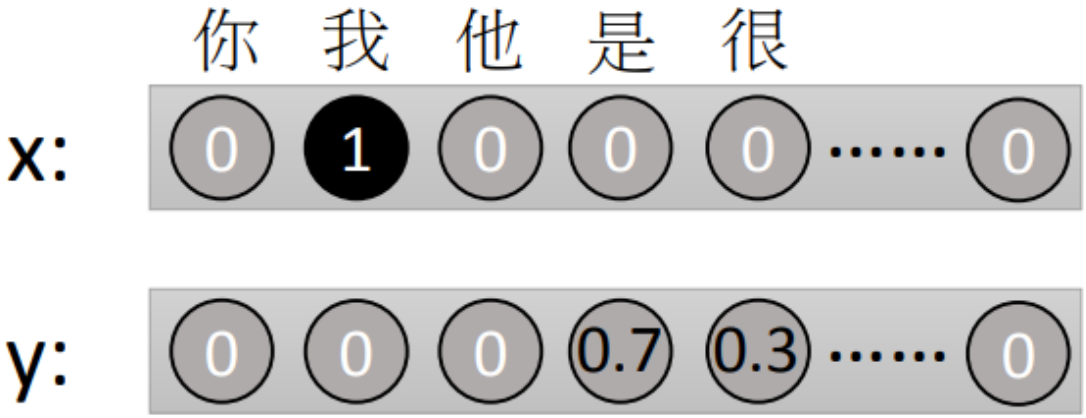
Generation

Generating a structured object component-by-component

Text

- Sentences are composed of characters/words
- Generating a character/word at each time by RNN
- Input : The token generated at the last time step
- Output : Distribution over the token



- Start from . Generate until is generated.

