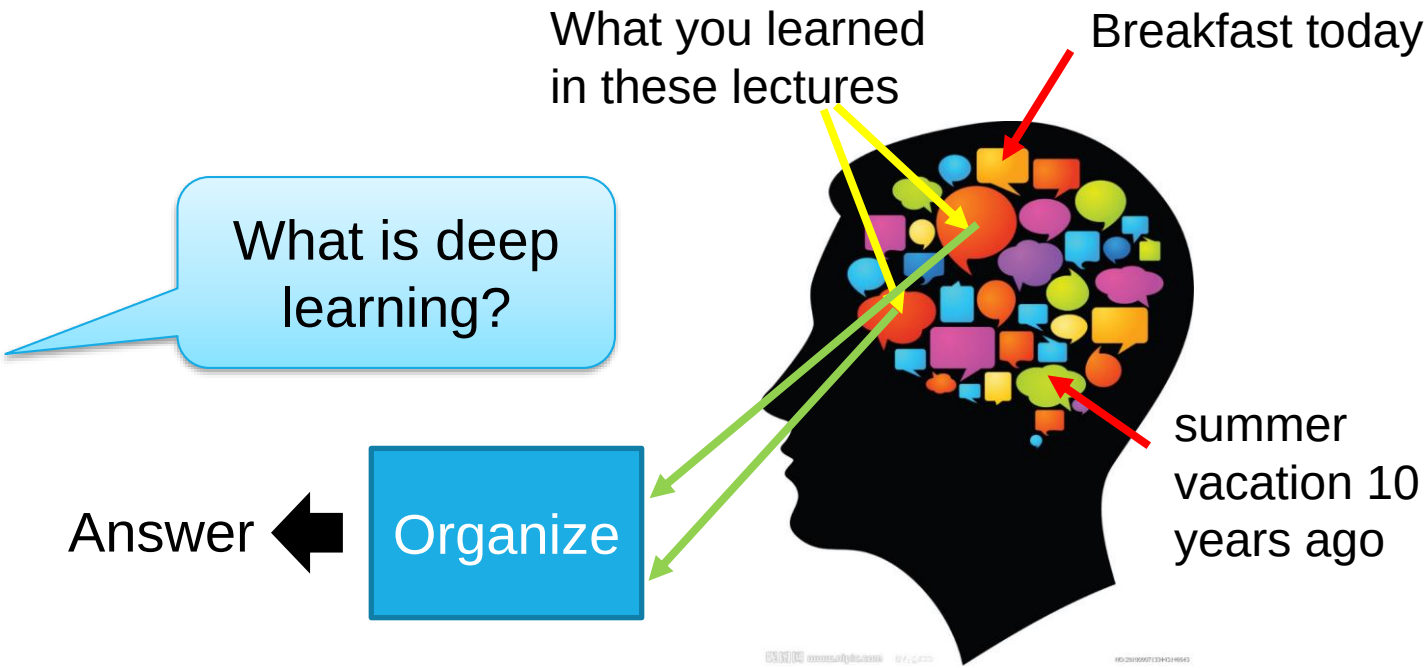
Applied Deep Learning

Attention Mechanism

March 24th, 2020 http://adl.miulab.tw

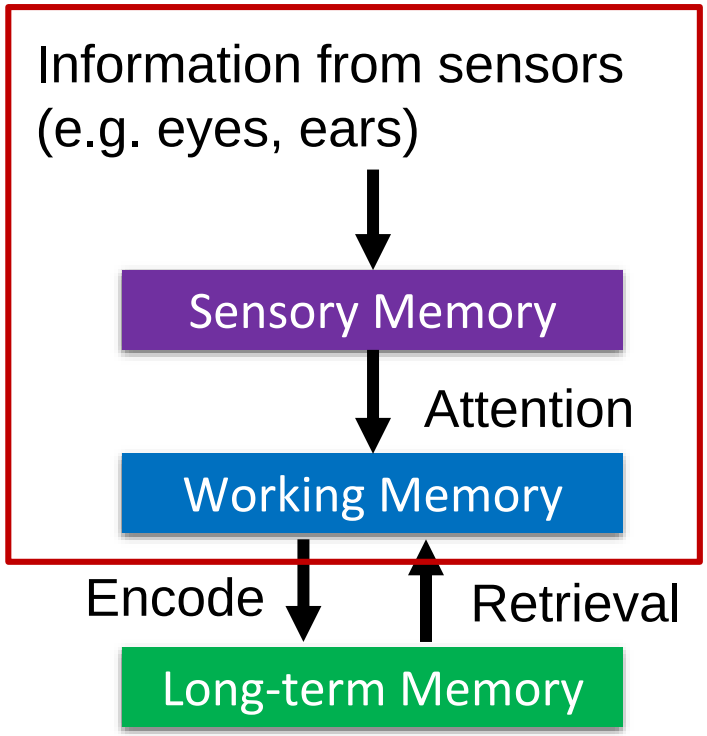
National Taiwan University

Attention and Memory



http://henrylo1605.blogspot.tw/2015/05/blog-post_56.html

Attention and Memory

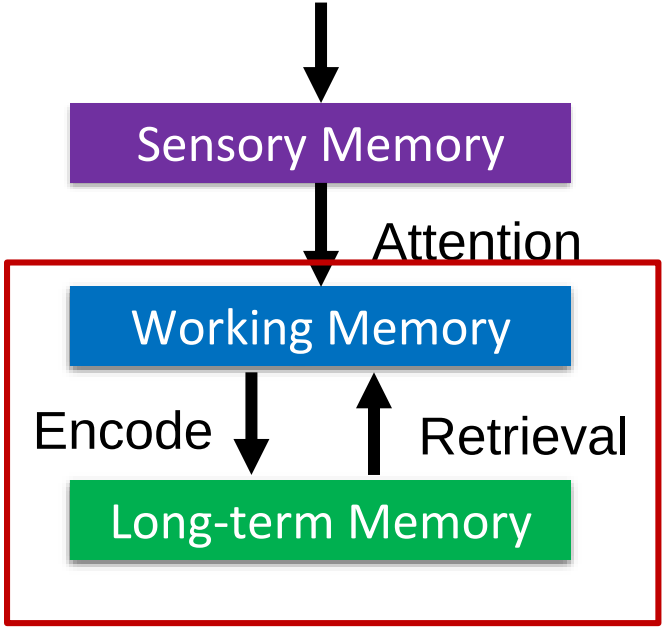


Problem: very long sequence or an image

Solution: pay attention on the **partial** input object each time

Attention and Memory

Information from sensors
(e.g. eyes, ears)



