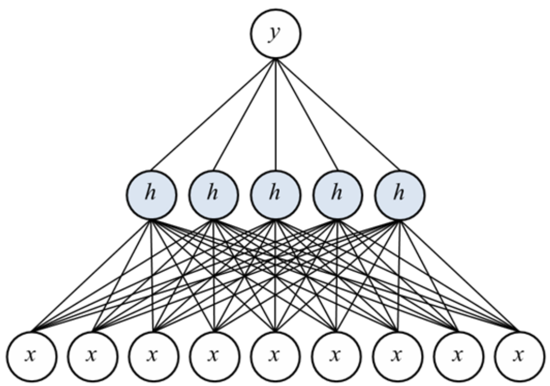




Single Layer

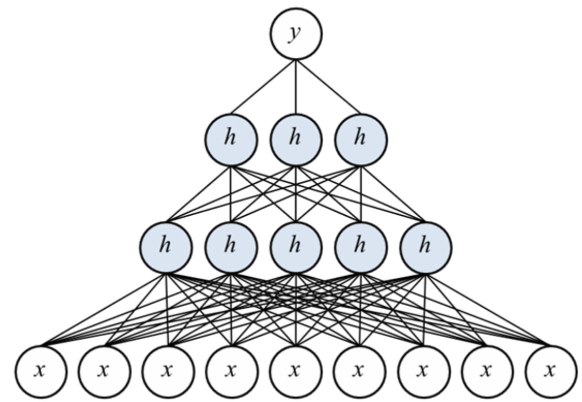
Seq2Seq Encoding and Decoding

V → V' → 1



Single Layer

V →V' →1



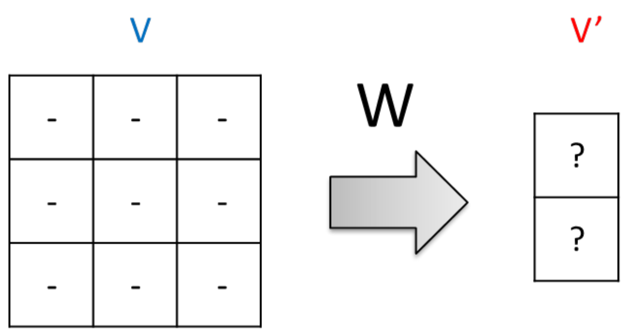
Multilayer

V →V' →V'' →1

4 Seq2Seq Encoding and Decoding

Seq2Seq Encoding

Single Item Summarisation



Multiple Item Summarisation

?

Multiple Item Summarisation

10	2	8
2	15	3
5	1	5

Data 1

13	4	8
4	5	2
1	45	31

Data 2

6	3	4
1	7	1
3	4	0

Data 3

?	?	?
?	?	?
?	?	?

V

?
?

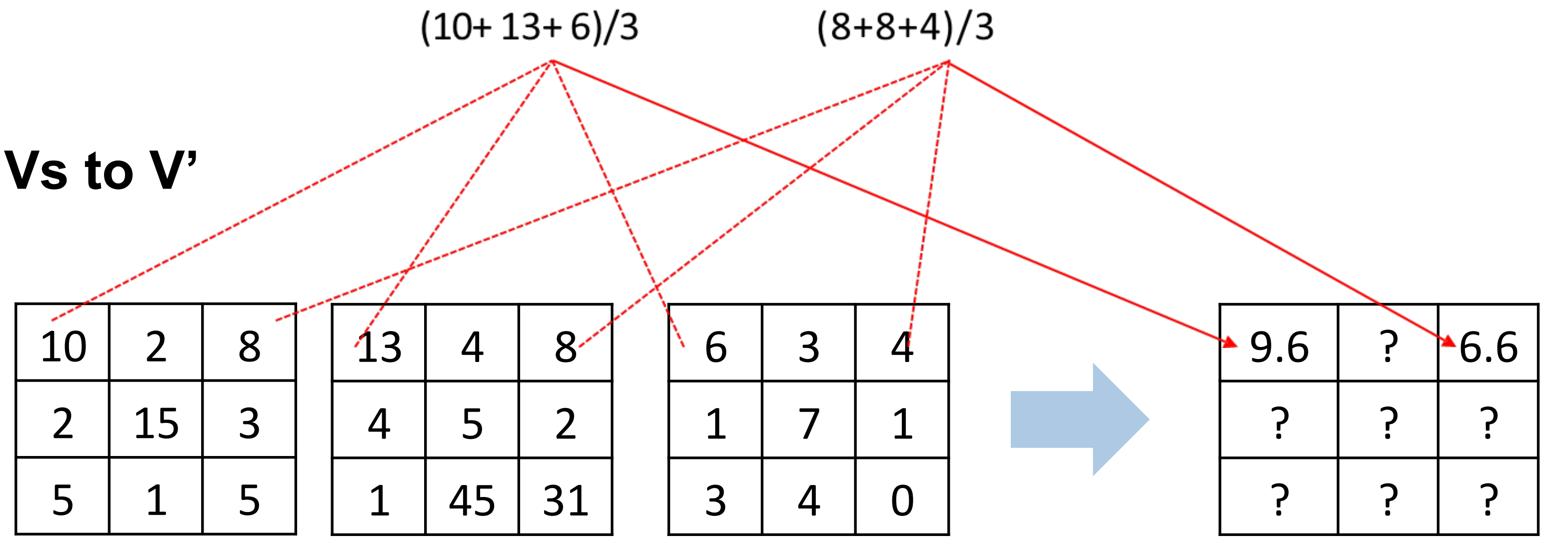
V'

?

1

4 Seq2Seq Encoding and Decoding

Vs to V’



Element-wise Average

Seq2Seq Encoding and Decoding

Vs to V'

$$\begin{bmatrix} 10 & 2 & 8 \\ 2 & 15 & 3 \\ 5 & 1 & 5 \end{bmatrix} \quad \begin{bmatrix} 13 & 4 & 8 \\ 4 & 5 & 2 \\ 1 & 45 & 31 \end{bmatrix} \quad \begin{bmatrix} 6 & 3 & 4 \\ 1 & 7 & 1 \\ 3 & 4 & 0 \end{bmatrix}$$

x $w^1 = 0.2$ x $w^2 = 0.4$ x $w^3 = 0.4$

Element-wise multiplication

$$\begin{bmatrix} 2 & 0.4 & 1.6 \\ 0.4 & 3 & 0.6 \\ 1 & 0.2 & 1.0 \end{bmatrix} + \begin{bmatrix} 5.2 & 1.6 & 3.2 \\ 1.6 & 2 & 0.8 \\ 0.4 & 18 & 12.4 \end{bmatrix} + \begin{bmatrix} 2.4 & 1.2 & 1.6 \\ 0.4 & 2.8 & 0.4 \\ 1.2 & 1.6 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 9.6 & 3.2 & 6.4 \\ 2.4 & 7.8 & 1.8 \\ 2.6 & 19.8 & 13.4 \end{bmatrix}$$

Element-wise summation

Seq2Seq Encoding and Decoding

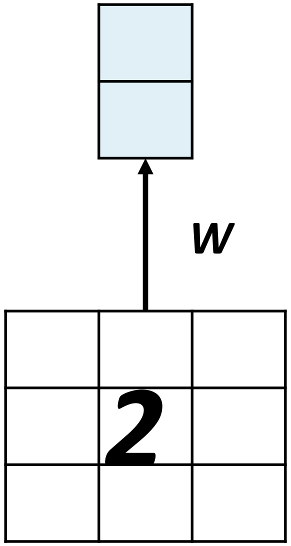
Temporal Summarisation

Context

How to include Temporal information?

