# CS 291A: Deep Learning for NLP

## Neural Networks: Attention and Memory

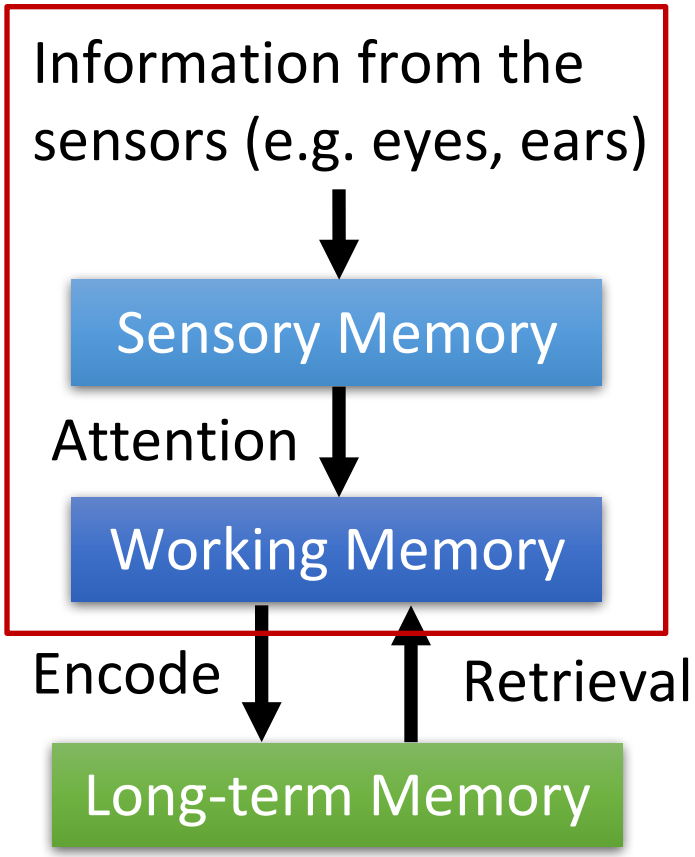
William Wang

UCSB Computer Science

william@cs.ucsb.edu

Slides adapted from V. Chen and H. Lee.

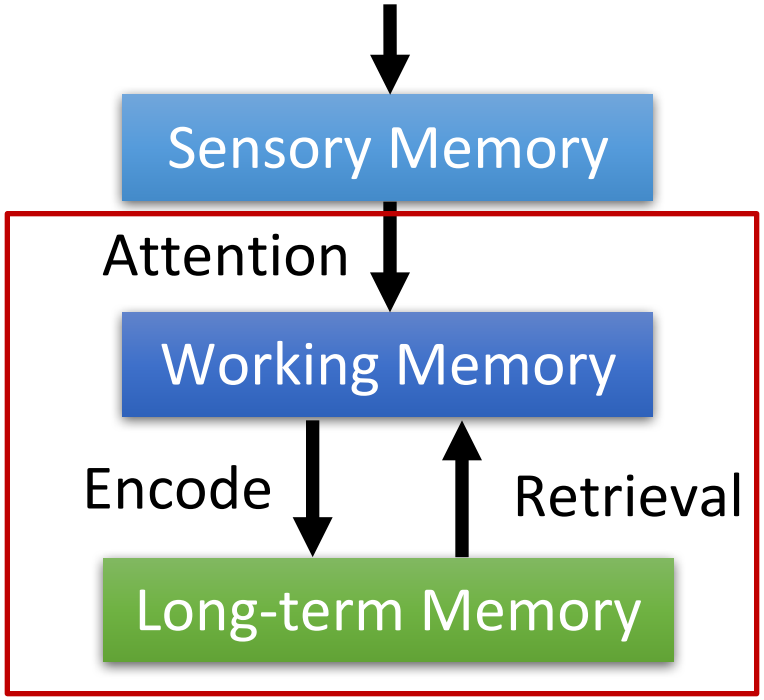
# Attention and Memory




When the input is a very long sequence or an image

→ Pay attention on partial of the input object each time

# Attention and Memory



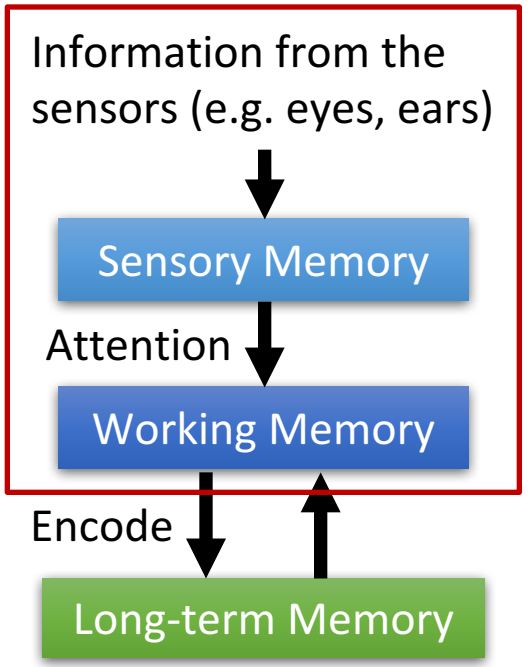

When the input is a very long sequence or an image

→ Pay attention on partial of the input object each time

In RNN/LSTM, larger memory implies more parameters

→ Increasing memory size will not increasing parameters

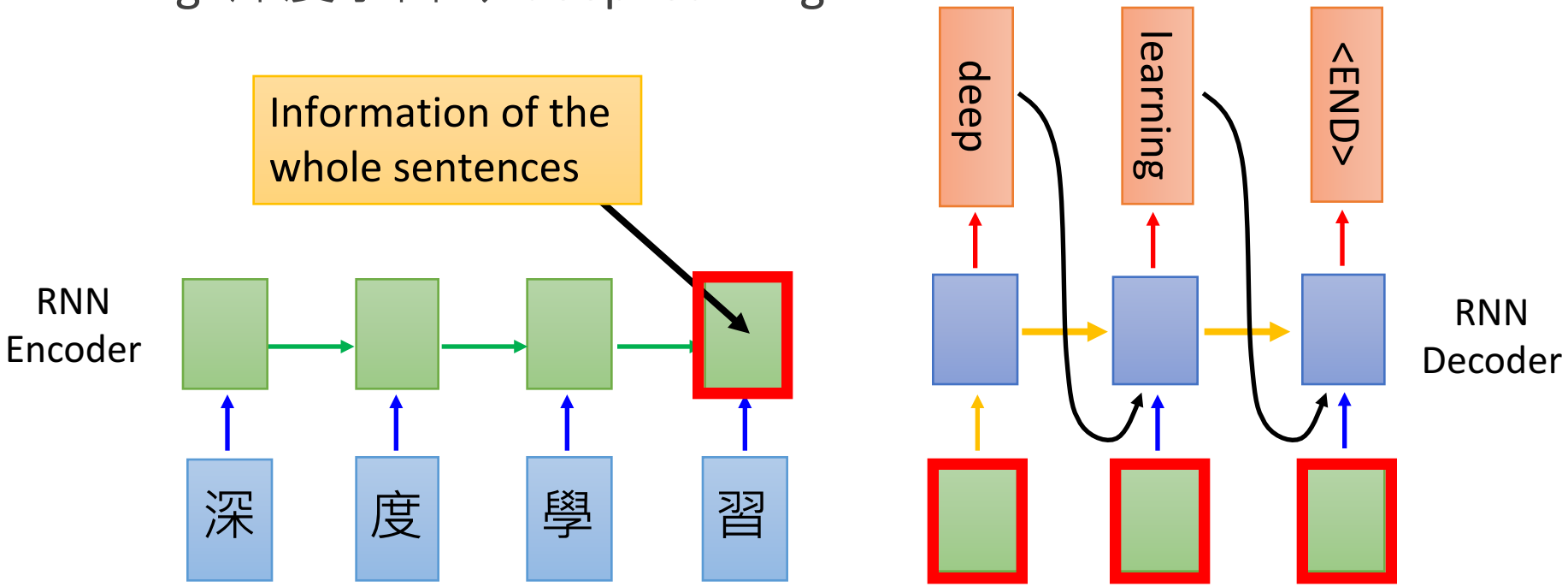
# Attention on Sensory Info



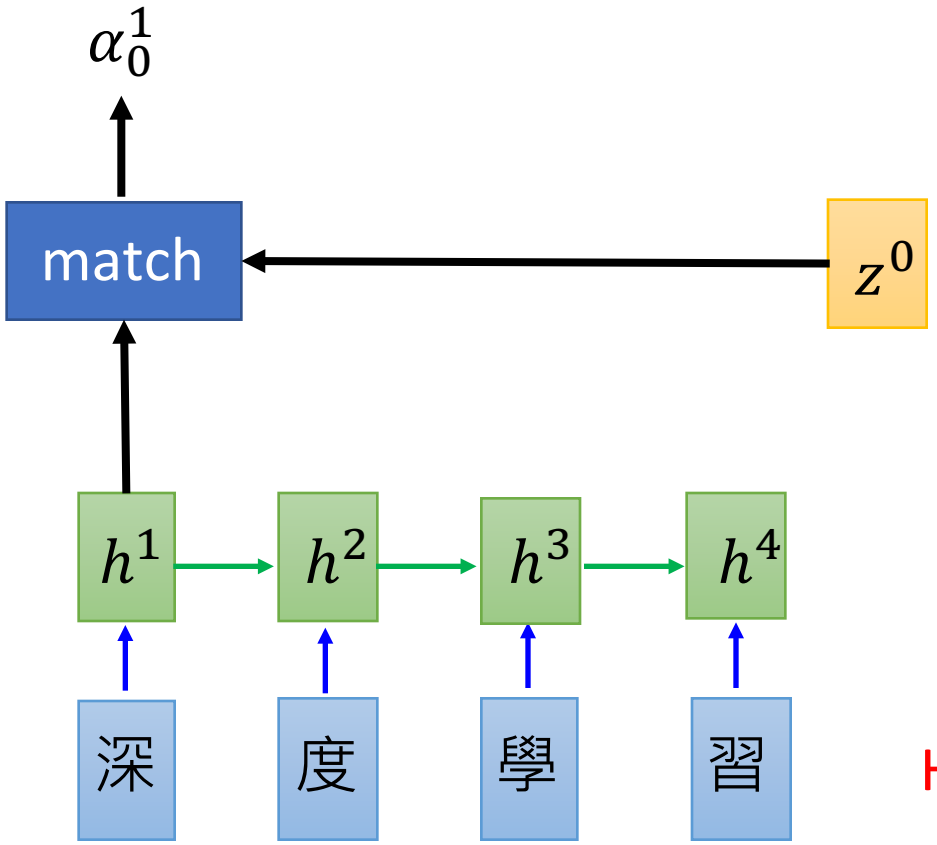
# Machine Translation

Sequence-to-sequence learning: both input and output are both sequences ***with different lengths***.