Problem: very long sequence or an image

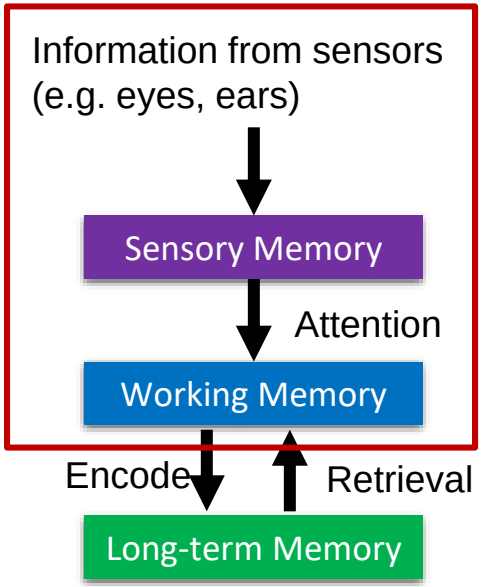
Solution: pay attention on the **partial** input object each time

Problem: larger memory implies more parameters in RNN

Solution: long-term memory increases memory size without increasing parameters

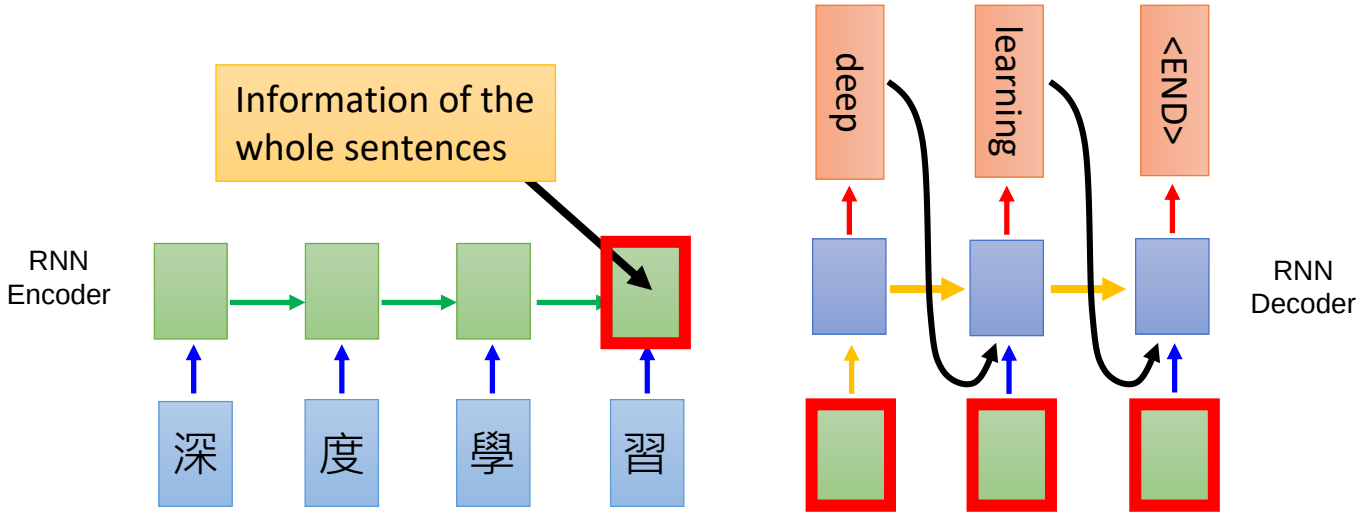
5

Attention on Sensory Info

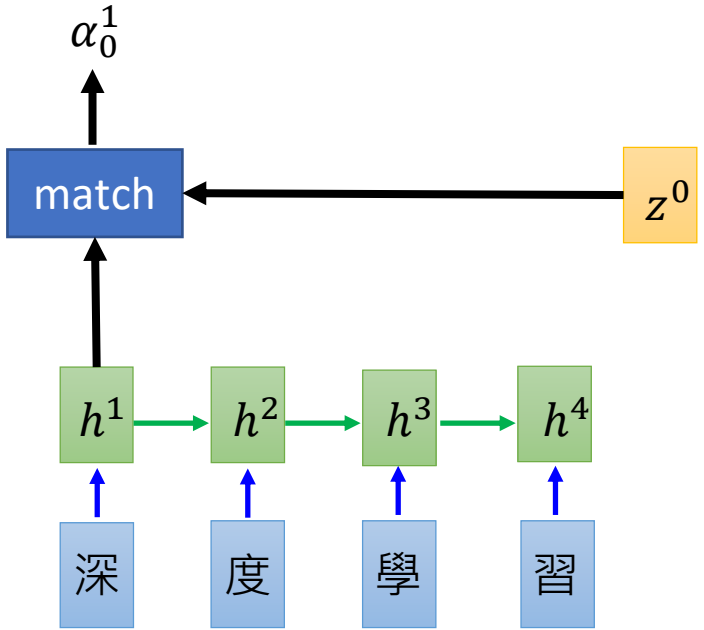


Machine Translation

- Sequence-to-sequence learning: both input and output are both sequences ***with different lengths***.
- E.g. 深度學習 → deep learning



Machine Translation with Attention

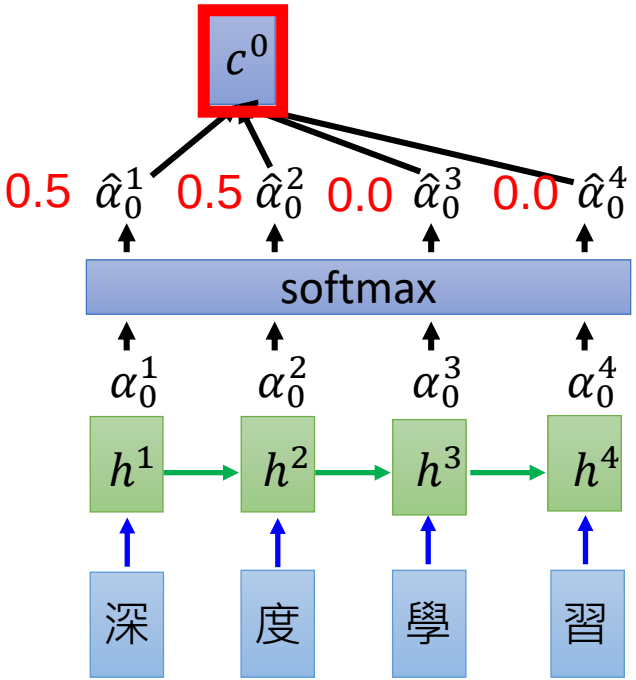


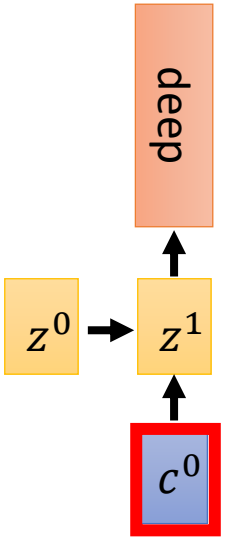
What is match ?

- Cosine similarity of z and h
- Small NN whose input is z and h, output a scalar
- $\alpha = h^T W z$

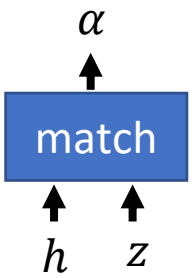
How to learn the parameters?