$$y^1: P(w|<BOS>)$$

$$y^2: P(w|<BOS>,床)$$

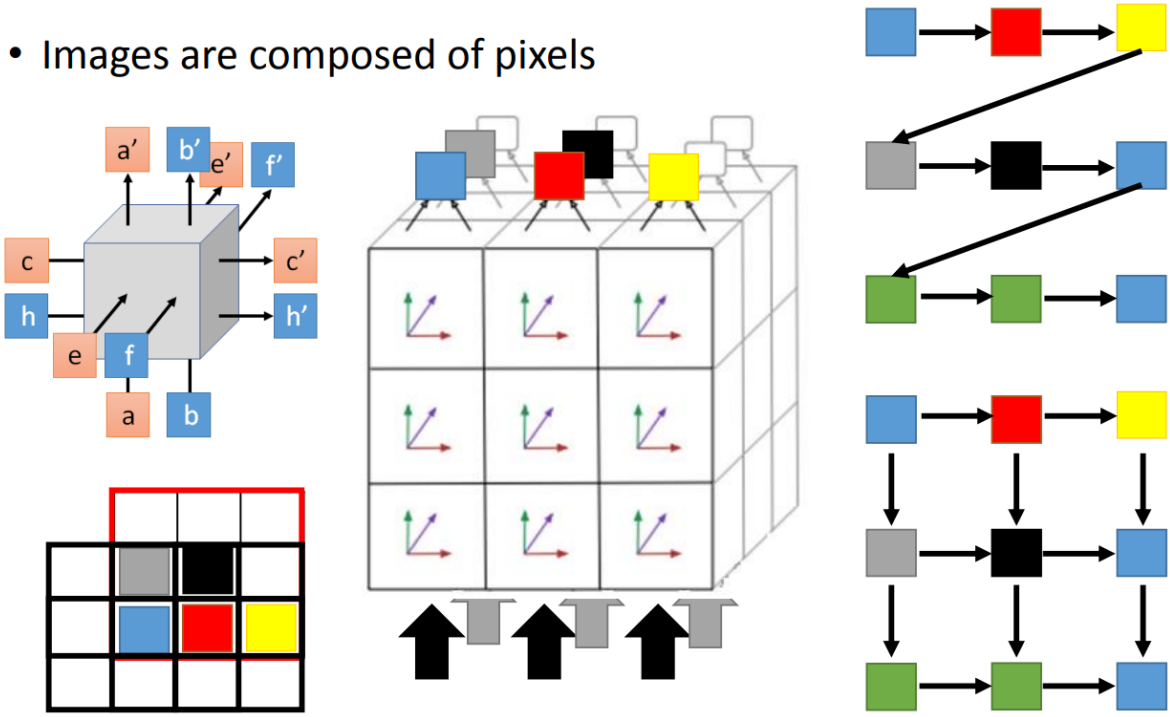
$$y^3: P(w|<BOS>,床,前)$$

- Minimize cross entropy during training

Image

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

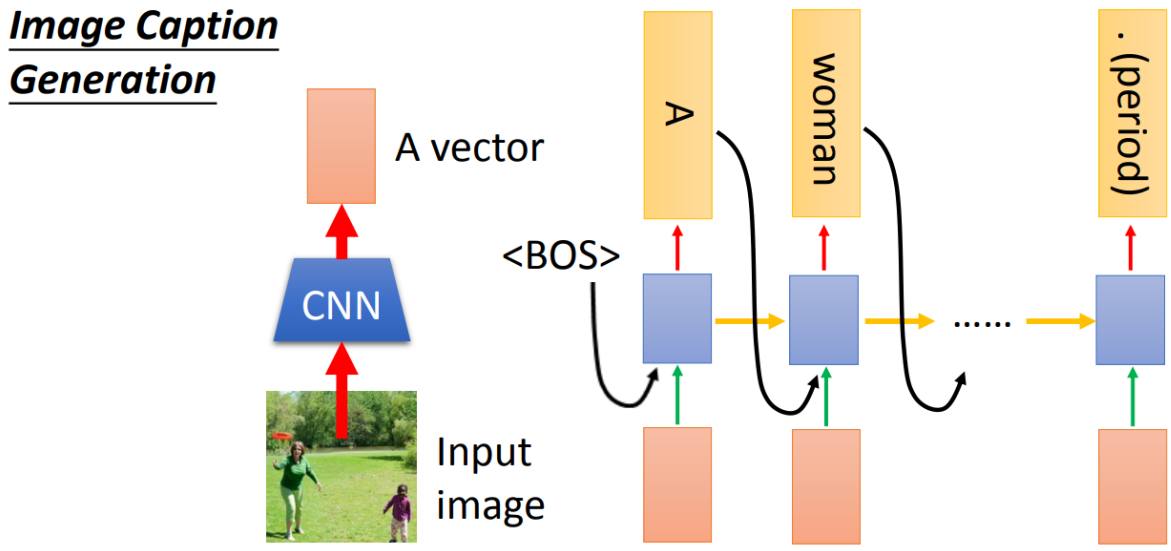
- Images are composed of pixels



Conditional Generation

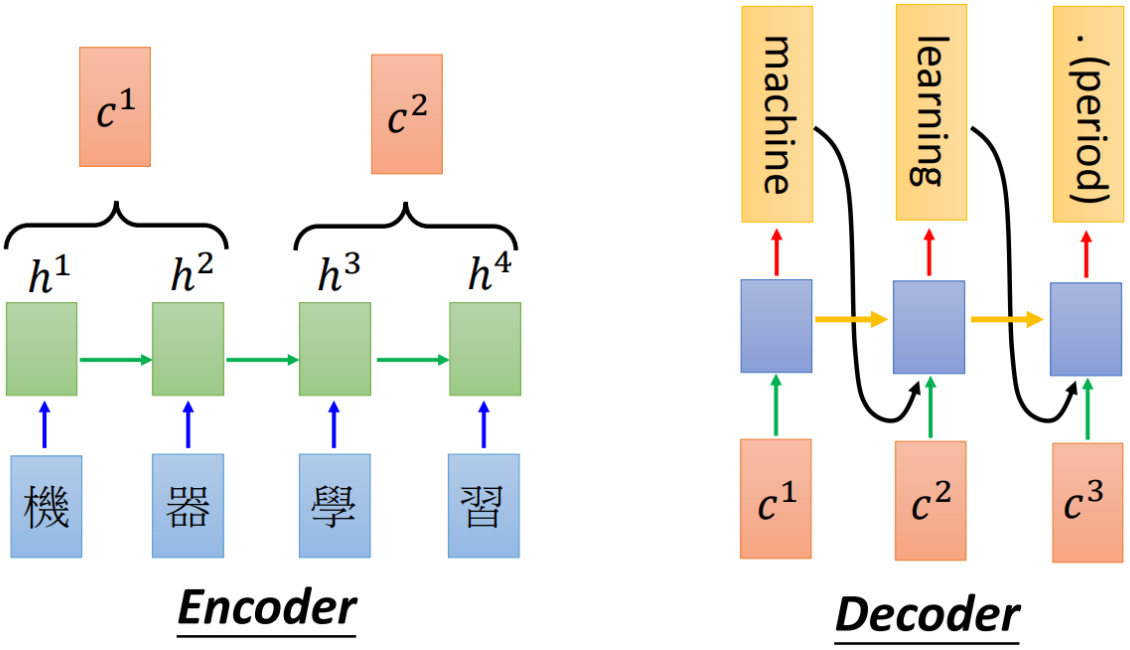
- Represent the input condition as a vector, and consider the vector as the input of RNN generator