E.g. 深度學習 → deep learning



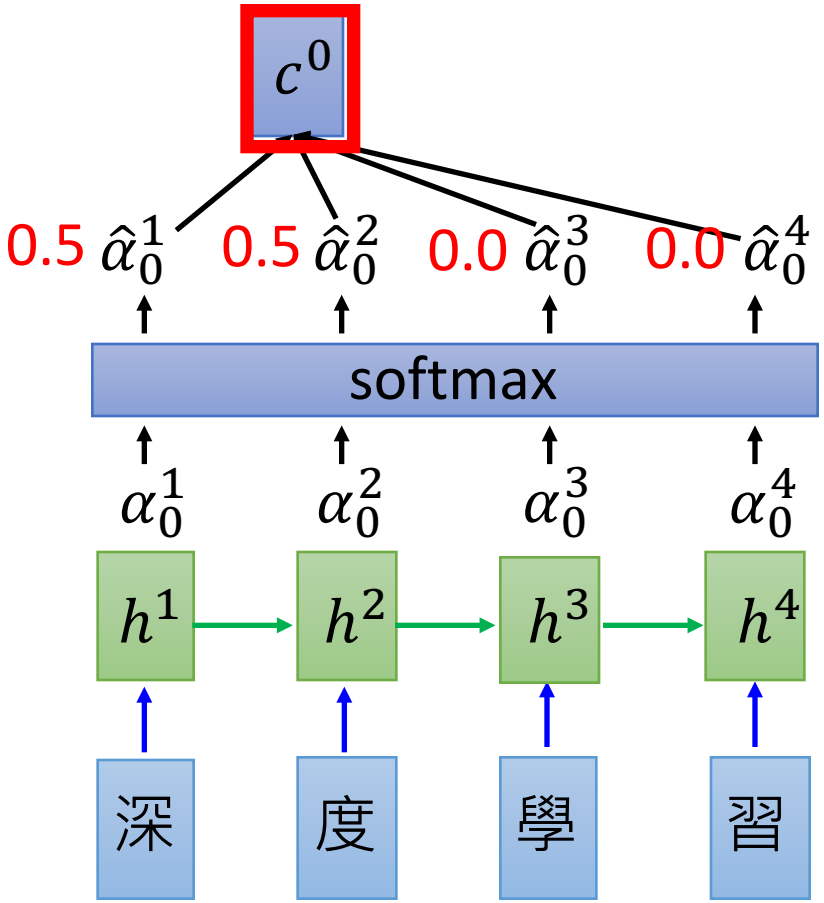
# Machine Translation with Attention

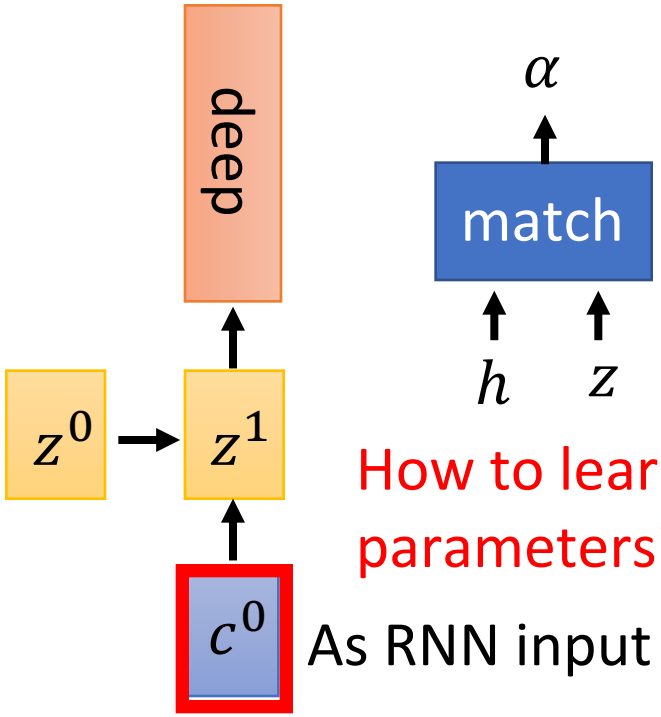



What is match ?

How to learn the parameters?

# Machine Translation with Attention



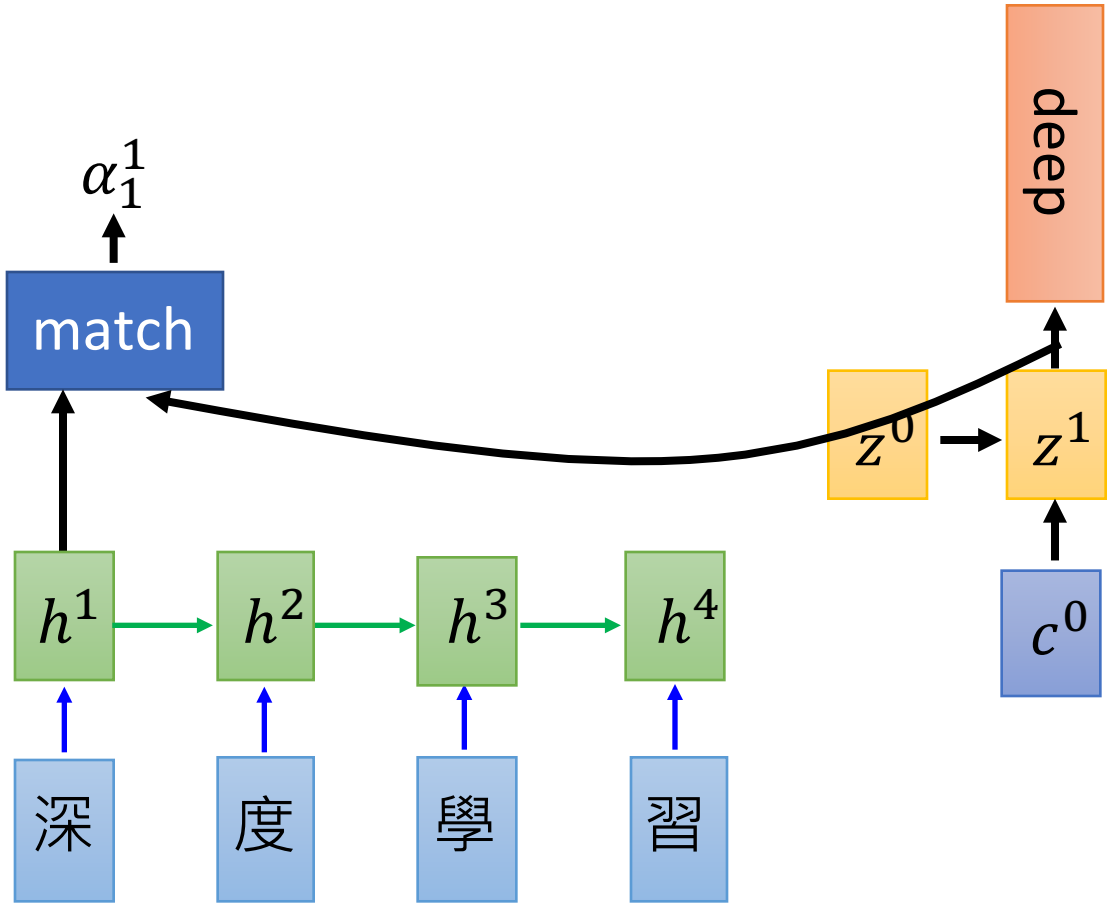




How to learn the parameters?

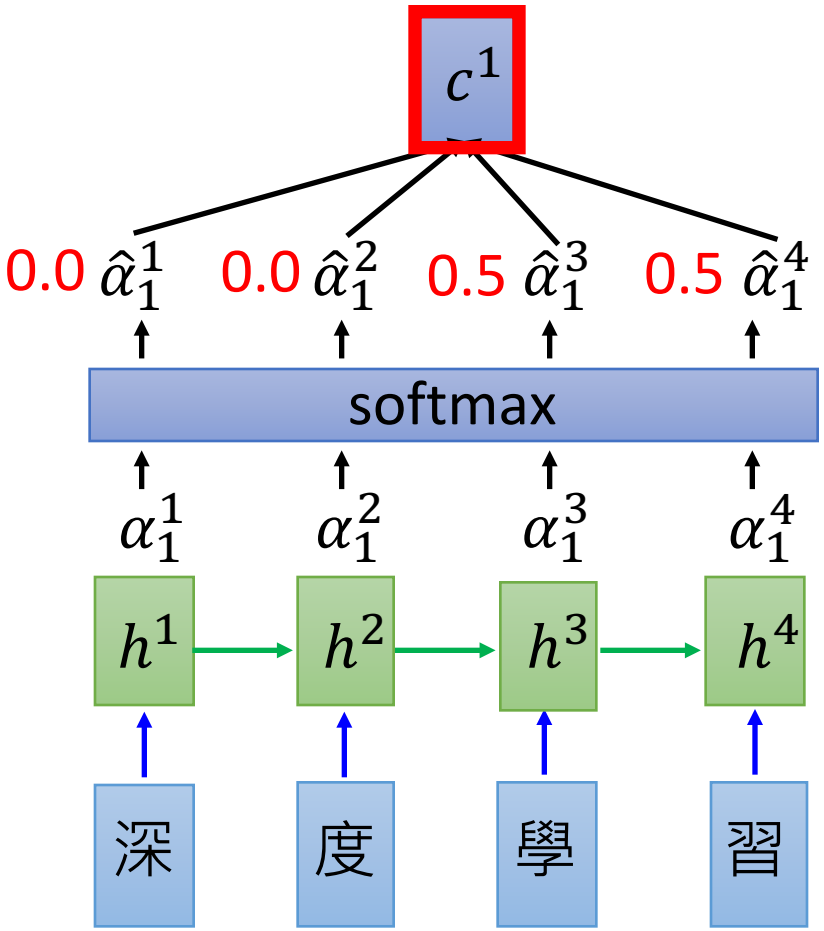
As RNN input

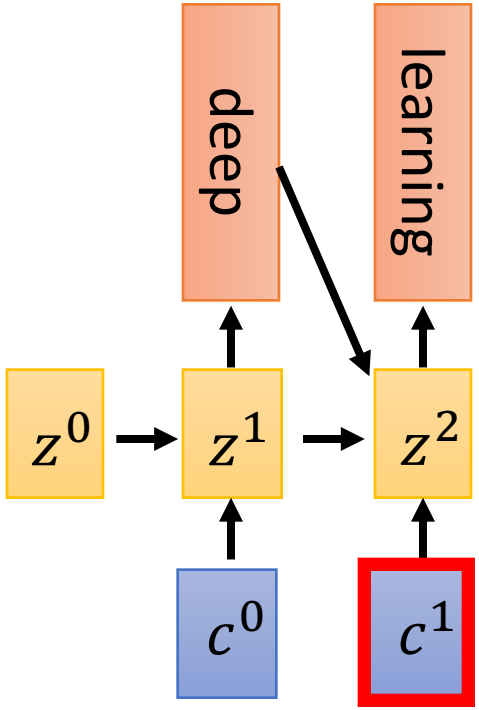
$$c^0 = \sum \hat{\alpha}_0^i h^i = 0.5h^1 + 0.5h^2$$