4 Seq2Seq Encoding and Decoding

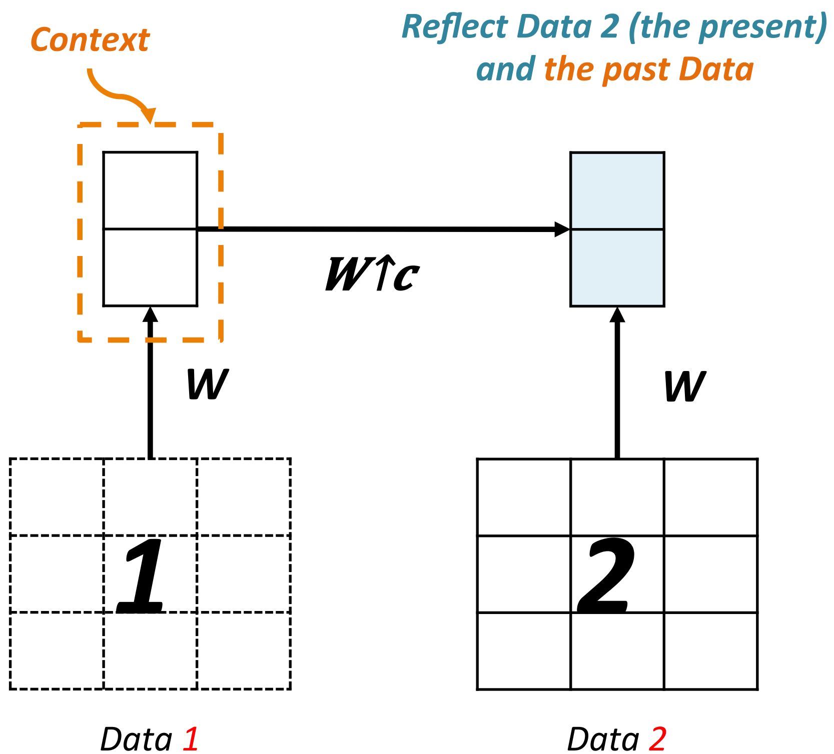
Vs → V’s → V’



Data 2

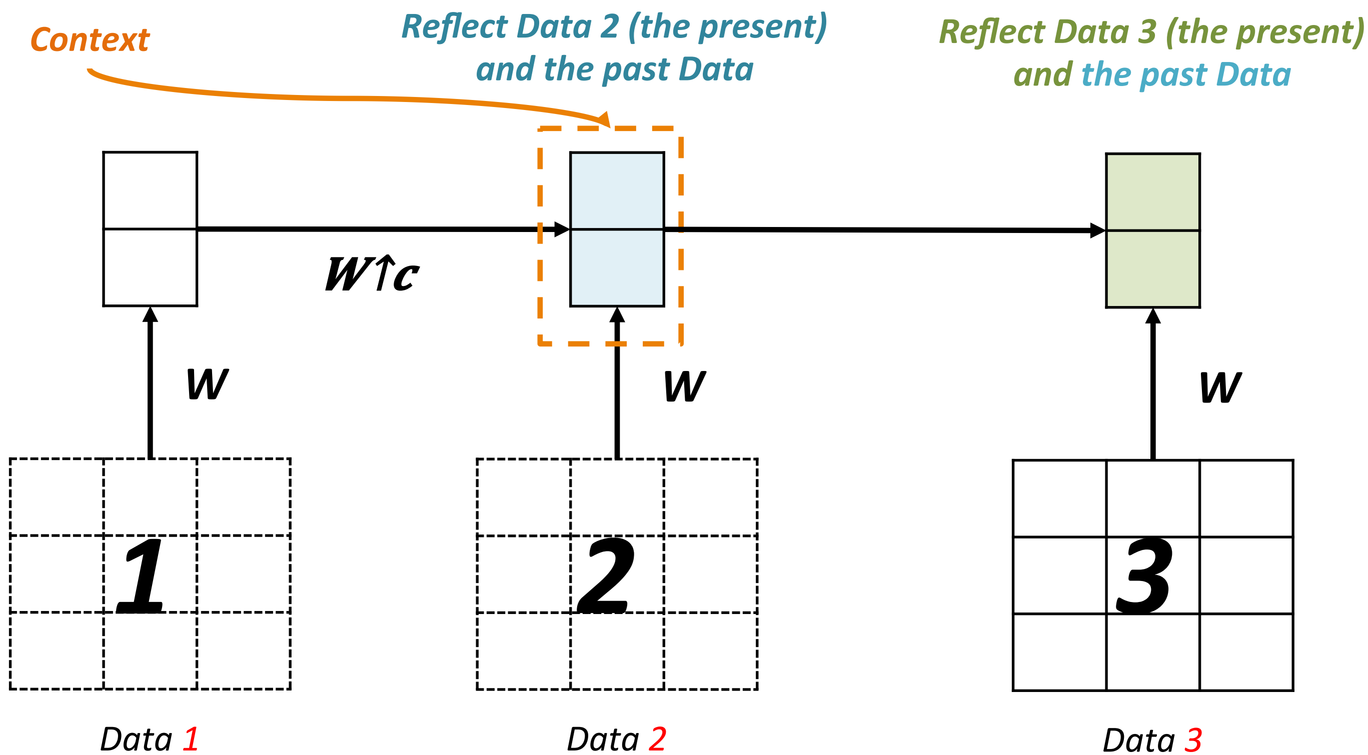
4 Seq2Seq Encoding and Decoding

Vs → V's → V'



