Conditional Generation by RNN & Attention

Hung-yi Lee
李宏毅

Outline

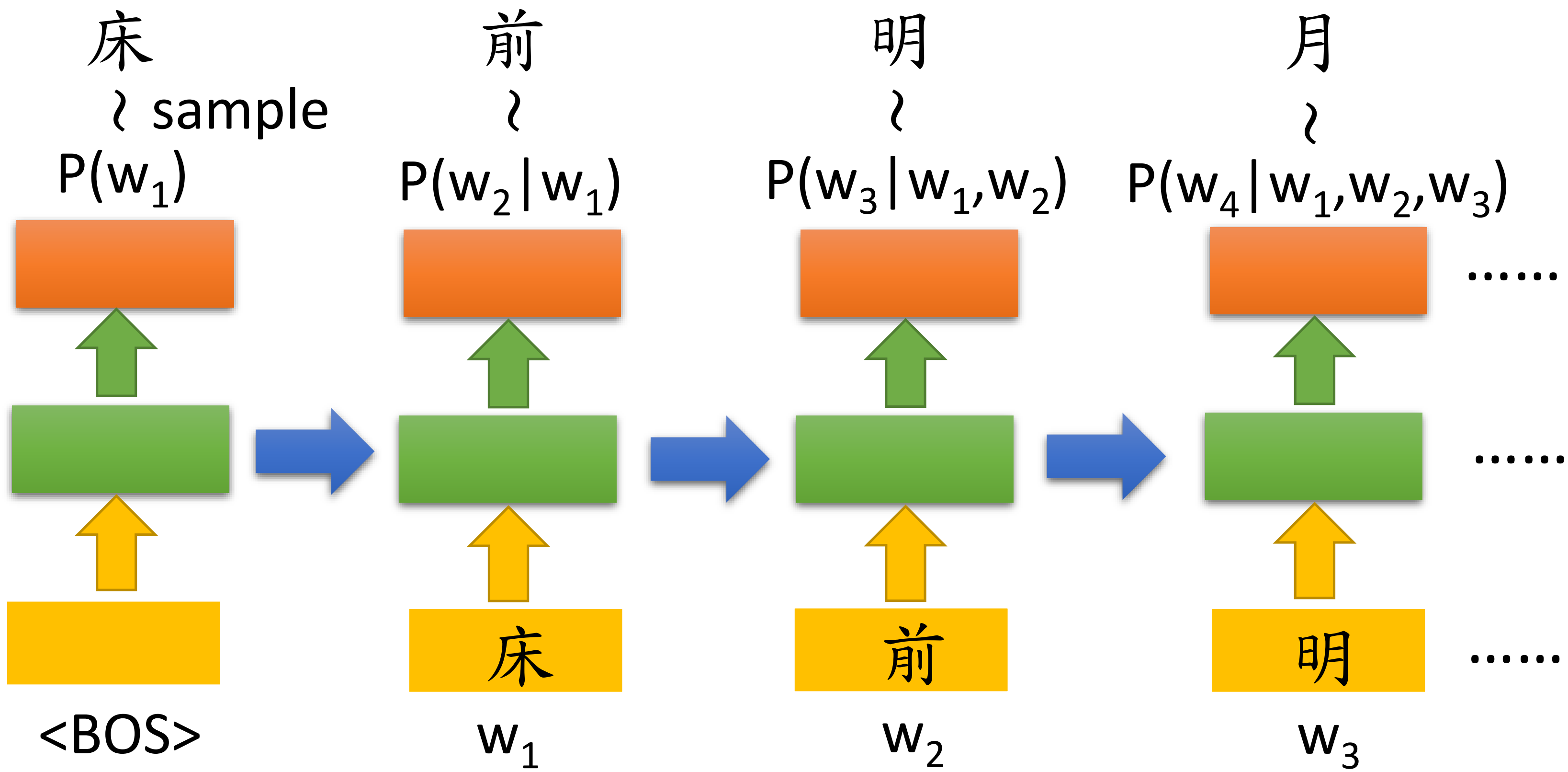
- Generation
- Attention
- Tips for Generation
- Pointer Network

Generation

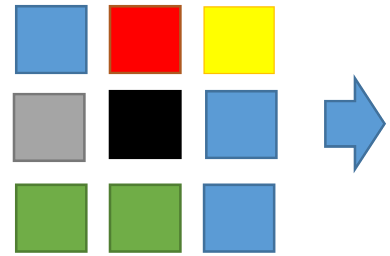
Generating a structured object component-by-component http://youtien.pixnet.net/blog/post/4604096-%E6%8E%A8%E6%96%87%E6%8E%A5%E9%BE%8D%E4%B9%8B%E5%B0%8D%E8%81%AF%E9%81%8A%E6%88%B2

Generation

- Sentences are composed of characters/words
  - Generating a character/word at each time by RNN



Generation

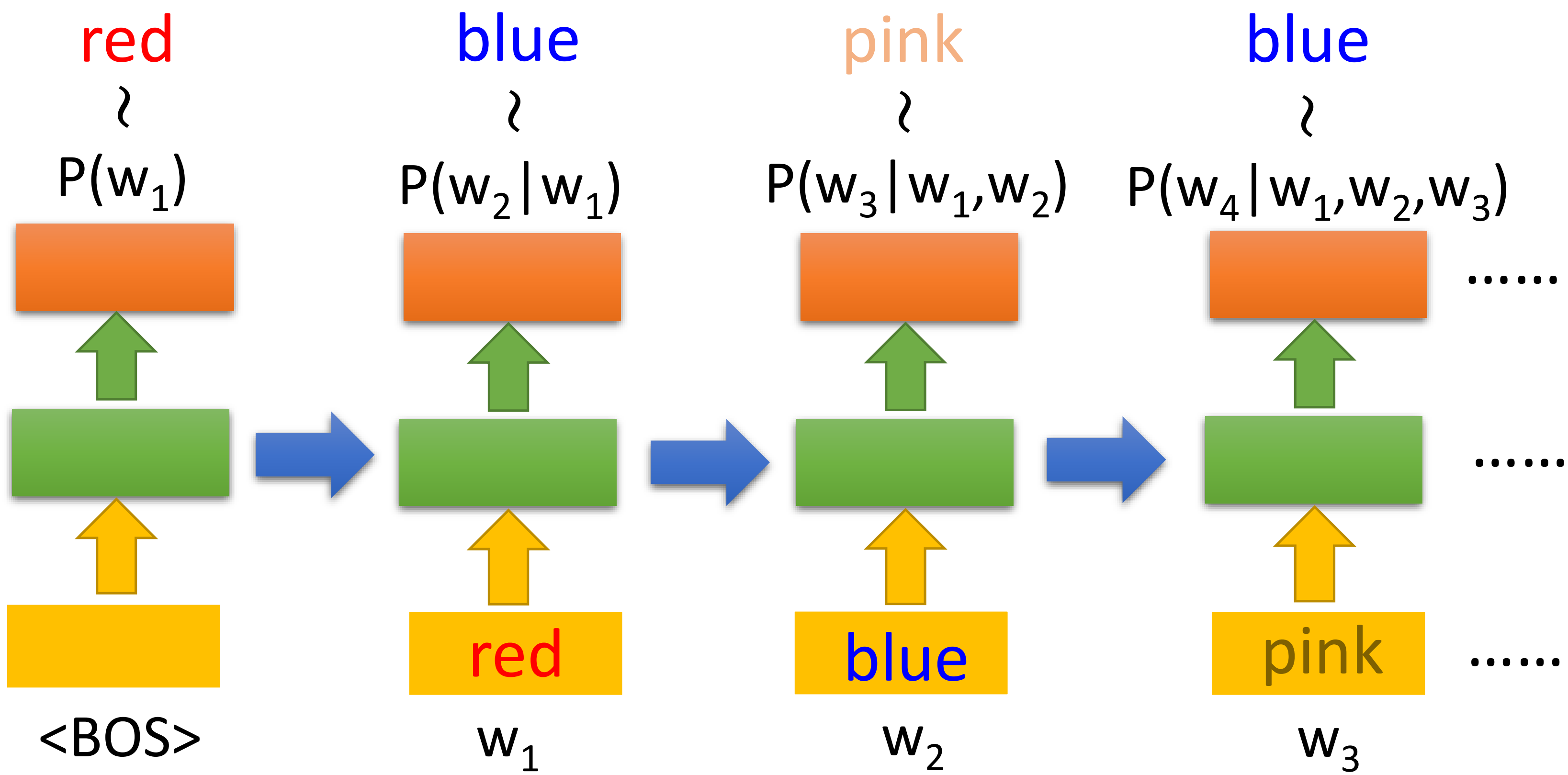


Consider as a sentence

blue red yellow gray ……

Train a language model based on the “sentences”

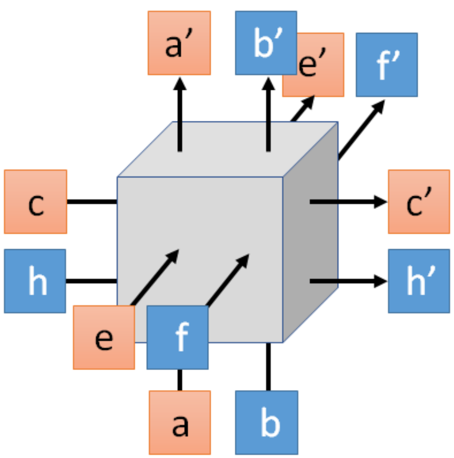
- Images are composed of pixels
  - Generating a pixel at each time by RNN

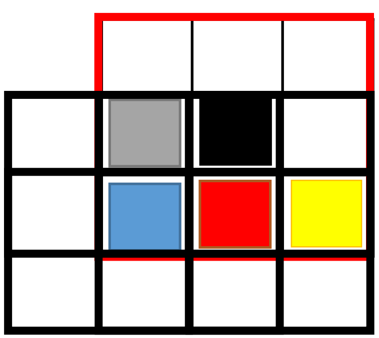


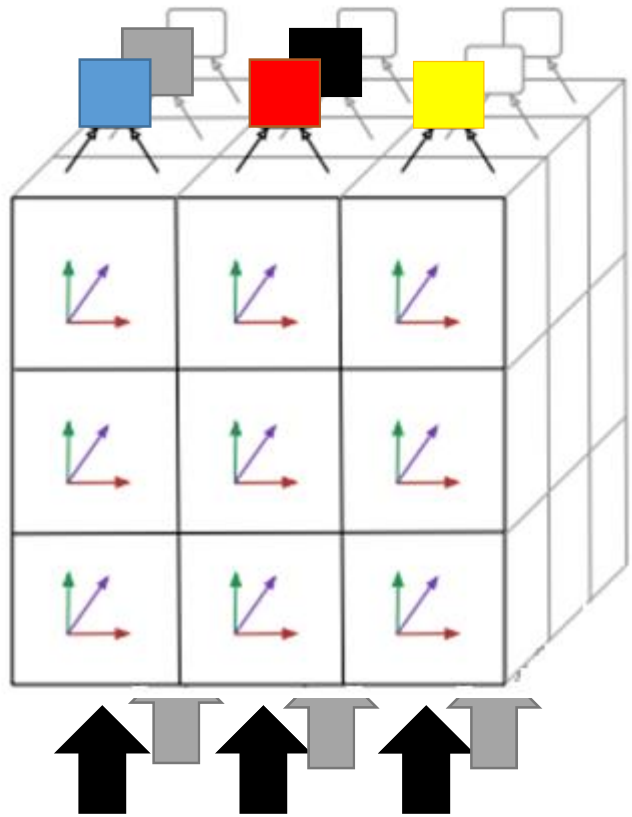
Generation

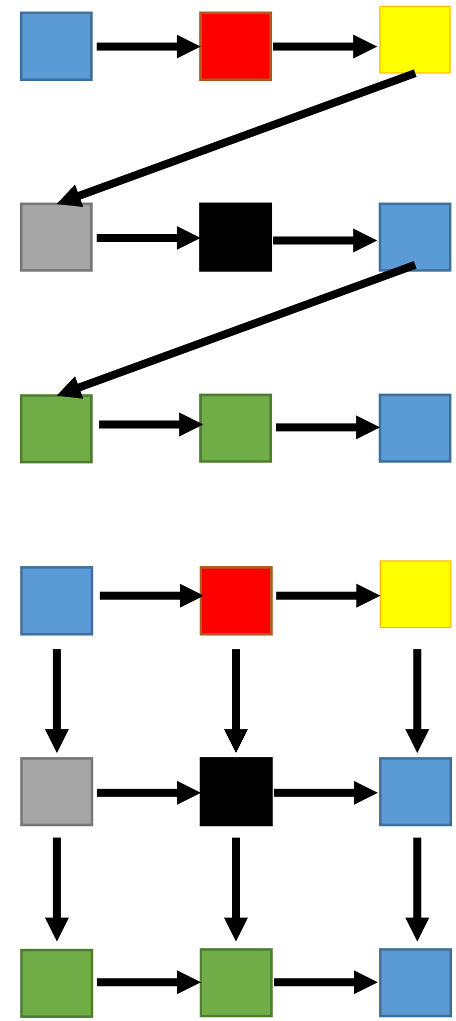
3 x 3 images

- Images are composed of pixels









Generation

- Image
  - Aaron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu, Pixel Recurrent Neural Networks, arXiv preprint, 2016
  - Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, Koray Kavukcuoglu, Conditional Image Generation with PixelCNN Decoders, arXiv preprint, 2016
- Video
  - Aaron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu, Pixel Recurrent Neural Networks, arXiv preprint, 2016
- Handwriting
  - Alex Graves, Generating Sequences With Recurrent Neural Networks, arXiv preprint, 2013
- Speech
  - Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior,Koray Kavukcuoglu, WaveNet: A Generative Model for Raw Audio, 2016

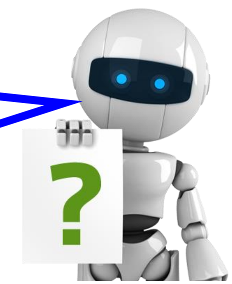
Conditional Generation

- We don't want to simply generate some random sentences.
- Generate sentences based on conditions:

Caption Generation

Given condition:

"A young girl is dancing."

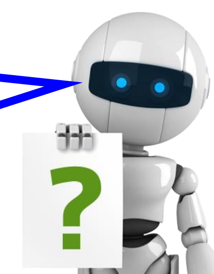


Chat-bot

Given condition:

"Hello"

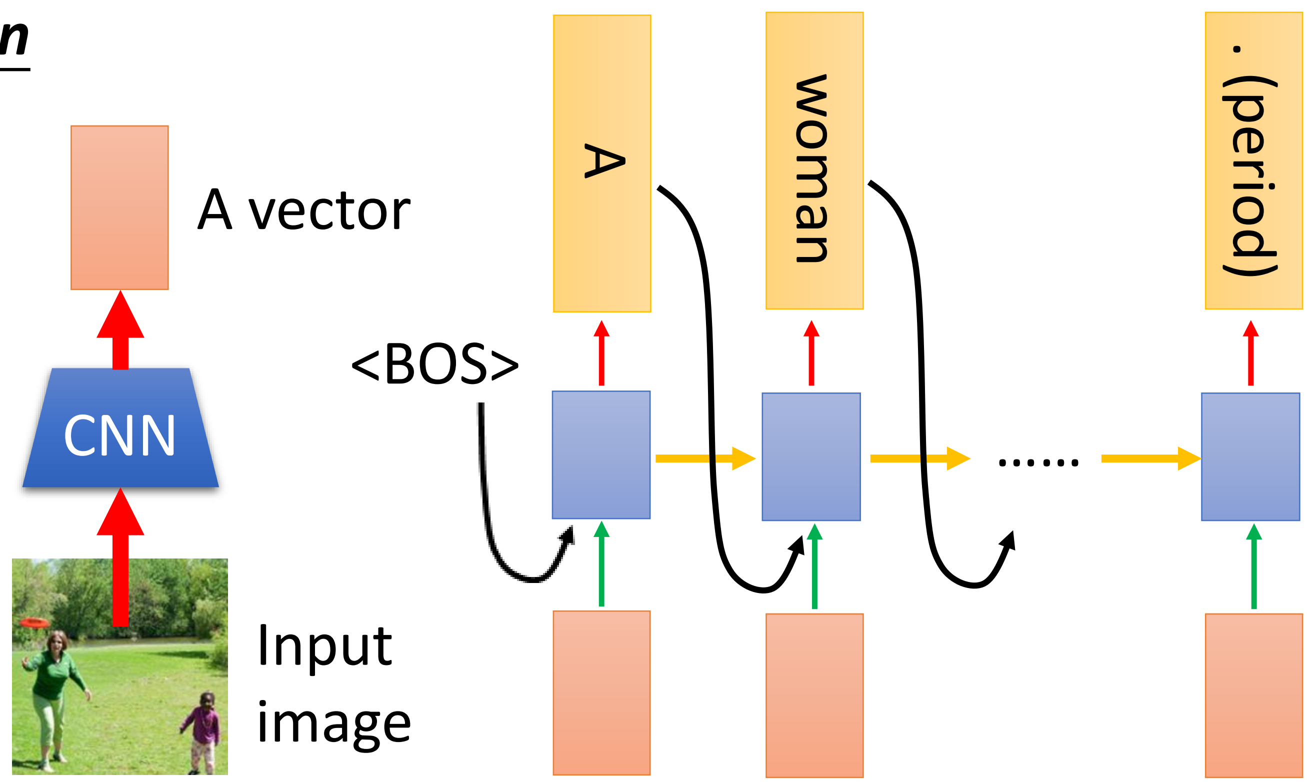
"Hello. Nice to see you."



Conditional Generation

- Represent the input condition as a vector, and consider the vector as the input of RNN generator

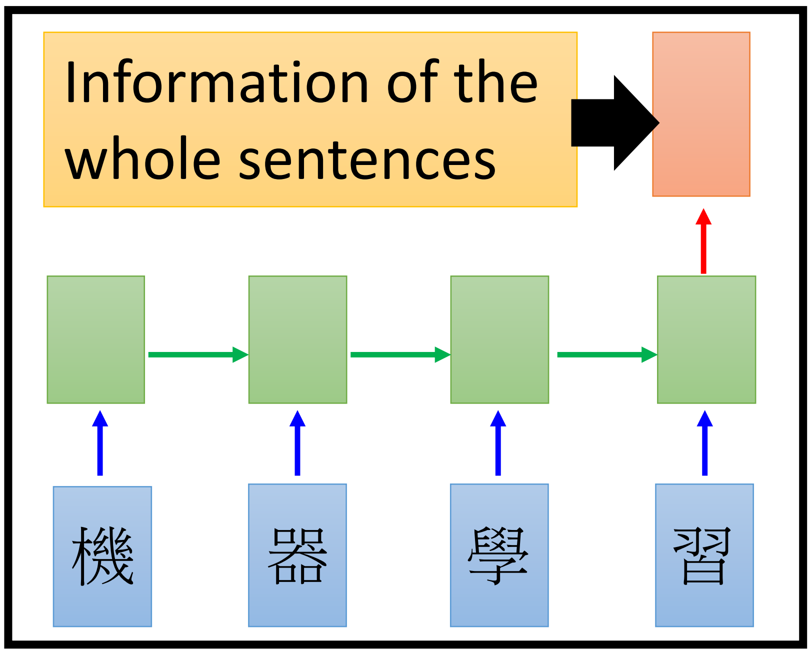
<u>Image Caption Generation</u>

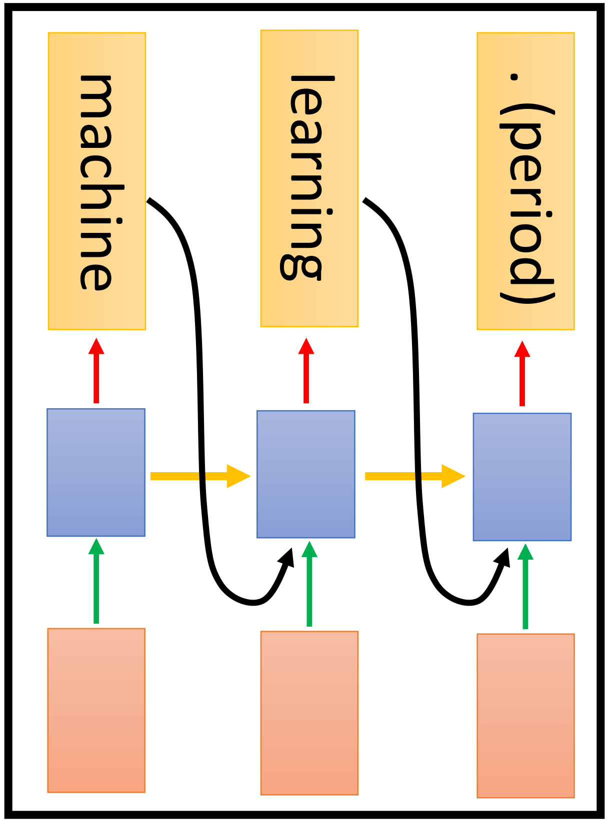


Conditional Generation

Sequence-to-sequence learning

- Represent the input condition as a vector, and consider the vector as the input of RNN generator
- E.g. Machine translation / Chat-bot