CV Task



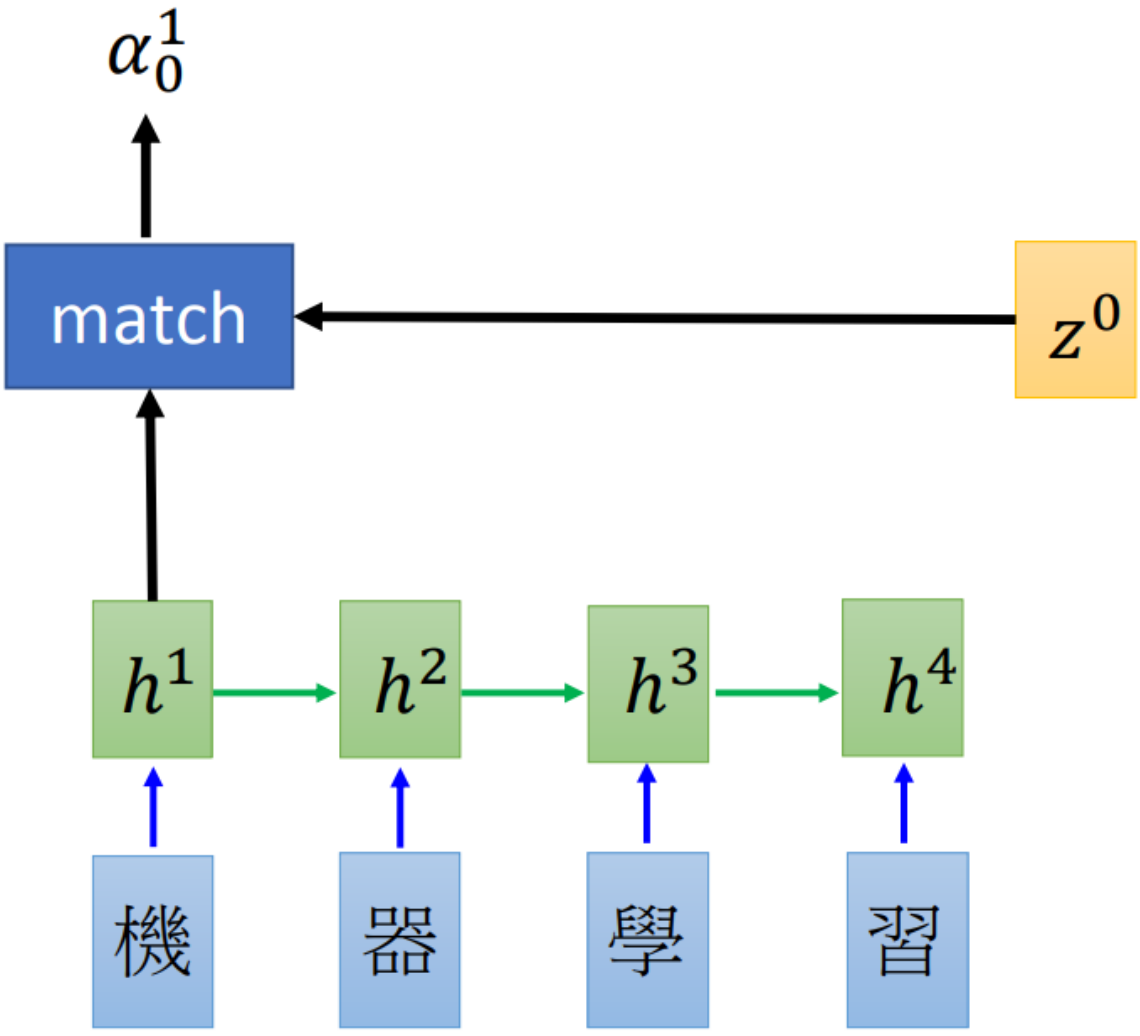
NLP Task

- Sequence-to-Sequence Learning
- Jointly train encoder & decoder
- Need to consider longer context during chatting



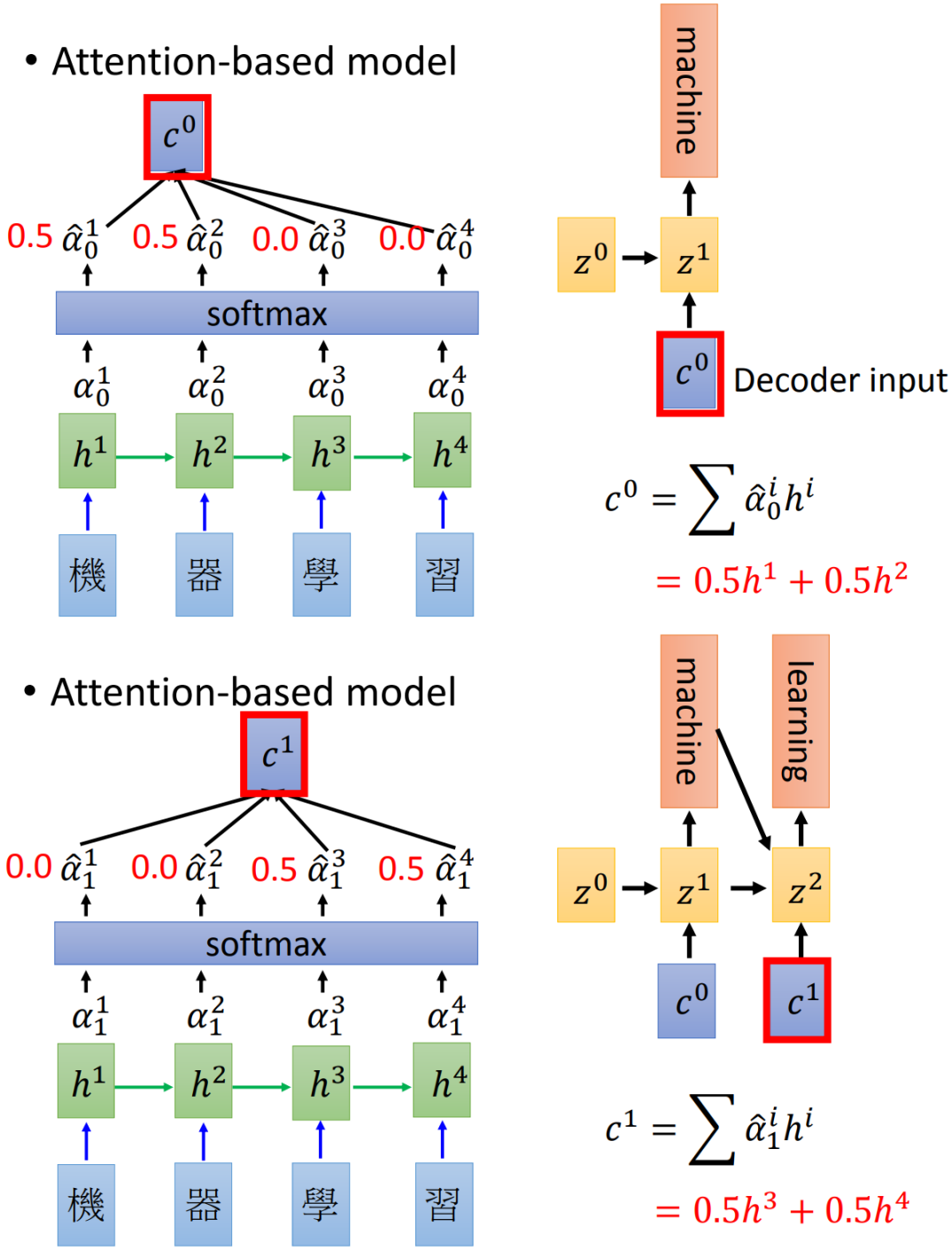
Dynamic Conditional Generation - Attention