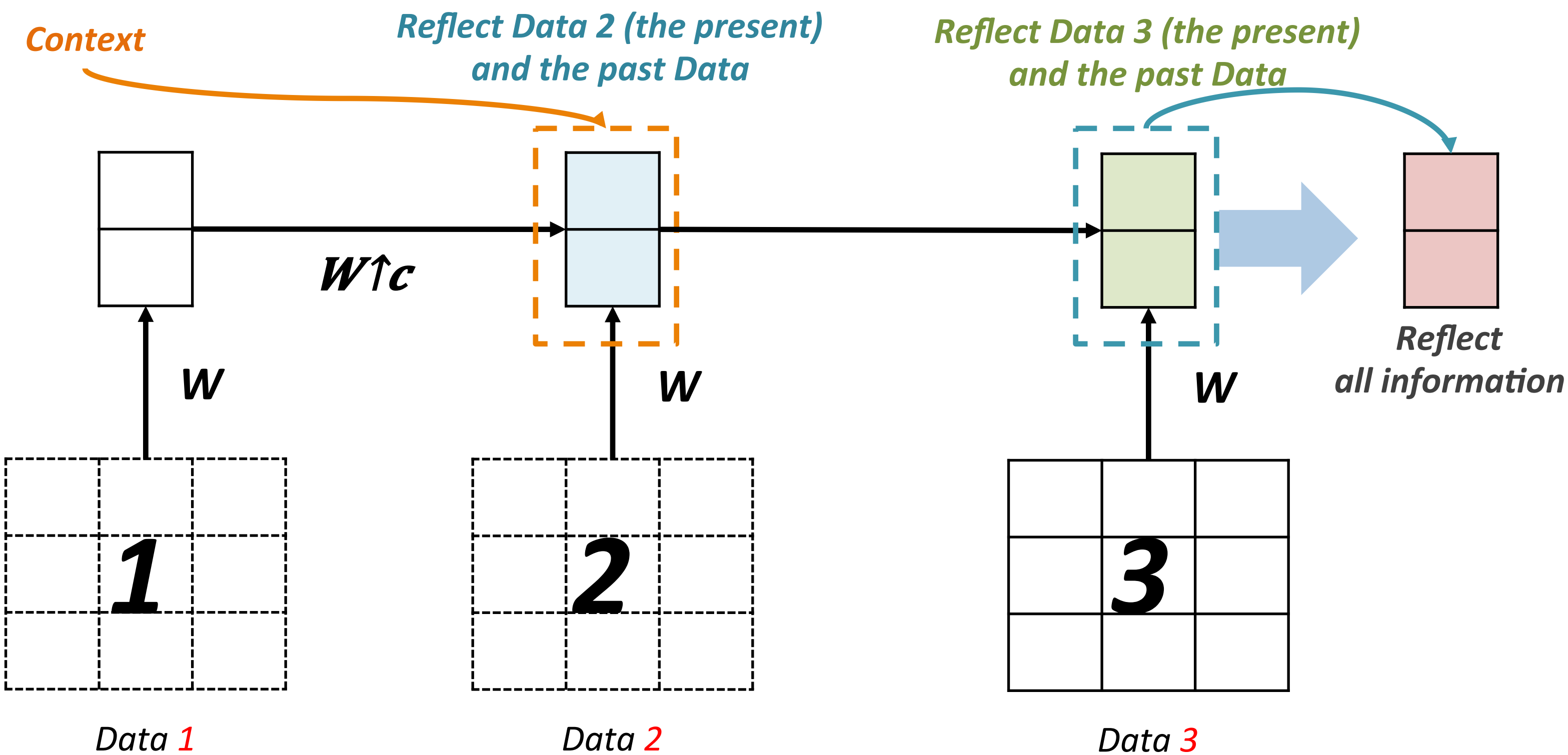
4 Seq2Seq Encoding and Decoding

Vs → V's → V'



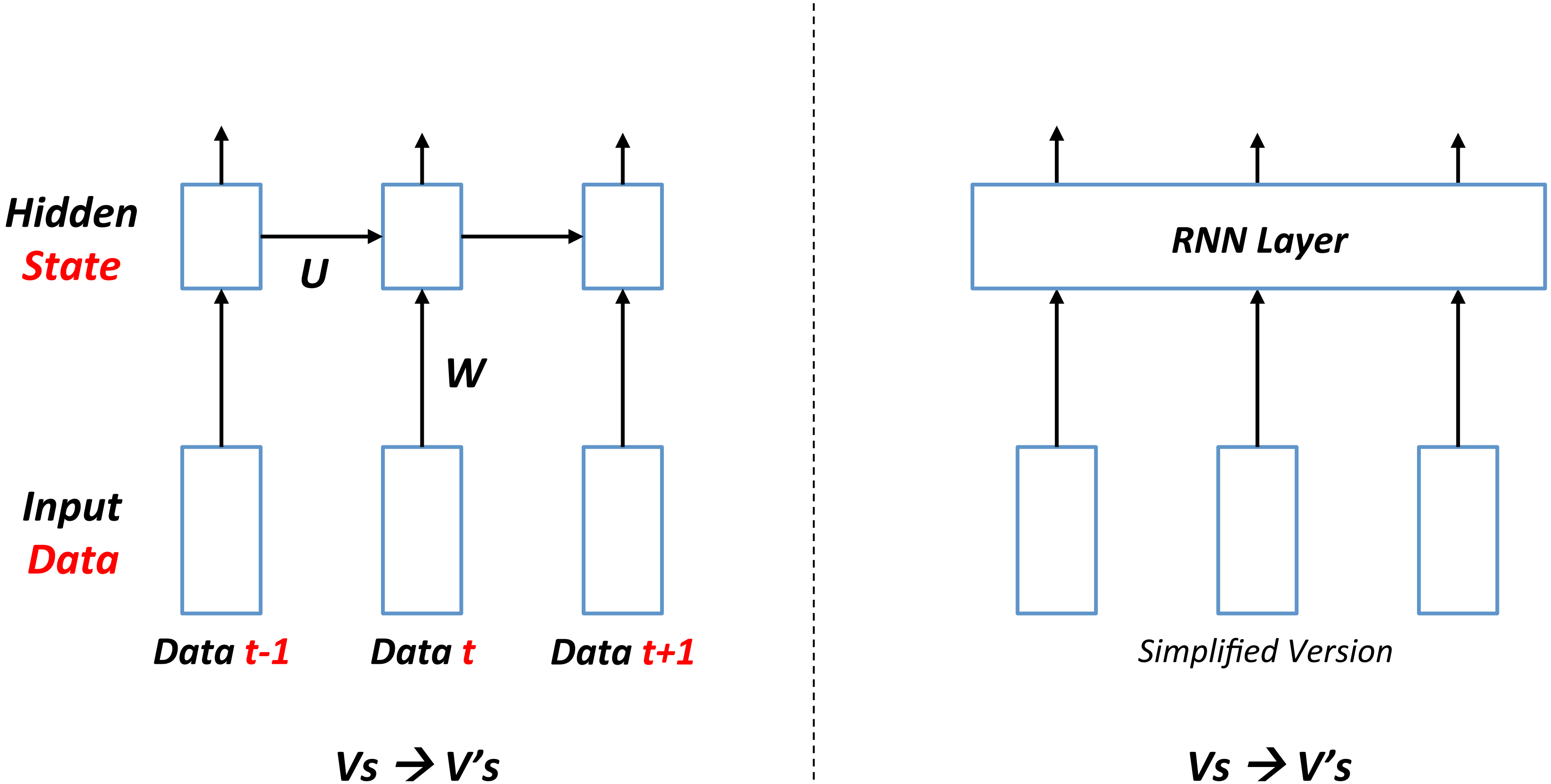
5 Seq2Seq Encoding and Decoding

Vs → V's → V'