Encoder ← Jointly train → ***Decoder***

Conditional Generation

M: Hello

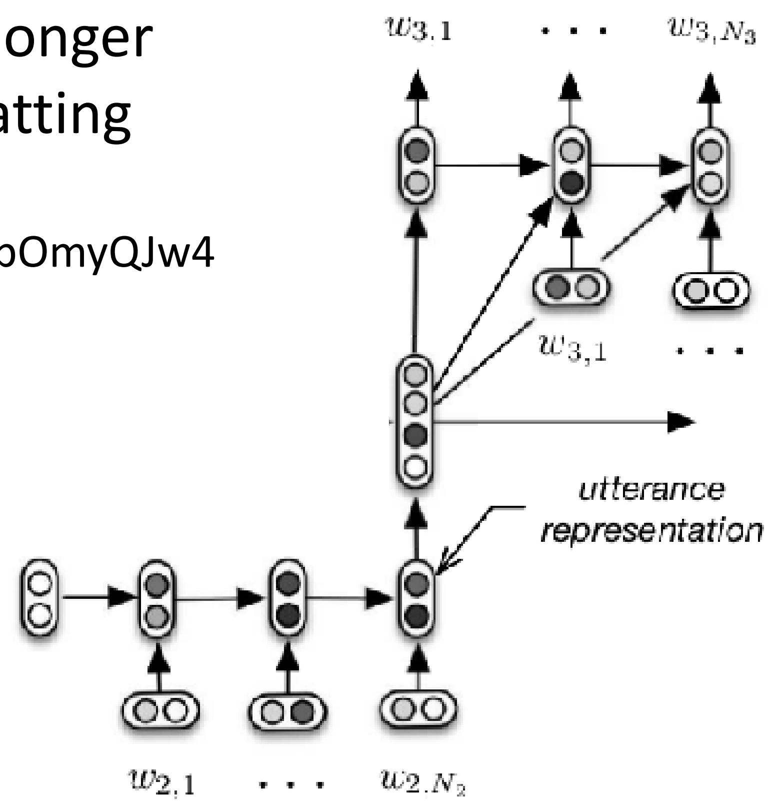
U: Hi

M: Hi

Need to consider longer context during chatting

https://www.youtube.com/watch?v=e2MpOmyQJw4

M: Hi



M: Hello

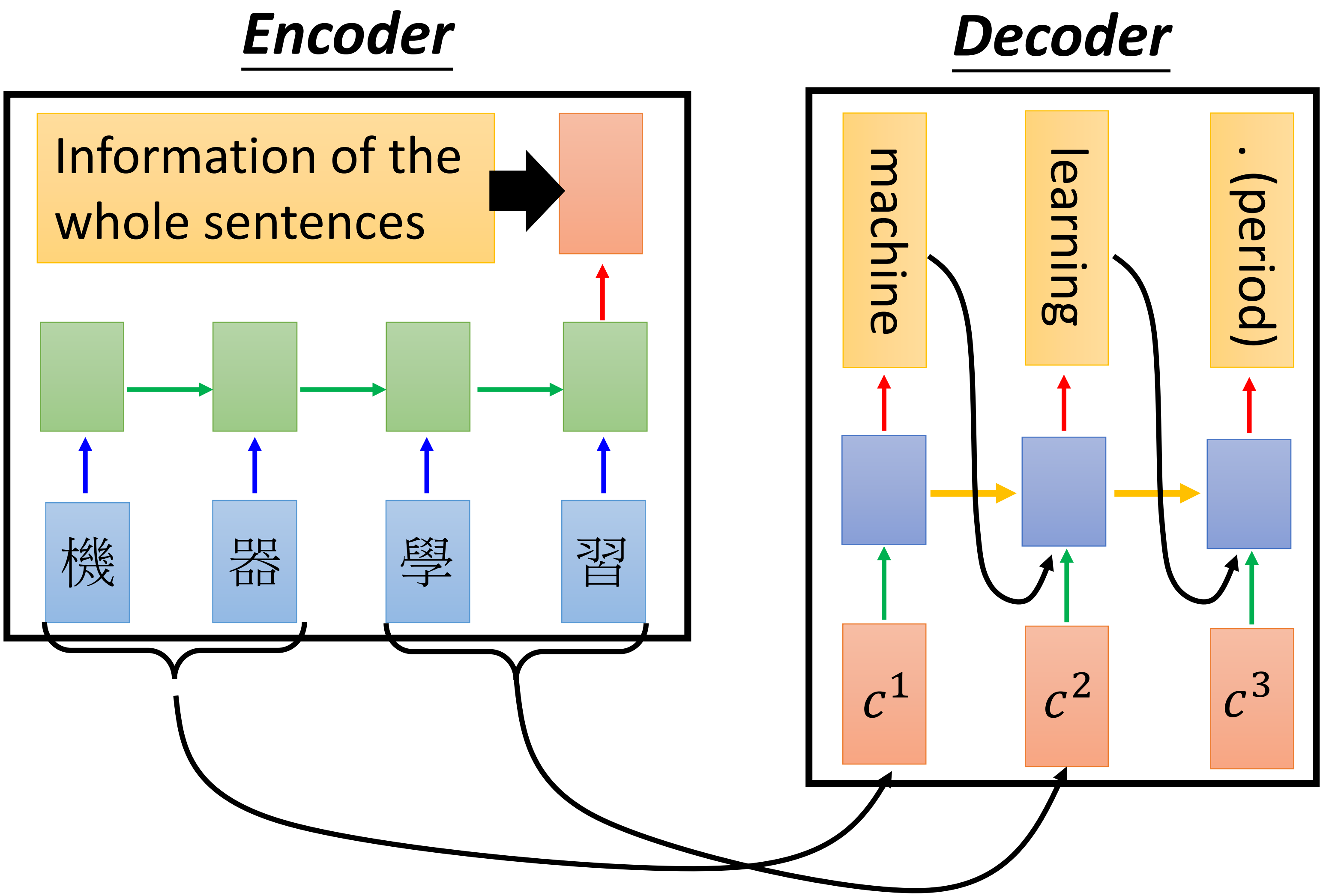
U: Hi

Serban, Iulian V., Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau, 2015 "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.

Attention

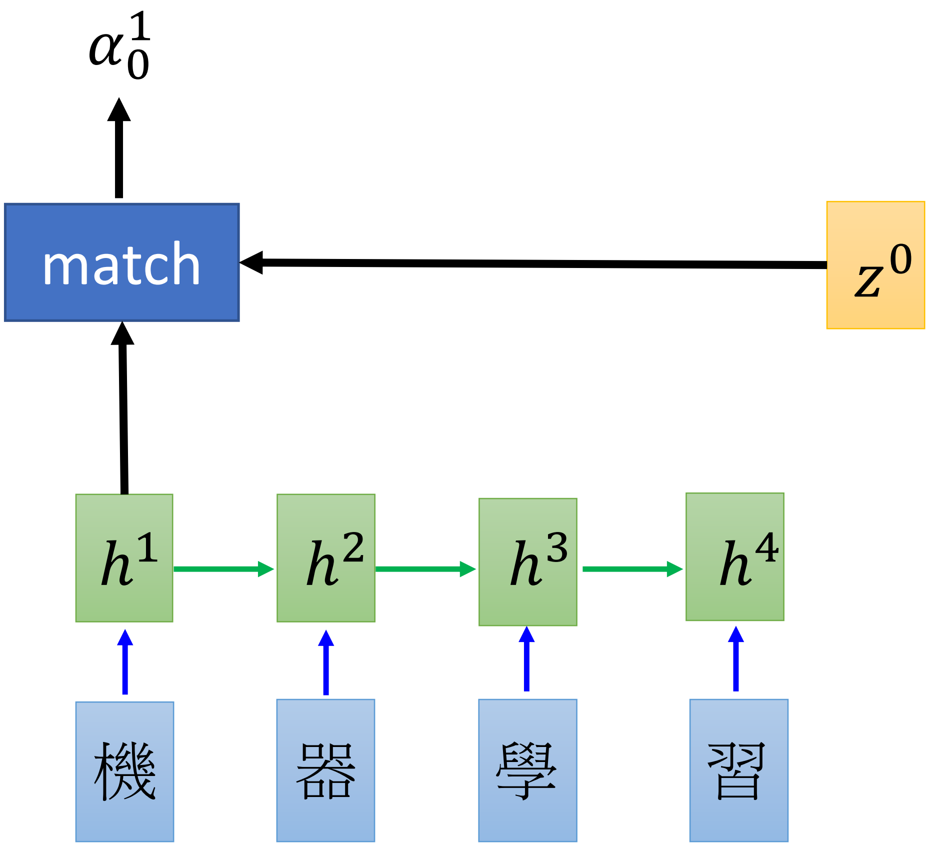
Dynamic Conditional Generation

Dynamic Conditional Generation

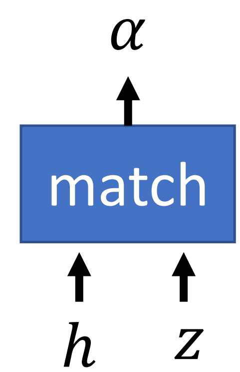


Machine Translation

- Attention-based model



Jointly learned with other part of the network



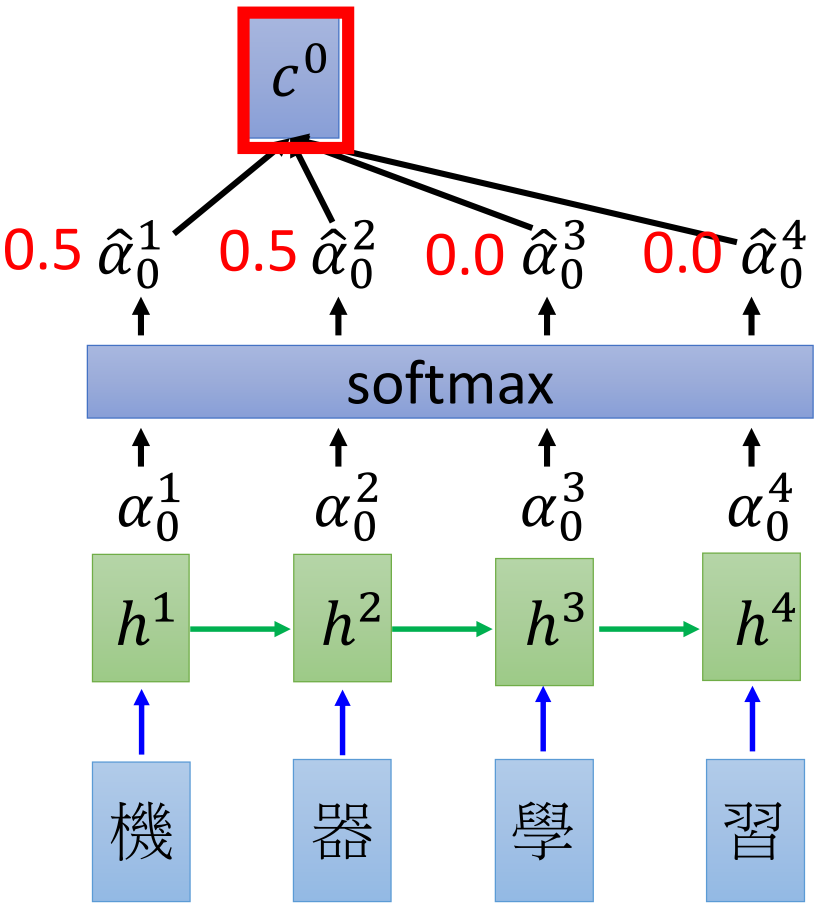
What is match ?