# Machine Translation with Attention



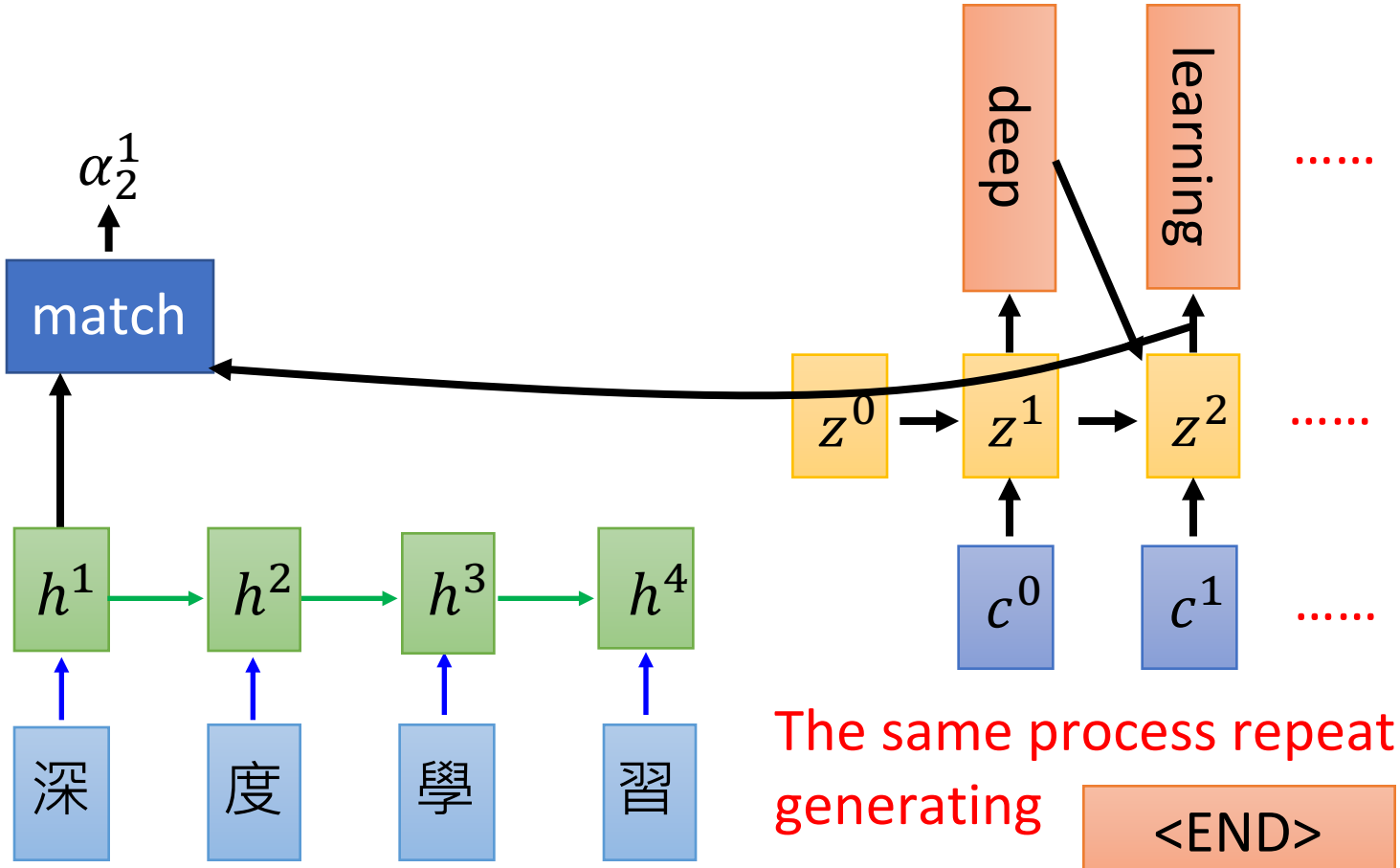
# Machine Translation with Attention



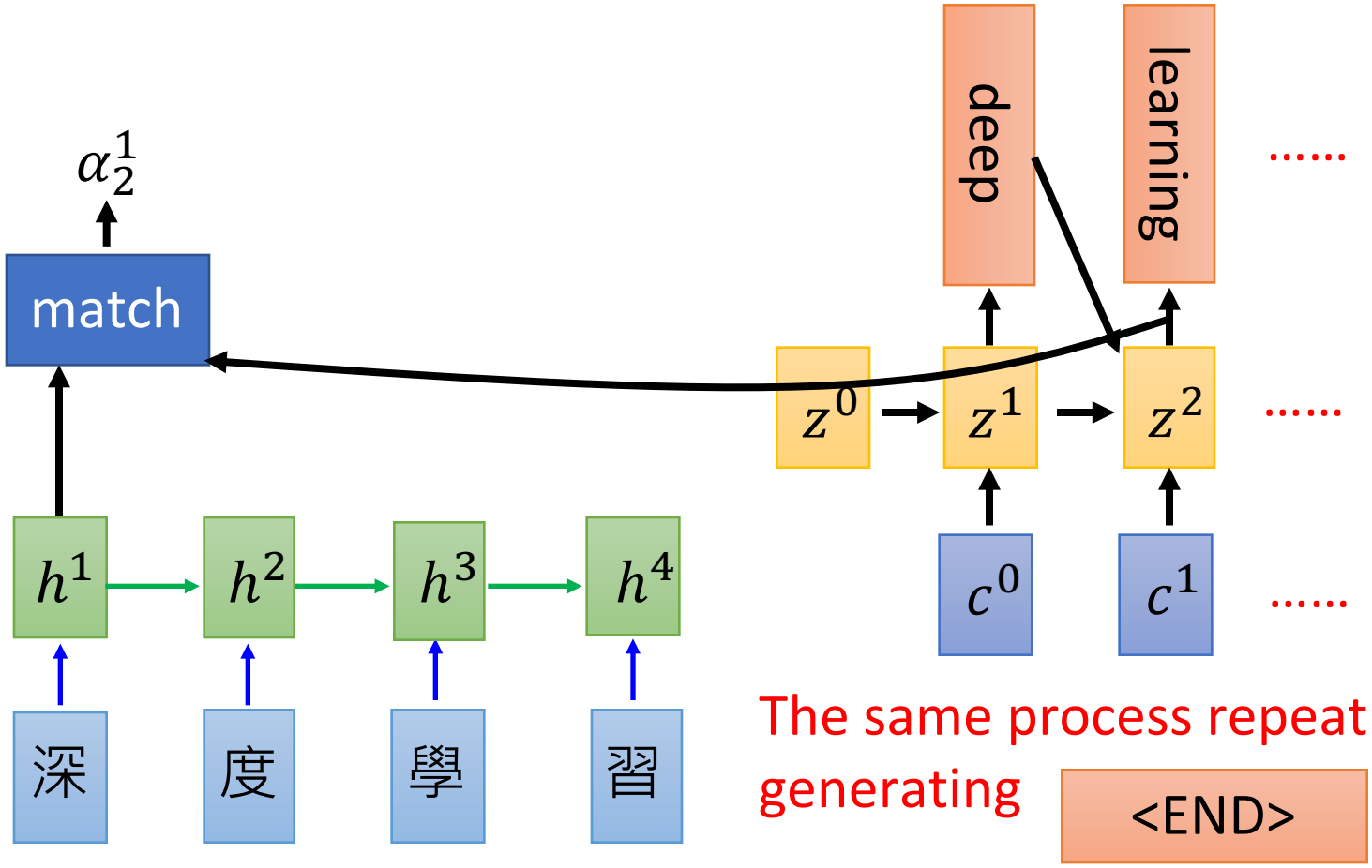




$$c^1 = \sum \hat{\alpha}_1^i h^i = 0.5h^3 + 0.5h^4$$

# Machine Translation with Attention



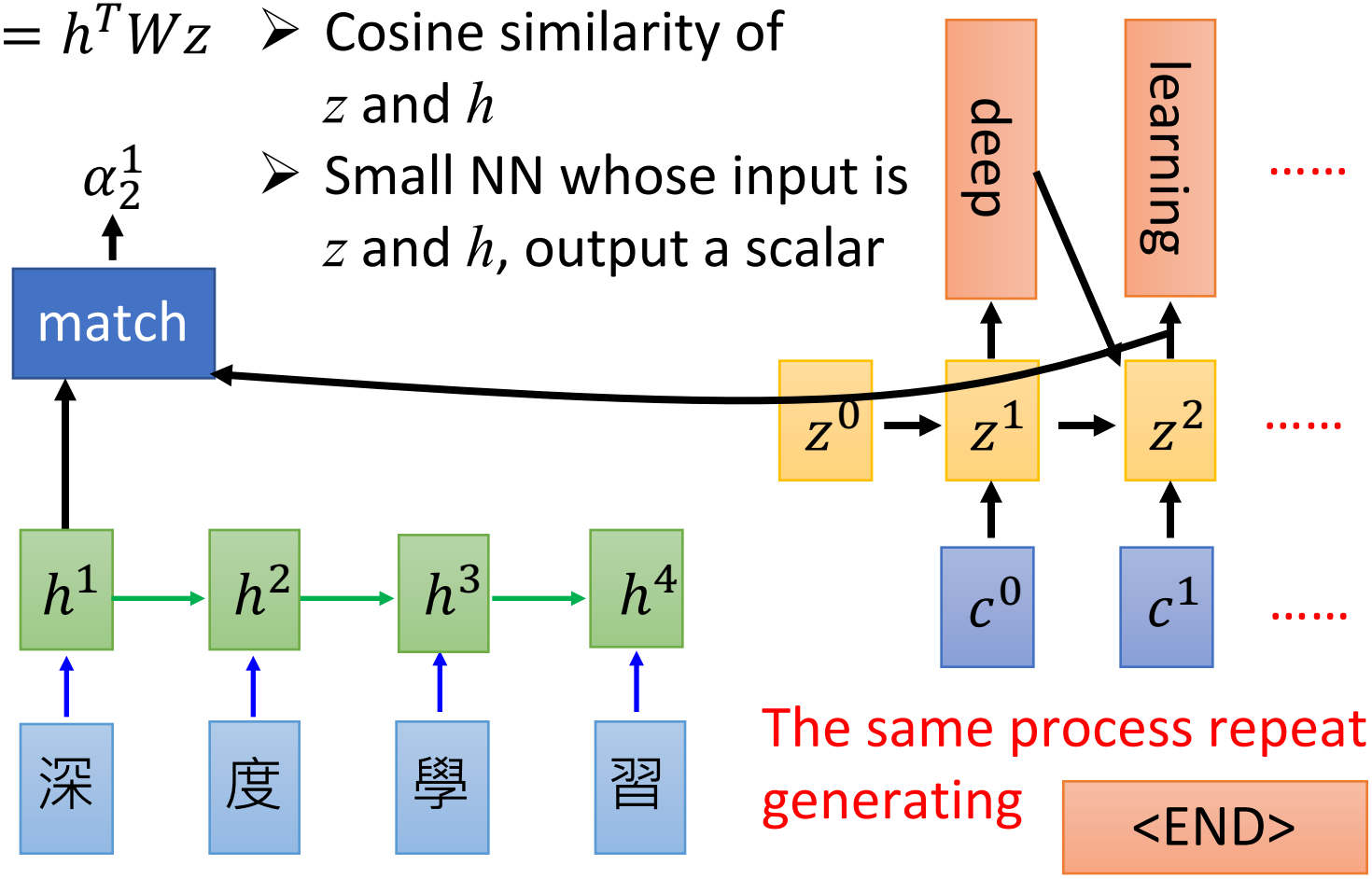
# Group Discussion: how would you derive the attention weight $\alpha_2^1$?