Machine Translation



Match Function

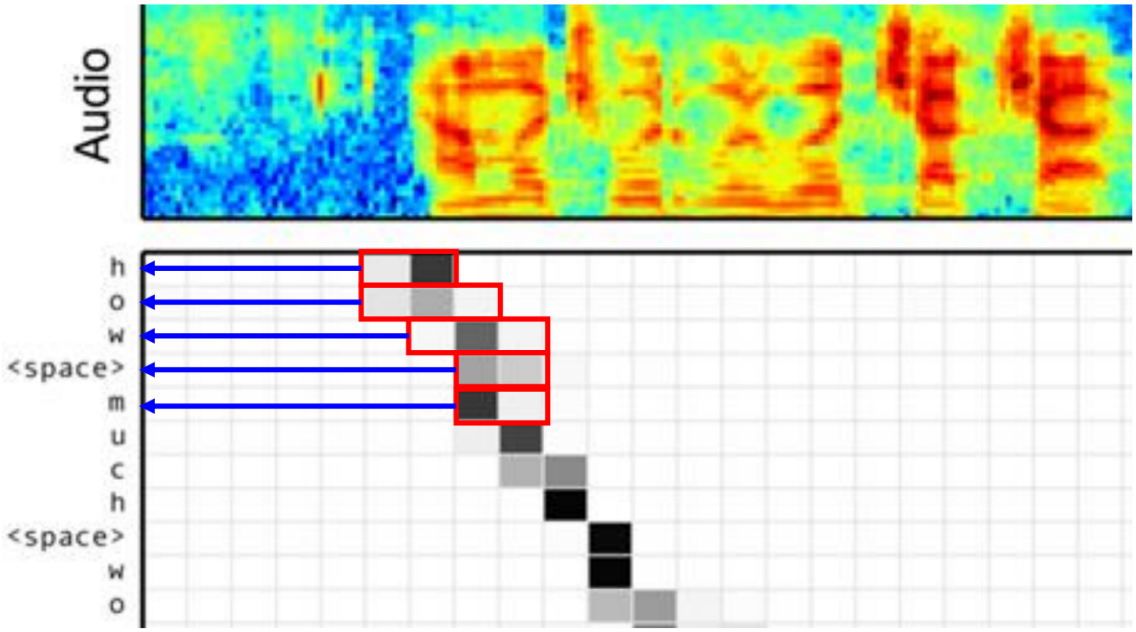
- Cosine similarity of z and h
- Small NN whose input is z and h, output a scalar
- $$\alpha = h^T W z$$



- The same process repeat until is generated

Speech Recognition

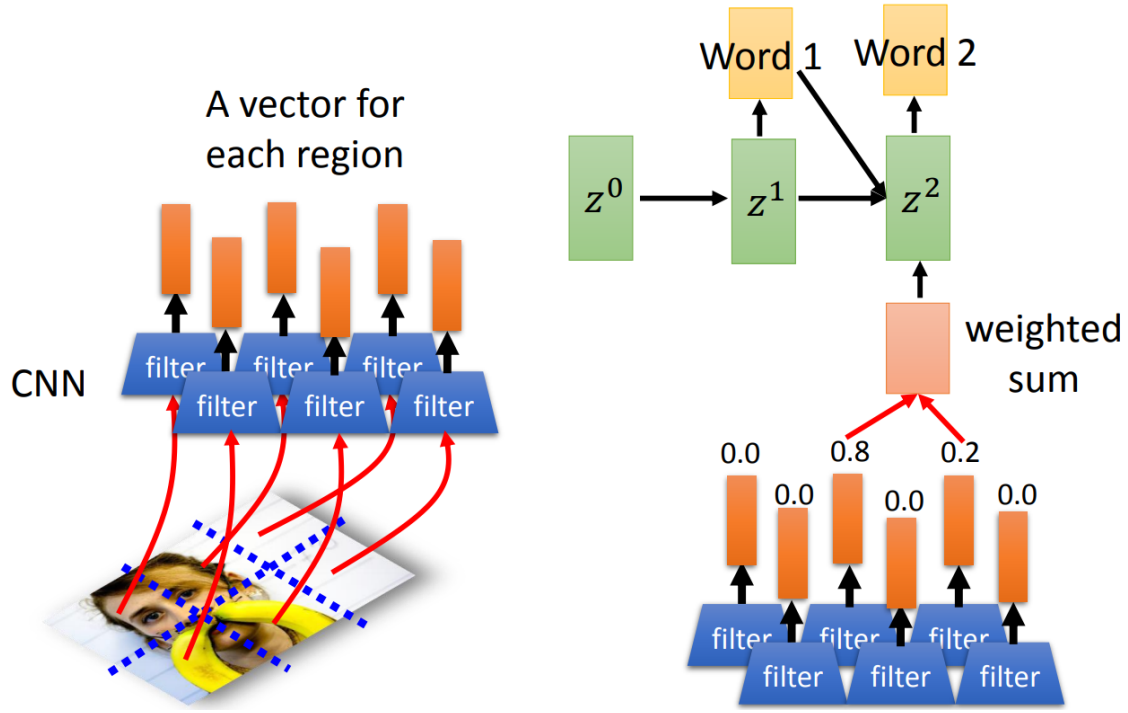
> William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, “Listen, Attend and Spell”, ICASSP, 2016