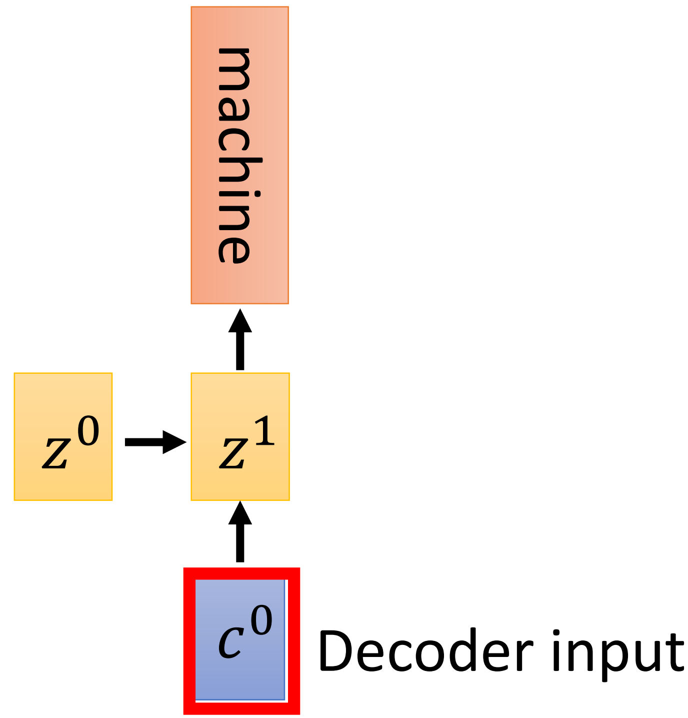
Design by yourself

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

Machine Translation

- Attention-based model



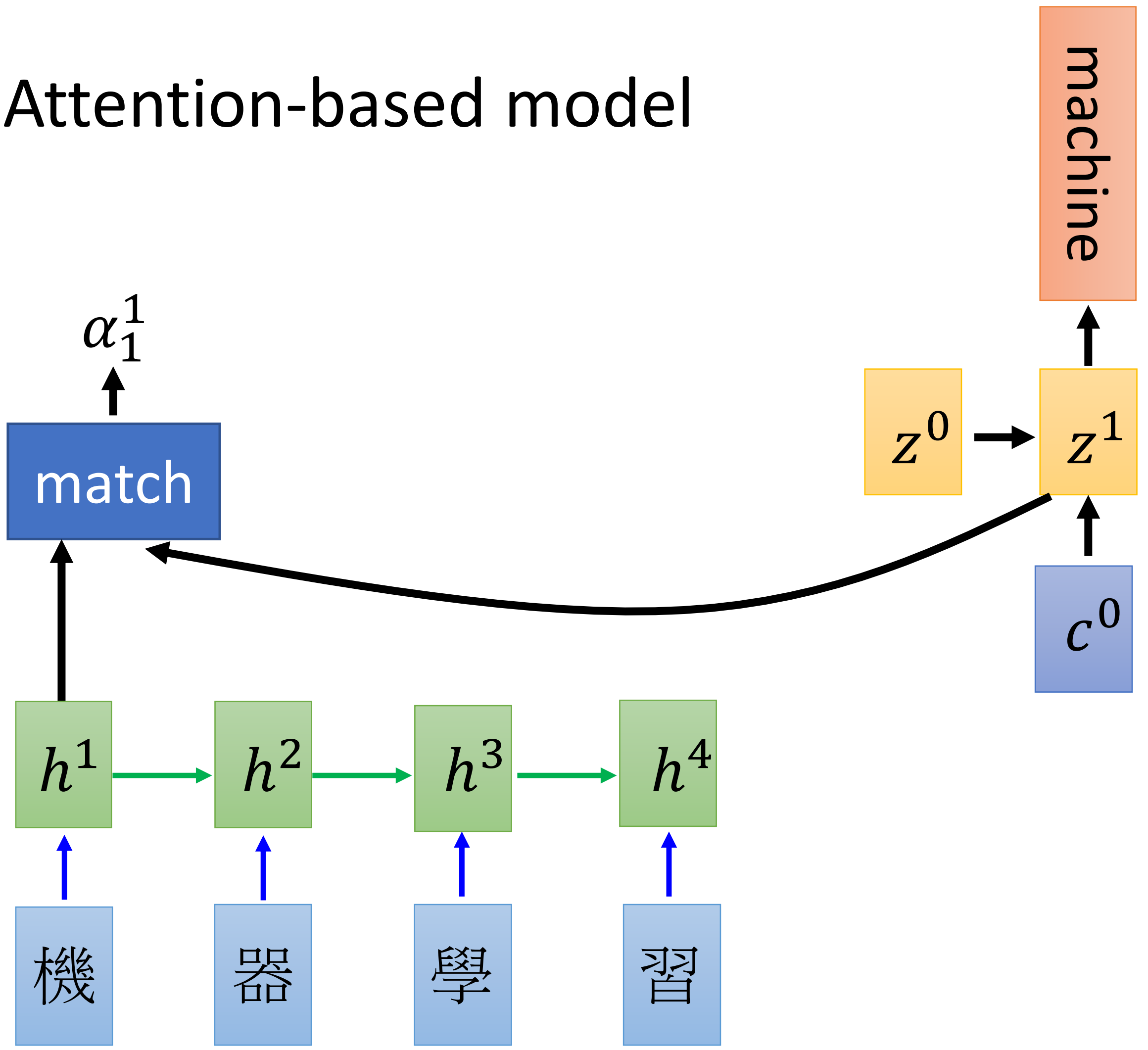


$$c^0 = \sum \hat{\alpha}_0^i h^i$$

$$= 0.5h^1 + 0.5h^2$$

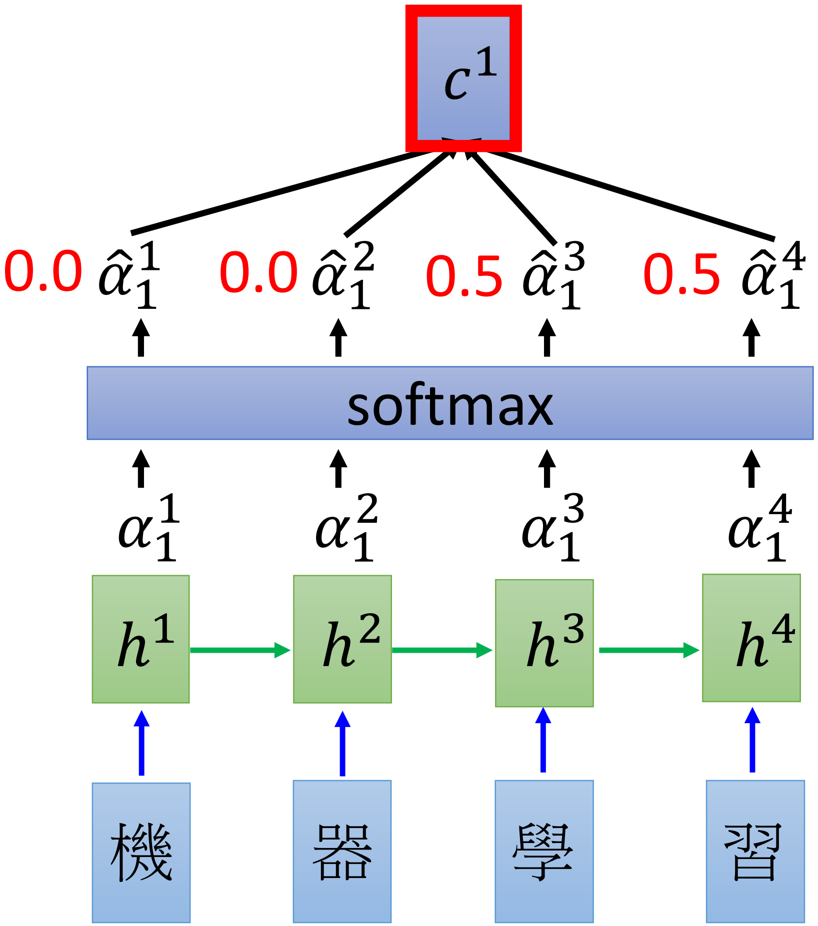
Machine Translation

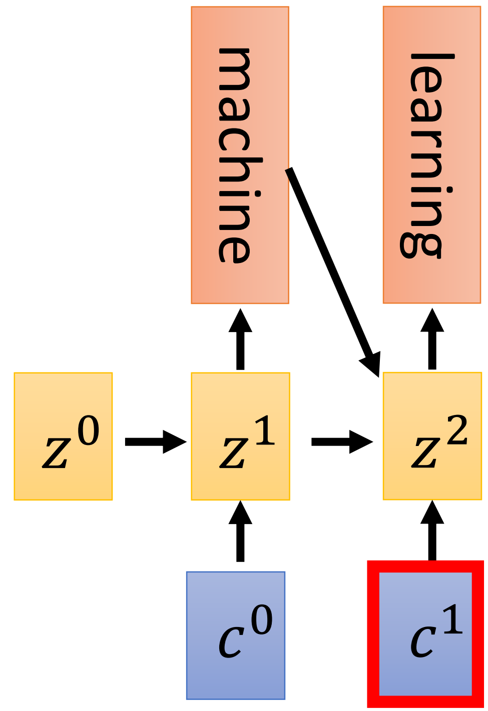
- Attention-based model



Machine Translation

- Attention-based model



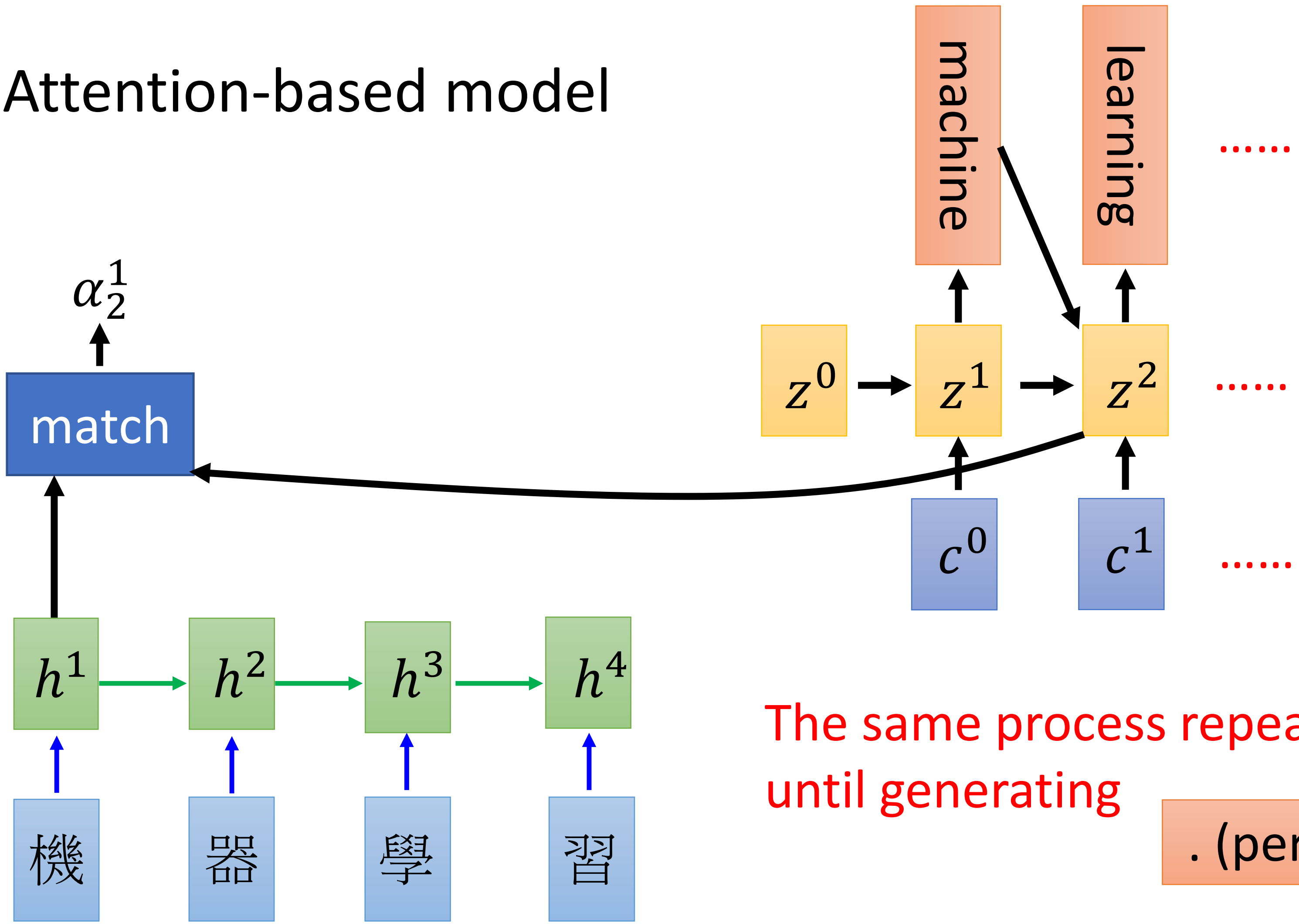


$$c^1 = \sum \hat{\alpha}_1^i h^i$$

$$= 0.5h^3 + 0.5h^4$$

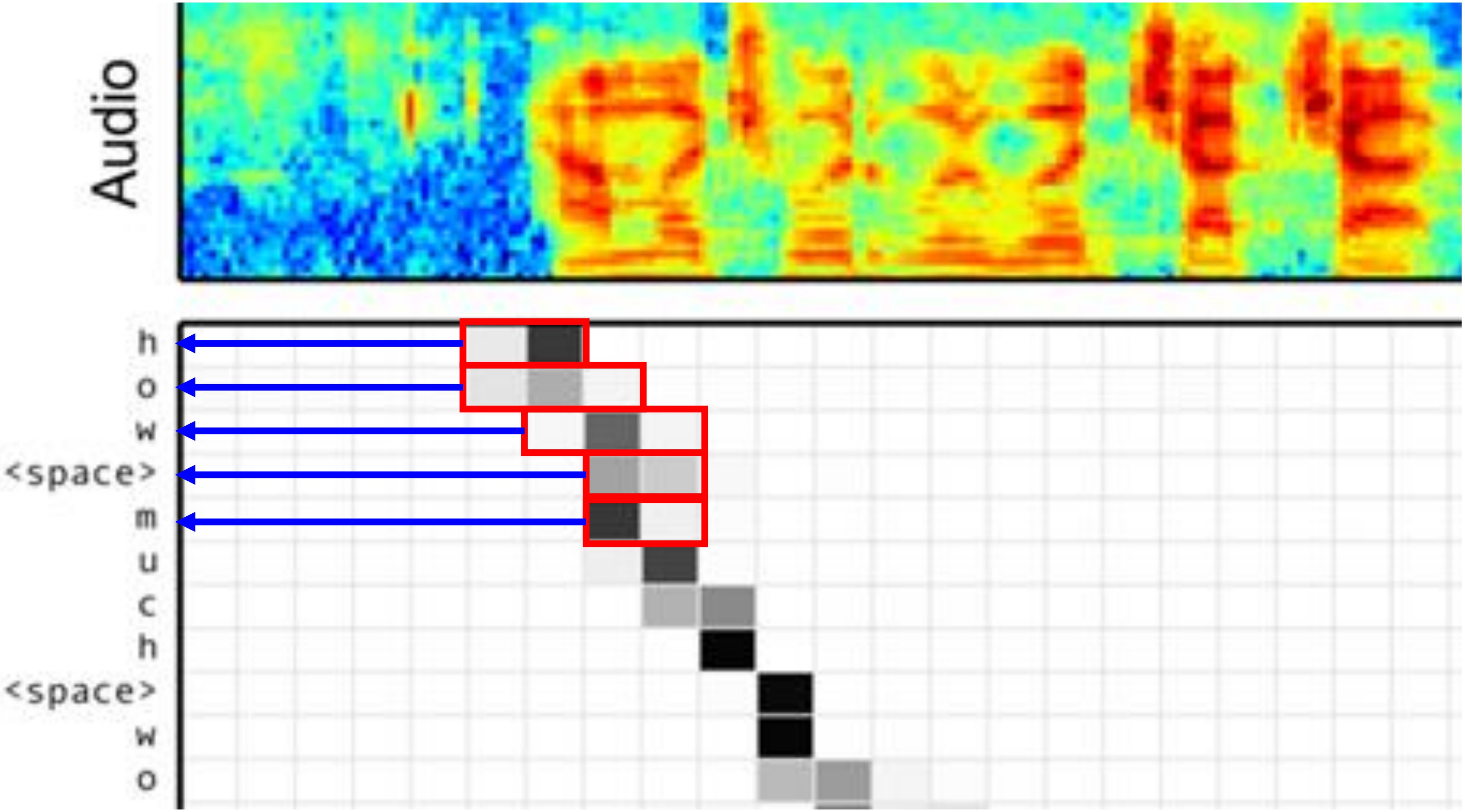
Machine Translation

- Attention-based model



The same process repeat until generating . (period)

Speech Recognition



Model	Clean WER	Noisy WER
CLDNN-HMM [22]	8.0	8.9
LAS	14.1	16.5
LAS + LM Rescoring	10.3	12.0

William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, “Listen, Attend and Spell”, ICASSP, 2016

Image Caption Generation

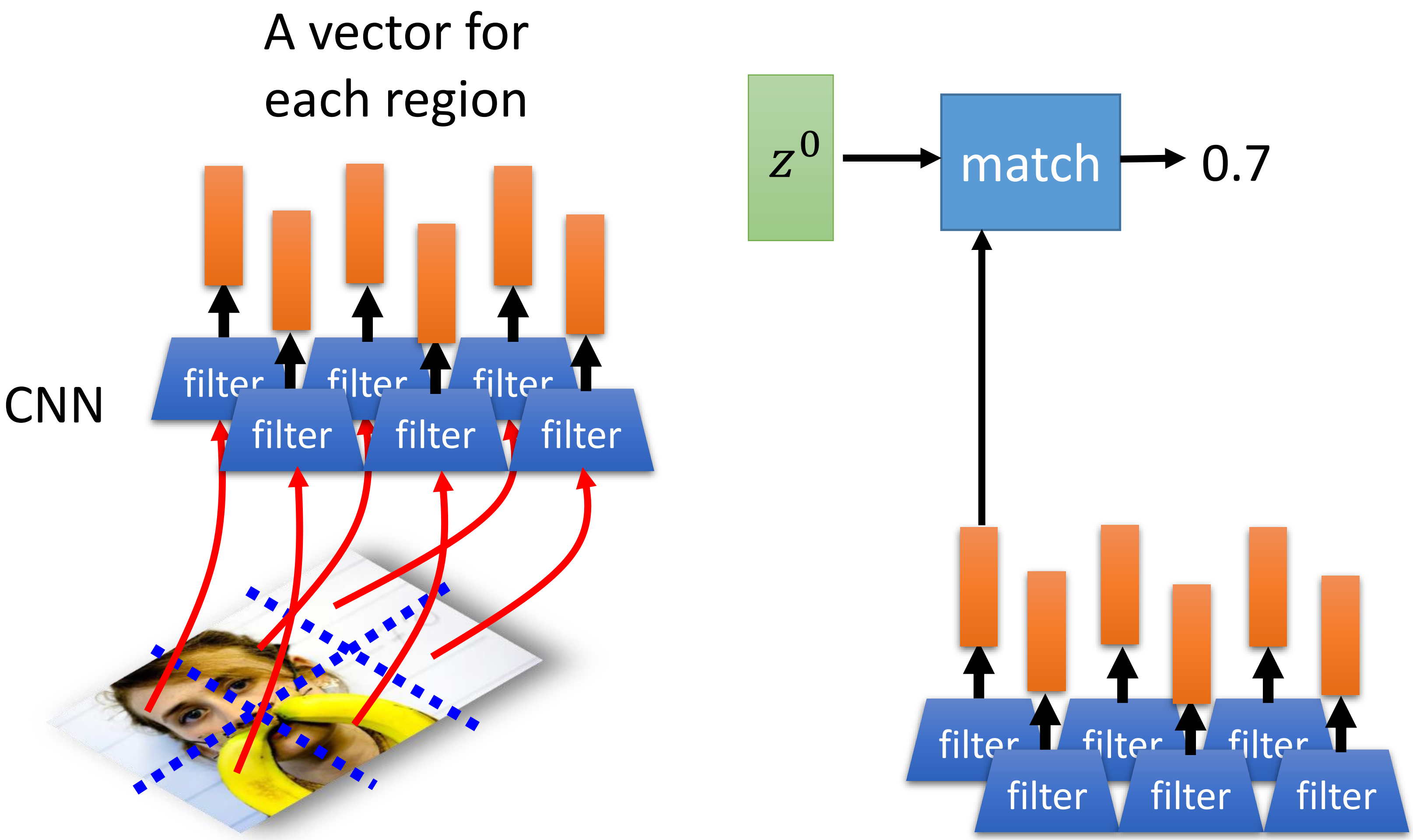


Image Caption Generation

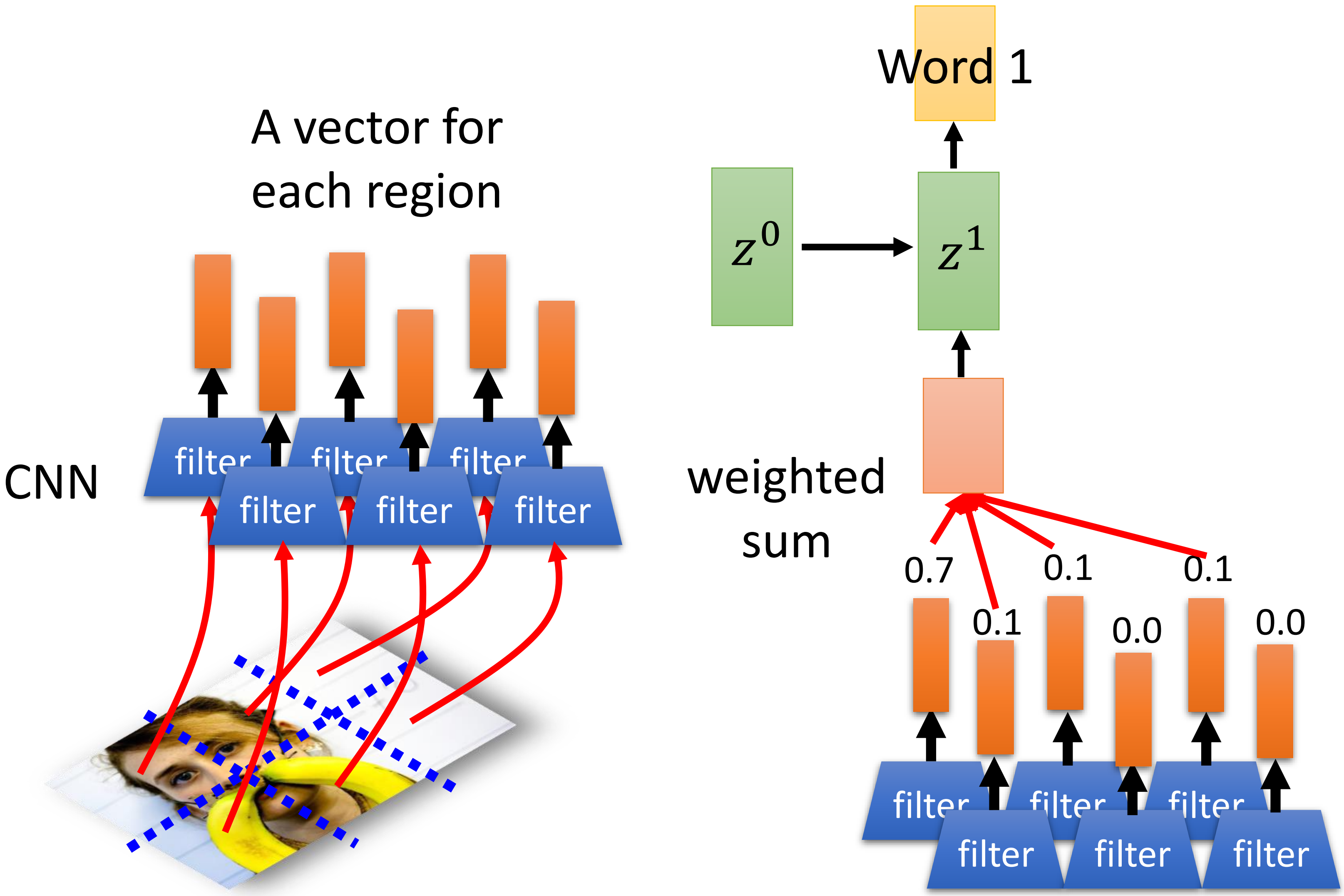


Image Caption Generation

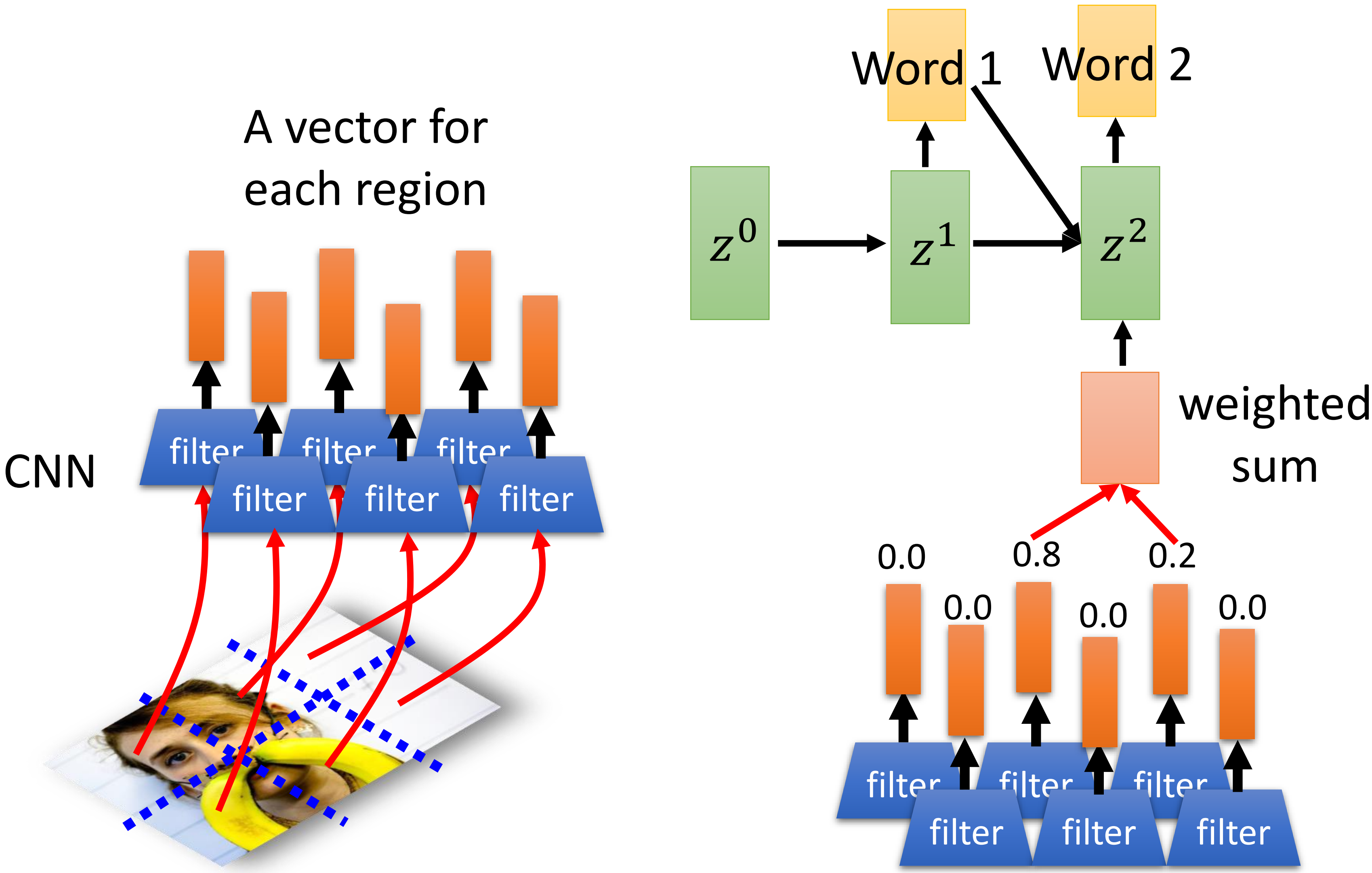


Image Caption Generation

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML, 2015

Image Caption Generation

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML, 2015

Ref: A man and a woman ride a motorcycle

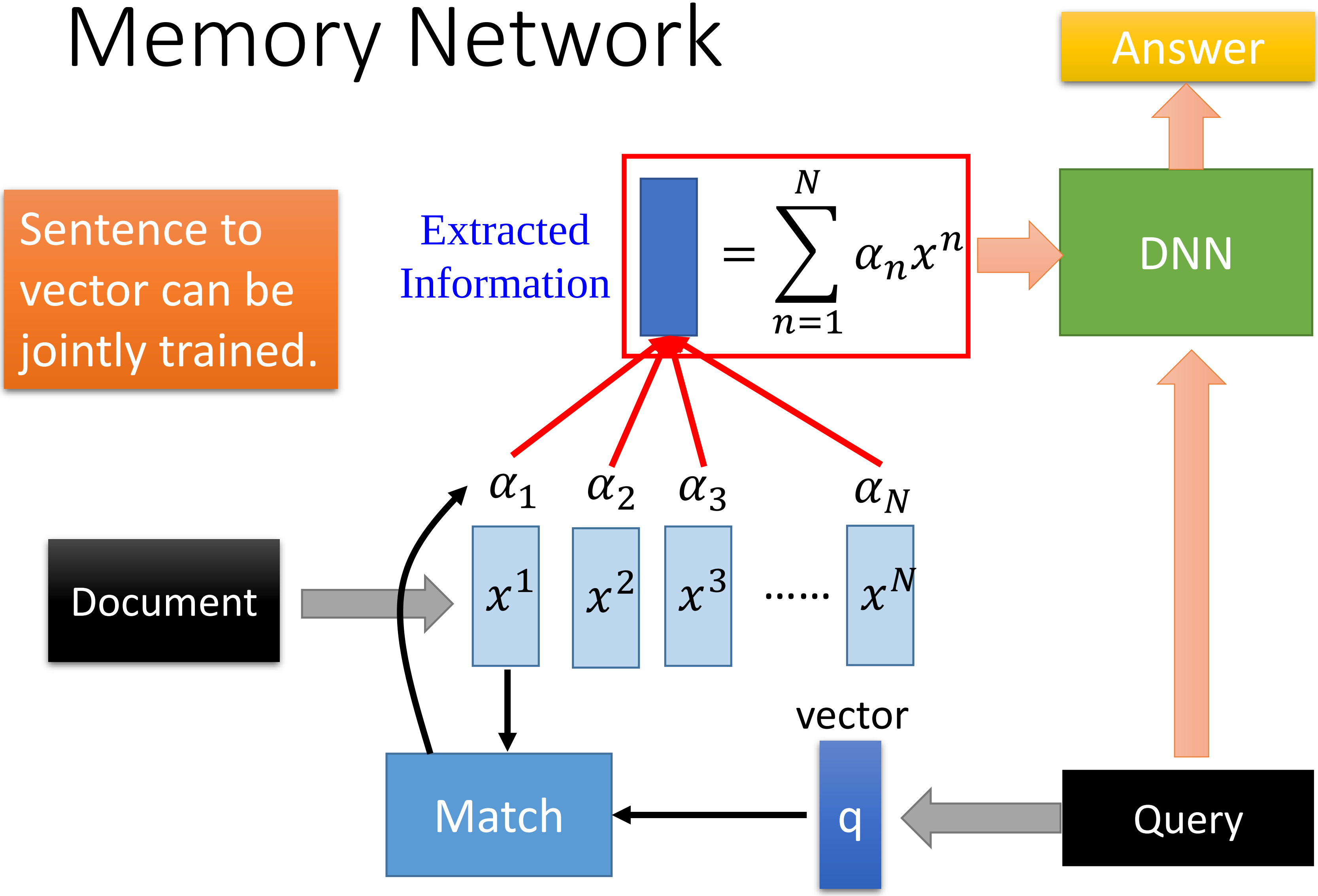
A **man** and a **woman** are **talking** on the **road**