Machine Translation with Attention





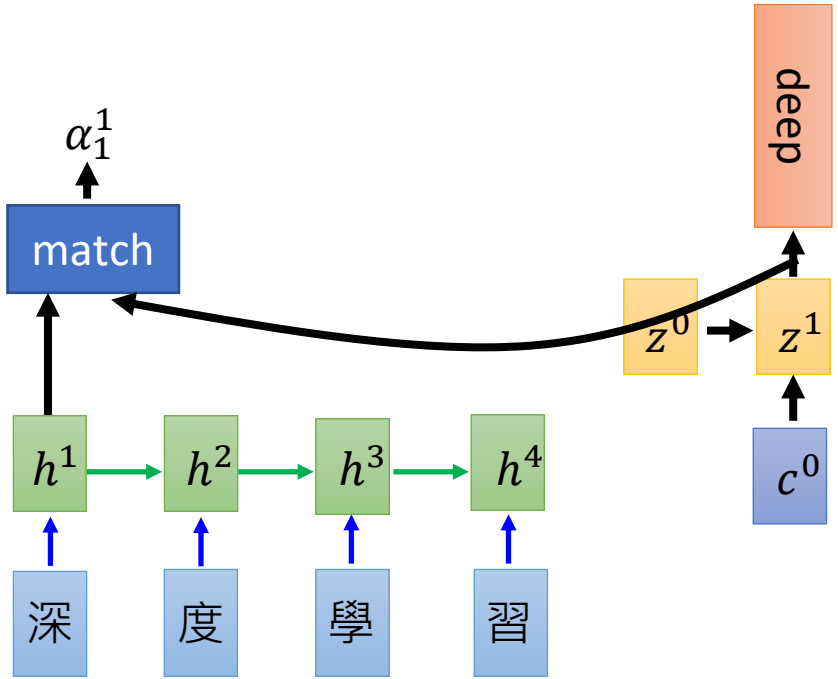
as RNN input



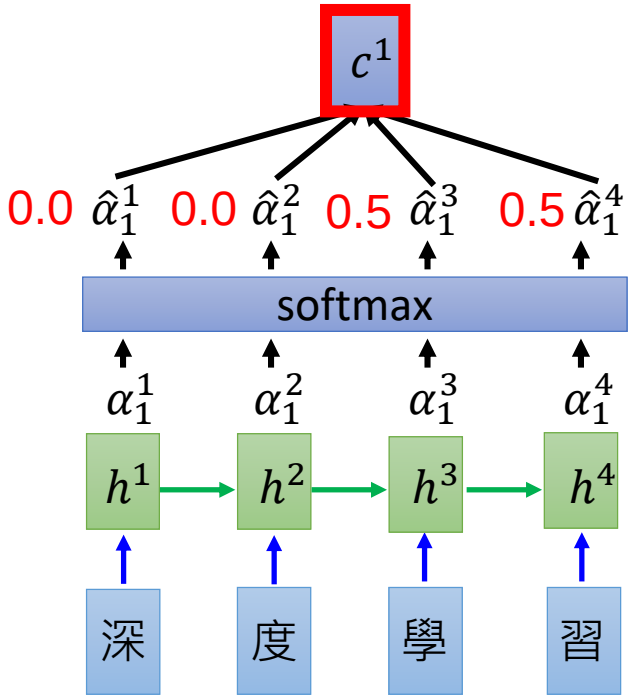
How to learn the parameters?

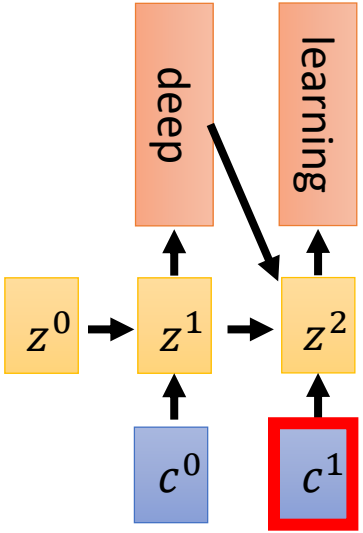
$$c^0 = \sum \hat{\alpha}_0^i h^i = 0.5h^1 + 0.5h^2$$