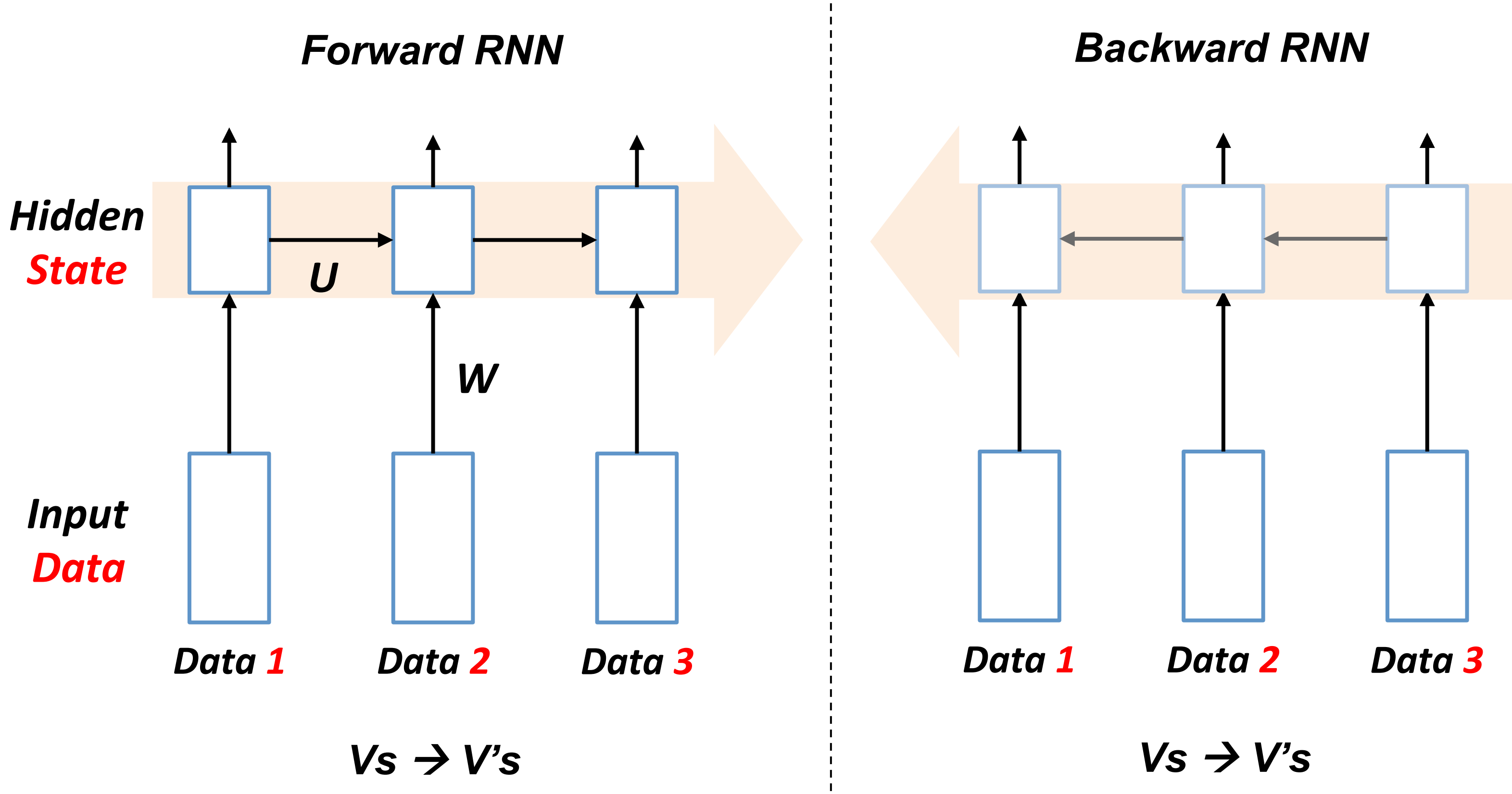
5 # Recurrent Neural Network

Graphical Notation

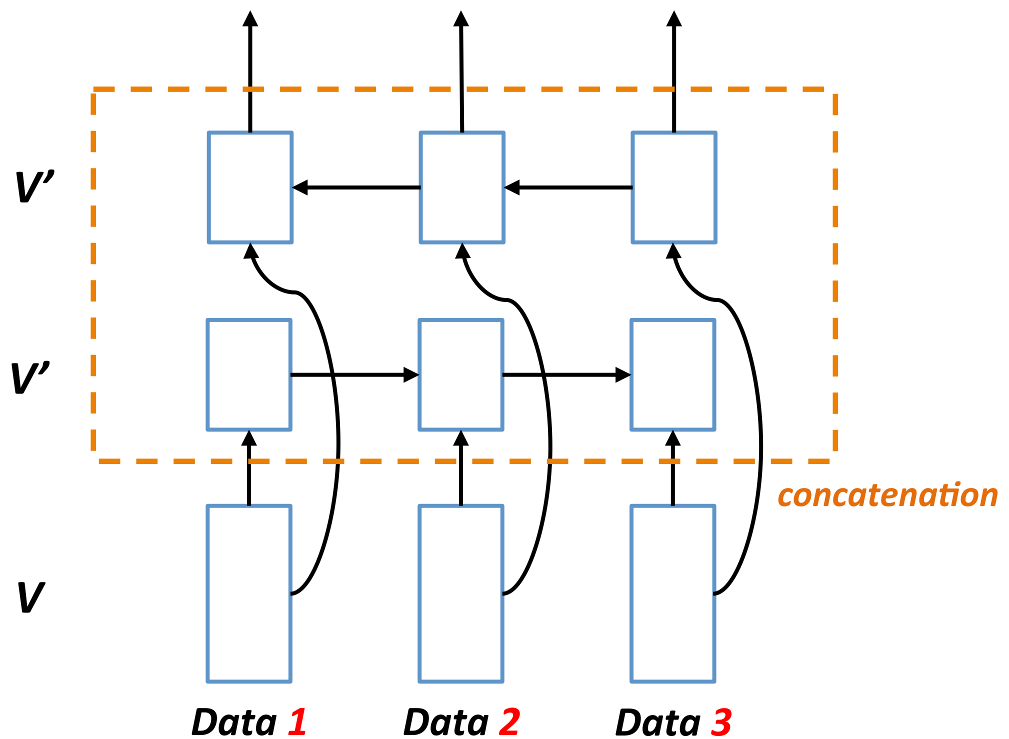


5 Recurrent Neural Network

Forward/Backward RNN



Bidirectional RNN

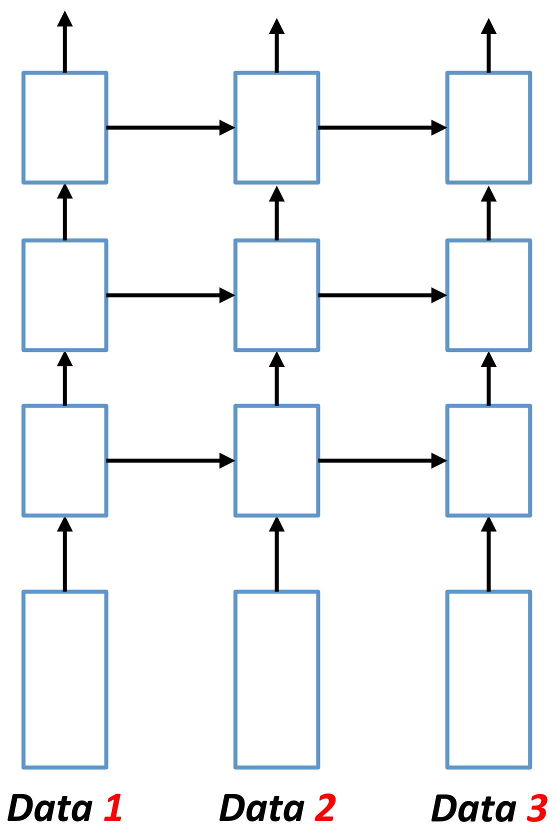


Vs → (2*V')s

5 Recurrent Neural Network

THE UNIVERSITY OF SYDNEY

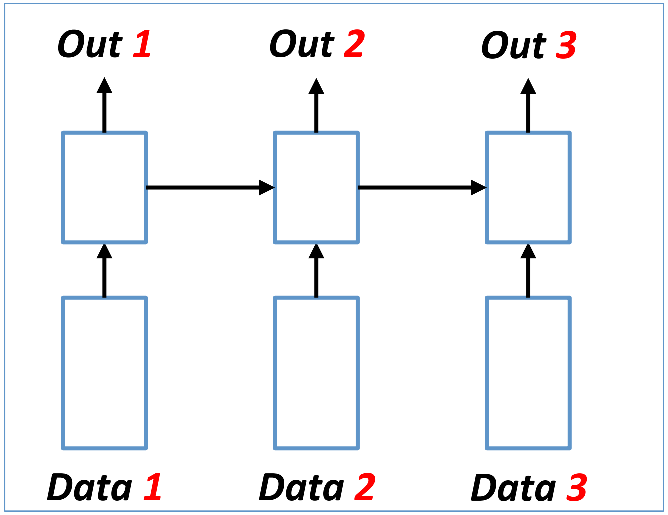
Stacking RNN



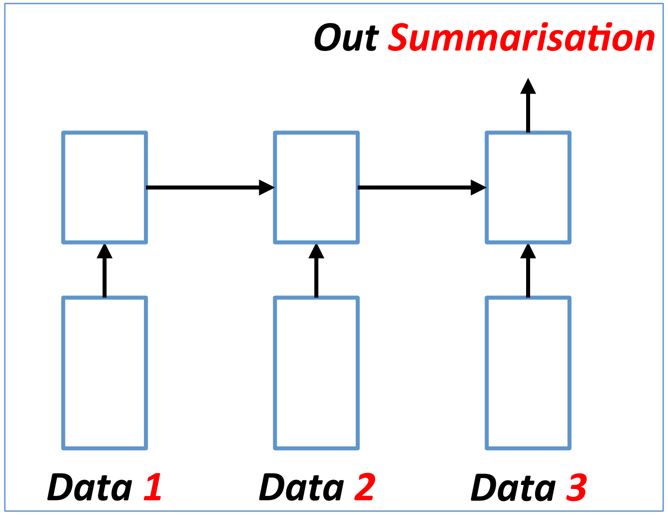
Vs → V's

5 Recurrent Neural Network

RNN: Input and Output



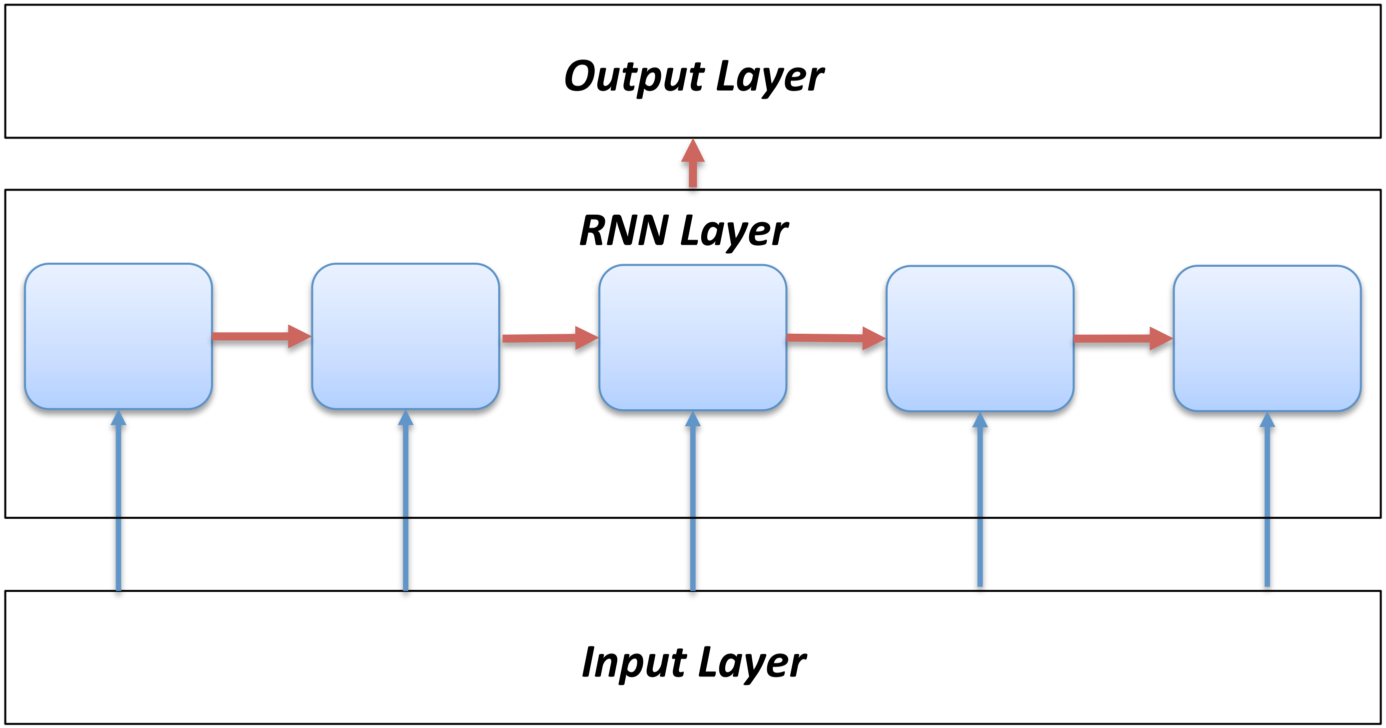
- ✓ *Vs→ V's*
- ✓ *Len(Vs)→ Len(V's)*



- ✓ *Vs → 1*

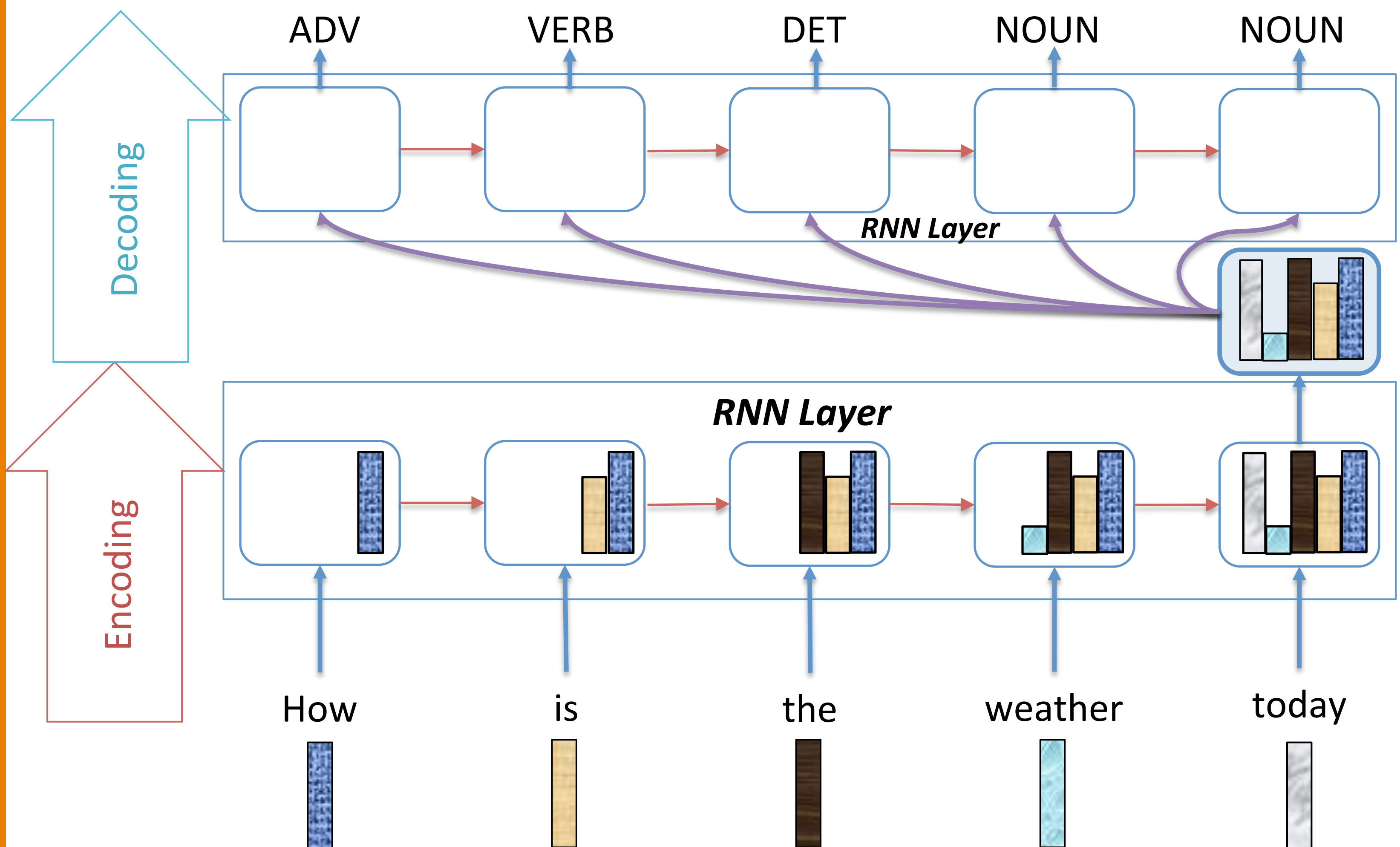
4 Seq2Seq Encoding and Decoding

Recurrent Neural Network



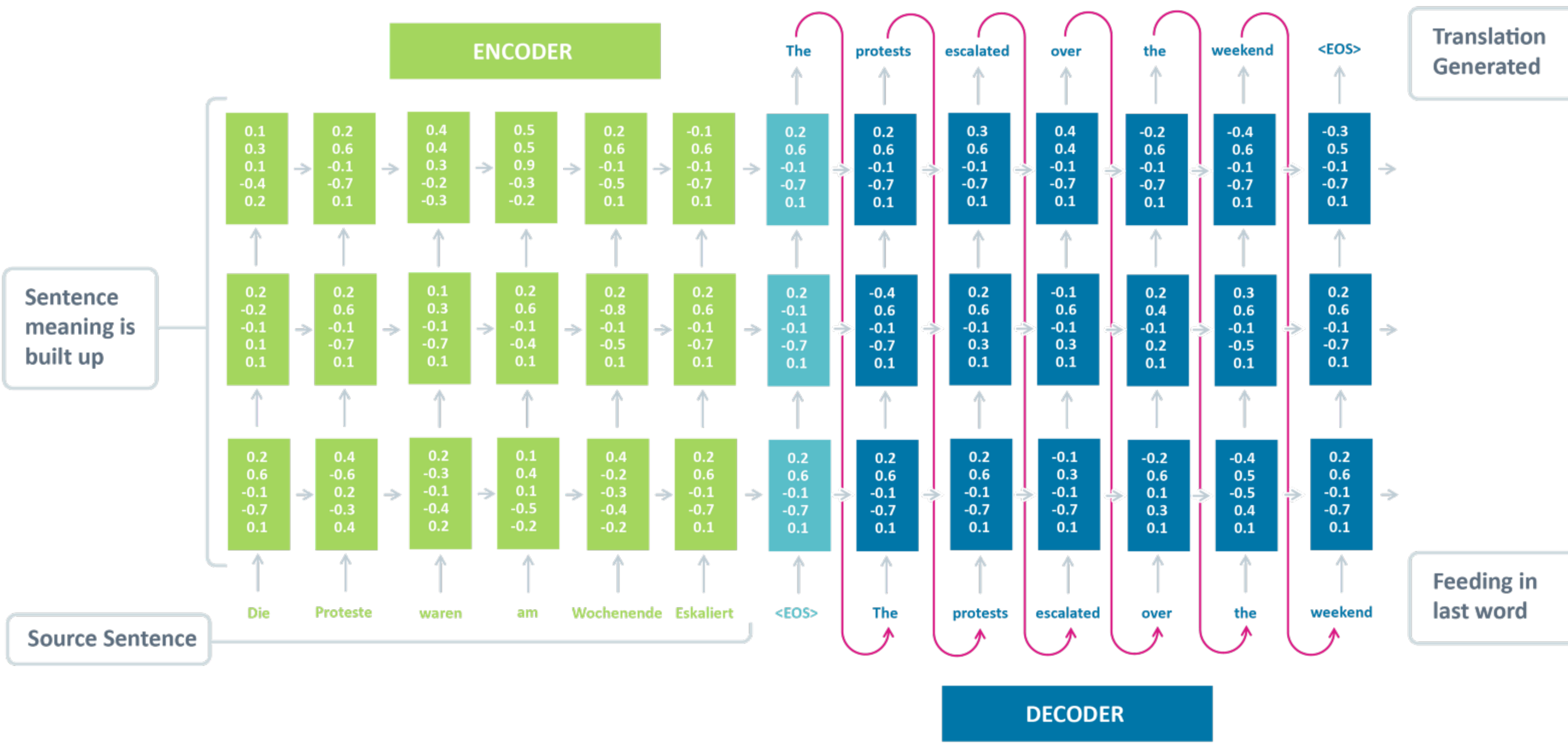
4 Seq2Seq Encoding and Decoding

Seq2Seq Encoding and Decoding Approach