Ref: A woman is frying food

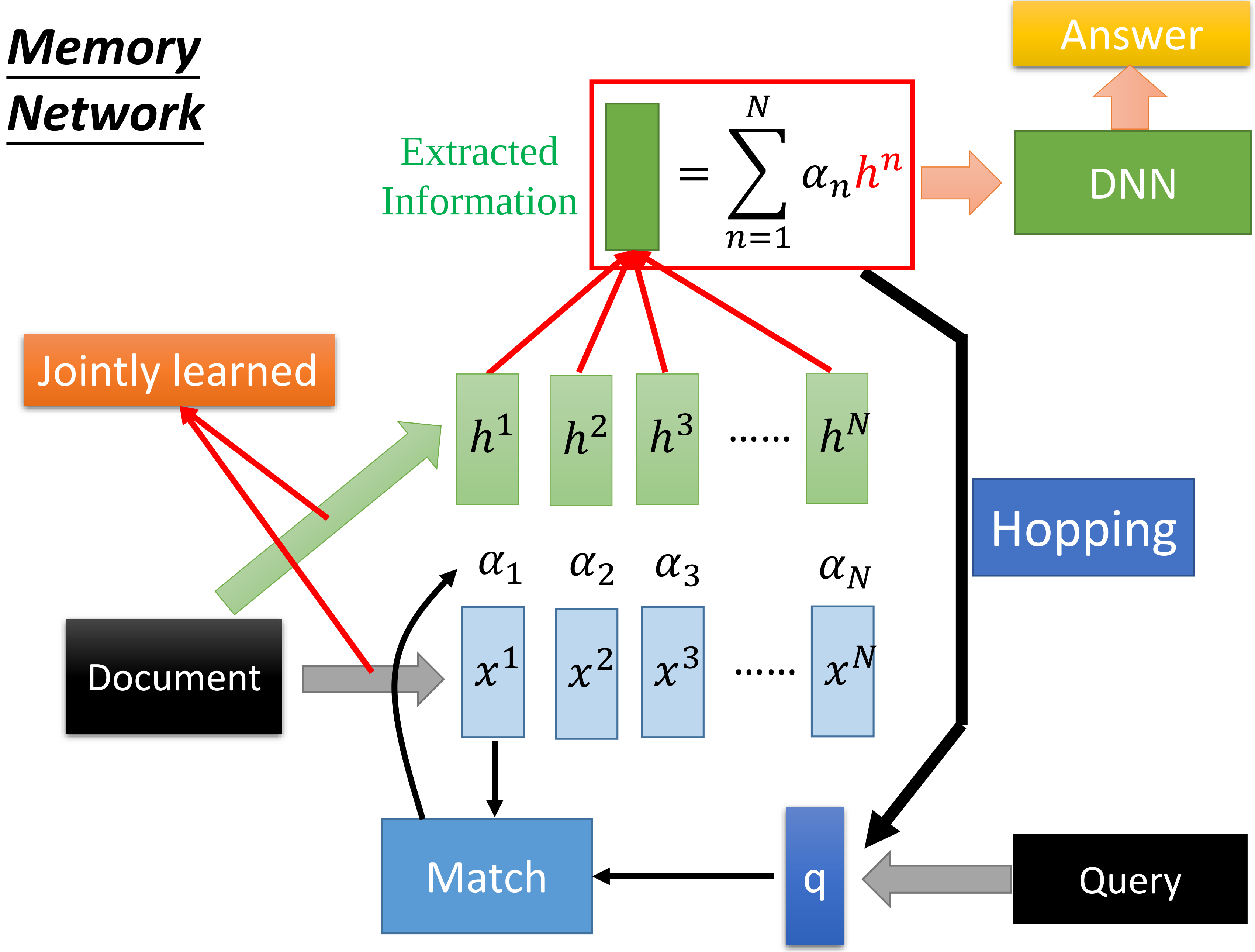
Someone is **frying** a **fish** in a **pot**

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville, "Describing Videos by Exploiting Temporal Structure", ICCV, 2015

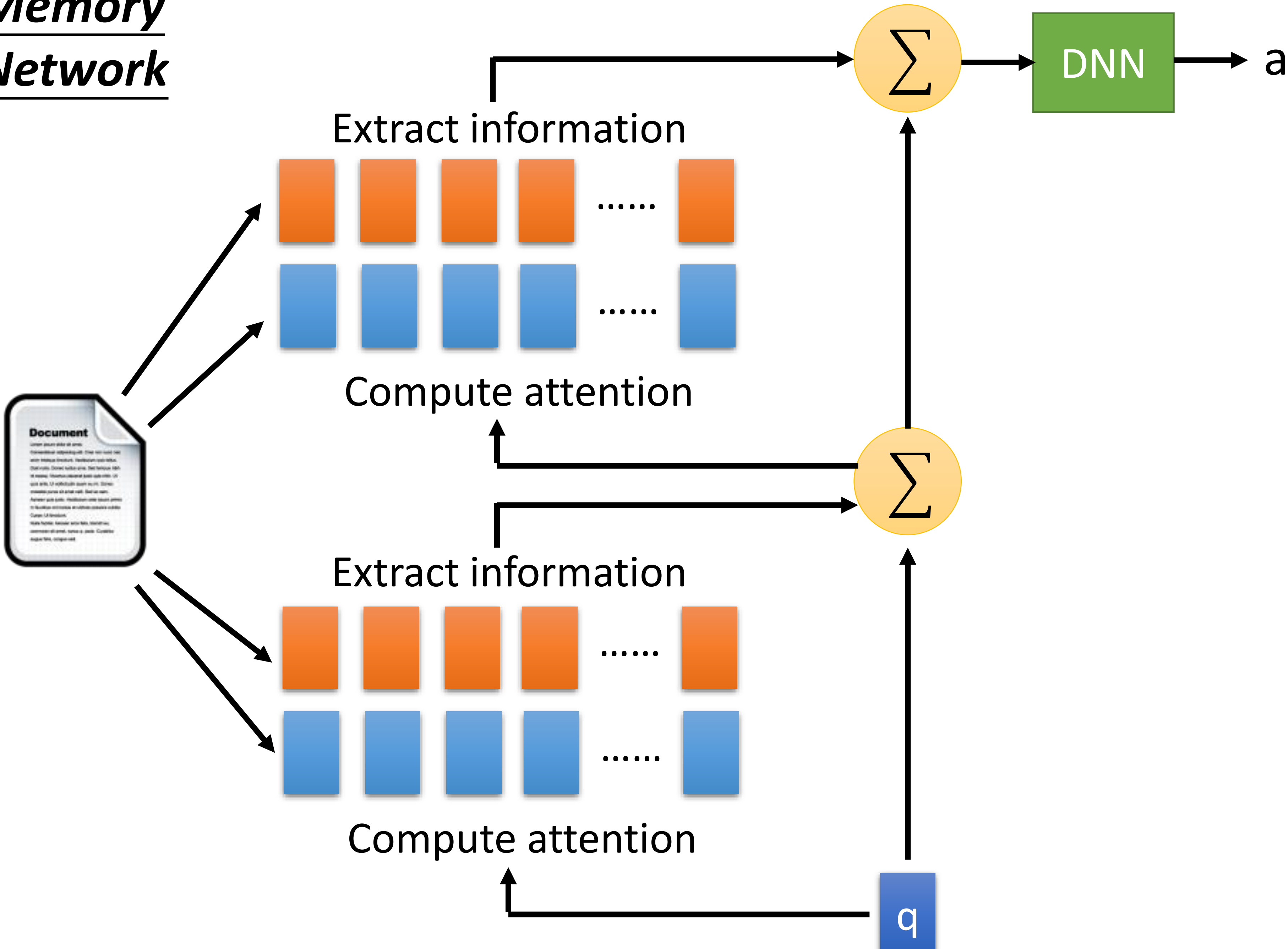
Memory Network

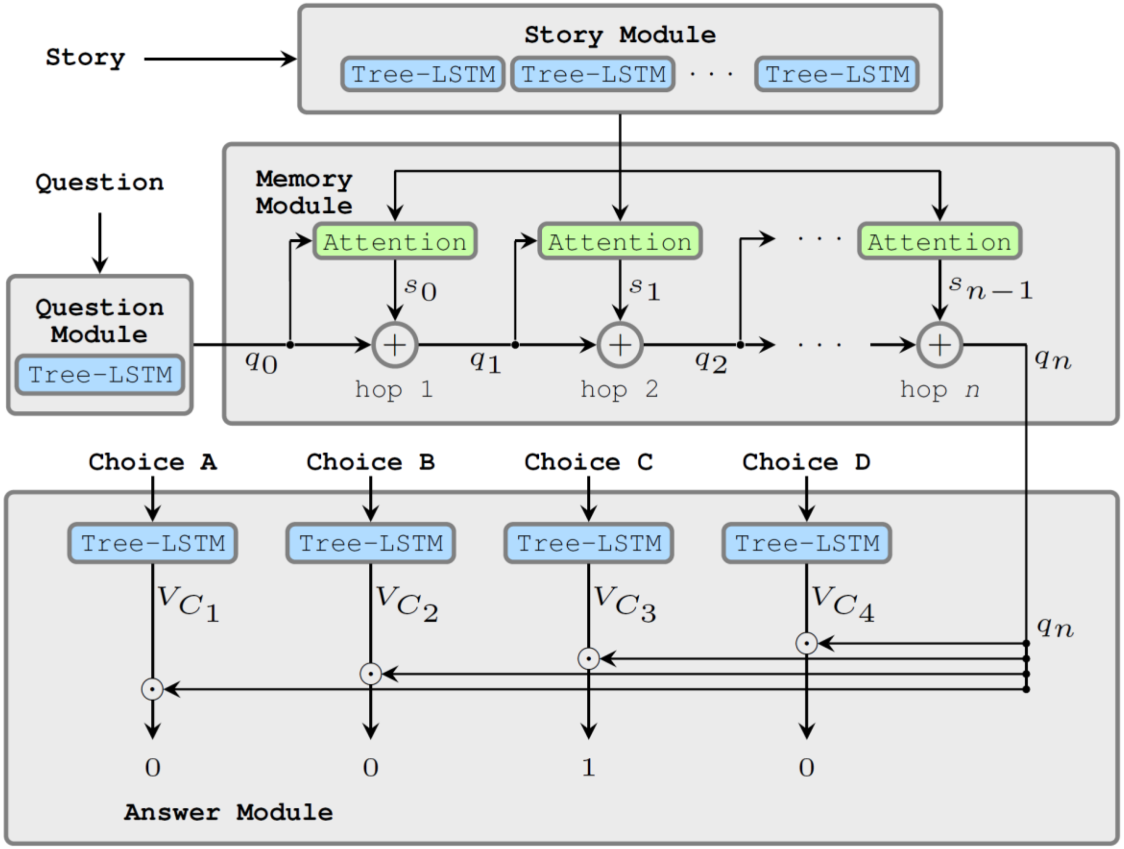


Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus, "End-To-End Memory Networks", NIPS, 2015

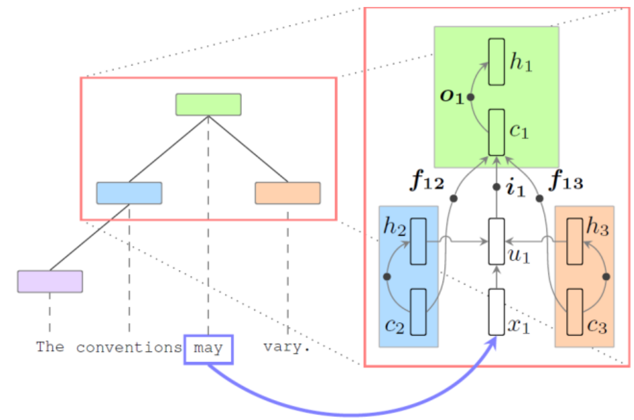

Memory Network
Extracted Information
$= \sum_{n=1}^{N} \alpha_n h^n$
Answer
DNN
Jointly learned
h^1 h^2 h^3 …… h^N
Hopping
α_1 α_2 α_3 α_N
Document
x^1 x^2 x^3 …… x^N
Match
q
Query

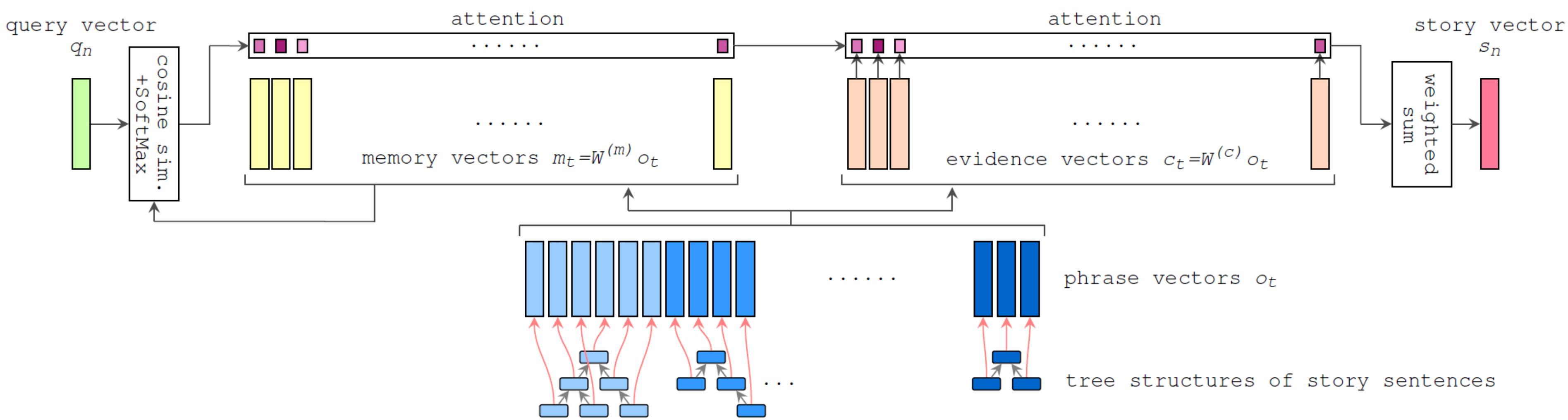
Memory Network





Wei Fang, Juei-Yang Hsu, Hung-yi Lee, Lin-Shan Lee, "Hierarchical Attention Model for Improved Machine Comprehension of Spoken Content", SLT, 2016



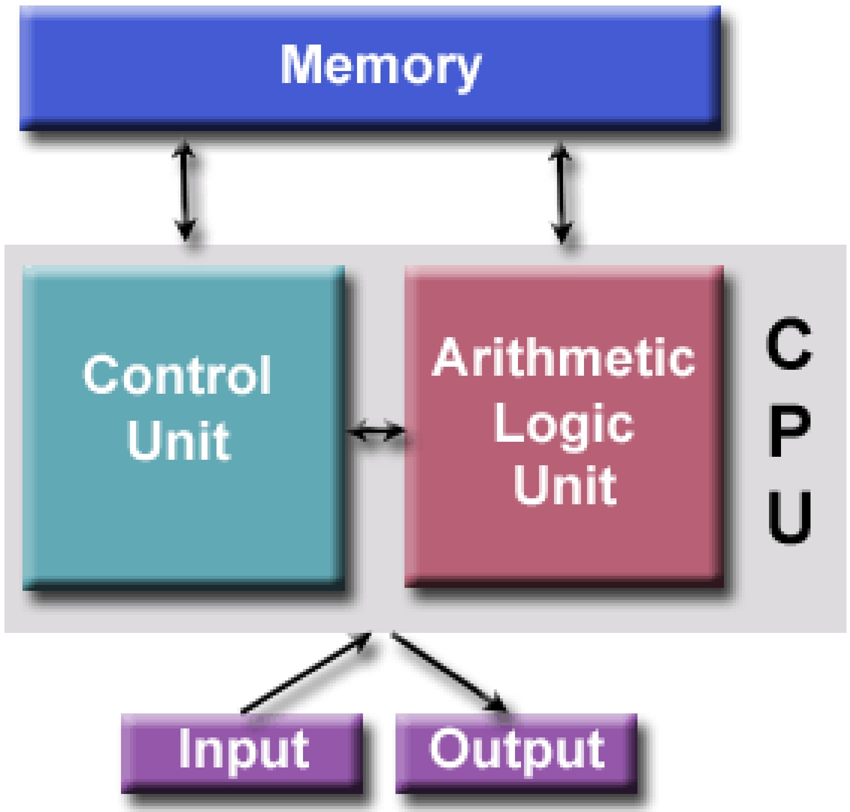


Neural Turing Machine

- von Neumann architecture

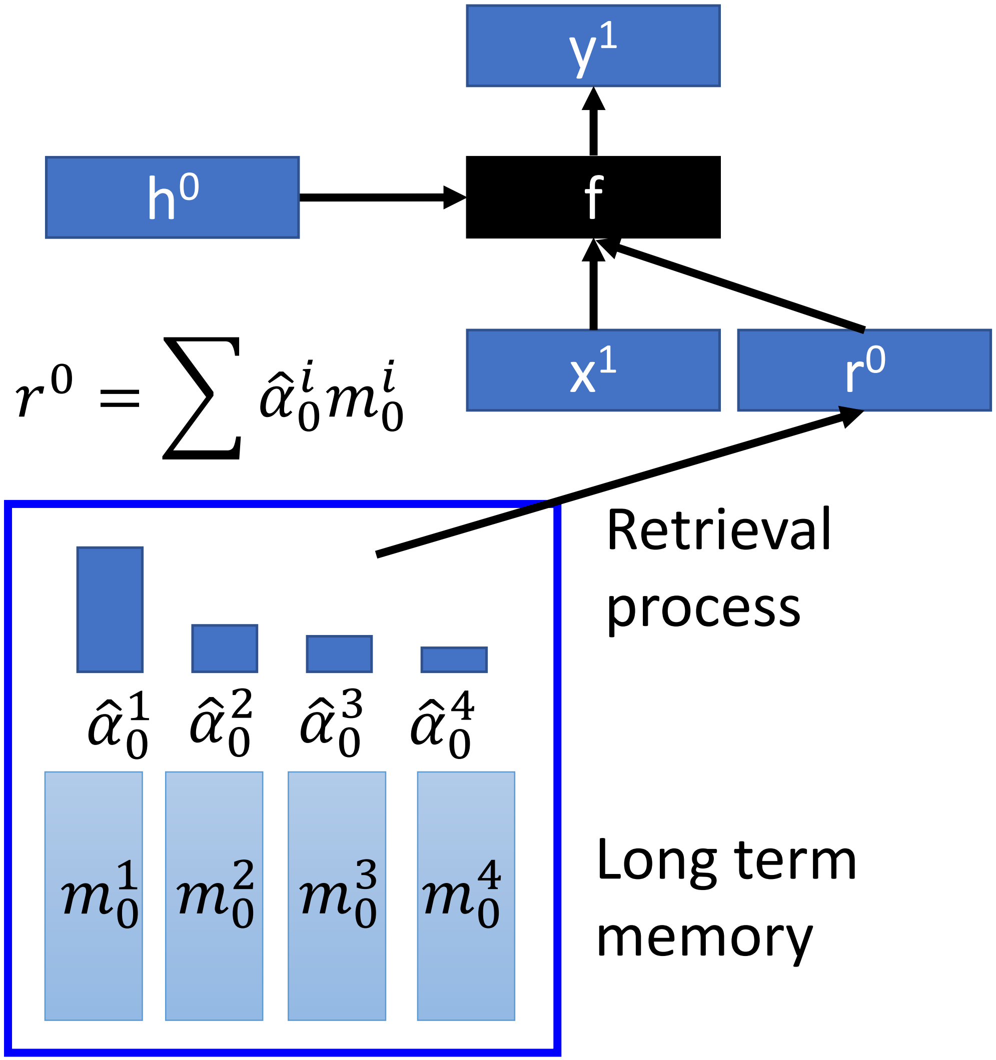
  Neural Turing Machine not only read from memory

  Also modify the memory through attention

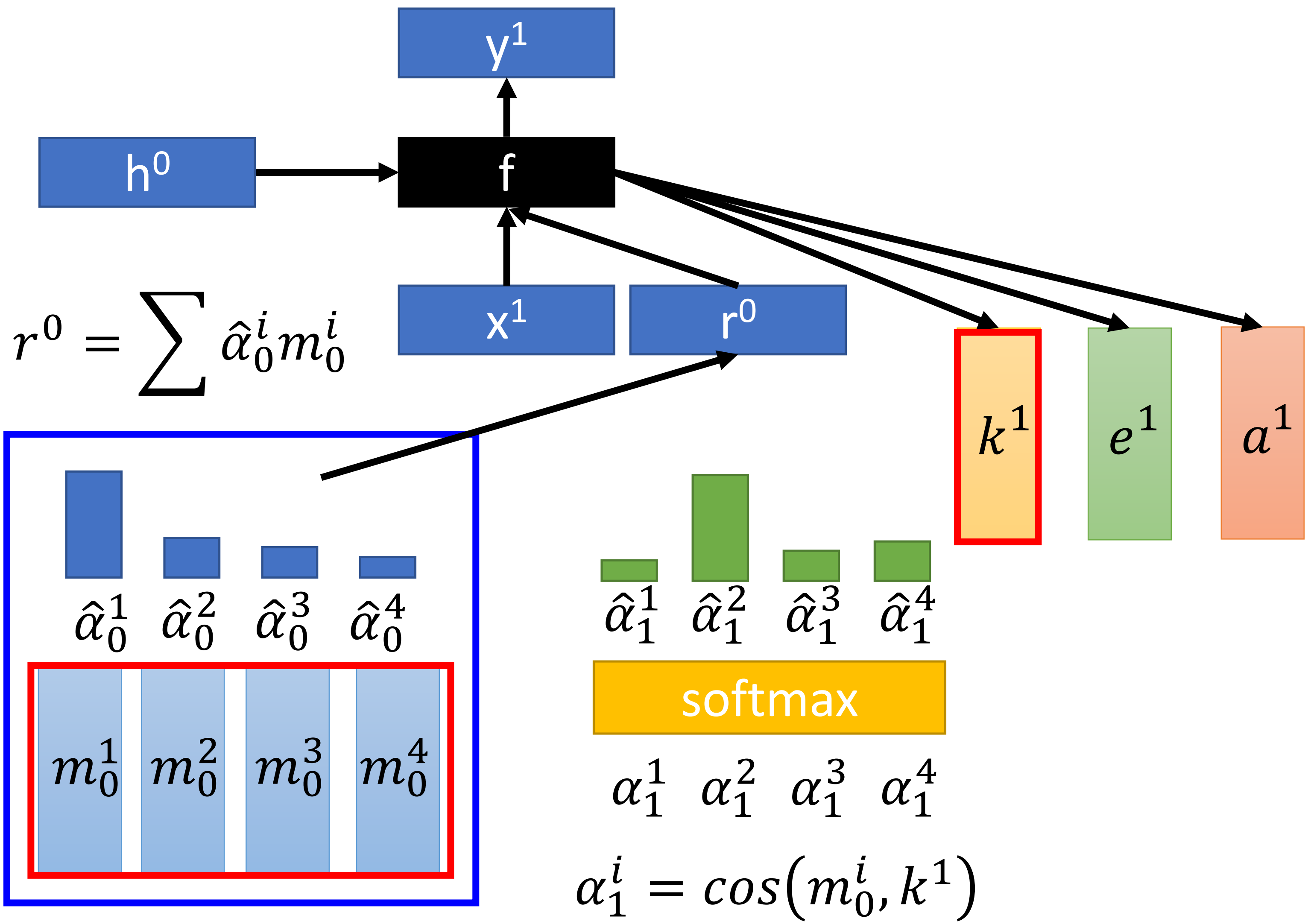


https://www.quora.com/How-does-the-Von-Neumann-architecture-provide-flexibility-for-program-development