Machine Translation with Attention



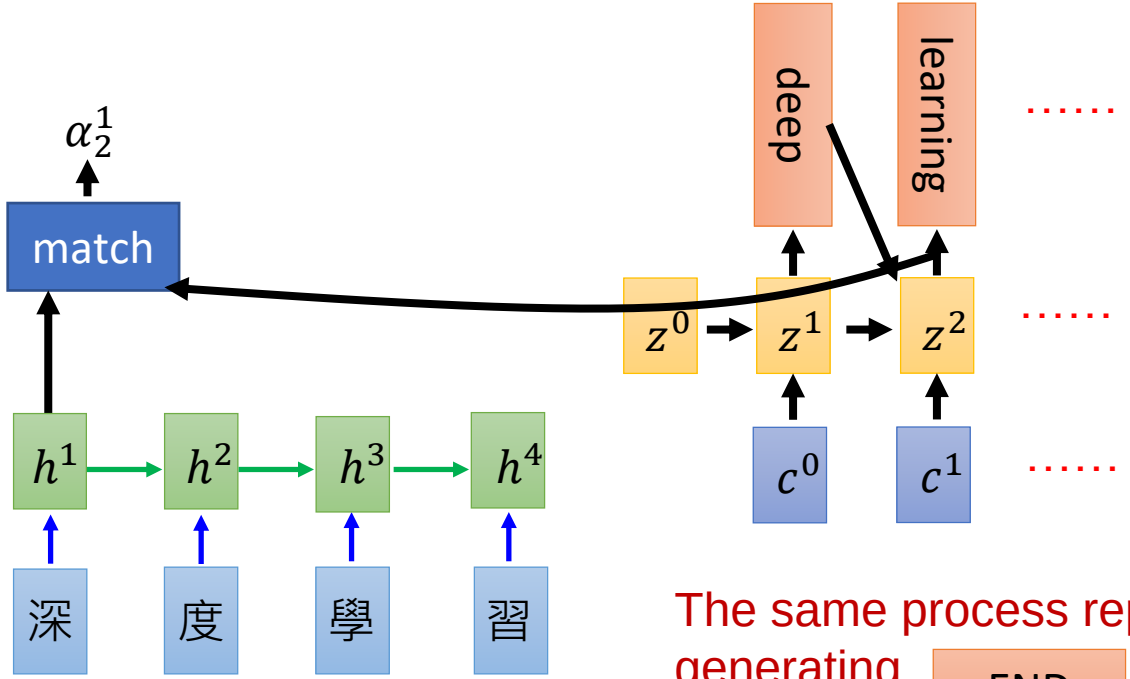
Machine Translation with Attention





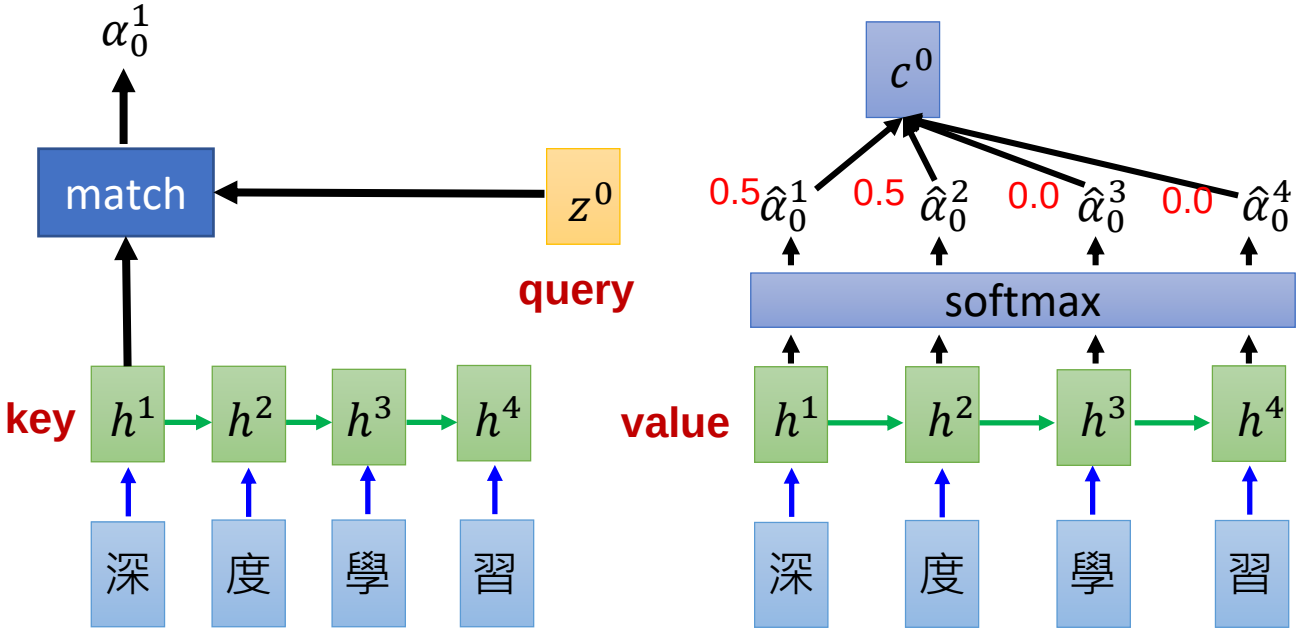
$$c^1 = \sum \hat{\alpha}_1^i h^i = 0.5h^3 + 0.5h^4$$

Machine Translation with Attention



The same process repeat until generating <END>

Machine Translation with Attention



Dot-Product Attention

- Input: a query q and a set of key-value (k-v) pairs to an output
- Output: weighted sum of values

Inner product of
query and corresponding key

$$A(q, K, V) = \sum_i \frac{\exp(q \cdot k_i)}{\sum_j \exp(q \cdot k_j)} v_i$$

  - Query q is a d_k-dim vector
  - Key k is a d_k-dim vector
  - Value v is a d_v-dim vector

Dot-Product Attention in Matrix

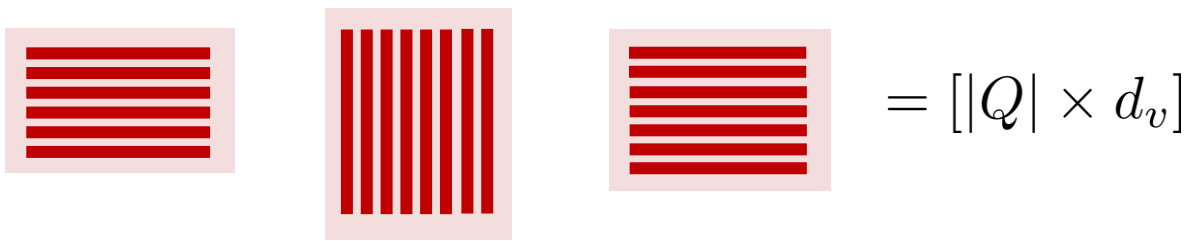
- Input: *multiple* queries q and a set of key-value (k-v) pairs to an output
- Output: a set of weighted sum of values

$$A(q, K, V) = \sum_i \frac{\exp(q \cdot k_i)}{\sum_j \exp(q \cdot k_j)} v_i$$

$$A(Q, K, V) = \text{softmax}(QK^T)V$$

$$[|Q| \times d_k] \times [d_k \times |K|] \times [|K| \times d_v]$$

softmax row-wise

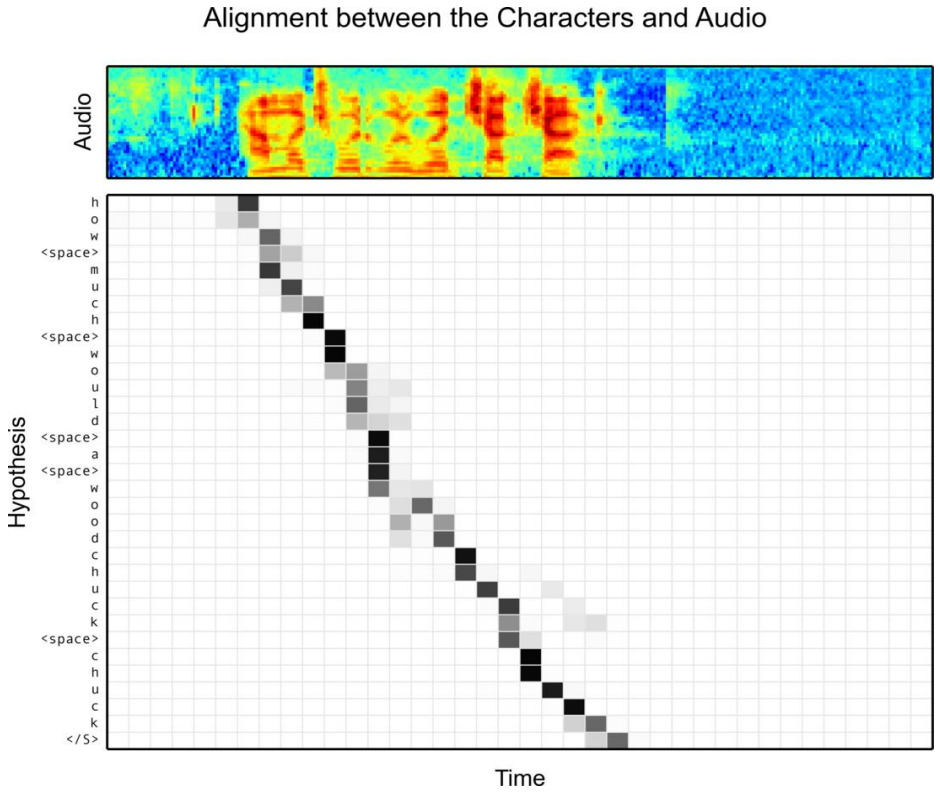


$$= [|Q| \times d_v]$$

Attention Applications

各種不同的應用都用得到 Attention

Speech Recognition with Attention



Chan et al., “Listen, Attend and Spell”, arXiv, 2015 .