The same process repeat until generating <END>

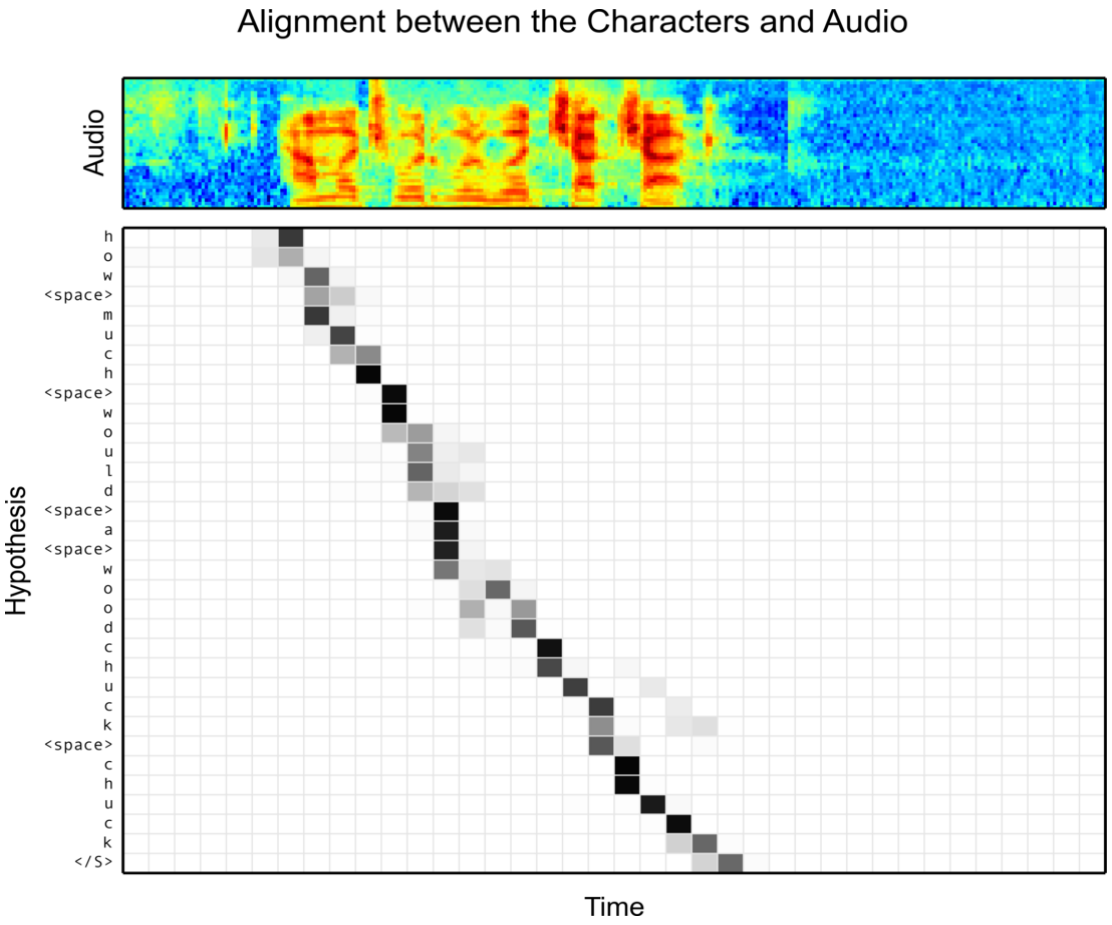
# Group Discussion: how would you derive the attention weight $\alpha_2^1$?