Neural Turing Machine

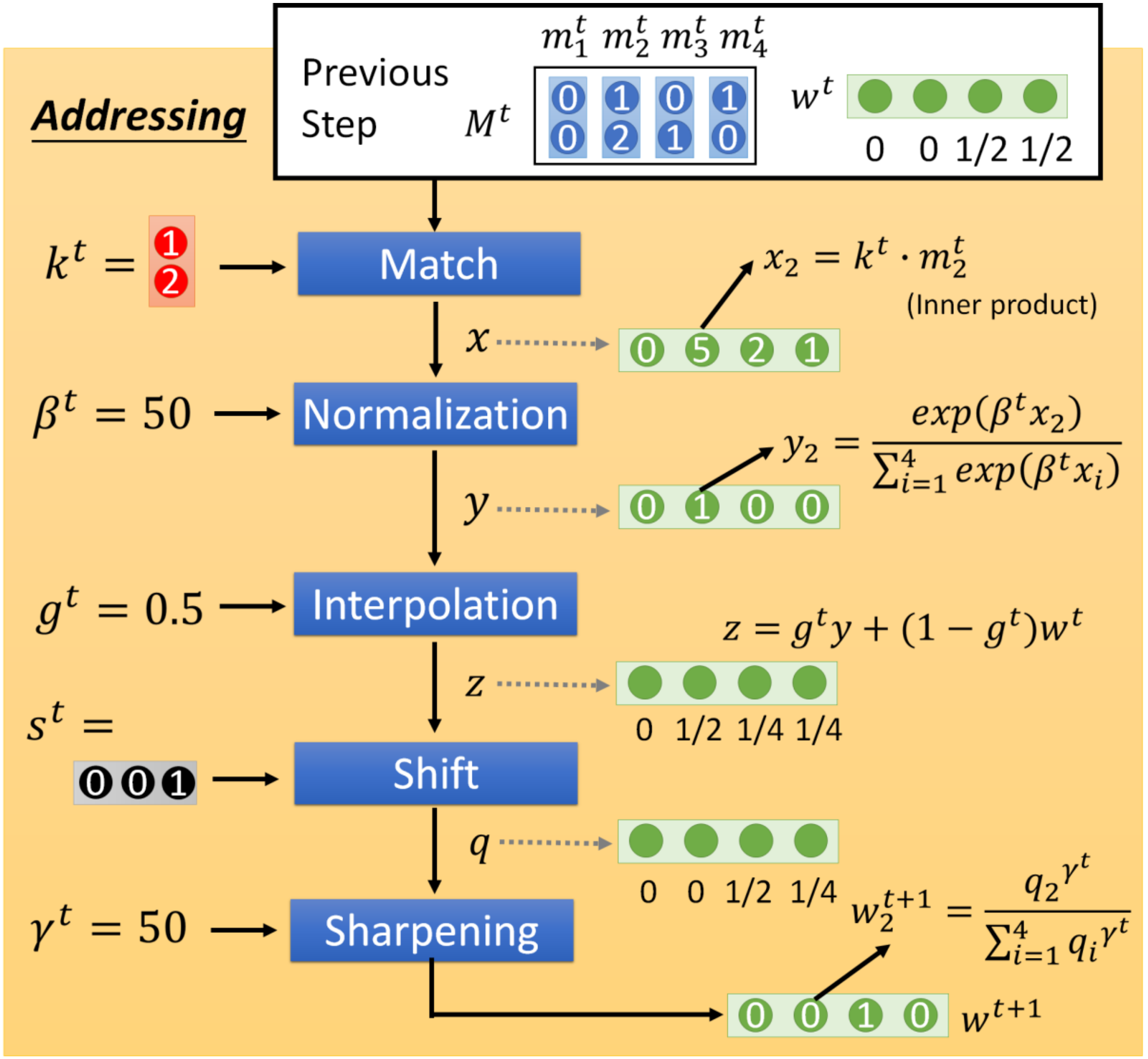


Neural Turing Machine



Neural Turing Machine

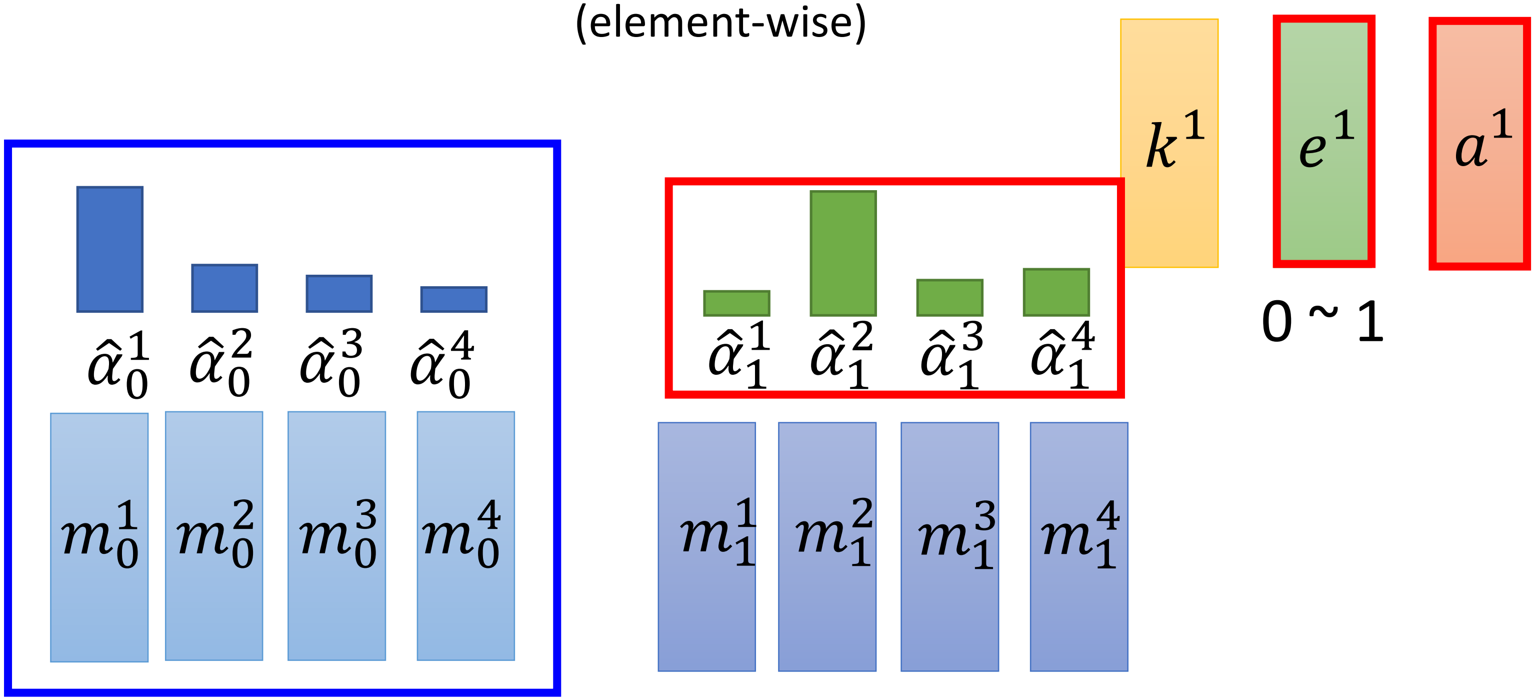
- Real version



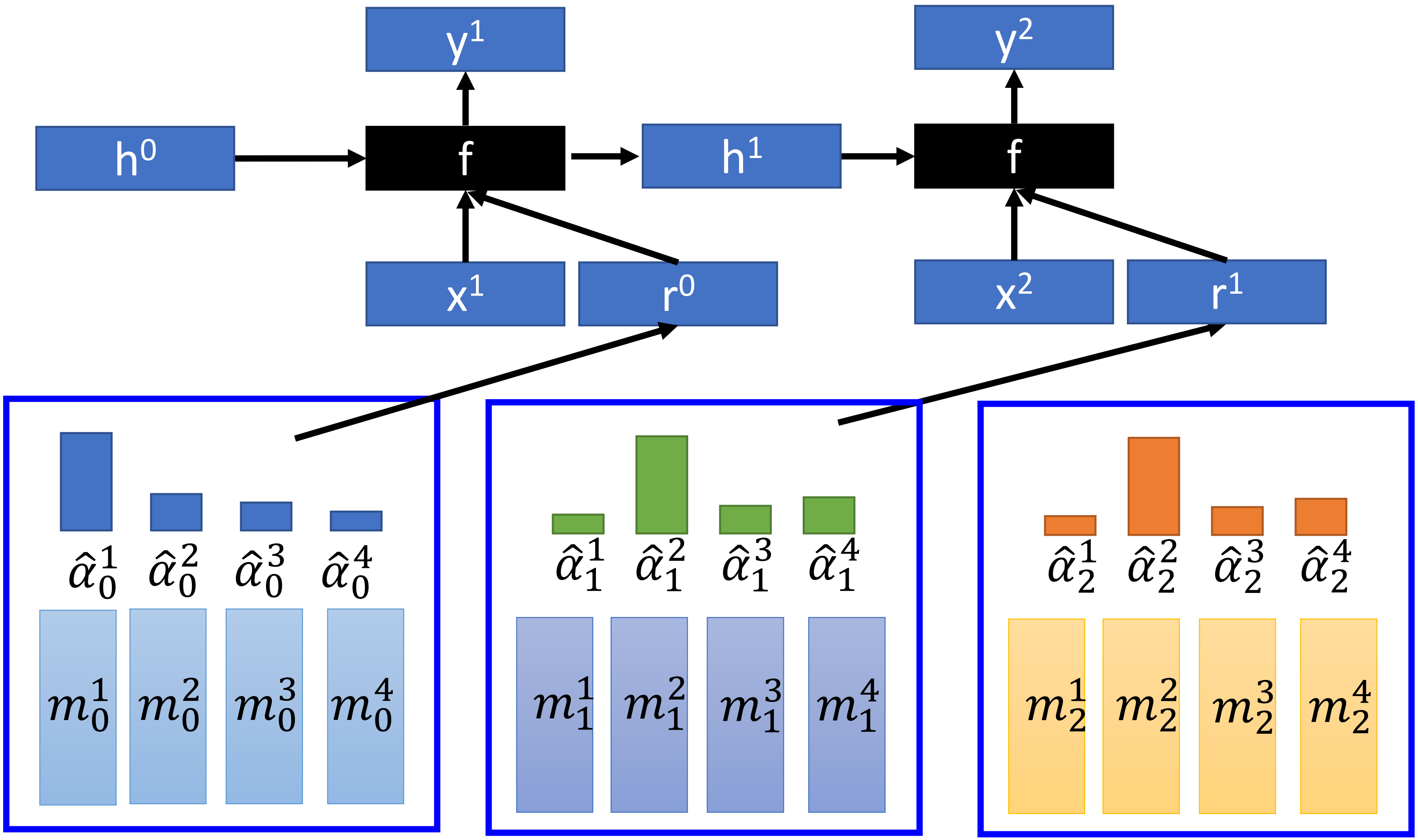
Neural Turing Machine

$$m_1^i = m_0^i - \hat{\alpha}_1^i \, e^1 \odot m_0^i + \hat{\alpha}_1^i \, a^1$$

(element-wise)



Neural Turing Machine



Tips for Generation

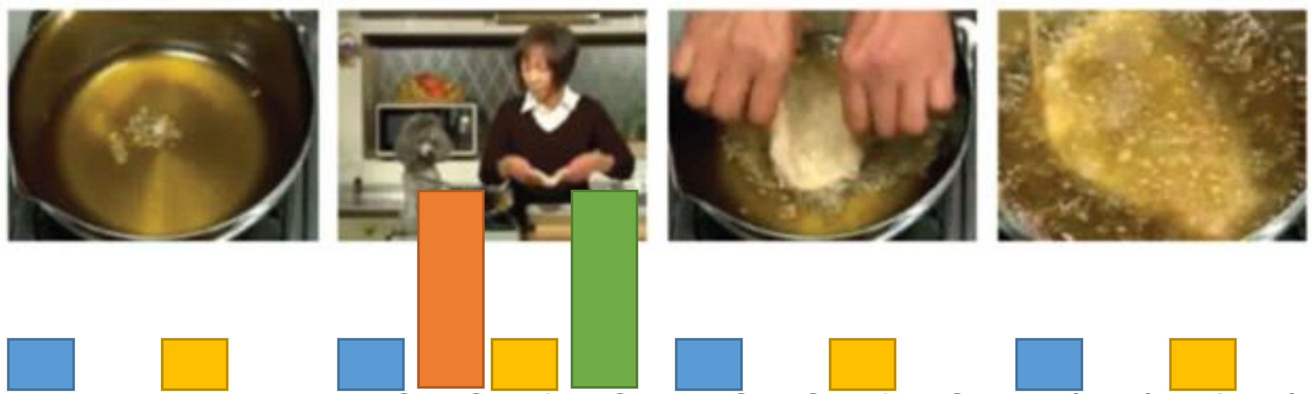
Attention

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML, 2015

component

α_t^i

time



Bad Attention

$\alpha_1^1 \alpha_2^1 \alpha_3^1 \alpha_4^1$ $\alpha_1^2 \alpha_2^2 \alpha_3^2 \alpha_4^2$ $\alpha_1^3 \alpha_2^3 \alpha_3^3 \alpha_4^3$ $\alpha_1^4 \alpha_2^4 \alpha_3^4 \alpha_4^4$

w_1 w_2 (woman) w_3 w_4 (woman) …… no cooking

Good Attention: each input component has approximately the same attention weight

E.g. Regularization term: $\sum_i \left(\tau - \sum_t \alpha_t^i\right)^2$

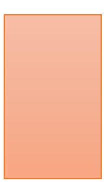
For each component Over the generation

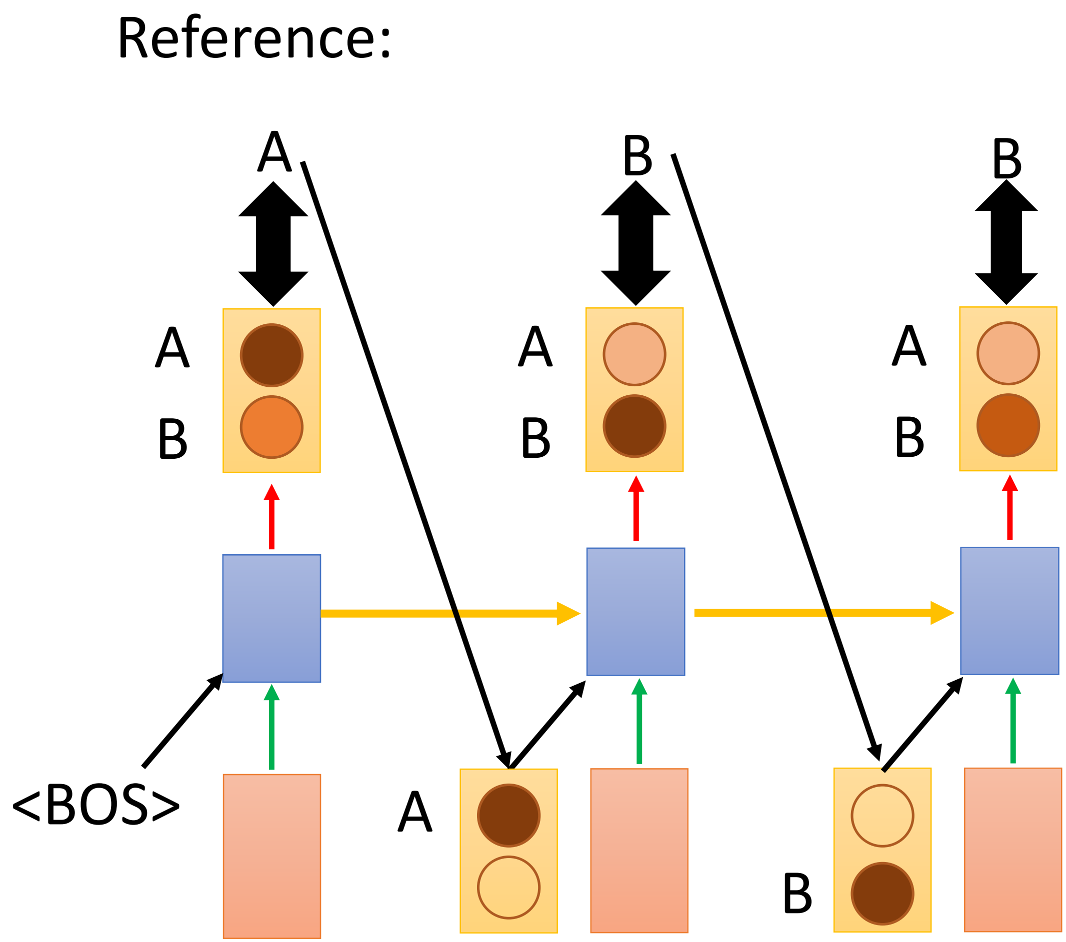
Mismatch between Train and Test

- ***Training***

$$C = \sum_t C_t$$

Minimizing cross-entropy of each component

 : condition



Mismatch between Train and Test

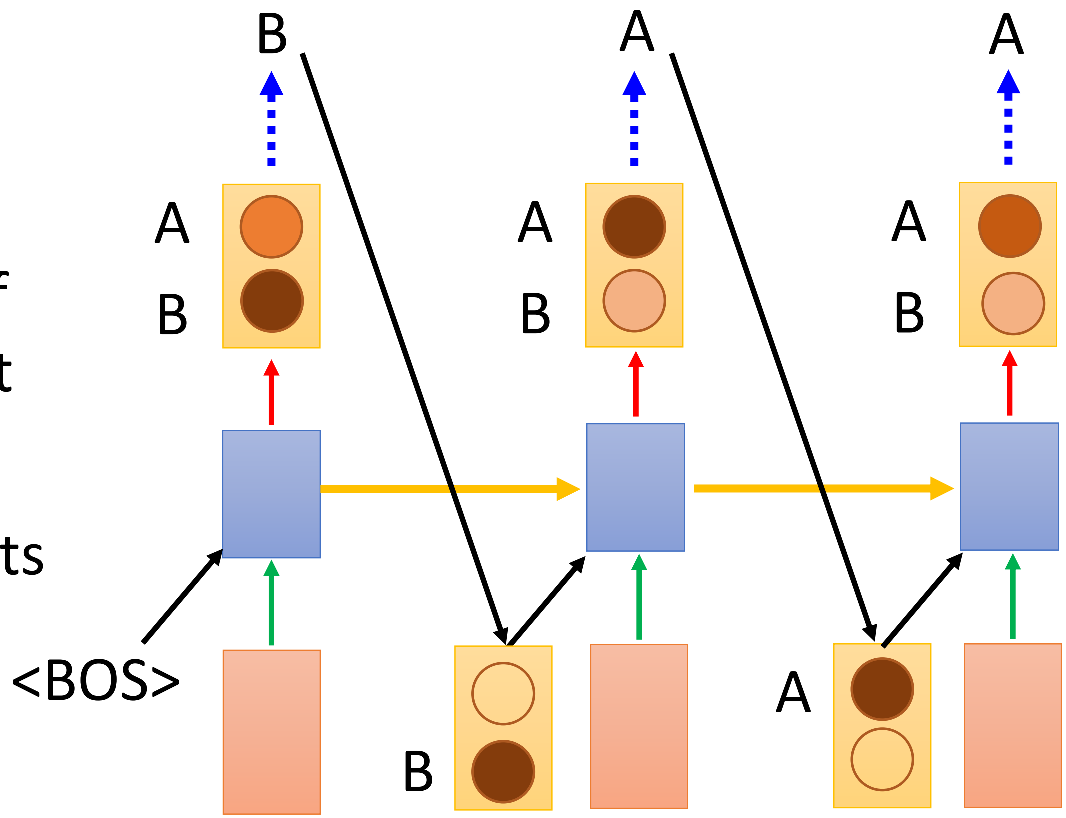
- ***Generation***

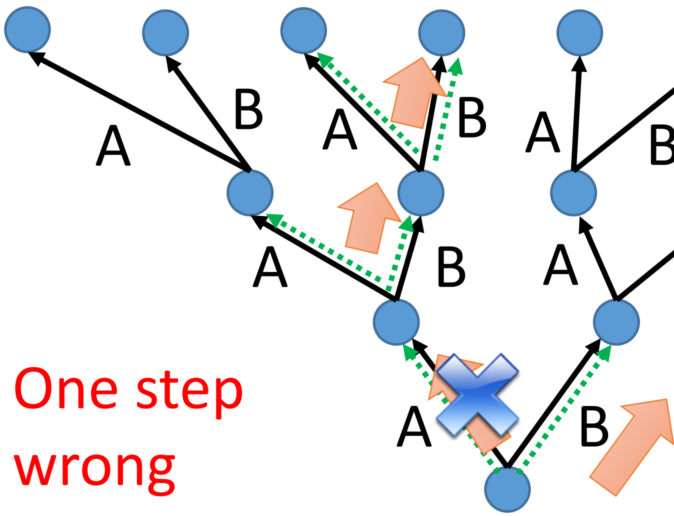
We do not know the reference

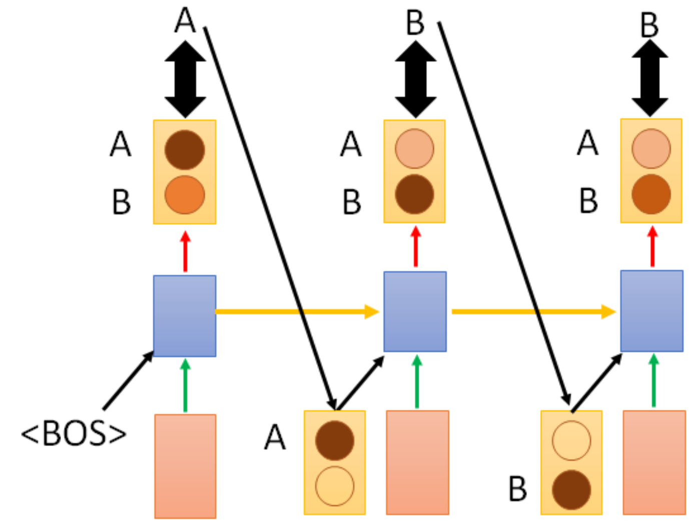
Testing: Output of model is the input of the next step.