Image Captioning

- Input: image
- Output: word sequence

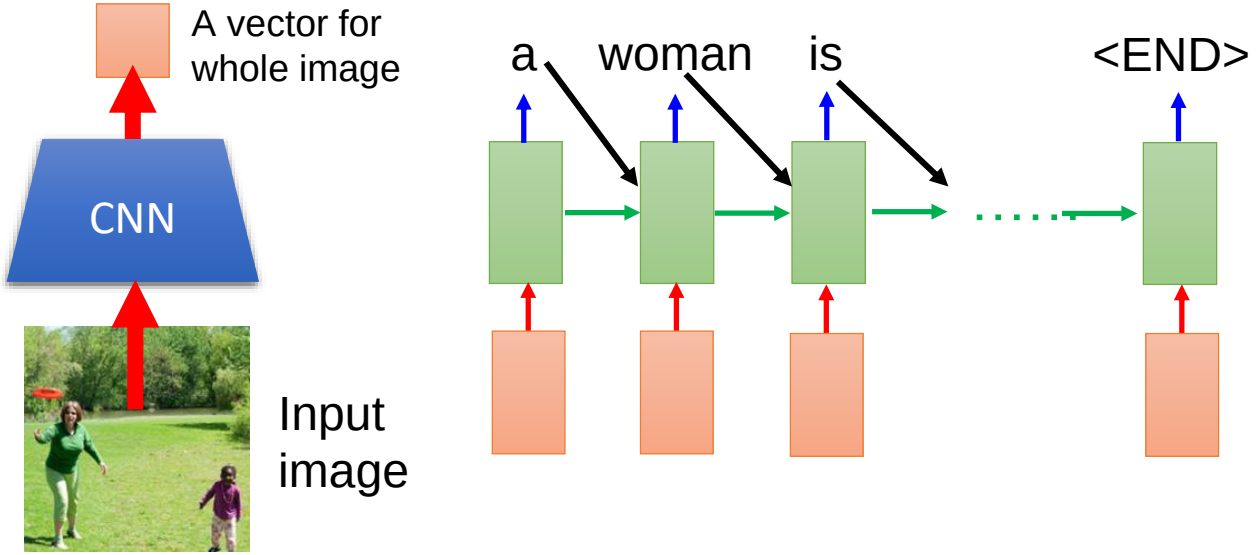


Image Captioning with Attention

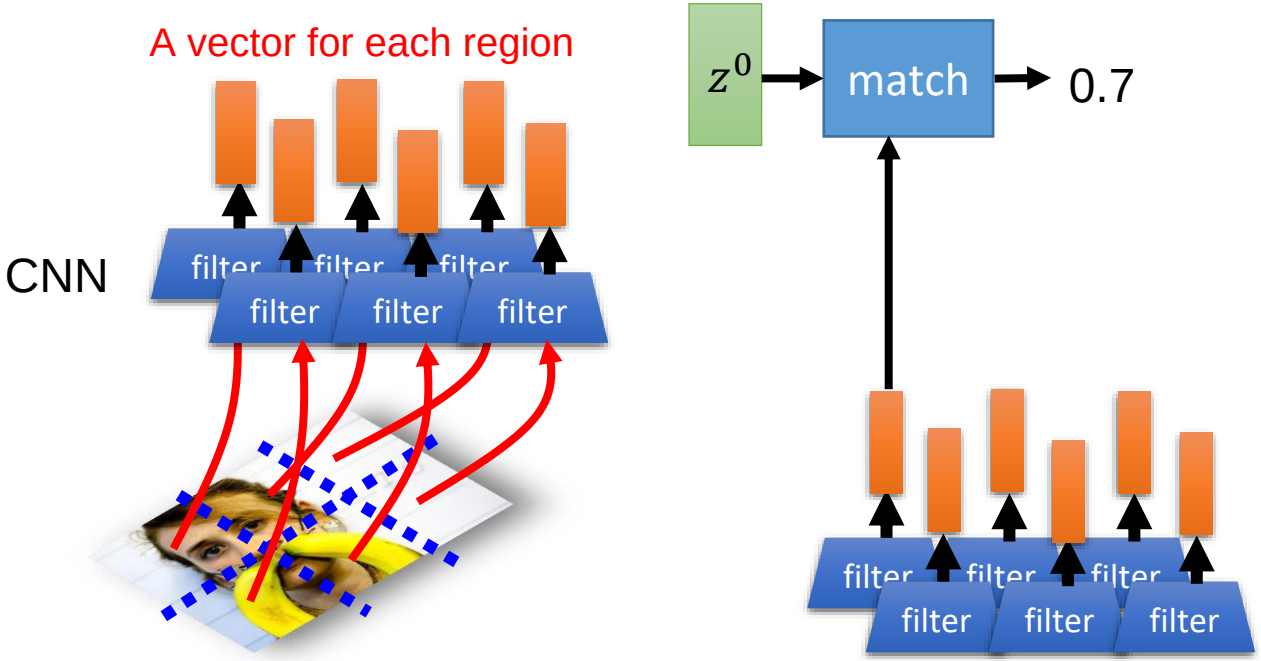


Image Captioning with Attention

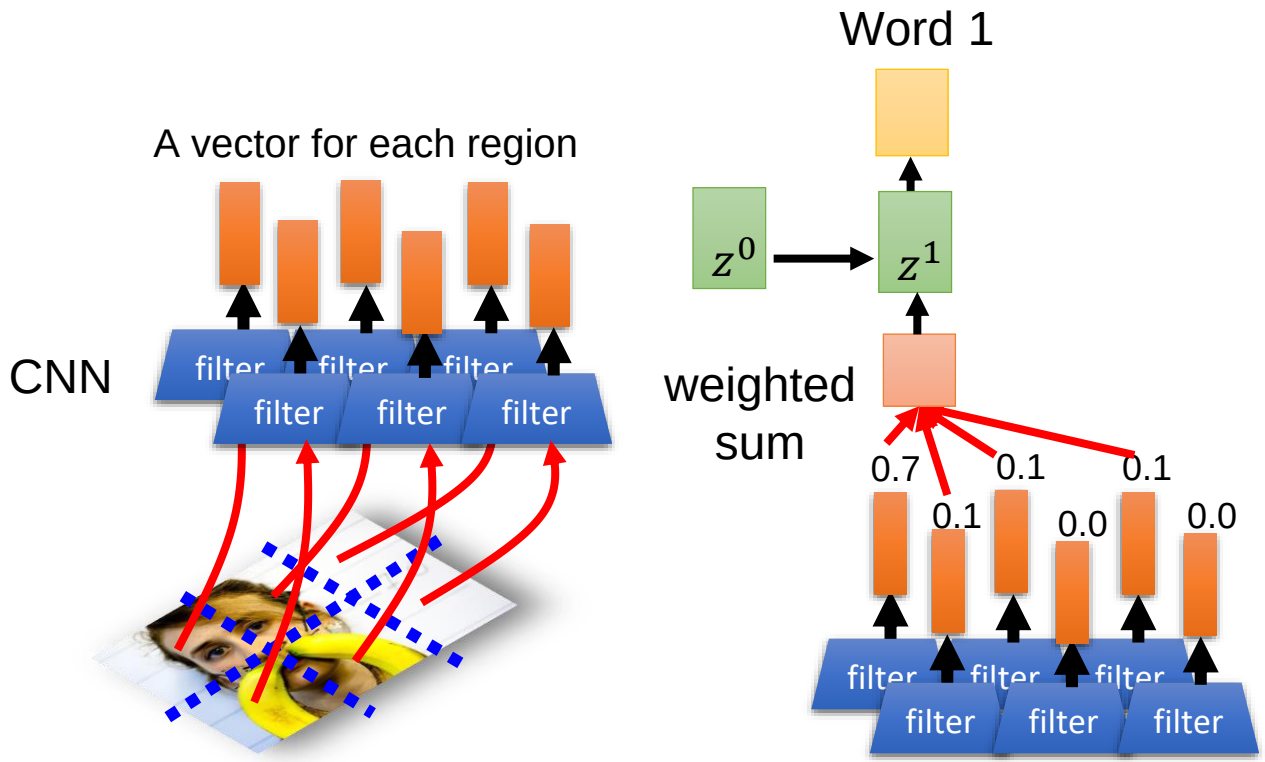


Image Captioning with Attention

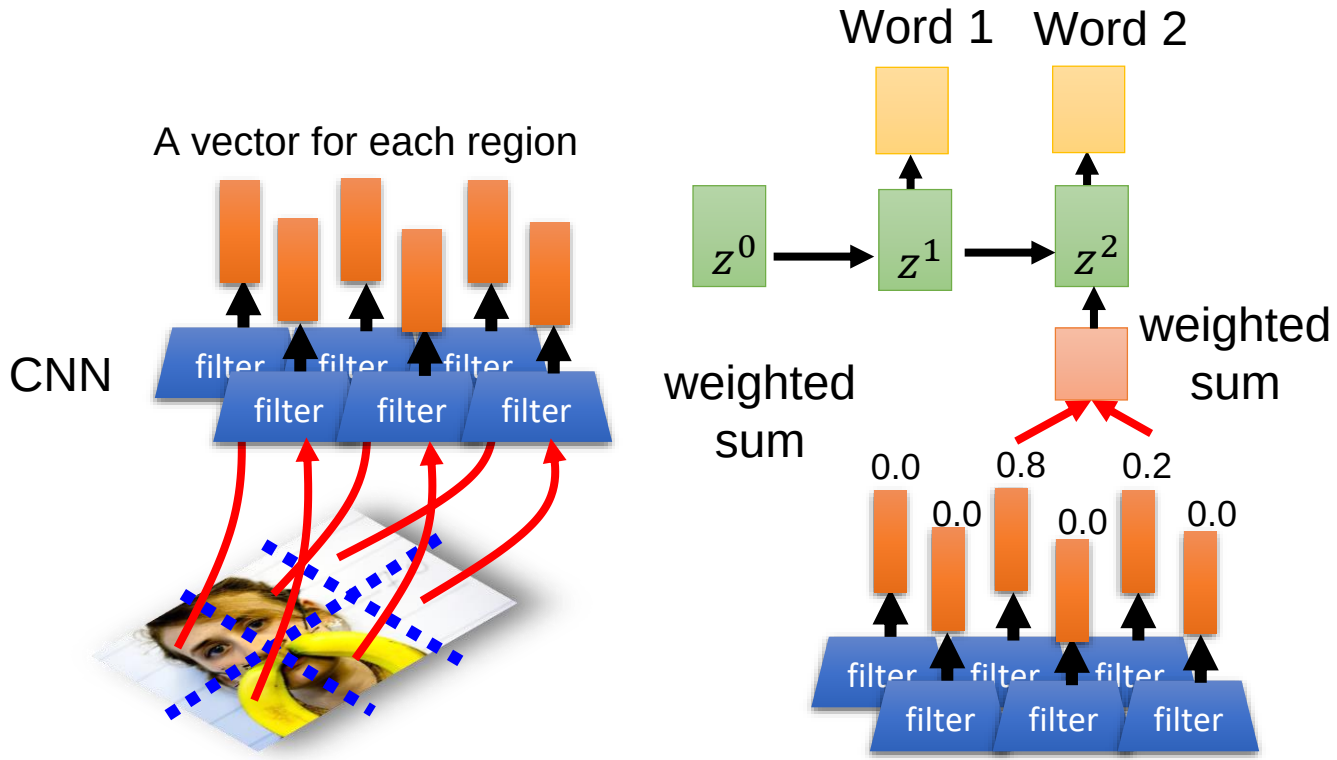