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




The same process repeat until generating <END>

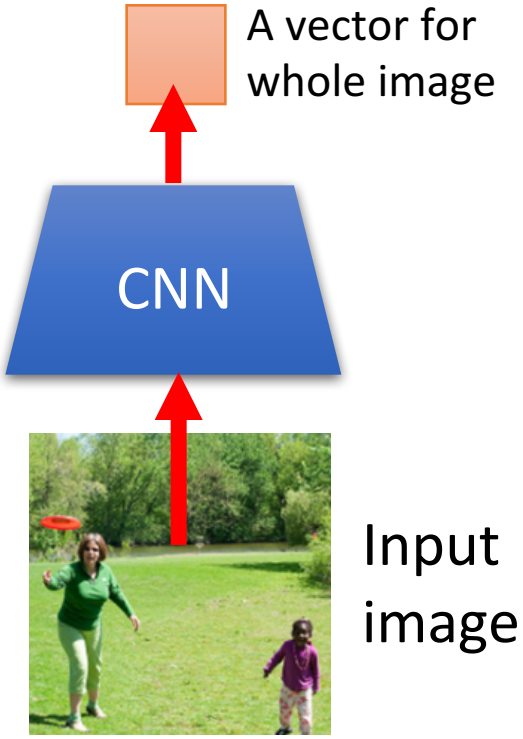
# Speech Recognition with Attention

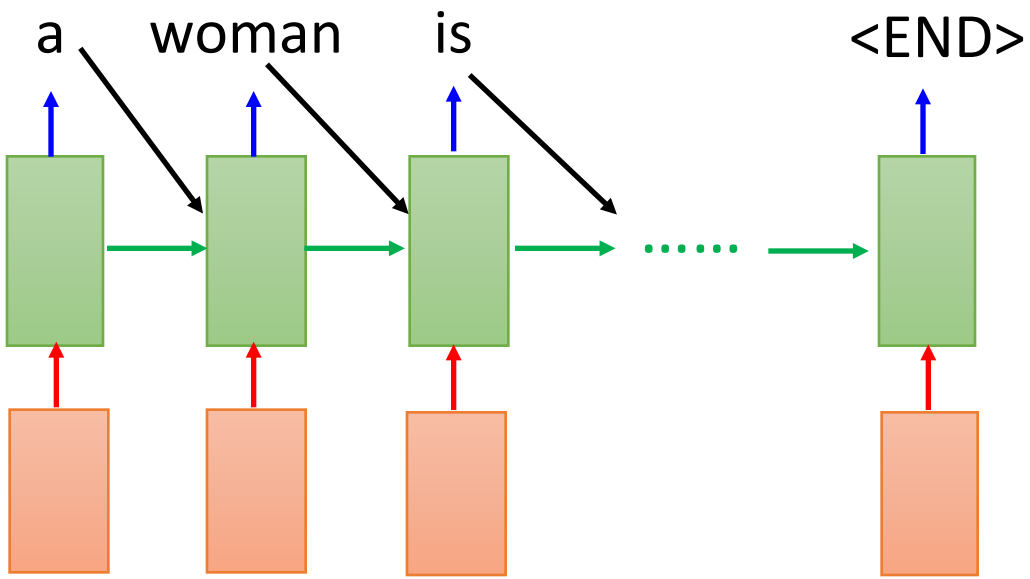


# Image Captioning

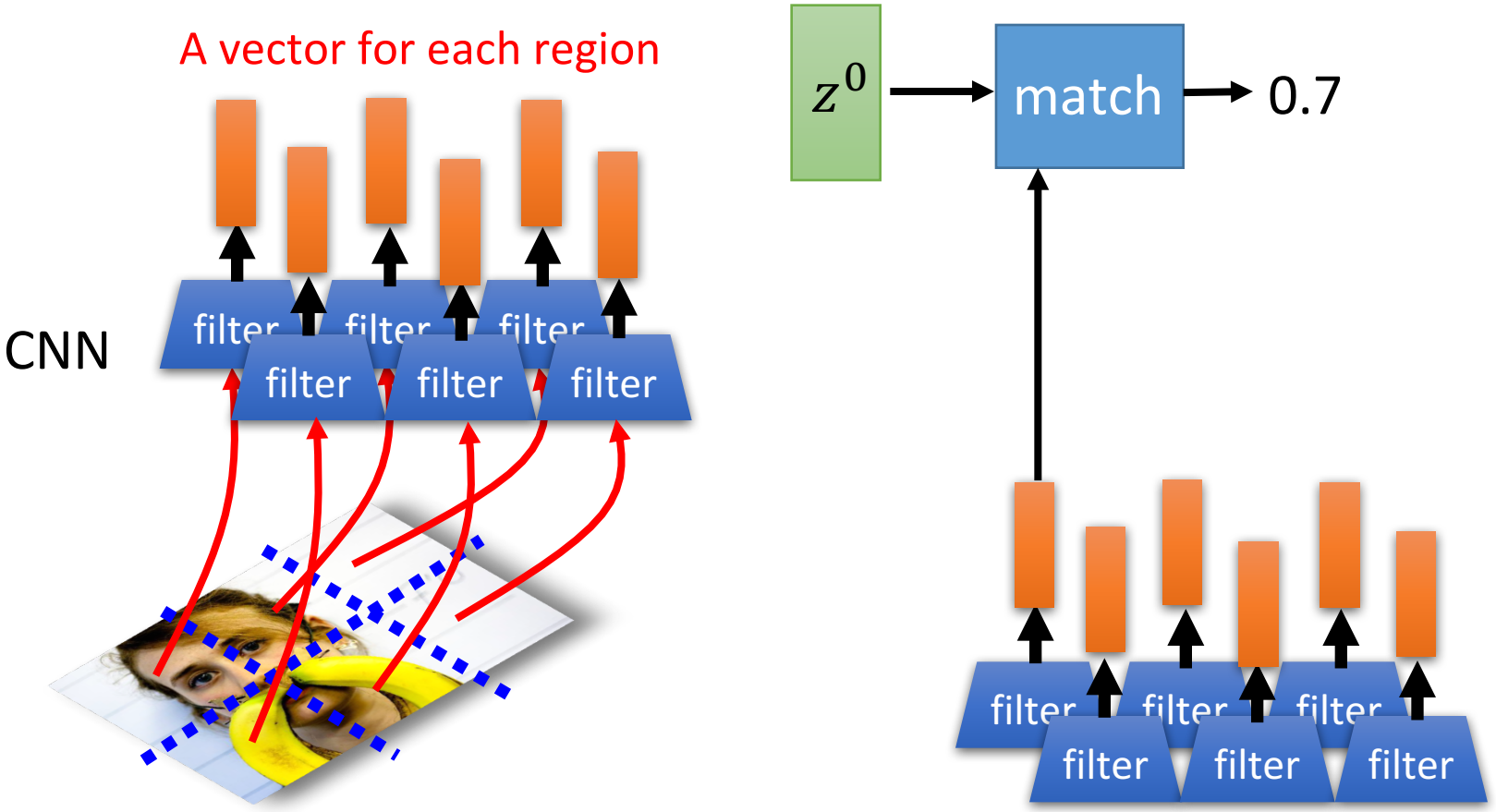
Input: image

Output: word sequence

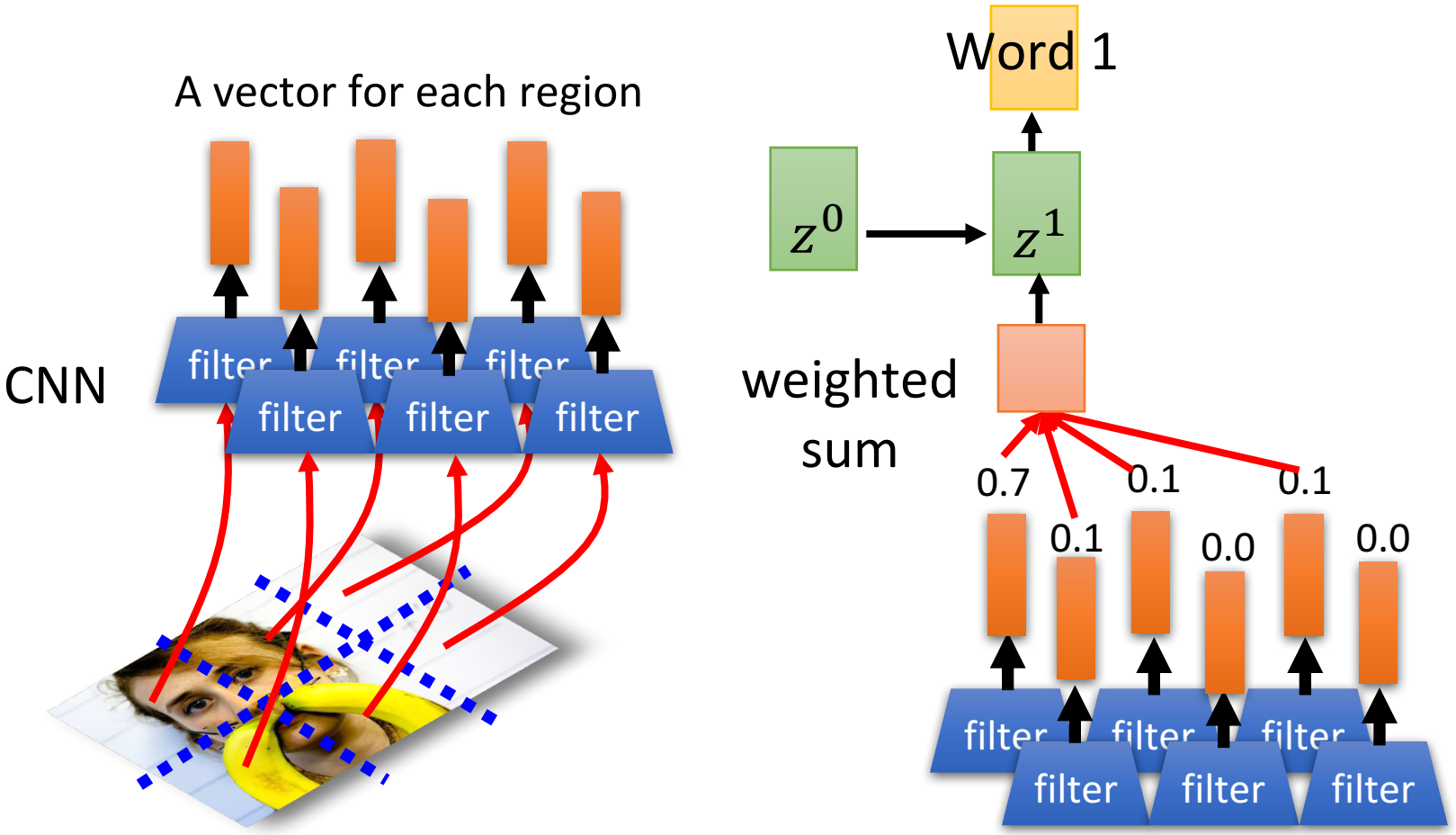


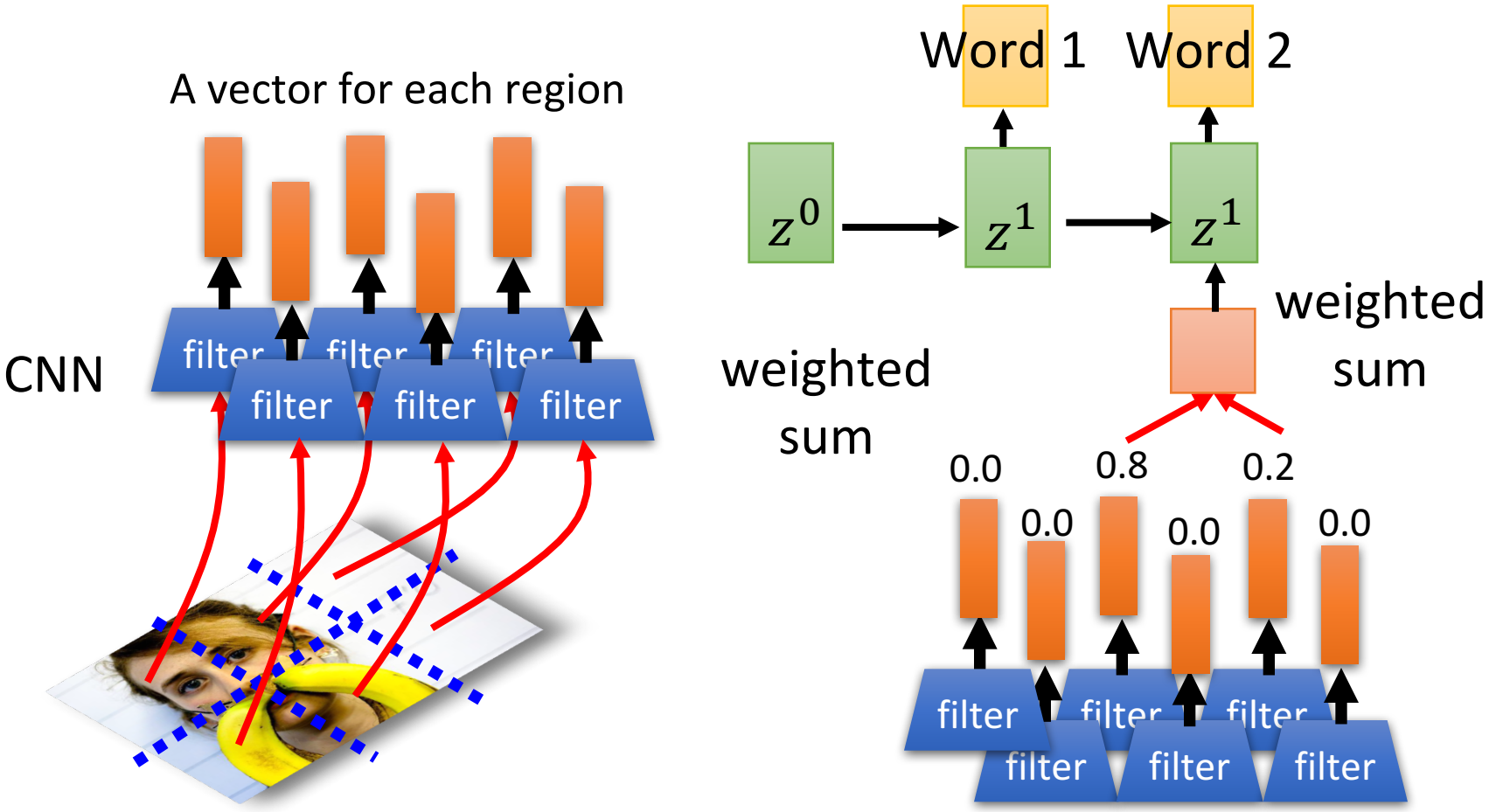
# Image Captioning with Attention



# 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

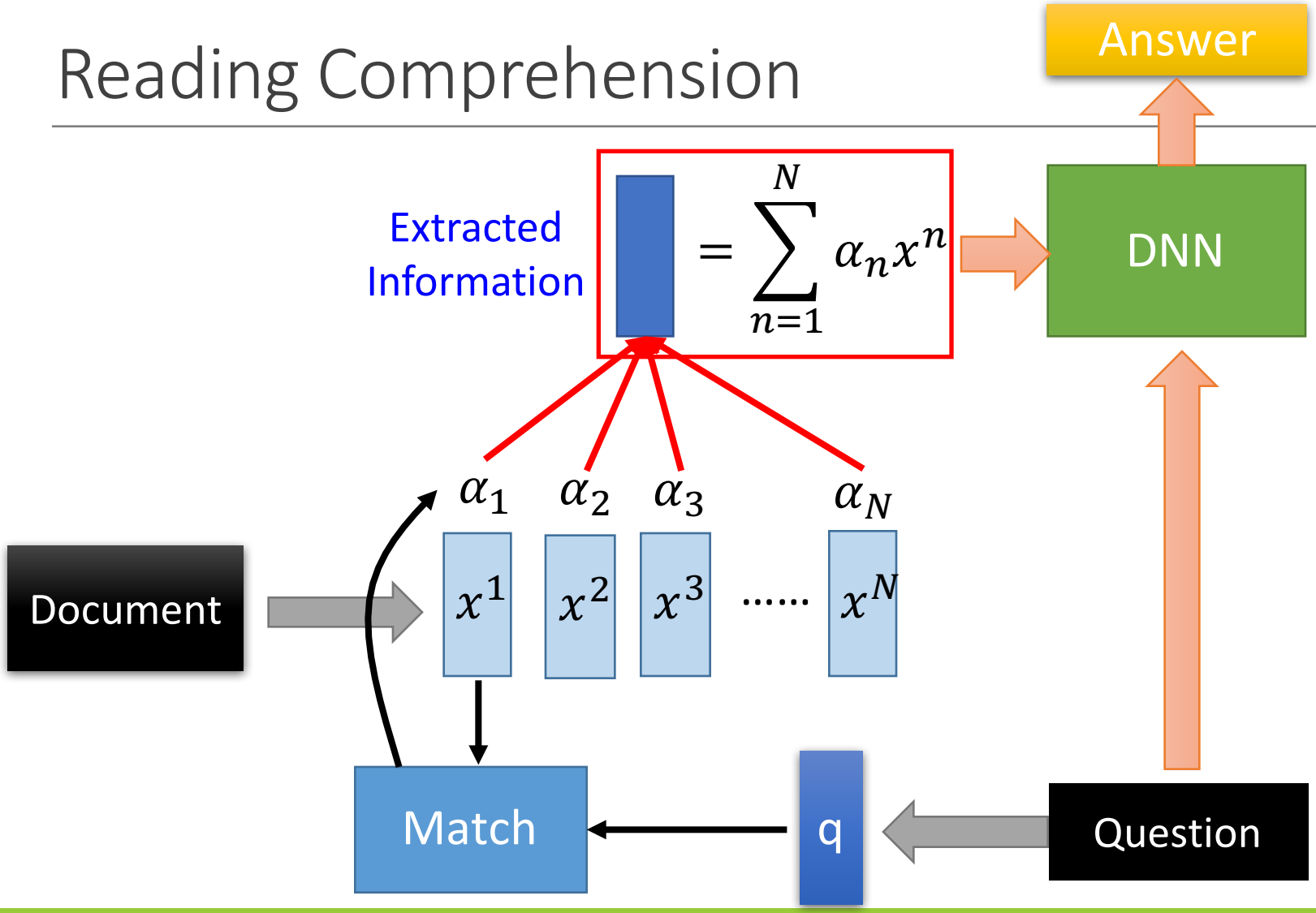
**Someone** is **frying** a **fish** in a **pot**