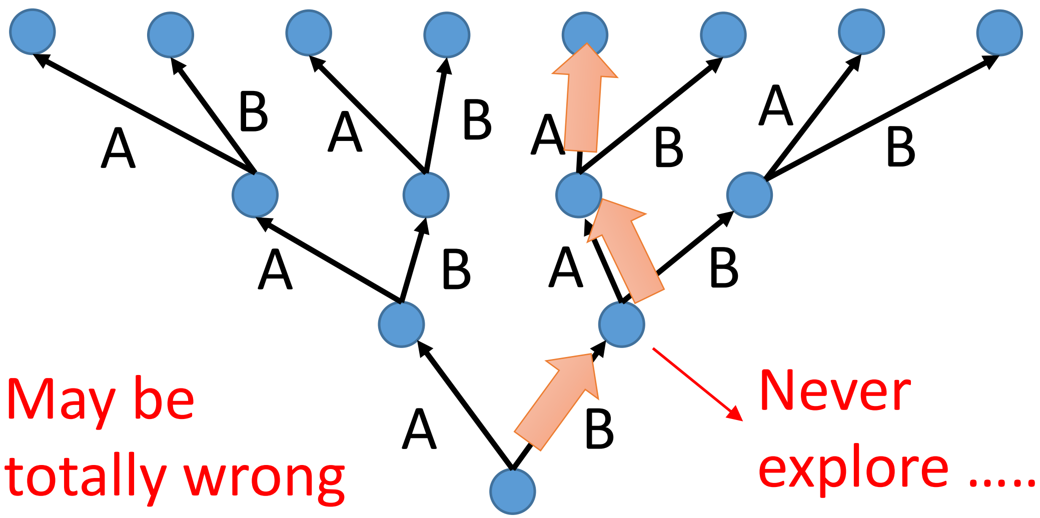
Training: the inputs are reference.

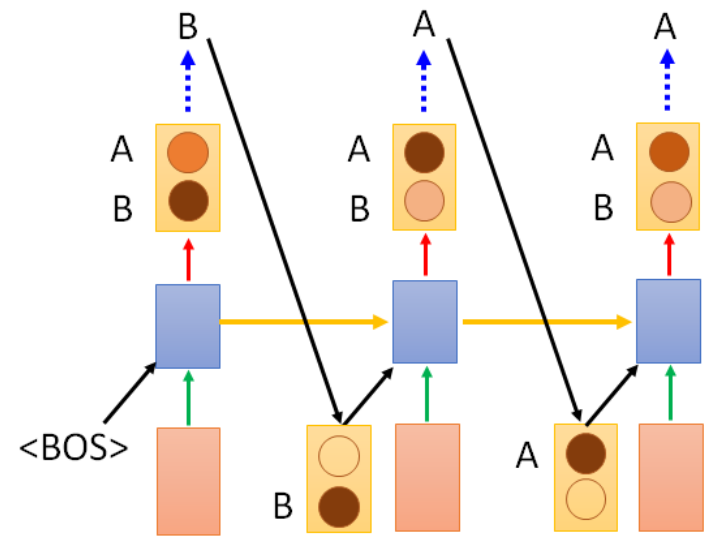
Exposure Bias











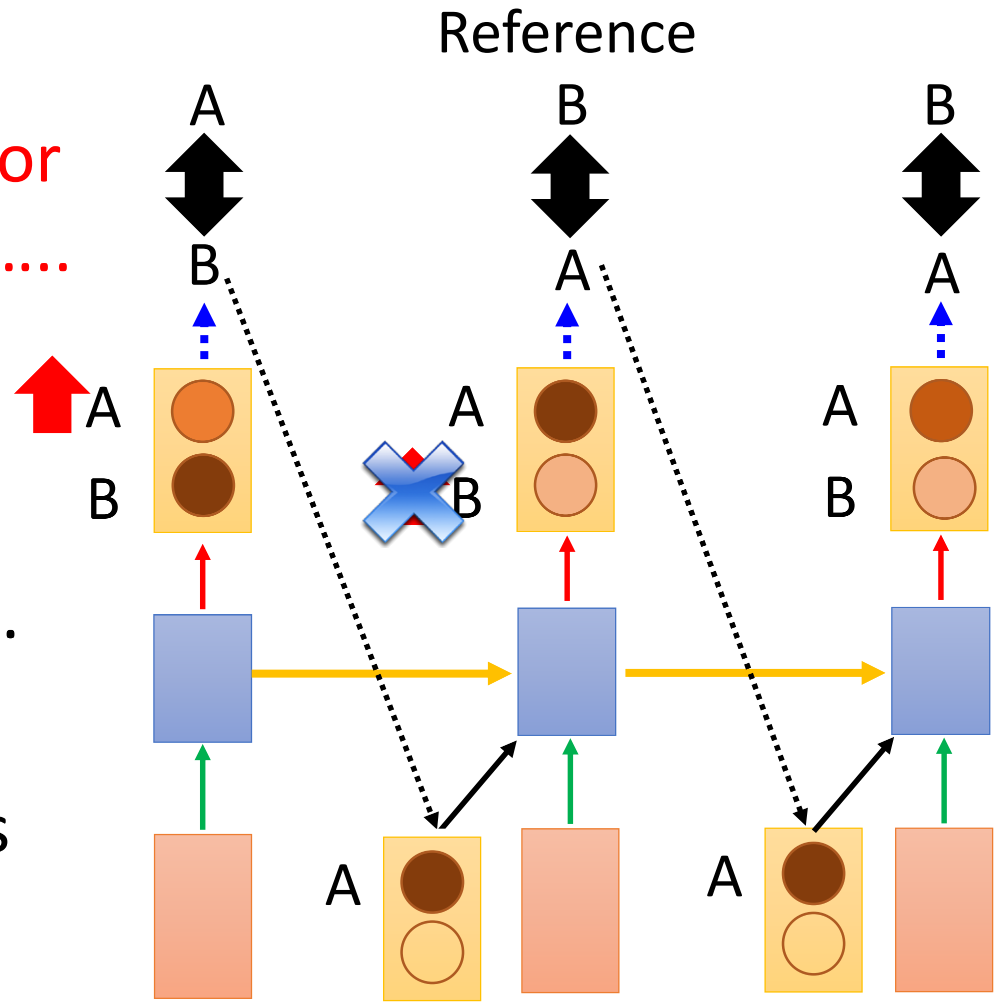
一步錯，步步錯

Modifying Training Process?

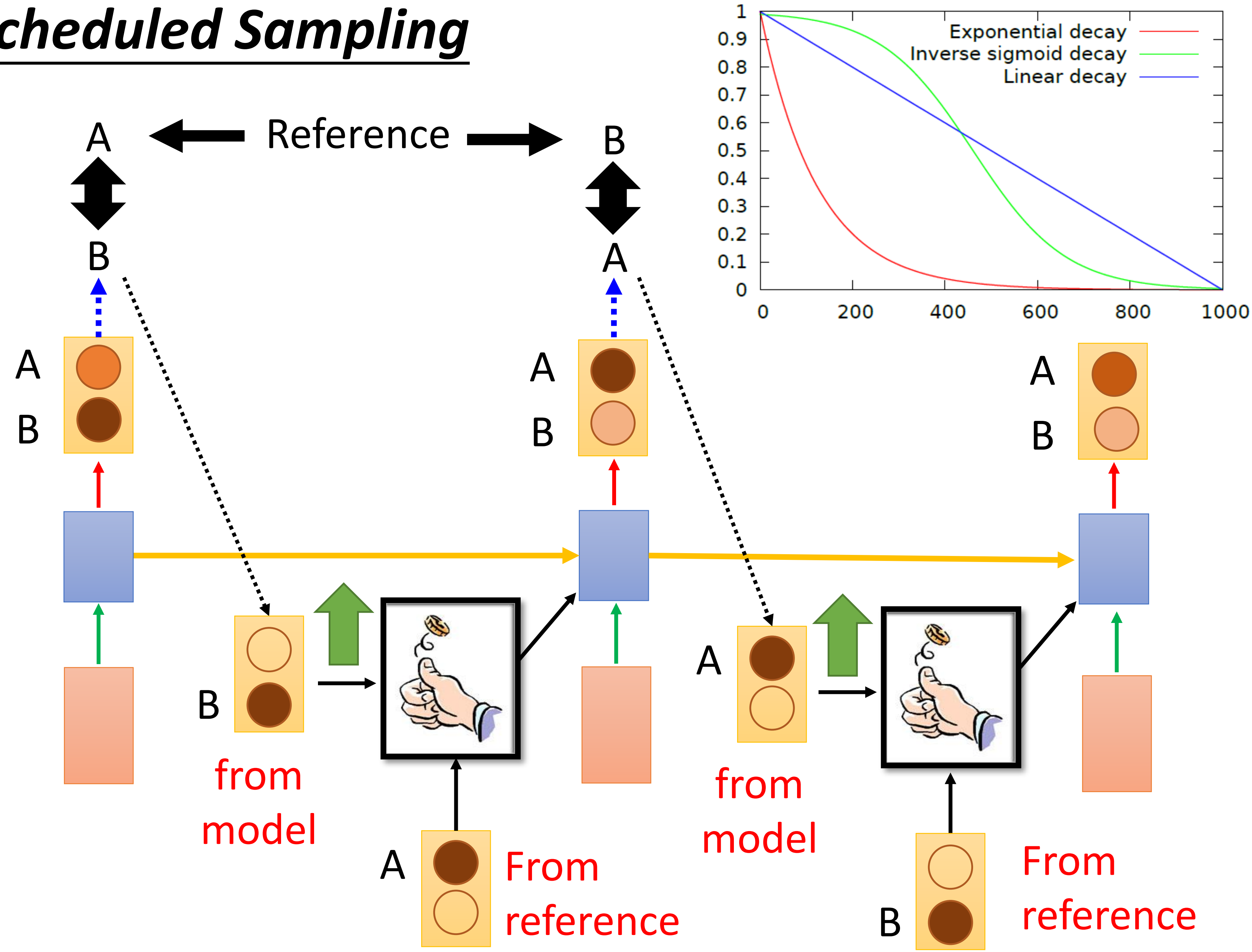
When we try to decrease the loss for both step 1 and 2 .....

Training is matched to testing.

In practice, it is hard to train in this way.



Scheduled Sampling



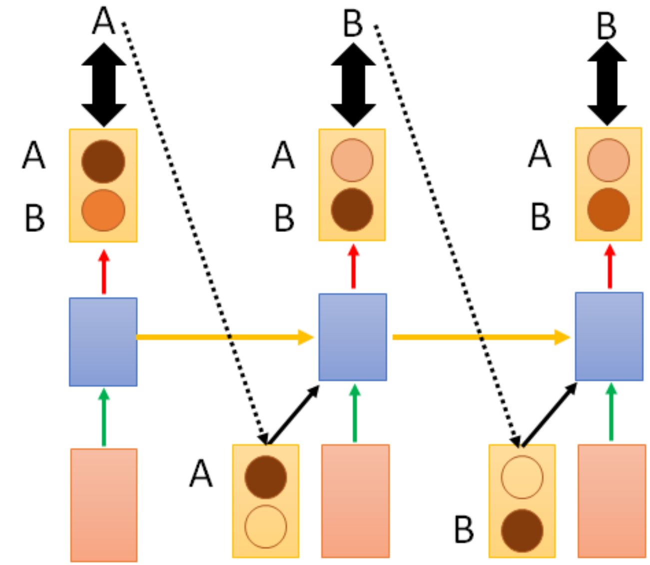
Scheduled Sampling

- Caption generation on MSCOCO

	BLEU-4	METEOR	CIDER
Always from reference	28.8	24.2	89.5
Always from model	11.2	15.7	49.7
Scheduled Sampling	30.6	24.3	92.1

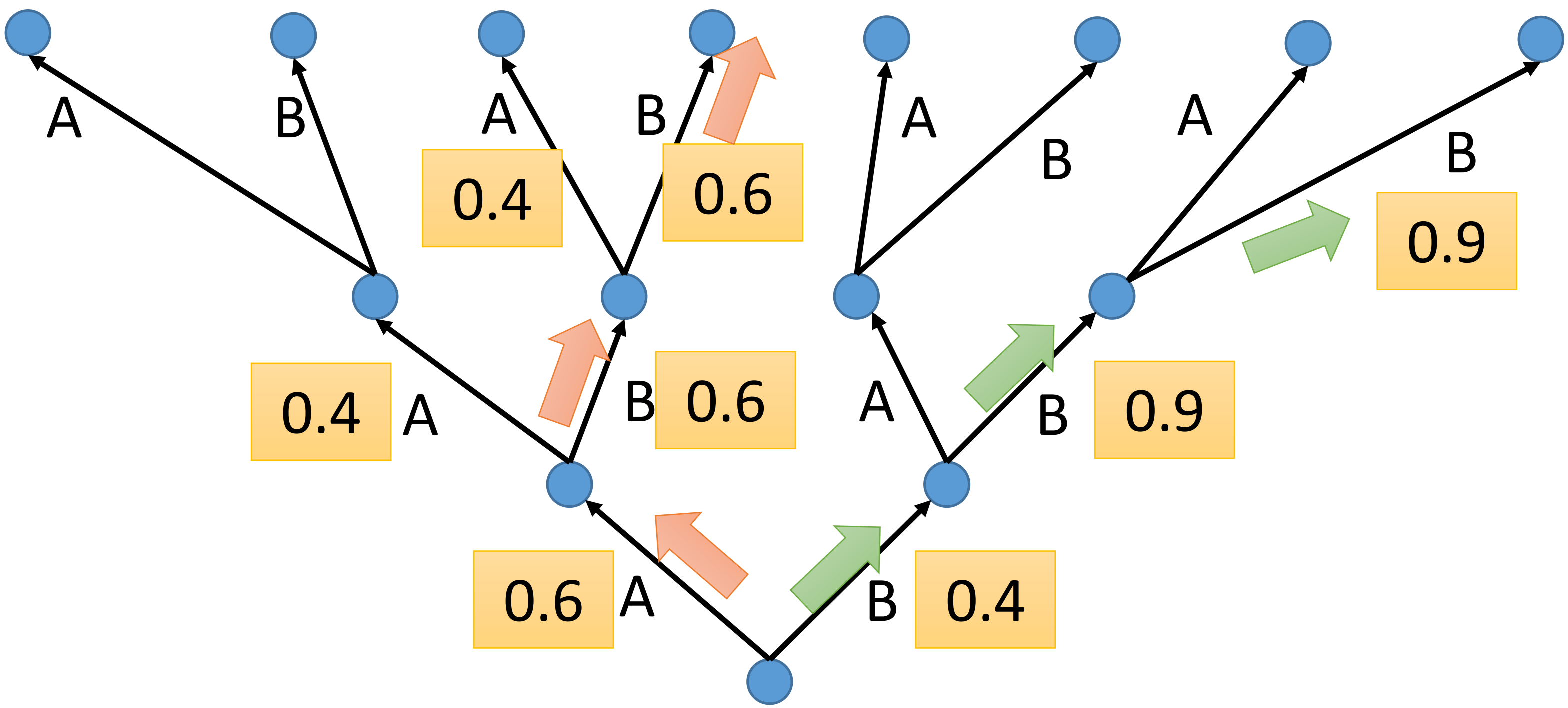
Samy Bengio, Oriol Vinyals, Navdeep Jaitly, Noam Shazeer, Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks, arXiv preprint, 2015

Beam Search

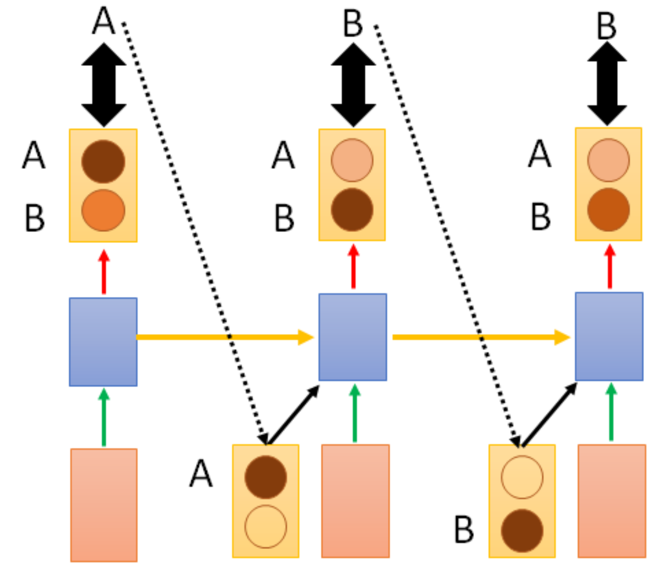


The green path has higher score.