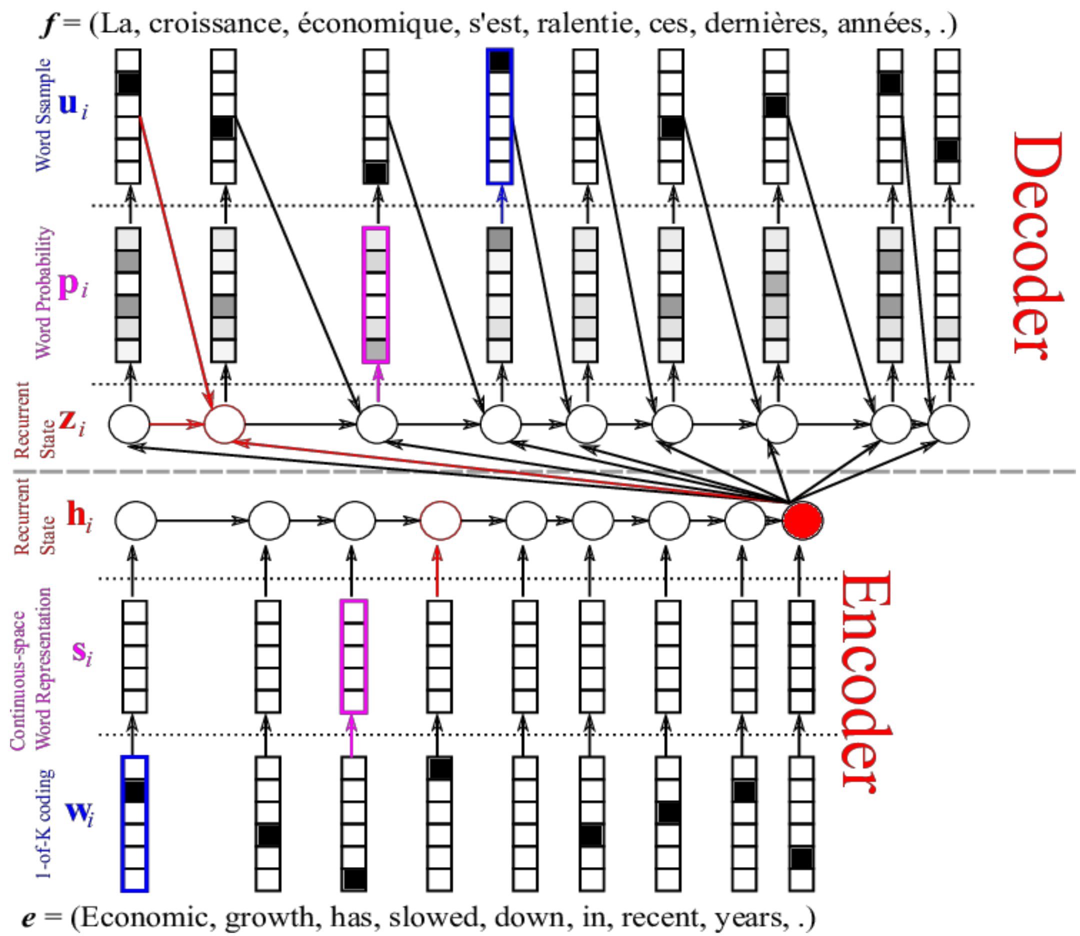


THE UNIVERSITY OF SYDNEY

Seq2Seq Encoding and Decoding- Machine Translation

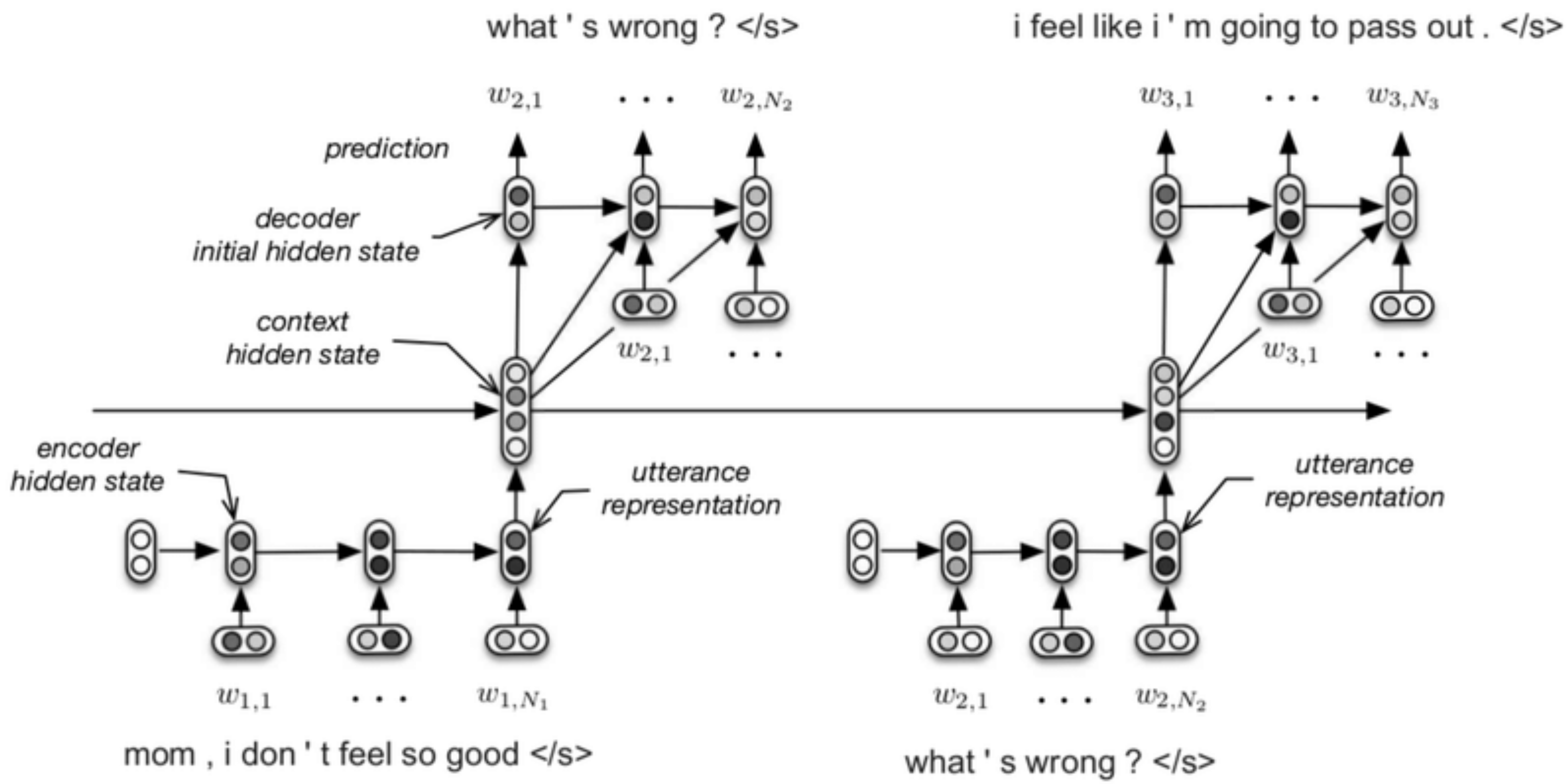


Seq2Seq Encoding and Decoding- Machine Translation



Google translate, 2014

4 Seq2Seq Encoding and Decoding



Sordoni et al., 2014

4 Seq2Seq Encoding and Decoding

Reference (U_1, U_2)	MAP	Target (U_3)
U_1: yeah , okay . U_2: well , i guess i ' ll be going now .	i ' ll see you tomorrow .	yeah .
U_1: oh . <continued_utterance> oh . U_2: what ' s the matter , honey ?	i don ' t know .	oh .
U_1: it ' s the cheapest . U_2: then it ' s the worst kind ?	no , it ' s not .	they ' re all good , sir .
U_1: <person> ! what are you doing ? U_2: shut up ! c ' mon .	what are you doing here ?	what are you that crazy ?