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



> 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

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

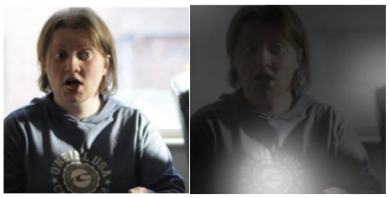
A stop sign is on a road with a mountain in the background.

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.

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.

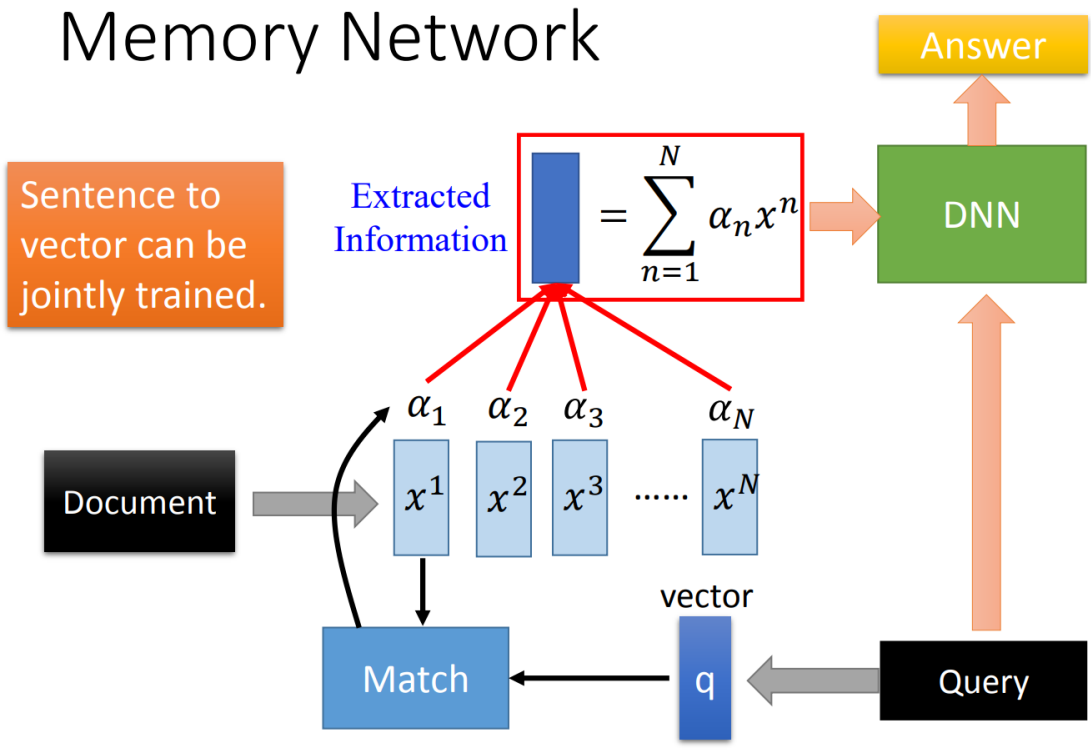
A person is standing on a beach with a surfboard.

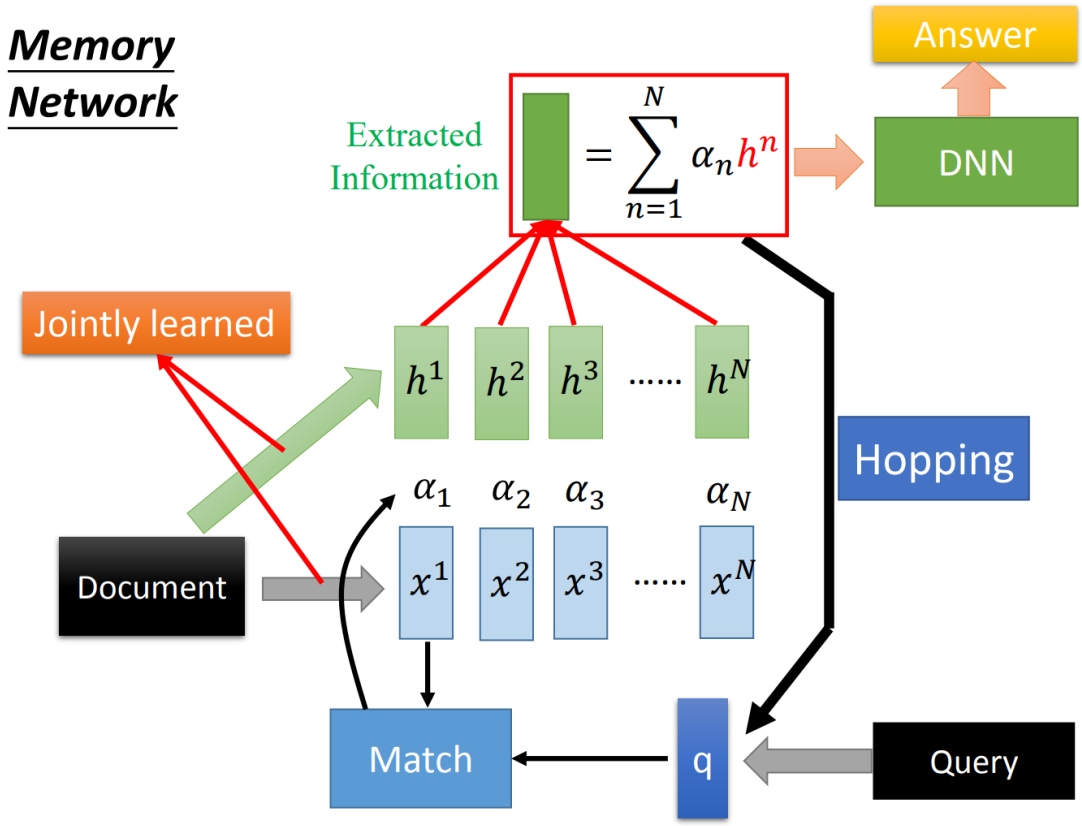
A woman is sitting at a table with a large pizza.