# Group Discussion:

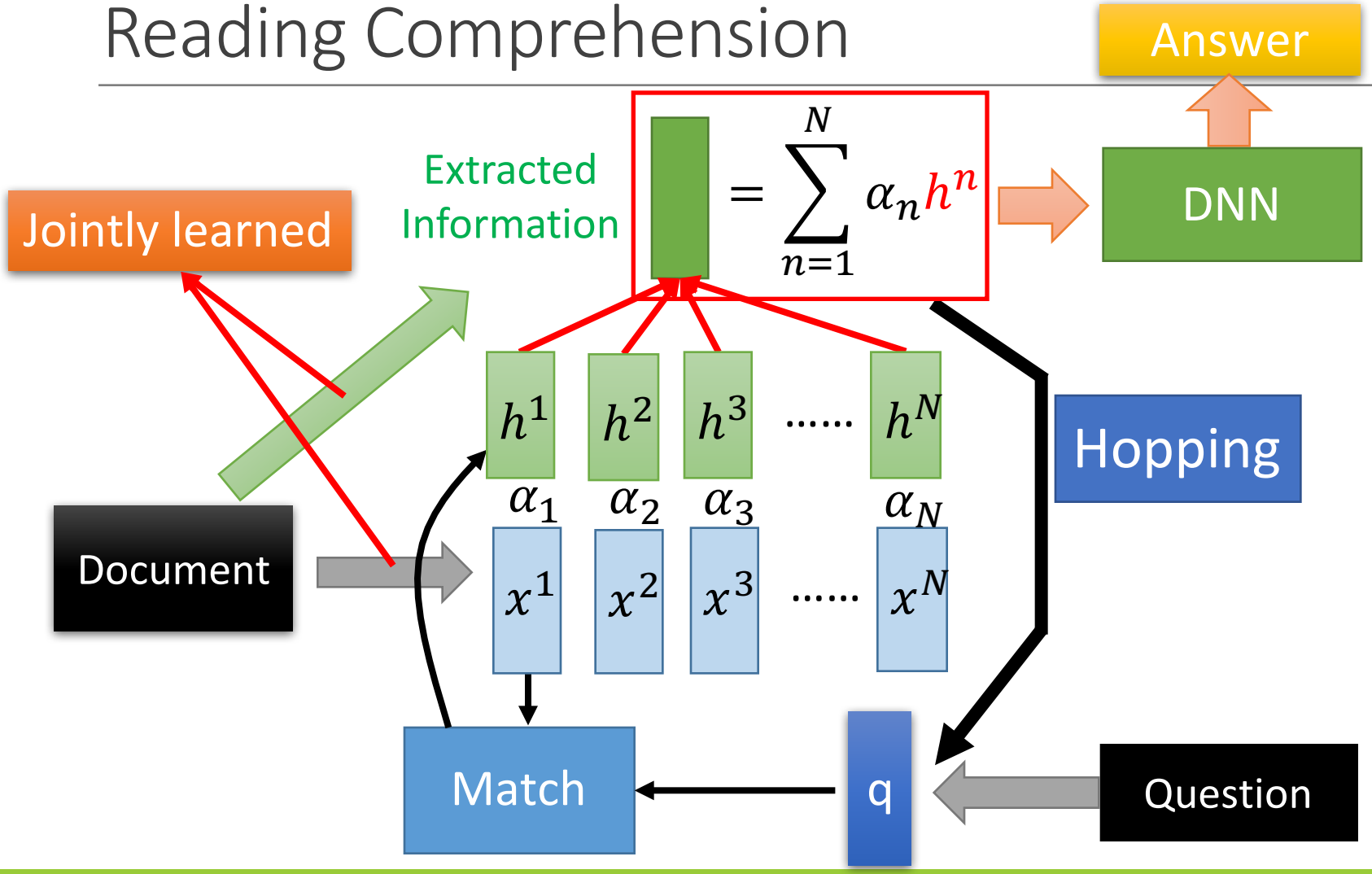
Learning attention weights vs using cosine similarity, which one is better? Why?