Not possible to check all the paths

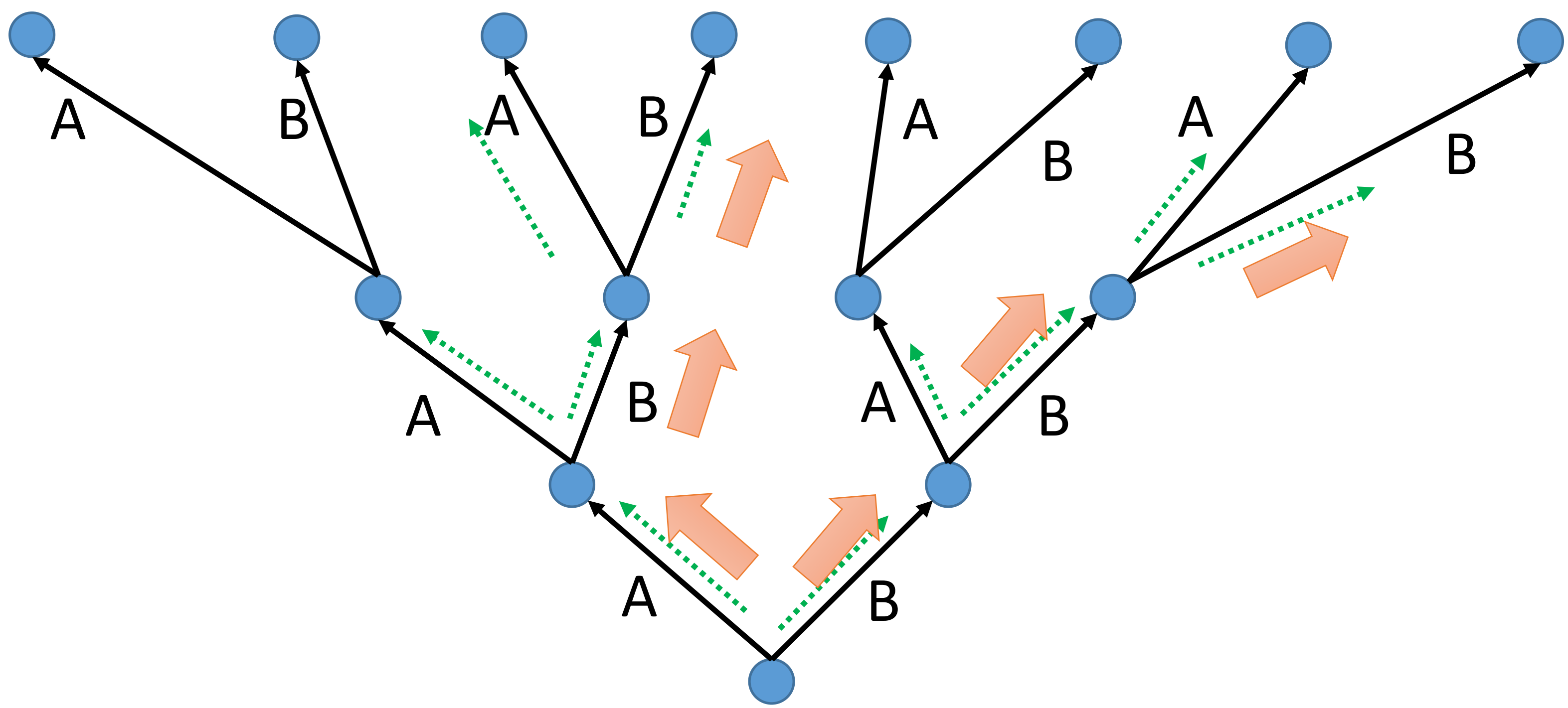


Beam Search

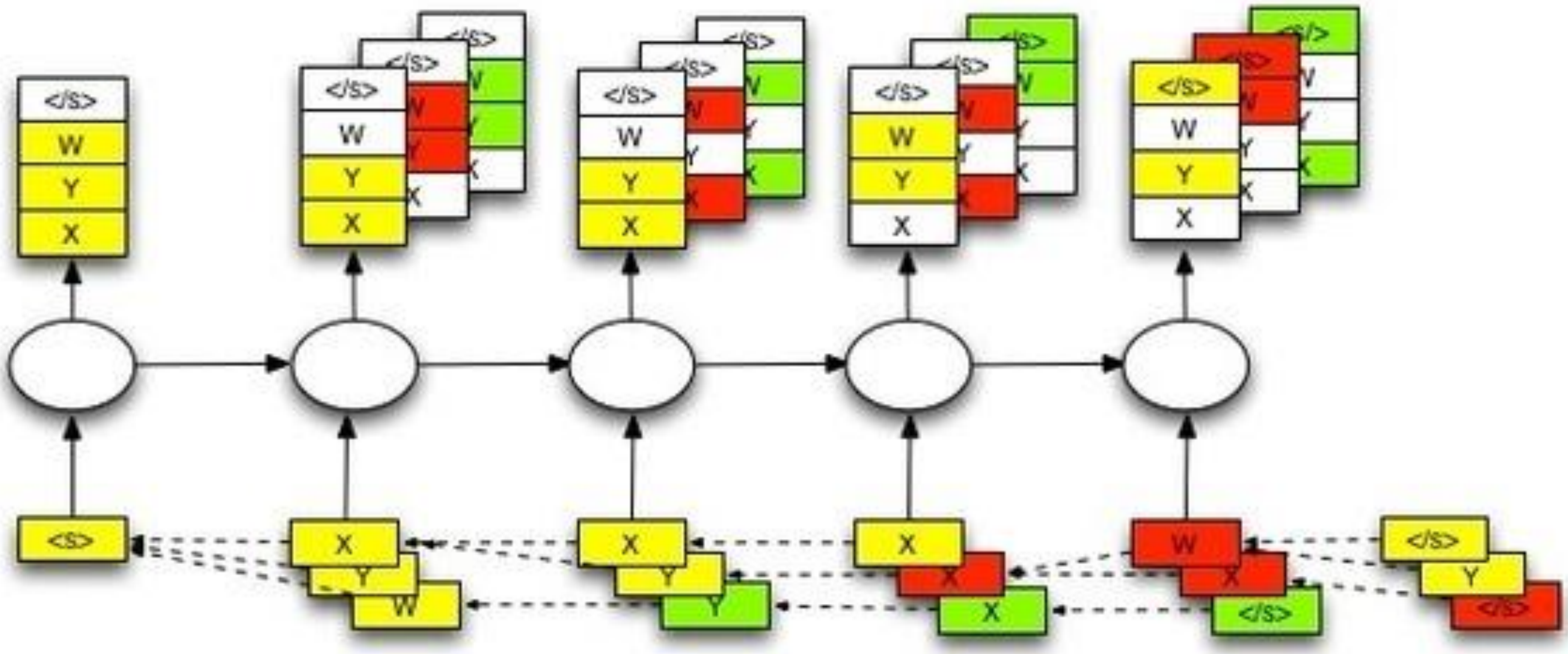


Keep several best path at each step

Beam size = 2



Beam Search



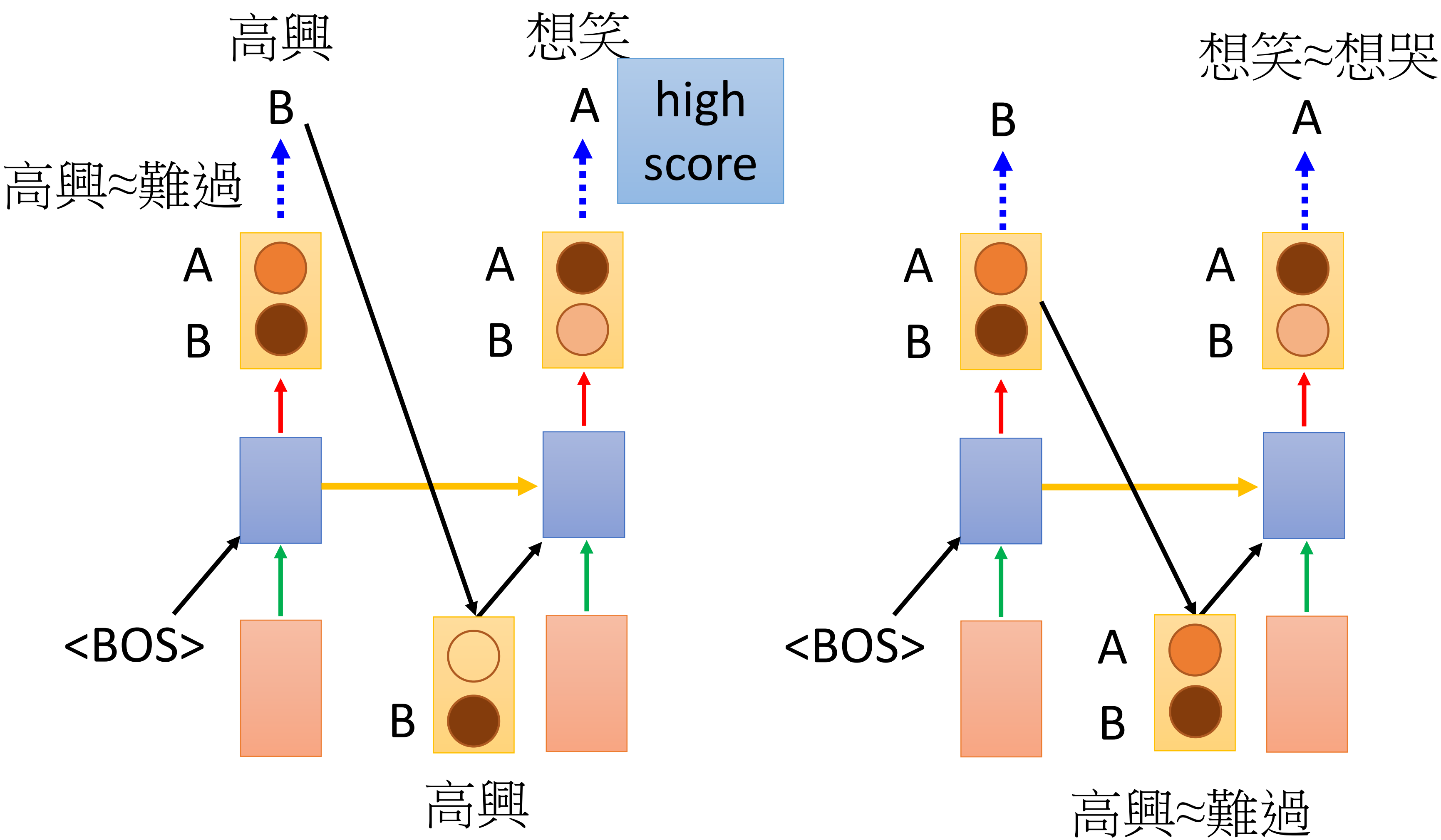
The size of beam is 3 in this example.

https://github.com/tensorflow/tensorflow/issues/654#issuecomment-169009989

Better Idea?

U: 你覺得如何?

M: 高興想笑 or 難過想哭



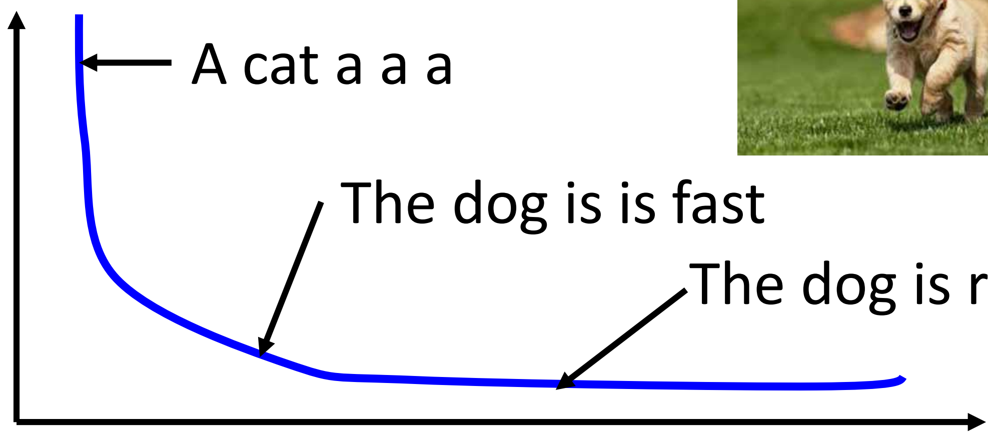
Object level v.s. Component level

- Minimizing the error defined on component level is not equivalent to improving the generated objects

Ref: The dog is running fast

$$C = \sum_t C_t$$

Cross-entropy of each step

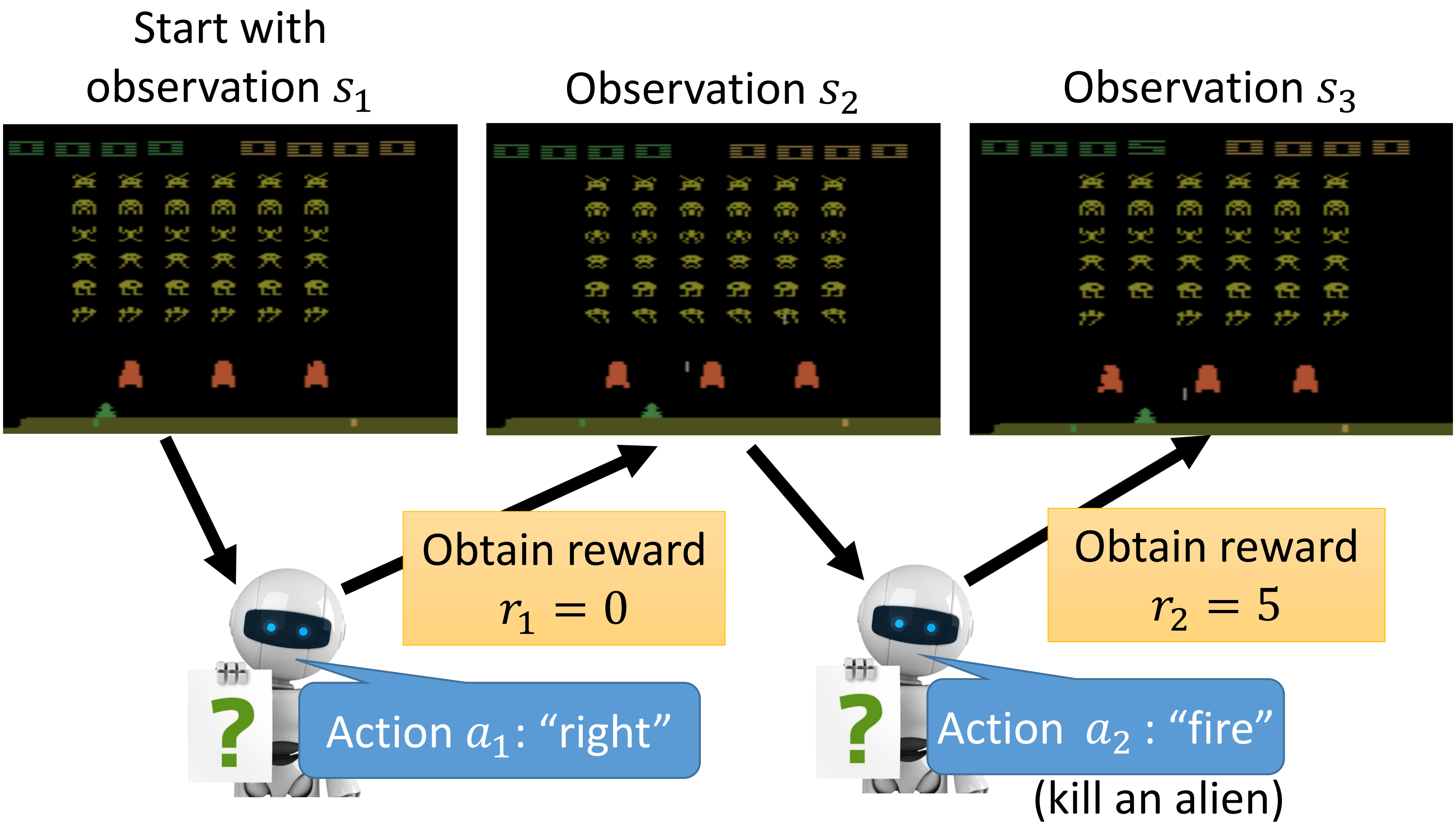


Optimize object-level criterion instead of component-level cross-entropy. object-level criterion: $R(y, \hat{y})$

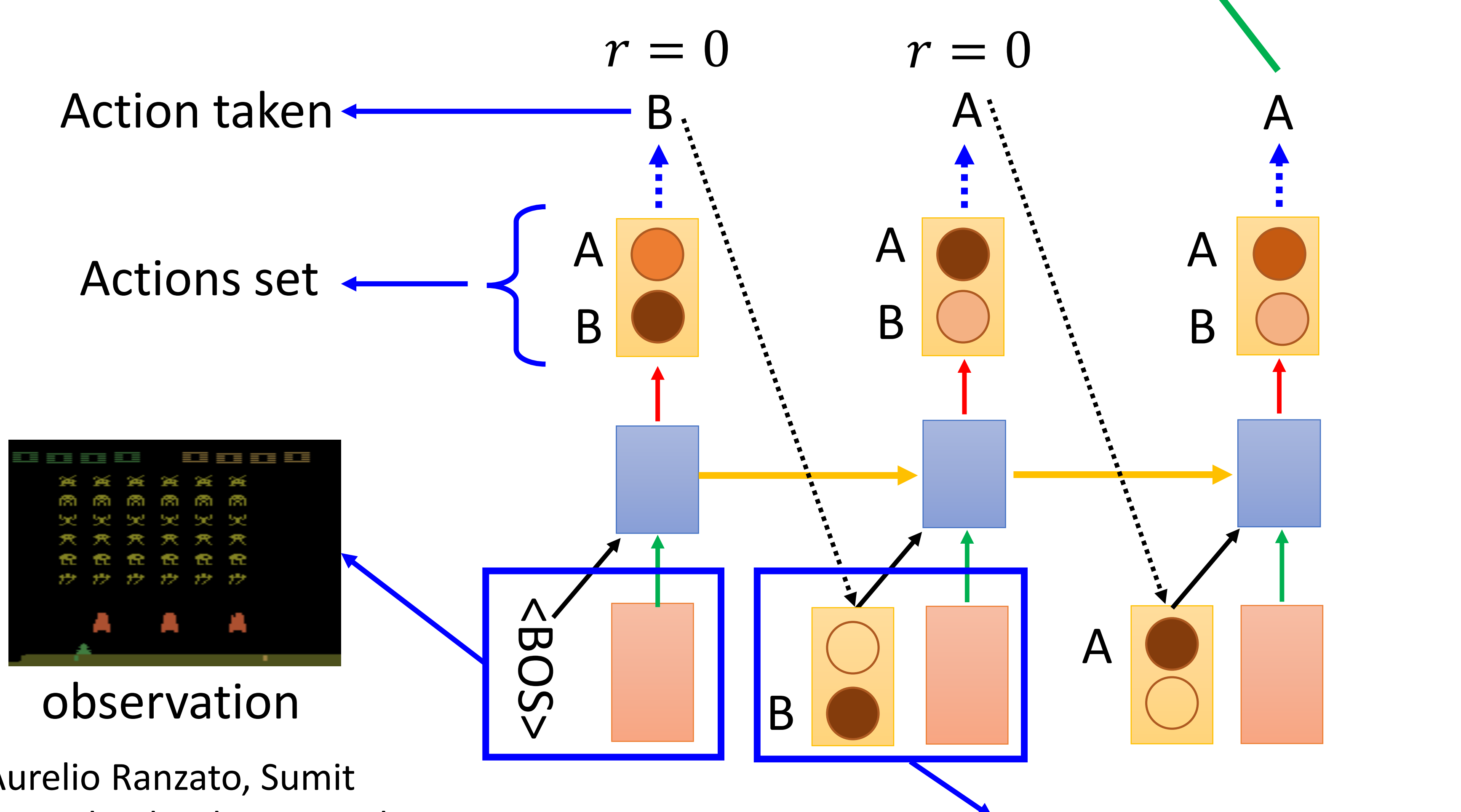
Gradient Descent?

y: generated utterance, $\hat{y}$: ground truth

Reinforcement learning?

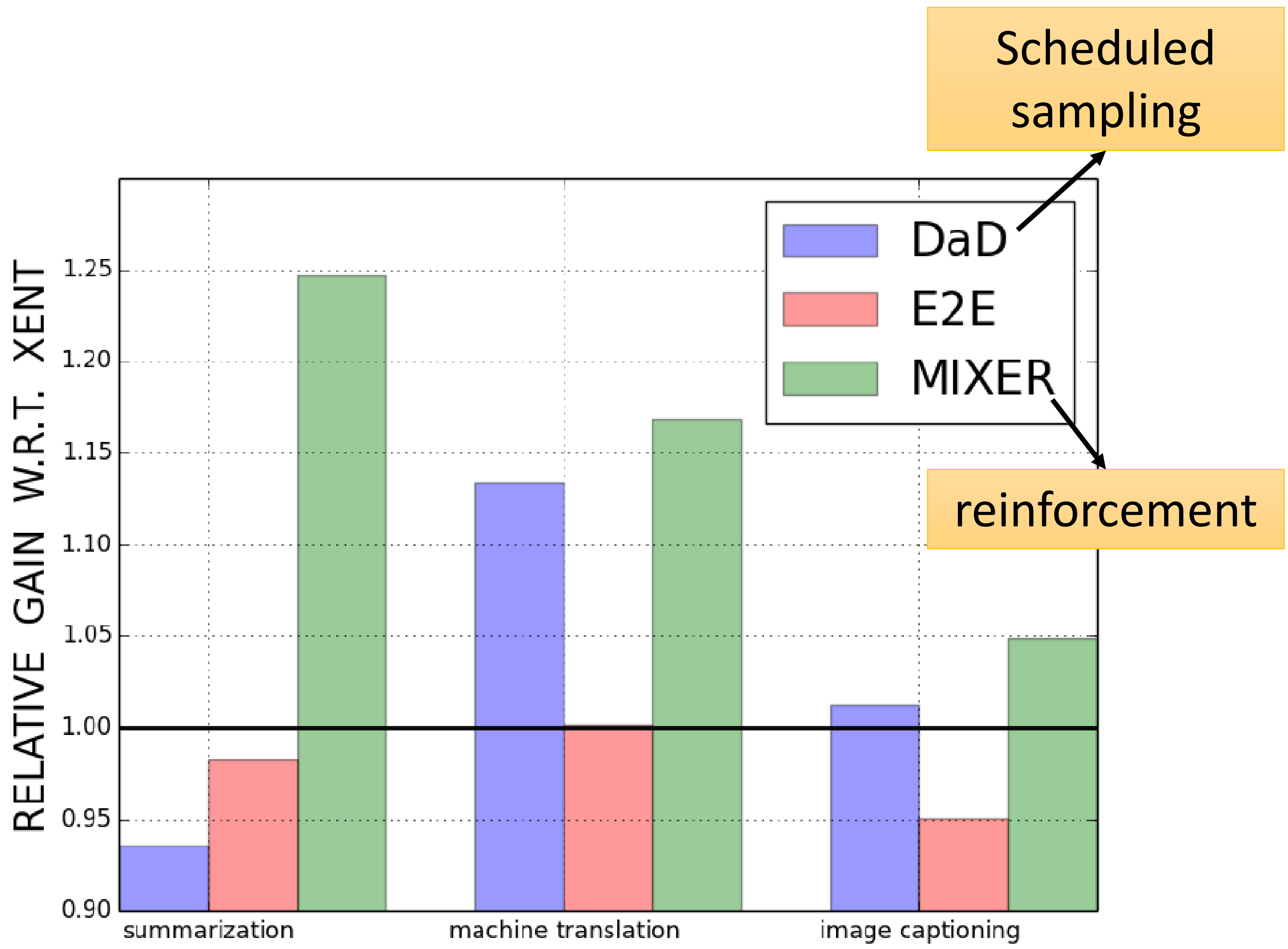


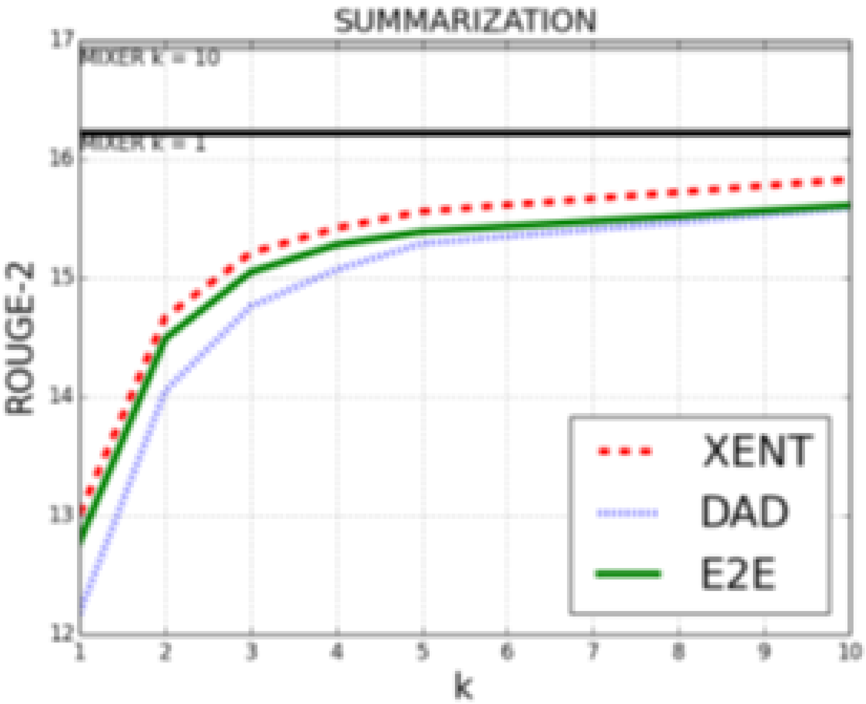
Reinforcement learning?