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

Memory Network

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

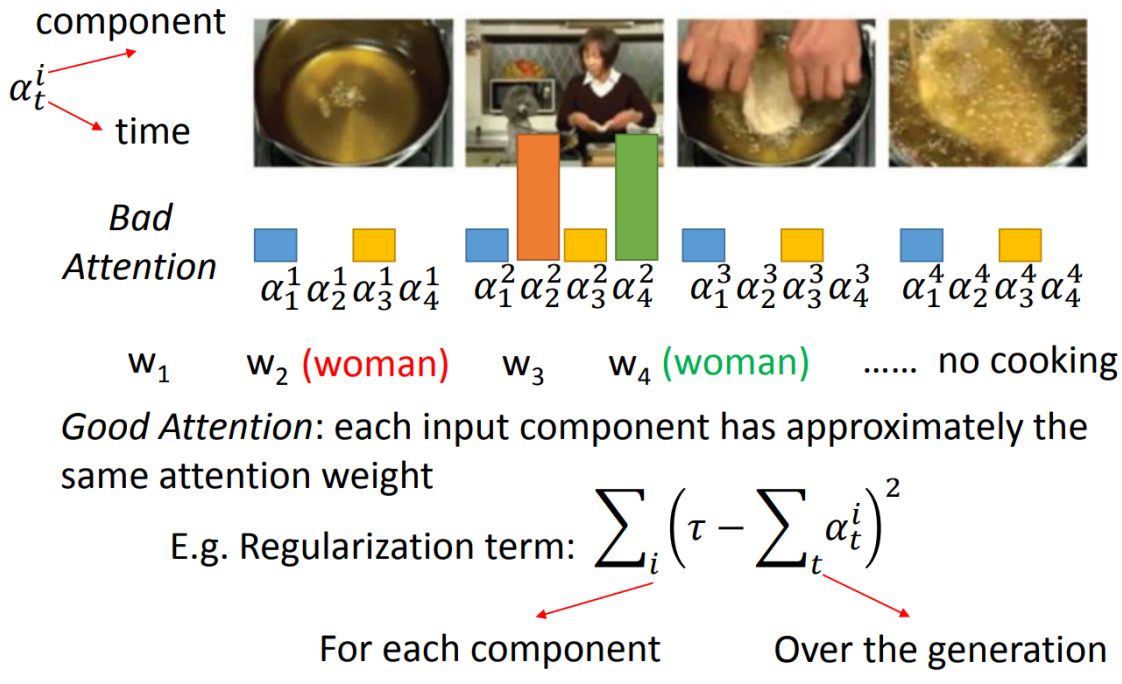




Tips for Generation

Good Attention & Bad 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



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

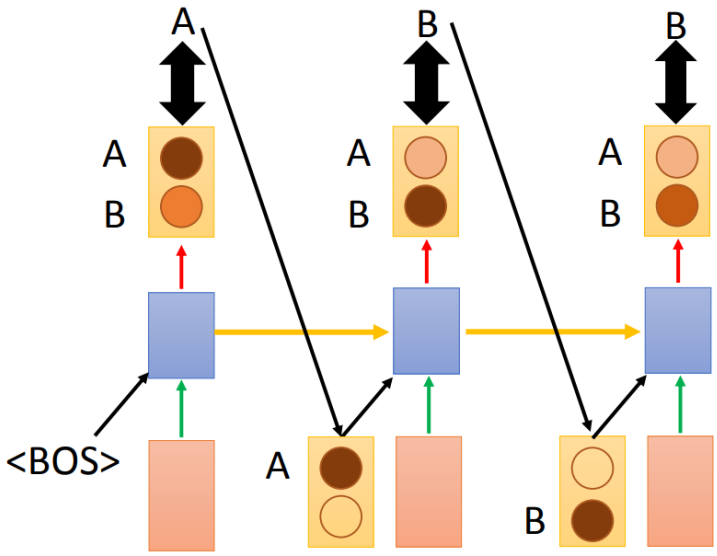
- ***Training***

$$C = \sum_t C_t$$

Minimizing cross-entropy of each component

: condition

Reference:



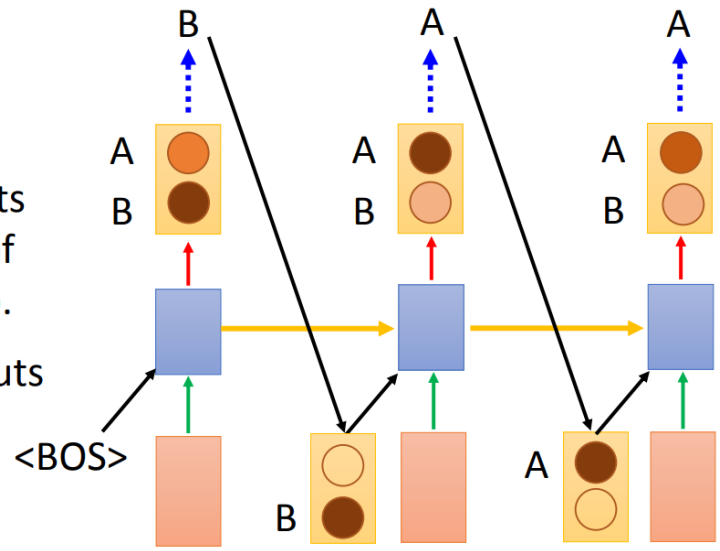
- ***Generation***

We do not know the reference

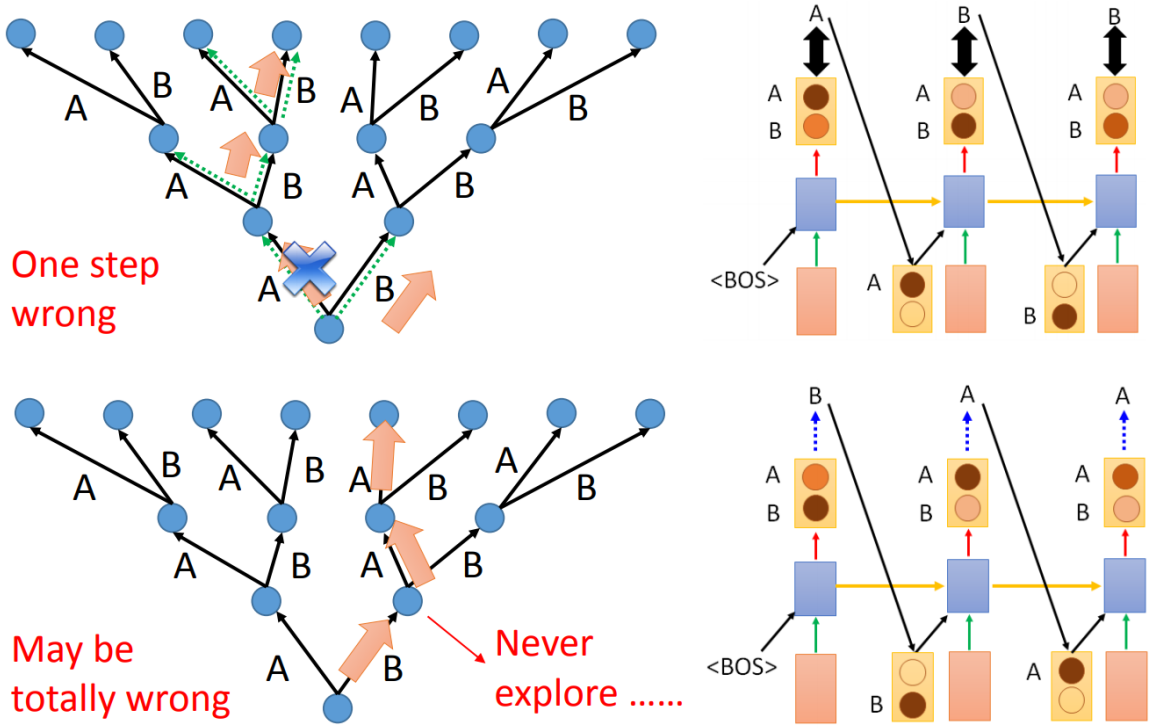
Testing: The inputs are the outputs of the last time step.

Training: The inputs are reference.