- $\alpha = h^T W z$
- Cosine similarity of $z$ and $h$

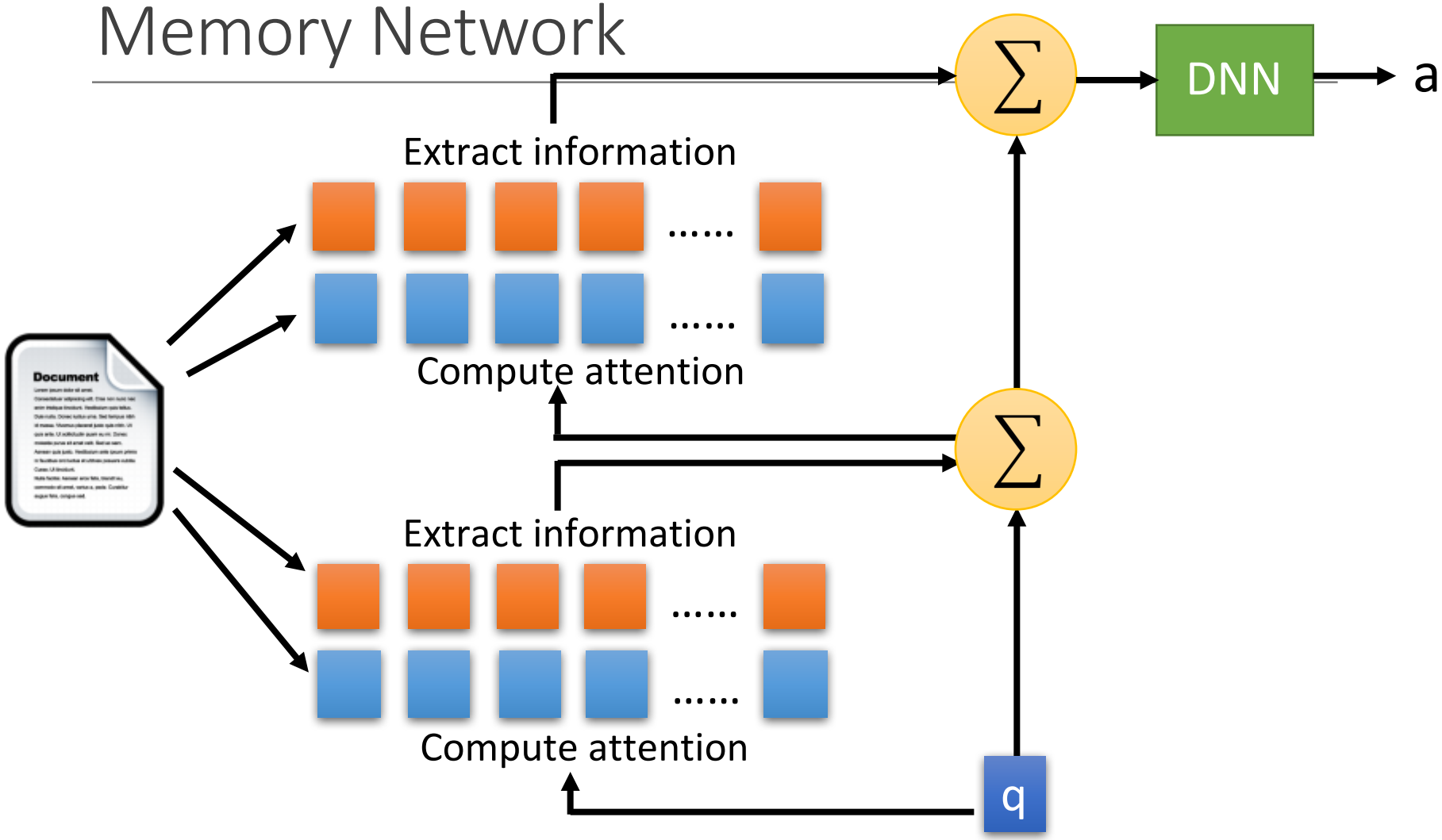
# Reading Comprehension



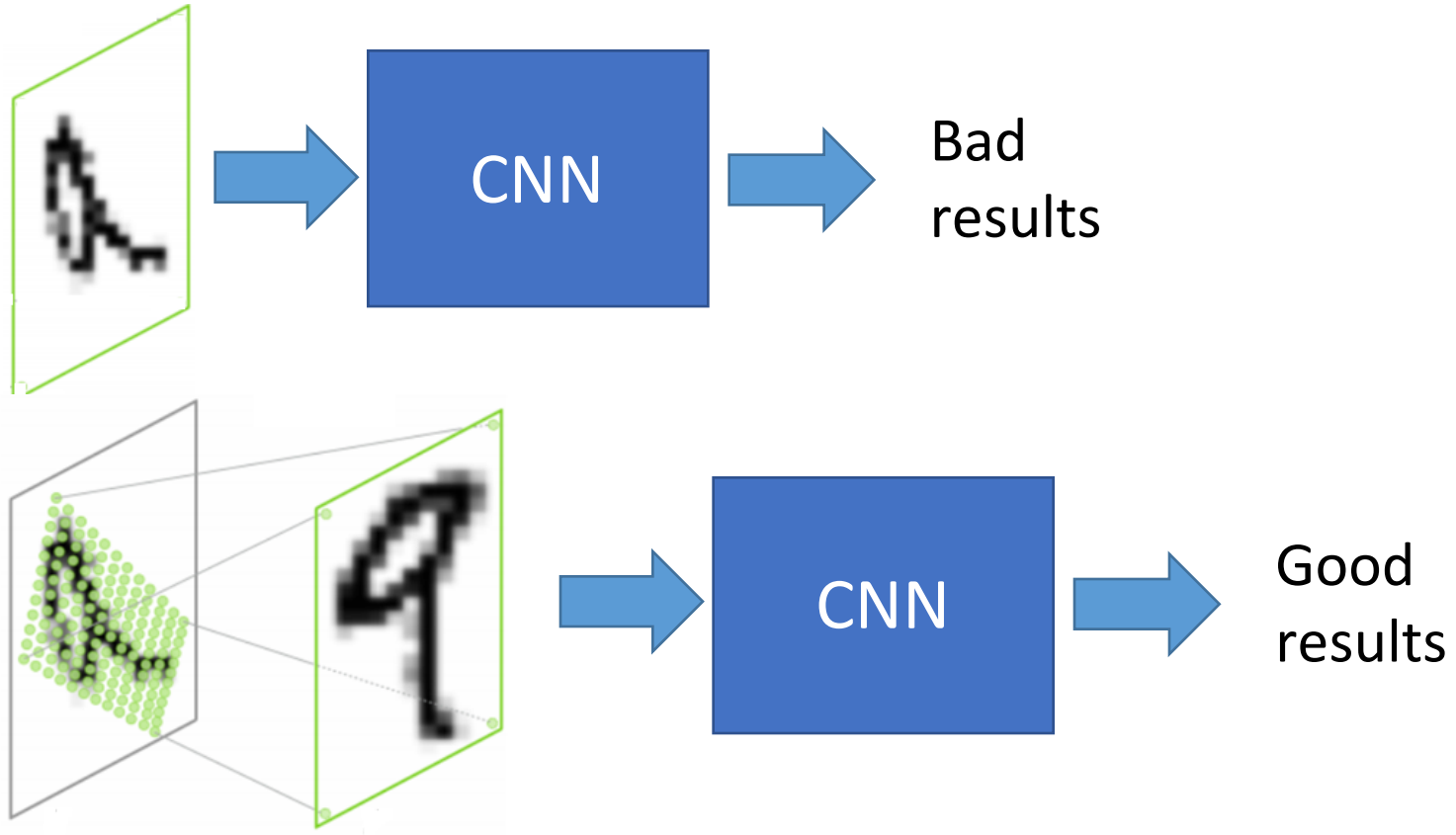
# Reading Comprehension



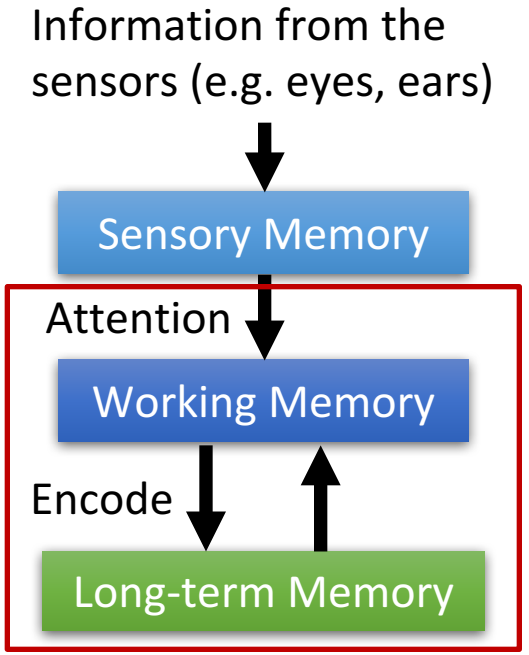
# Memory Network



# Special Attention: Spatial Transformers



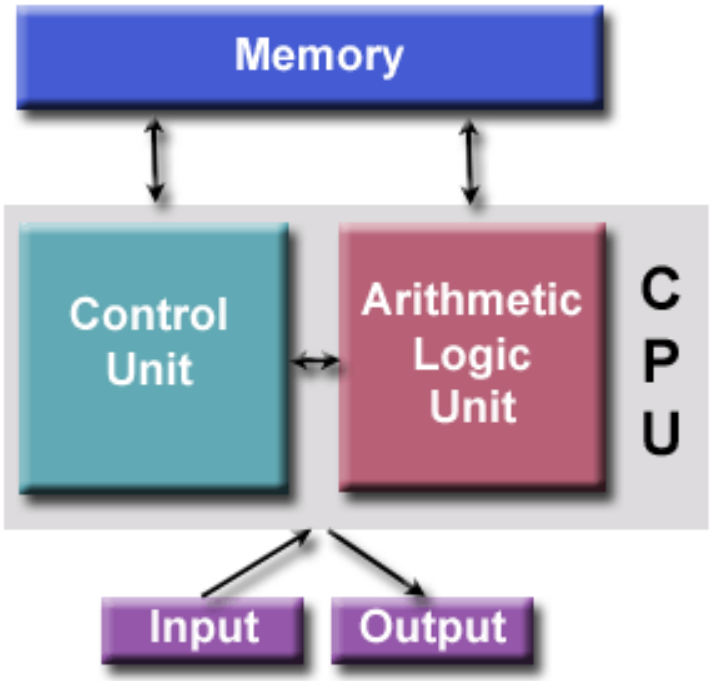
# Attention on Memory



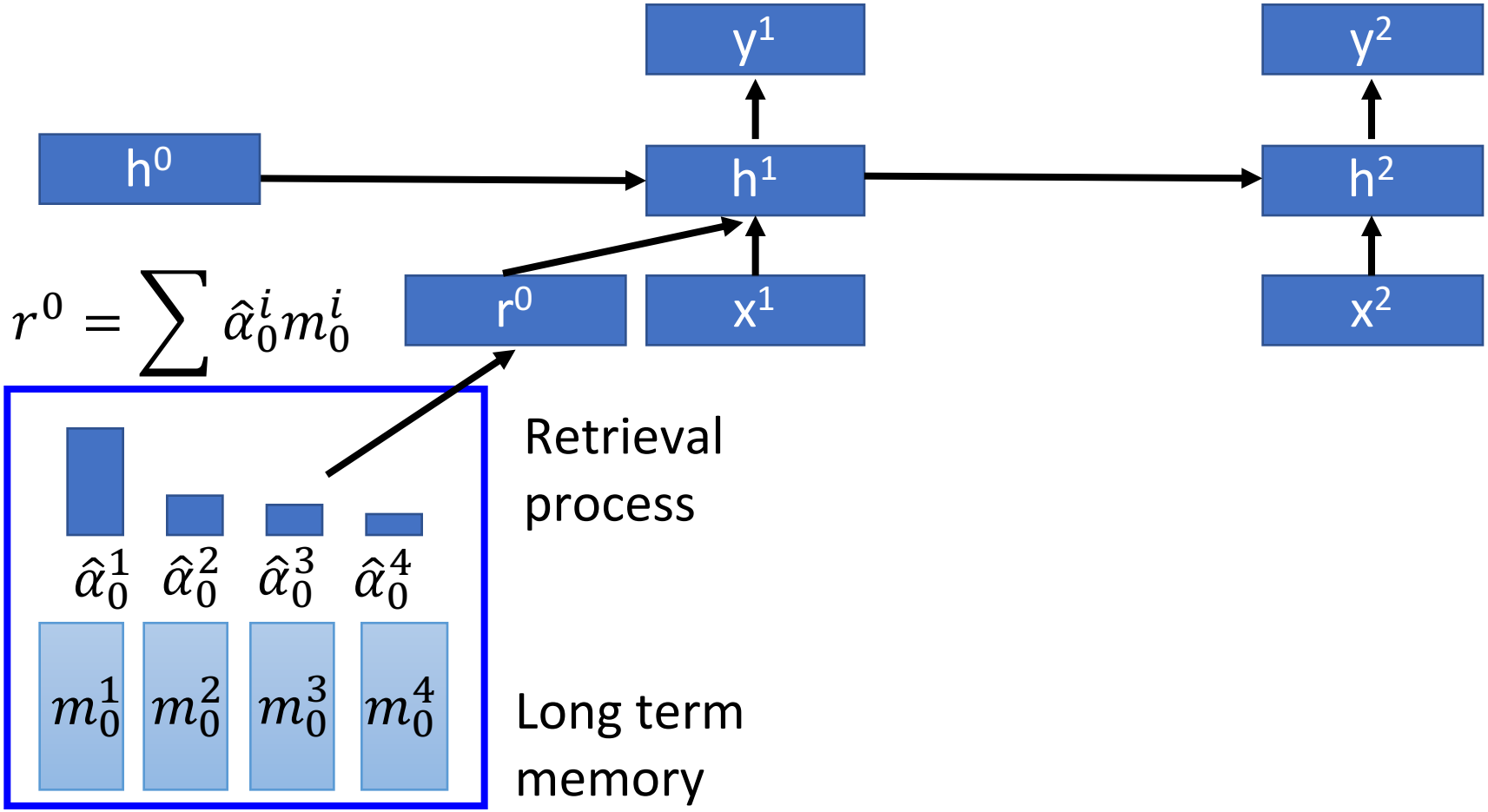
# Neural Turing Machine

## Von Neumann architecture

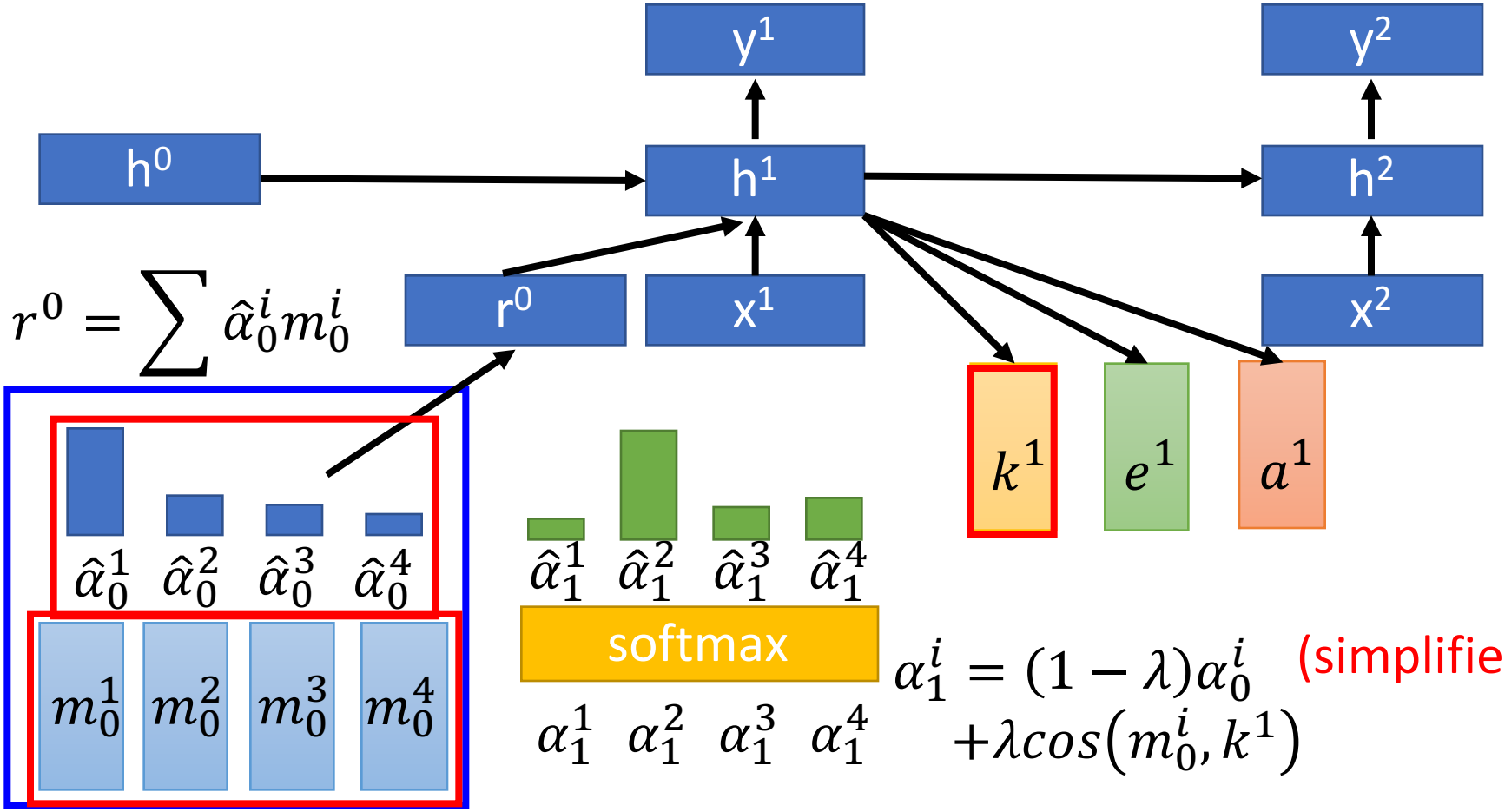
Neural Turing Machine is an advanced RNN/LSTM.