Image Captioning

Good examples

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Image Captioning

Bad examples

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

Video Captioning

Ref: A man and a woman ride a motorcycle

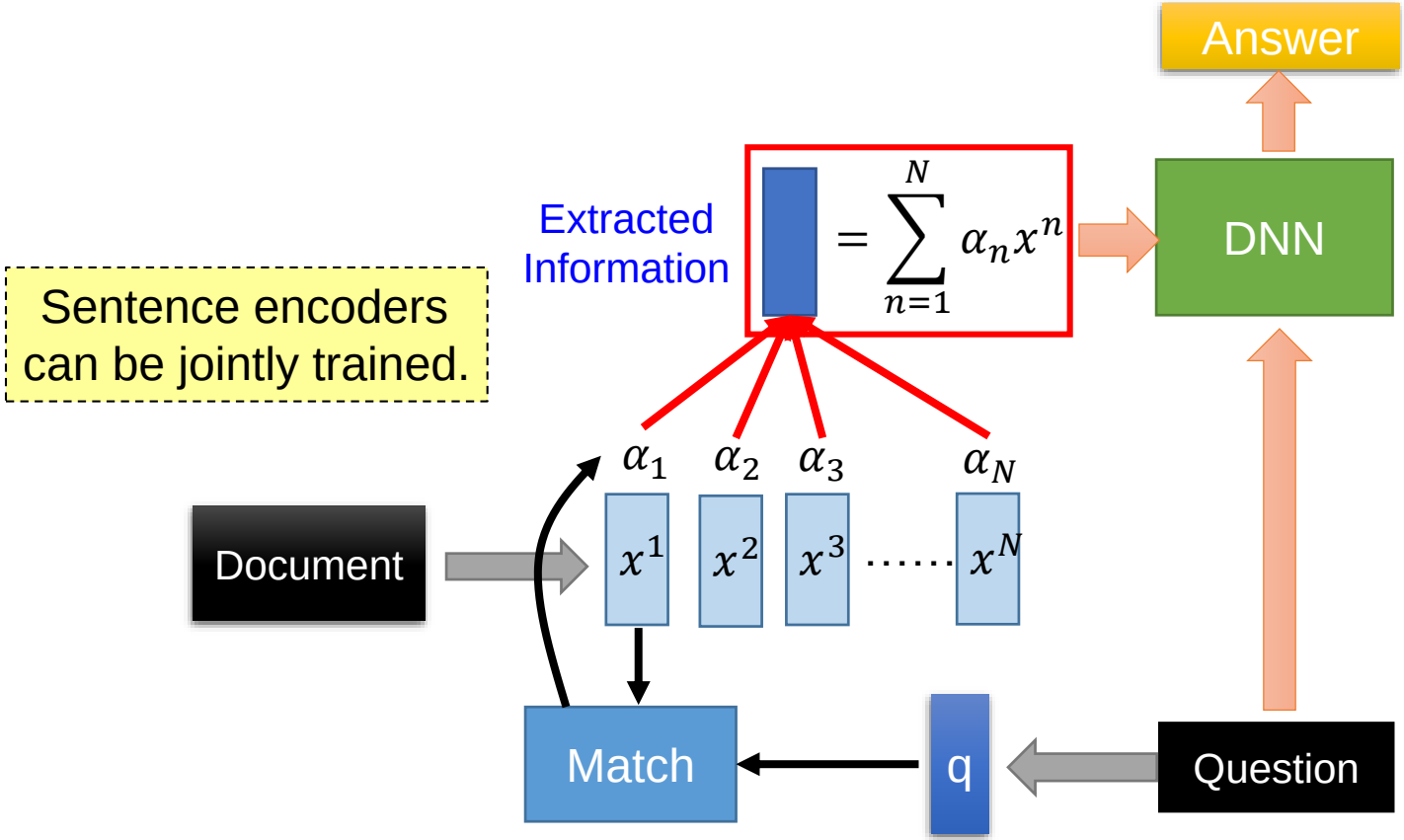
A **man** and a **woman** are **talking** on the **road**