Table 3: MAP outputs for HRED-Bidirectional bootstrapped from *SubTle* corpus. The first column shows the reference utterances, where U_1 and U_2 are respectively the first and second utterance in the test triple. The second column shows the MAP output produced by beam-search conditioned on U_1 and U_2. The third column shows the actual third utterance in the test triple.

Xiaoice conversation, 2014

The purpose of Natural Language Processing: Overview

Application

Understanding

- ***Sentiment Analysis***
- ***Topic Classification***
- ***Topic Modelling***
- ***....***

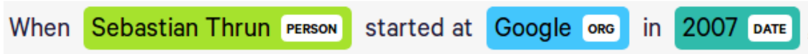
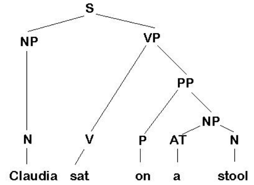
Dialog

Translation

Searching

- ***Search***
- ***....***

NLP Stack

Entity Extraction	When Sebastian Thrun …	
Parsing	Claudia sat on a stool	
PoS Tagging	She sells seashells	**[she/PRP] [sells/VBZ] [seashells/NNS]**
Stemming	Drinking, Drank, Drunk	**Drink**
Tokenisation	How is the weather today	**[How] [is] [the] [weather] [today]**

/ Reference

Reference for this lecture

- Deng, L., & Liu, Y. (Eds.). (2018). Deep Learning in Natural Language Processing. Springer.
- Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. " O'Reilly Media, Inc.".
- Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
- Blunsom, P 2017, Deep Natural Language Processing, lecture notes, Oxford University
- Manning, C 2017, Natural Language Processing with Deep Learning, lecture notes, Stanford University
- Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., & Nie, J. Y. (2015, October). A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 553-562). ACM.