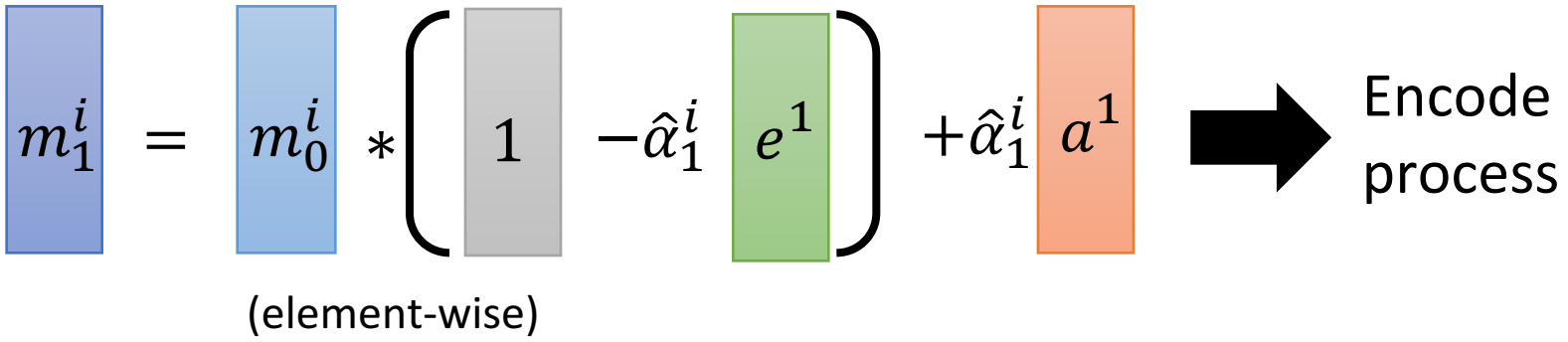
# Neural Turing Machine

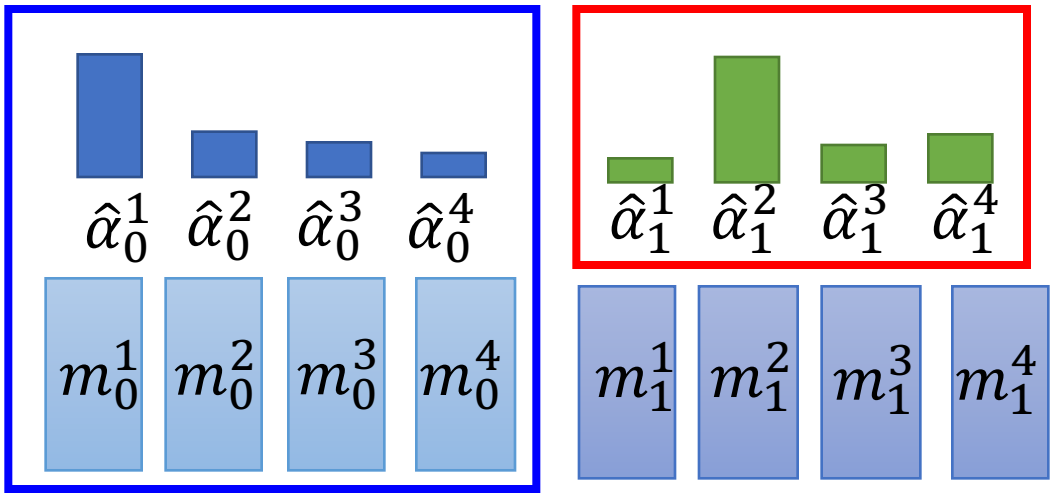


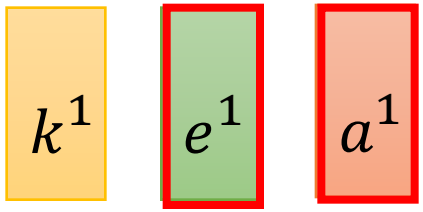
# Neural Turing Machine



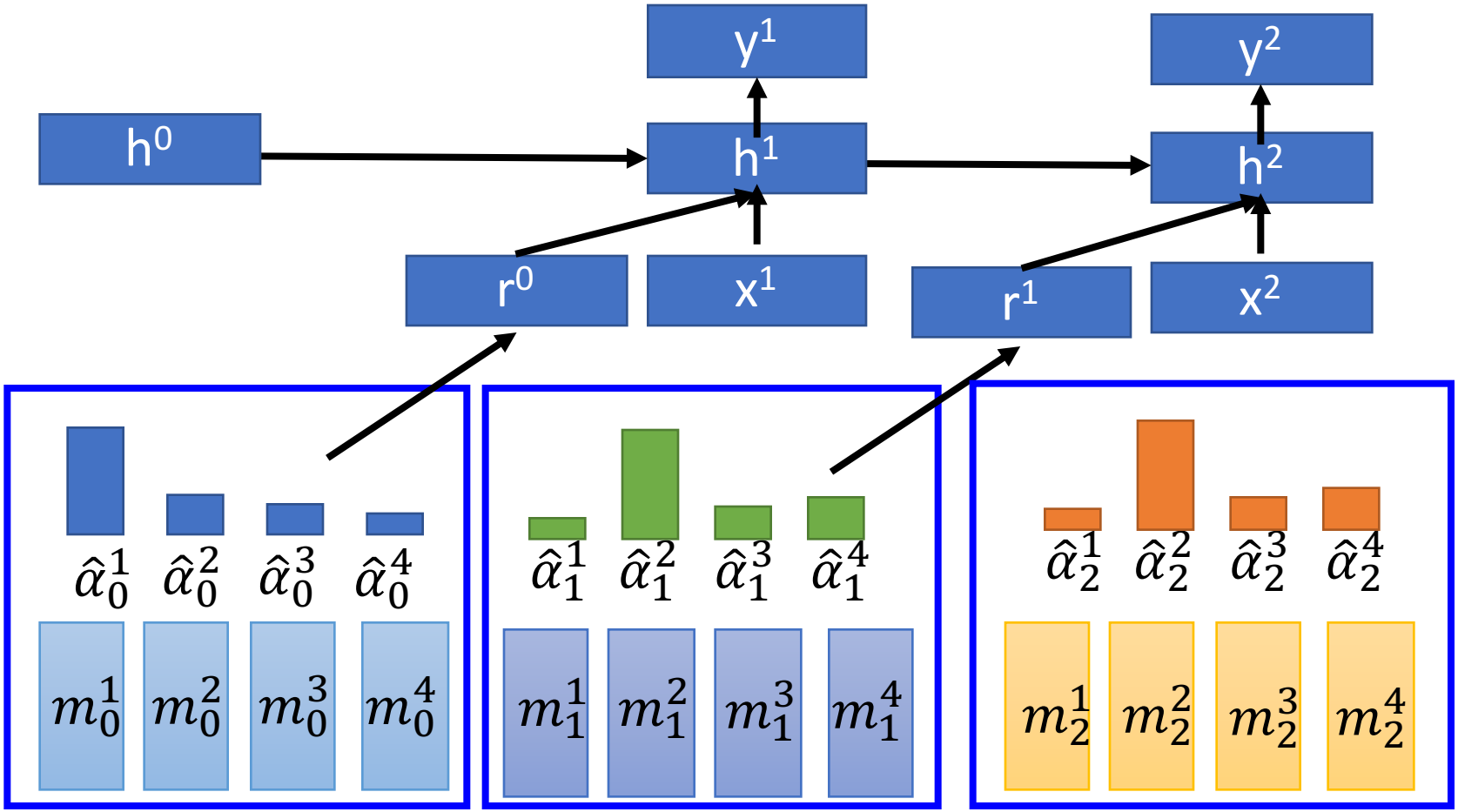
# Neural Turing Machine



# Neural Turing Machine



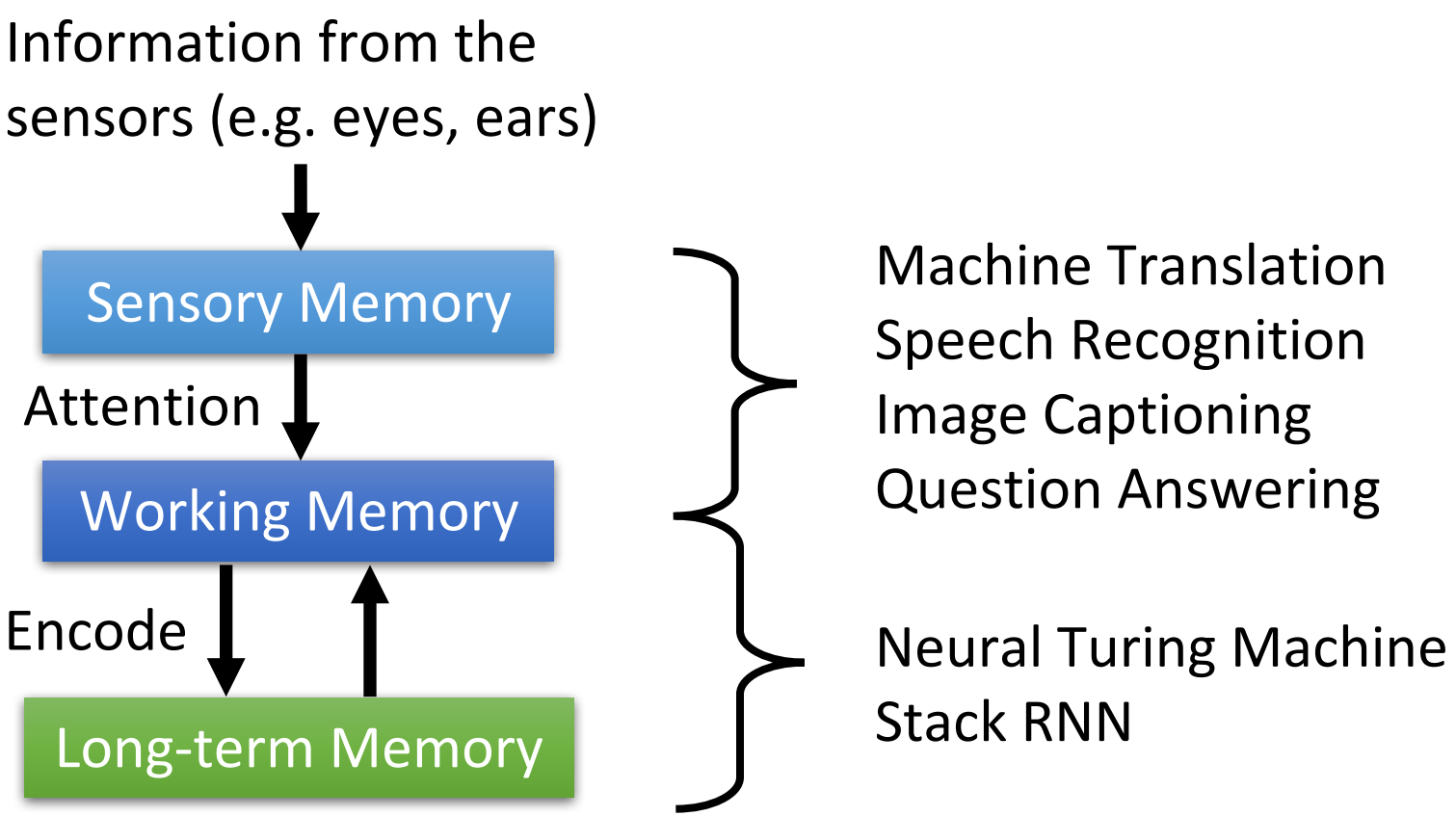

Zhang et al., "Structured Memory for Neural Turing Machines," arXiv, 2015.

# Stack RNN


stack
0.7
0.2
0.1
Push, Pop, Nothing
Information
to store
$y^t$
$h^t$
$x^t$
Push
X0.7
+
Pop
-1
-1
X0.2
+
Nothing
-1
X0.1

# Quick Question: why stack but not queue?

# Concluding Remarks



# Reference

End-To-End Memory Networks. S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus. arXiv Pre-Print, 2015.

Neural Turing Machines. Alex Graves, Greg Wayne, Ivo Danihelka. arXiv Pre-Print, 2014

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. Kumar et al. arXiv Pre-Print, 2015

Neural Machine Translation by Jointly Learning to Align and Translate. D. Bahdanau, K. Cho, Y. Bengio; International Conference on Representation Learning 2015.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Kelvin Xu et. al.. arXiv Pre-Print, 2015.

Attention-Based Models for Speech Recognition. Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio. arXiv Pre-Print, 2015.

A Neural Attention Model for Abstractive Sentence Summarization. A. M. Rush, S. Chopra and J. Weston. EMNLP 2015.