Video Captioning

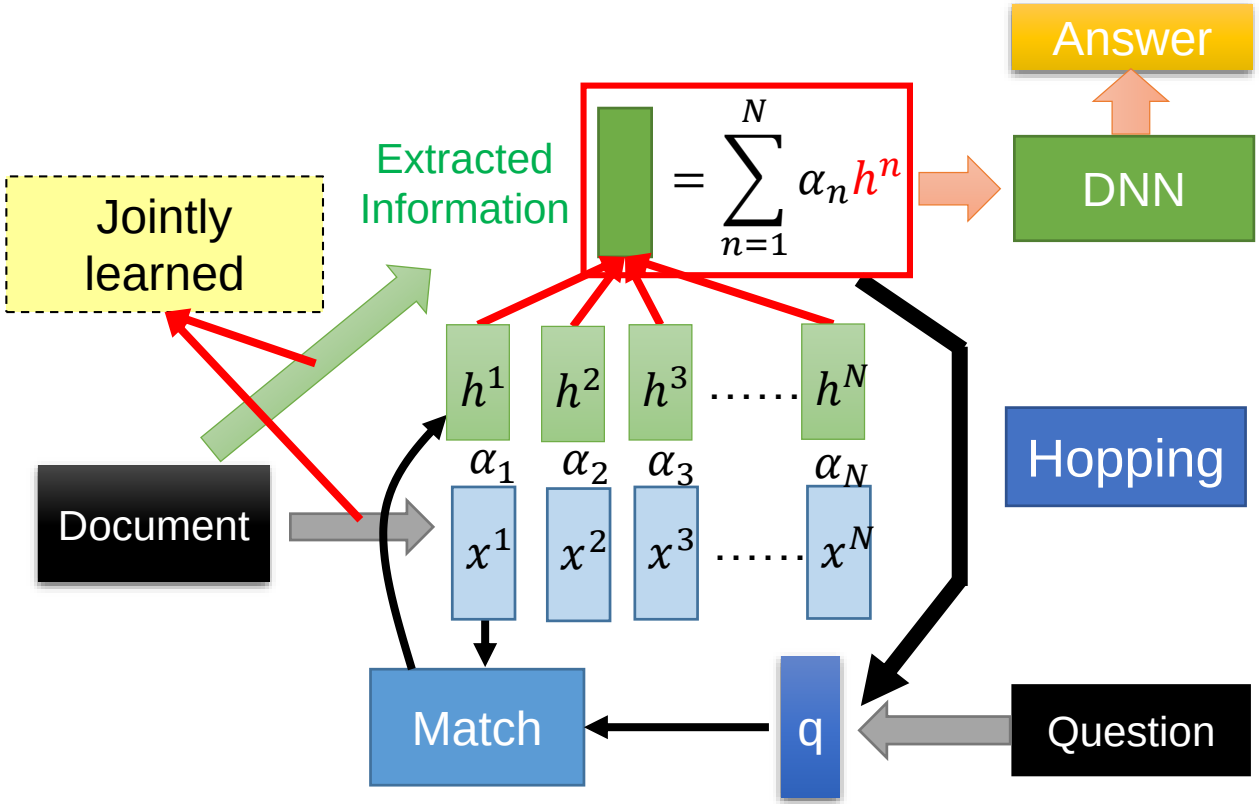
Ref: A woman is frying food

Someone is **frying** a **fish** in a **pot**

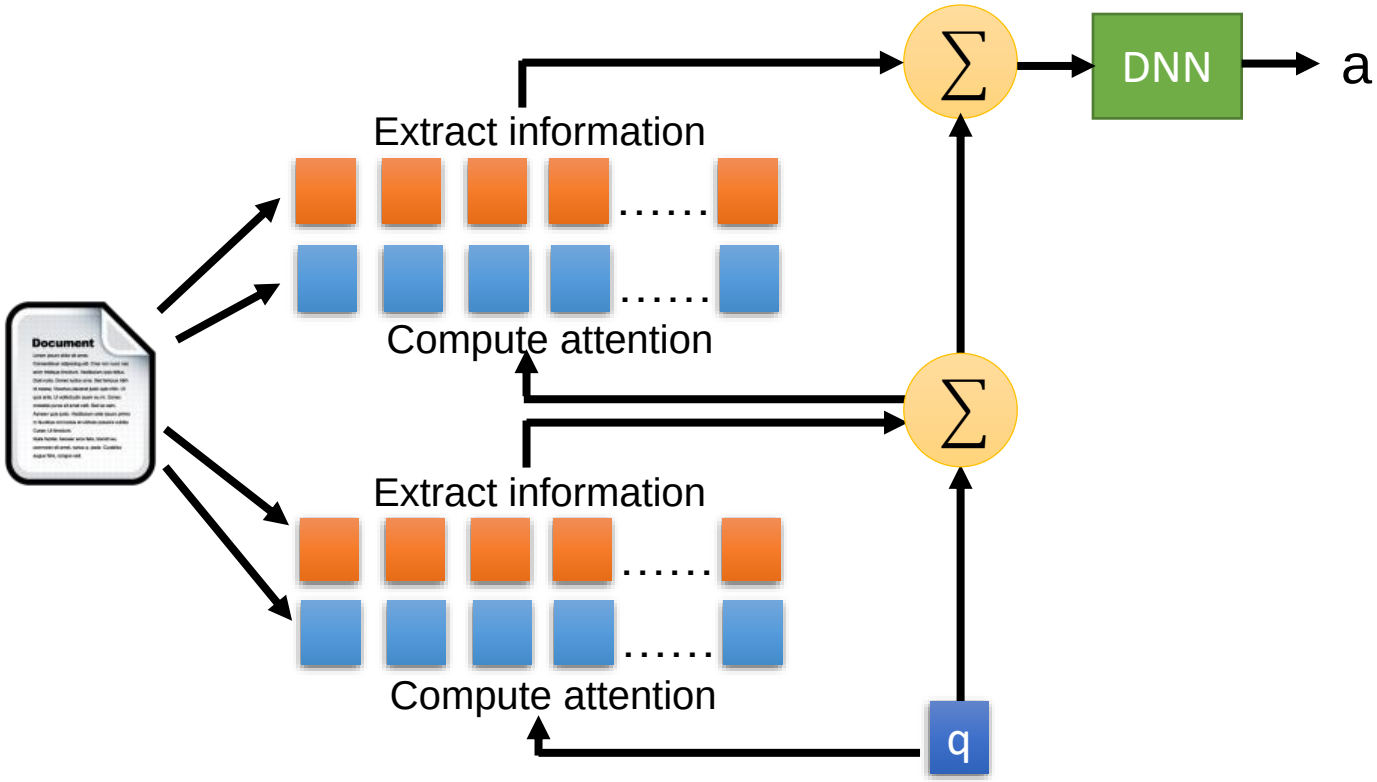
Reading Comprehension



Reading Comprehension



Memory Network



Memory Network

- Muti-hop performance analysis

Story (1: 1 supporting fact)	Support	Hop 1	Hop 2	Hop 3
Daniel went to the bathroom.		0.00	0.00	0.03
Mary travelled to the hallway.		0.00	0.00	0.00
John went to the bedroom.		0.37	0.02	0.00
John travelled to the bathroom.	yes	0.60	0.98	0.96
Mary went to the office.		0.01	0.00	0.00
Where is John? Answer: bathroom		**Prediction: bathroom**		

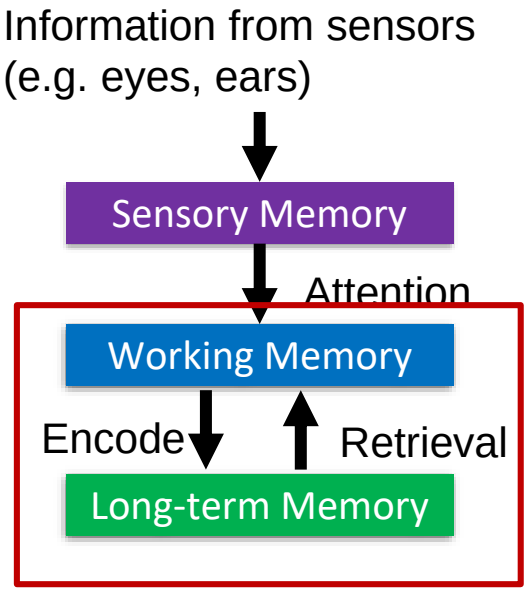
Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow		**Prediction: yellow**		

Conversational QA – CoQA, QuAC

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