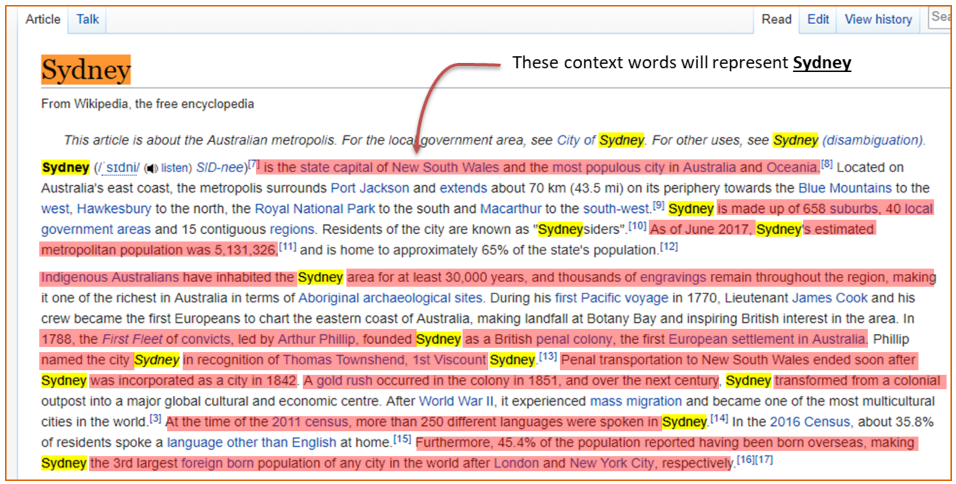
Figure Reference

- https://towardsdatascience.com/types-of-optimization-algorithms-used-in-neural-networks-and-ways-to-optimize-gradient-95ae5d39529f
- https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

Word2Vec with Neural Network: Review

Deep Learning for Word2Vec

Wikipedia: “Sydney is the state capital of NSW…”


Article Talk Read Edit View history

These context words will represent **Sydney**

Sydney

From Wikipedia, the free encyclopedia

This article is about the Australian metropolis. For the local government area, see City of Sydney. For other uses, see Sydney (disambiguation).

Sydney (/ˈsɪdni/ listen) *SID-nee*)[7] is the state capital of New South Wales and the most populous city in Australia and Oceania.[8] Located on Australia's east coast, the metropolis surrounds Port Jackson and extends about 70 km (43.5 mi) on its periphery towards the Blue Mountains to the west, Hawkesbury to the north, the Royal National Park to the south and Macarthur to the south-west.[9] Sydney is made up of 658 suburbs, 40 local government areas and 15 contiguous regions. Residents of the city are known as "Sydneysiders".[10] As of June 2017, Sydney's estimated metropolitan population was 5,131,326,[11] and is home to approximately 65% of the state's population.[12]

Indigenous Australians have inhabited the Sydney area for at least 30,000 years, and thousands of engravings remain throughout the region, making it one of the richest in Australia in terms of Aboriginal archaeological sites. During his first Pacific voyage in 1770, Lieutenant James Cook and his crew became the first Europeans to chart the eastern coast of Australia, making landfall at Botany Bay and inspiring British interest in the area. In 1788, the *First Fleet* of convicts, led by Arthur Phillip, founded Sydney as a British penal colony, the first European settlement in Australia. Phillip named the city *Sydney* in recognition of Thomas Townshend, 1st Viscount Sydney.[13] Penal transportation to New South Wales ended soon after Sydney was incorporated as a city in 1842. A gold rush occurred in the colony in 1851, and over the next century, Sydney transformed from a colonial outpost into a major global cultural and economic centre. After World War II, it experienced mass migration and became one of the most multicultural cities in the world.[3] At the time of the 2011 census, more than 250 different languages were spoken in Sydney.[14] In the 2016 Census, about 35.8% of residents spoke a language other than English at home.[15] Furthermore, 45.4% of the population reported having been born overseas, making Sydney the 3rd largest foreign born population of any city in the world after London and New York City, respectively.[16][17]

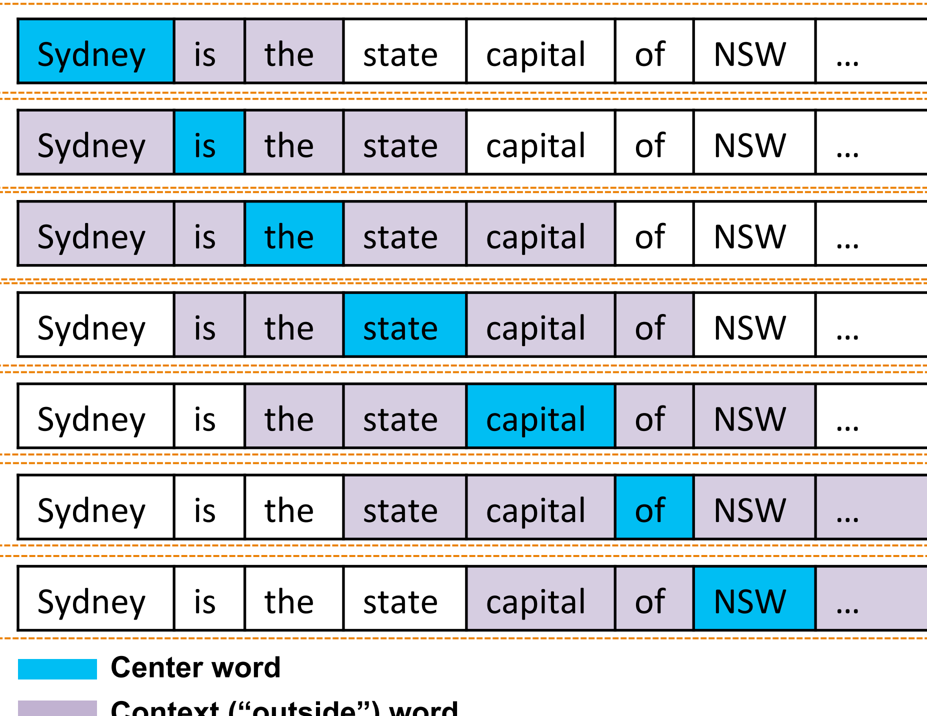
Word	One-hot vector	Index
Sydney	[1,0,0,0,0,0,0,…]	0
Is	[0,1,0,0,0,0,0,…]	1
the	[0,0,1,0,0,0,0,…]	2
state	[0,0,0,1,0,0,0,…]	3
capital	[0,0,0,0,1,0,0,…]	4
of	[0,0,0,0,0,1,0,…]	5
NSW	[0,0,0,0,0,0,1,…]	6
…	…	…

Word2Vec with Neural Network: Review

Deep Learning for Word2Vec

Wikipedia: “Sydney is the state capital of NSW…”

Center word	Context (“outside”) word
[1,0,0,0,0,0,0]	[0,1,0,0,0,0,0], [0,0,1,0,0,0,0]
[0,1,0,0,0,0,0]	[1,0,0,0,0,0,0], [0,0,1,0,0,0,0], [0,0,0,1,0,0,0]
[0,0,1,0,0,0,0]	[1,0,0,0,0,0,0], [0,1,0,0,0,0,0] [0,0,0,1,0,0,0], [0,0,0,0,1,0,0]
[0,0,0,1,0,0,0]	[0,1,0,0,0,0,0], [0,0,1,0,0,0,0] [0,0,0,0,1,0,0], [0,0,0,0,0,1,0]
[0,0,0,0,1,0,0]	[0,0,1,0,0,0,0], [0,0,0,1,0,0,0] [0,0,0,0,0,1,0], [0,0,0,0,0,0,1]
[0,0,0,0,0,1,0]	[0,0,0,1,0,0,0], [0,0,0,0,1,0,0] [0,0,0,0,0,0,1]
[0,0,0,0,0,0,1]	[0,0,0,0,1,0,0], [0,0,0,0,0,1,0]
…	…



/ Word2Vec with Neural Network: Review

Deep Learning for Word2Vec

Wikipedia: "Sydney is the state capital of NSW…"

{'Sydney', 'is', 'the', 'capital', 'of', 'NSW'}