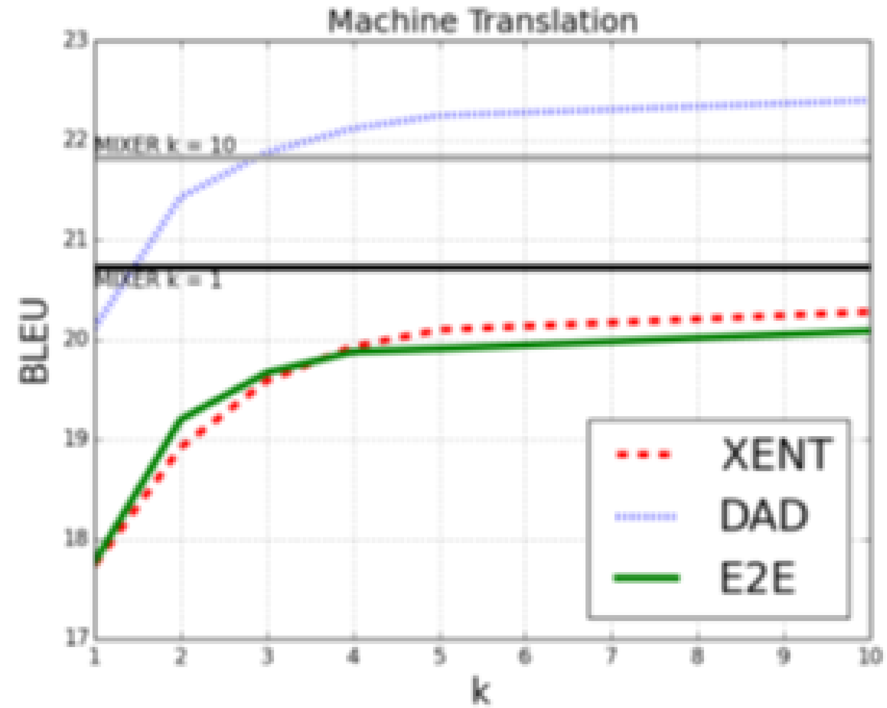


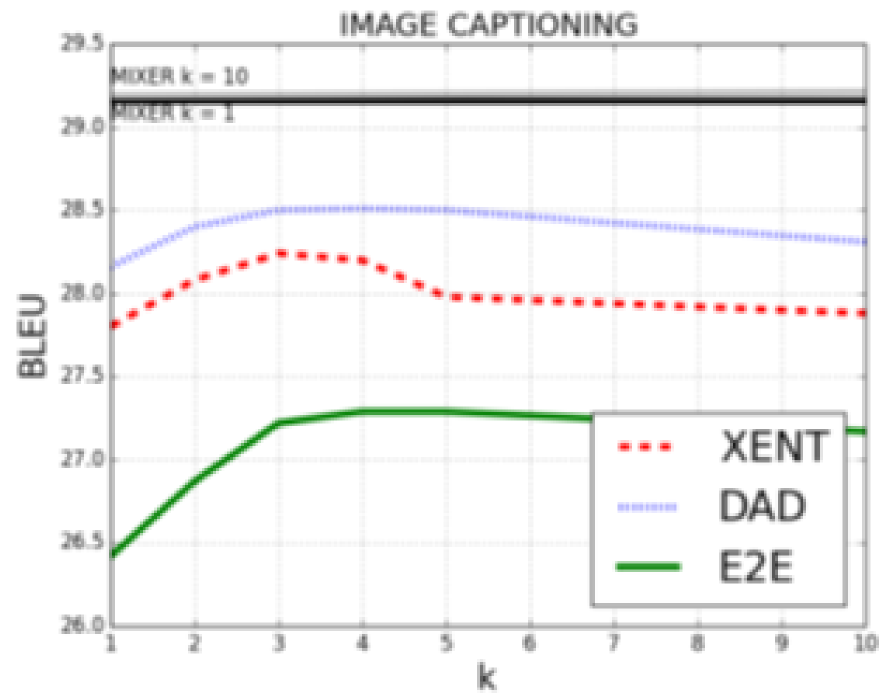
The action we take influence the observation in the next step

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, Wojciech Zaremba, "Sequence Level Training with Recurrent Neural Networks", ICLR, 2016


Scheduled sampling
reinforcement
DaD
E2E
MIXER
RELATIVE GAIN W.R.T. XENT
1.25
1.20
1.15
1.10
1.05
1.00
0.95
0.90
summarization
machine translation
image captioning



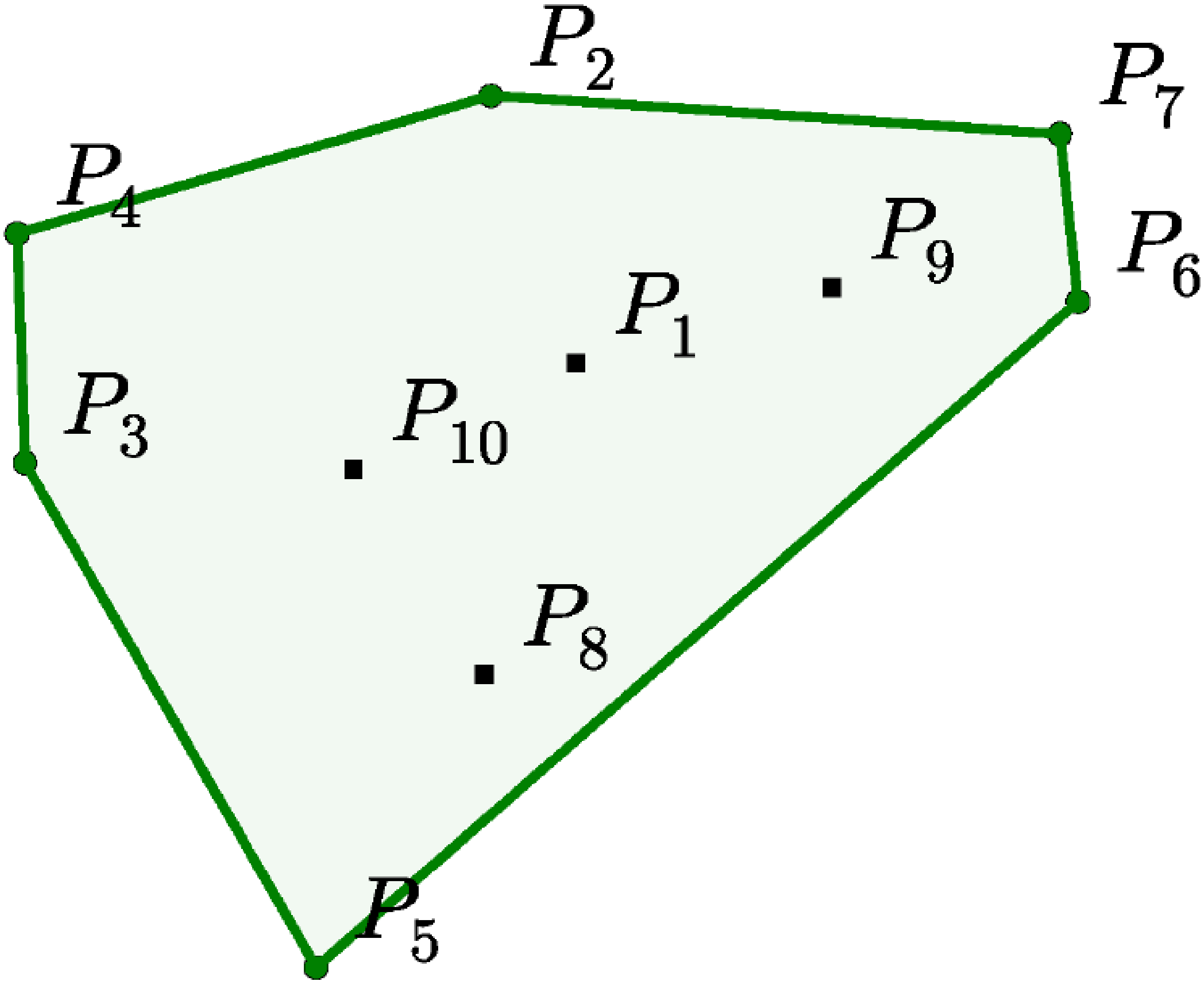




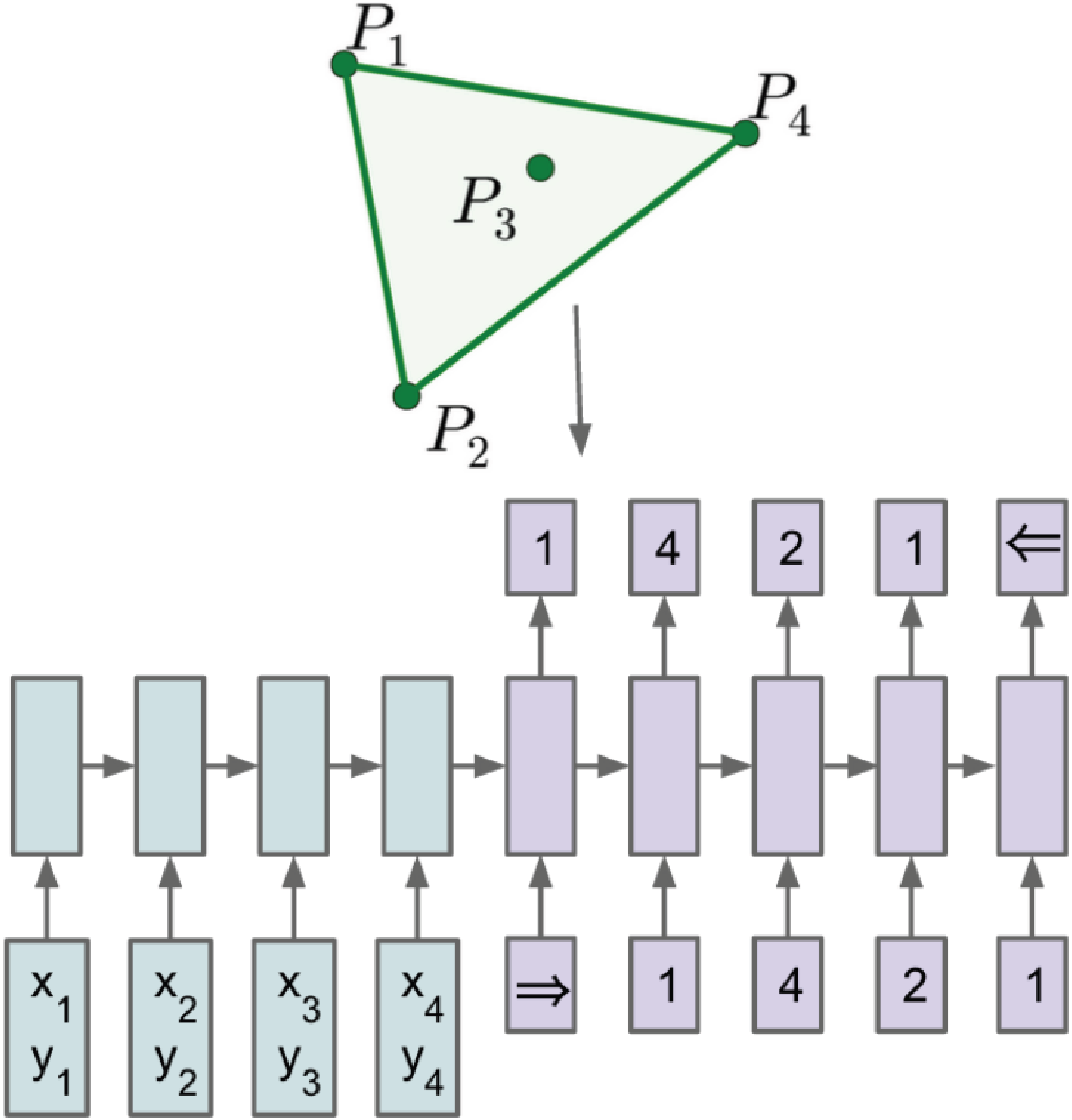
DAD: Scheduled Sampling
MIXER: reinforcement

Pointer Network

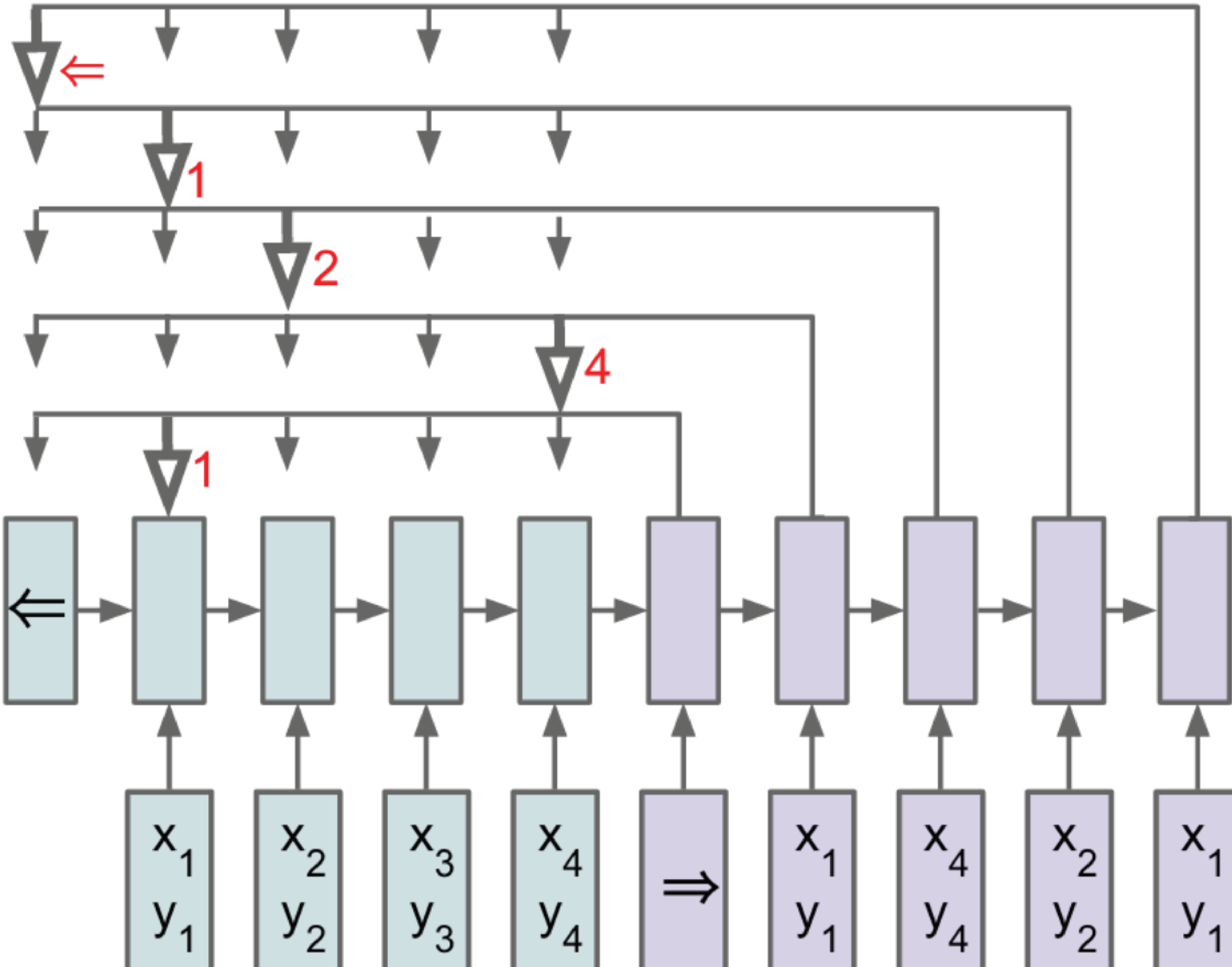
Oriol Vinyals, Meire Fortunato, Navdeep Jaitly, Pointer Network, NIPS, 2015



P_2
P_7
P_4
P_9
P_6
P_1
P_3
P_{10}
P_8
P_5

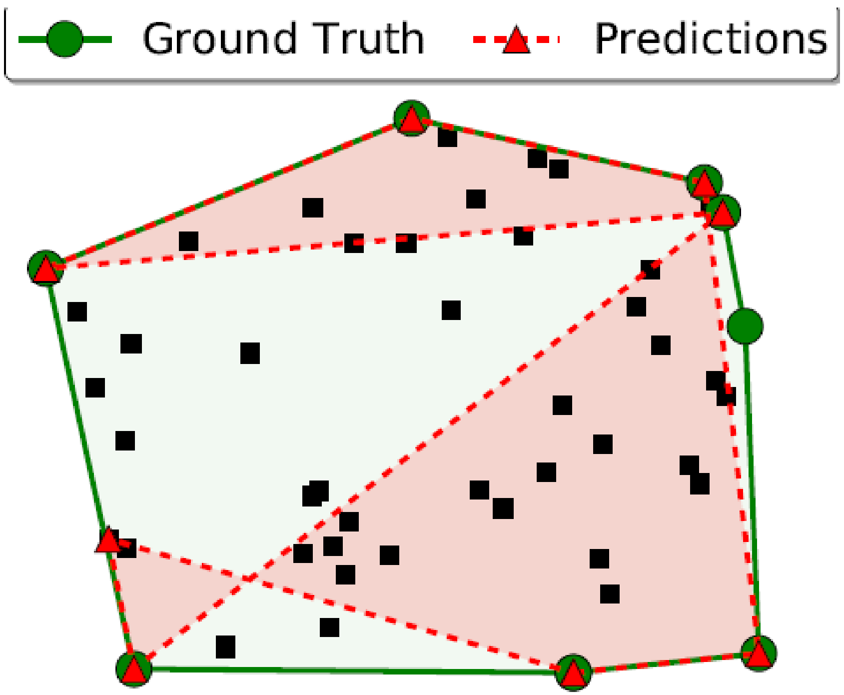


P_1
P_4
P_3
P_2
1
4
2
1
⇐
⇒
1
4
2
1
x_1 y_1
x_2 y_2
x_3 y_3
x_4 y_4

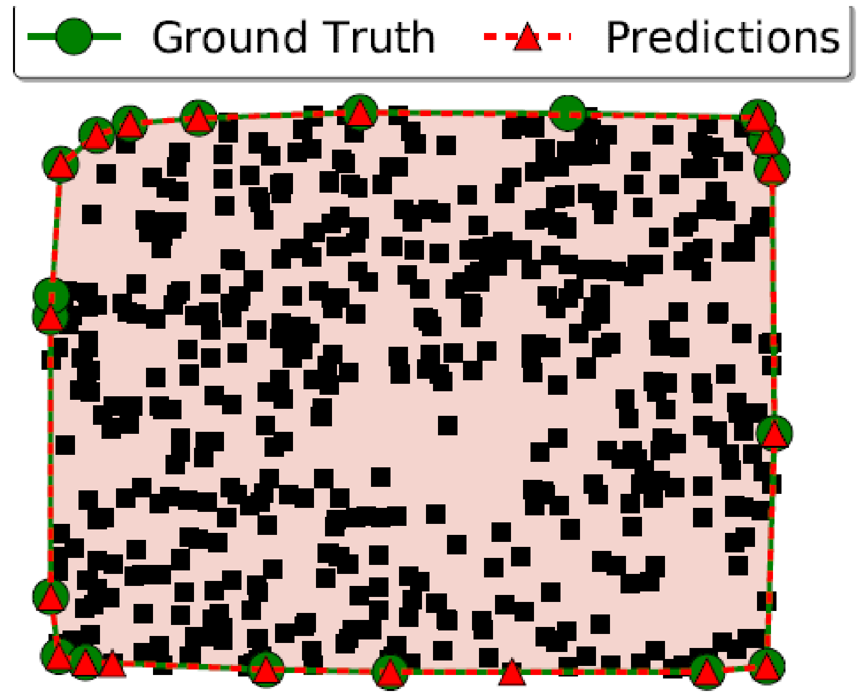


⇐
1
2
4
1
⇐
x1 y1
x2 y2
x3 y3
x4 y4
⇒
x1 y1
x4 y4
x2 y2
x1 y1

Method	trained n	n	Accuracy	Area
LSTM [1]	50	50	1.9%	FAIL
+attention [5]	50	50	38.9%	99.7%
Ptr-Net	50	50	72.6%	99.9%
LSTM [1]	5	5	87.7%	99.6%
Ptr-Net	5-50	5	92.0%	99.6%
LSTM [1]	10	10	29.9%	FAIL
Ptr-Net	5-50	10	87.0%	99.8%
Ptr-Net	5-50	50	69.6%	99.9%
Ptr-Net	5-50	100	50.3%	99.9%
Ptr-Net	5-50	200	22.1%	99.9%
Ptr-Net	5-50	500	1.3%	99.2%



(a) LSTM, m=50, n=50

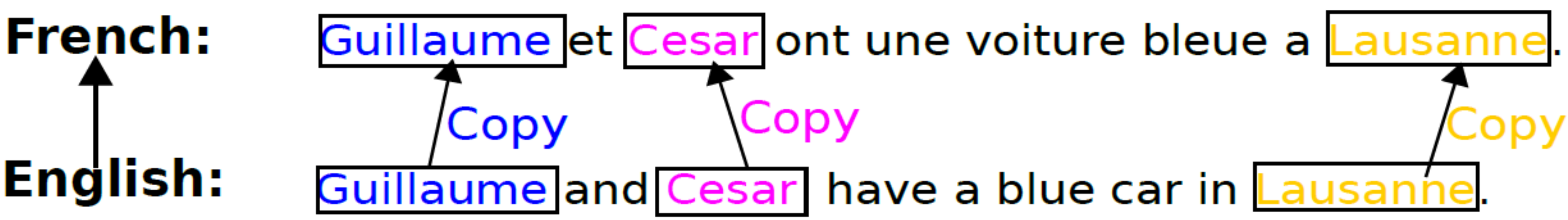


(d) Ptr-Net, m=5-50, n=500

Applications

Jiatao Gu, Zhengdong Lu, Hang Li, Victor O.K. Li, "Incorporating Copying Mechanism in Sequence-to-Sequence Learning", ACL, 2016
Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, Yoshua Bengio, "Pointing the Unknown Words", ACL, 2016

Machine Translation



Chat-bot

User: X寶你好，我是宏毅

Machine: 宏毅你好，很高興認識你