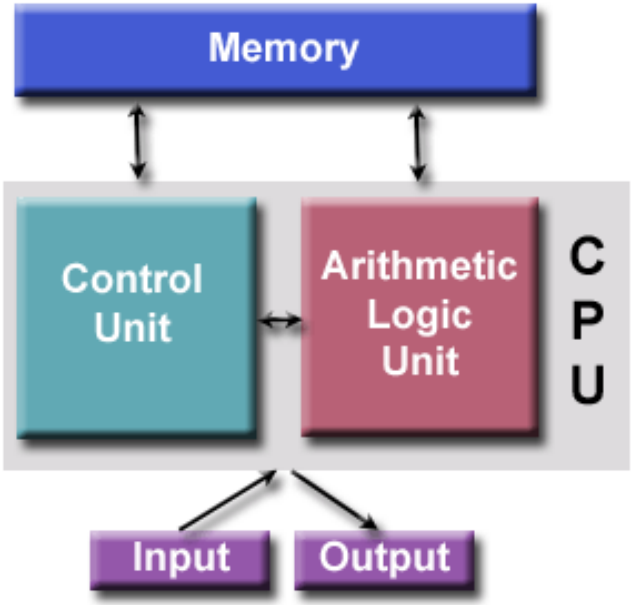
- **The QA pairs are conversational**
  - Q1: Who had a birthday?
  - A1: Jessica
  - Q2: How old would she be?
  - A2: 80
  - Q3: Did she plan to have any visitors?
  - A3: Yes
  - Q4: How many?
  - A4: Three
  - Q5: Who?
  - A5: Annie, Melanie, and Josh

Attention on Memory



Neural Turing Machine

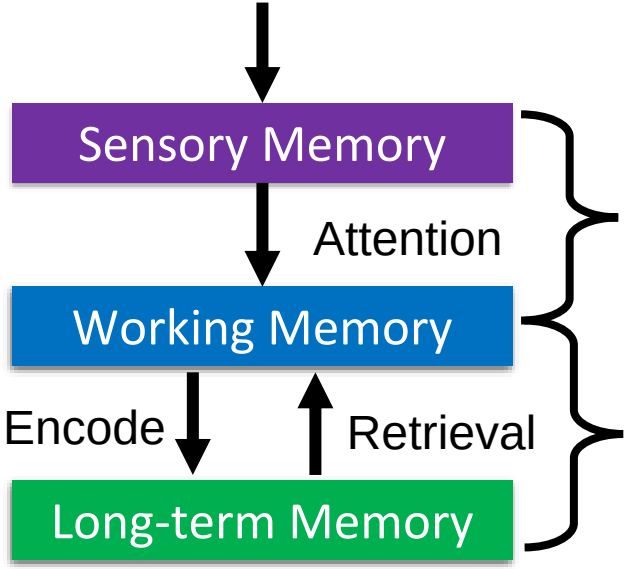
- Von Neumann architecture
- Neural Turing Machine is an advanced RNN/LSTM.



Zhang et al., “Structured Memory for Neural Turing Machines,” arXiv, 2015.

Concluding Remarks

Information from sensors (e.g. eyes, ears)



$$A(q, K, V) = \sum_i \frac{\exp(q \cdot k_i)}{\sum_j \exp(q \cdot k_j)} v_i$$

$$A(Q, K, V) = \text{softmax}(QK^T)V$$

Machine Translation

Speech Recognition

Image Captioning

Question Answering

Neural Turing

Machine Stack RNN