Exposure Bias

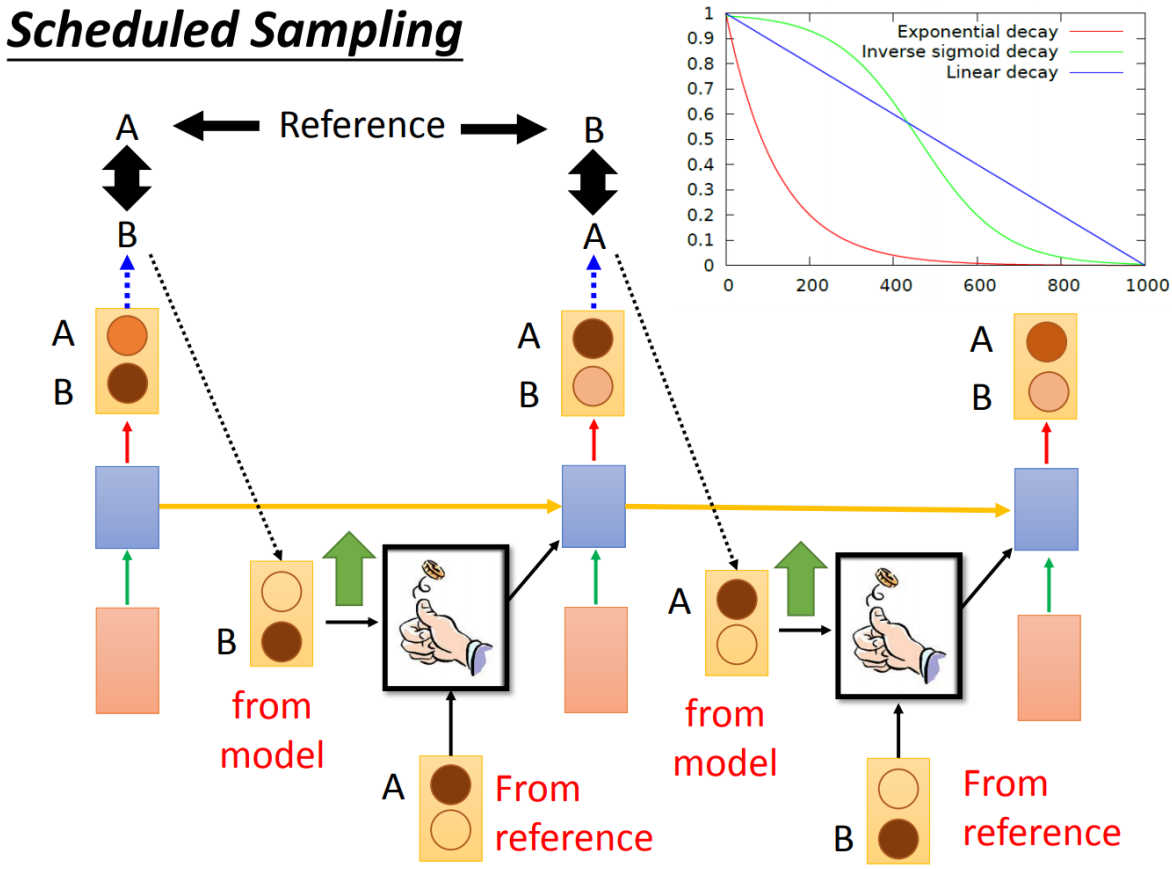


Difference on a Decision Tree



一步錯，步步錯

Scheduled Sampling



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

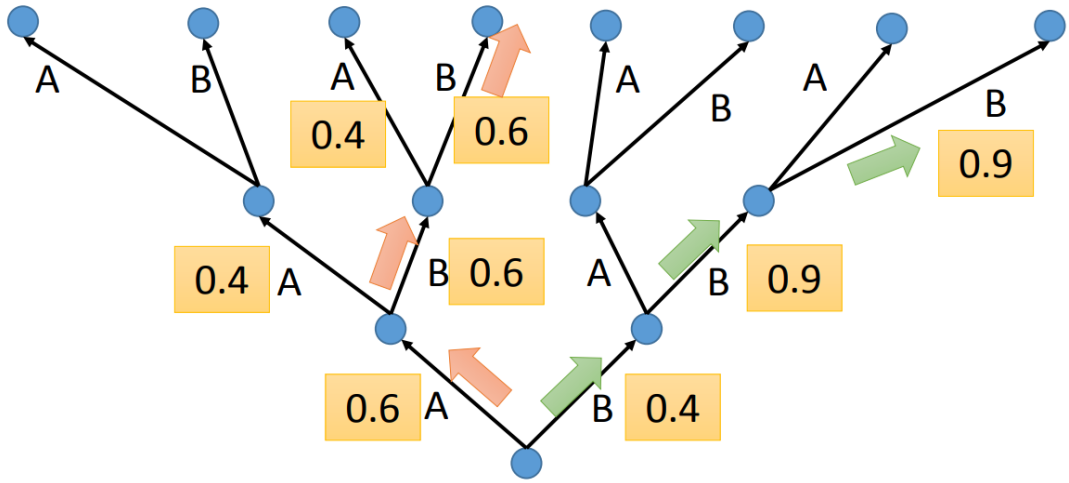
- Caption generation on MSCOCO

	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

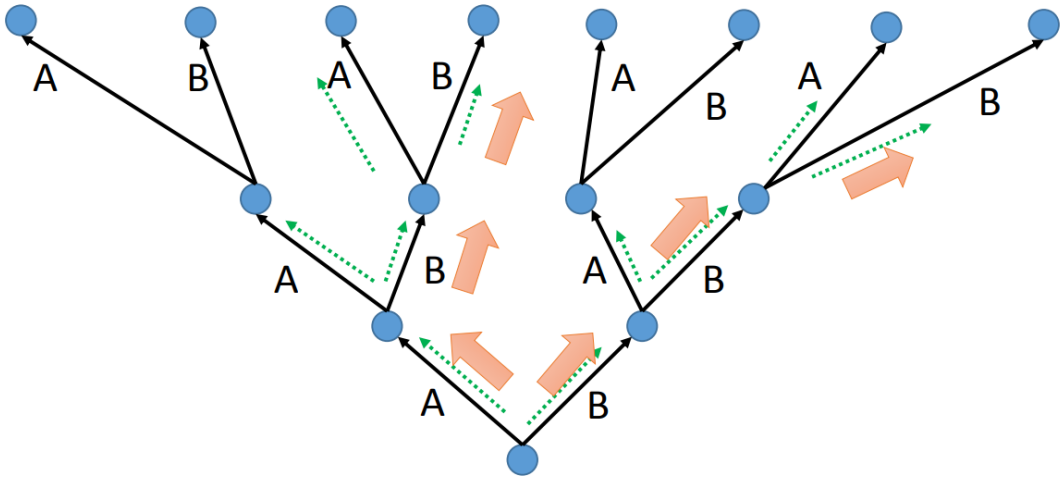
Beam Search

Higher each choice score ? Higher overall choice score ?

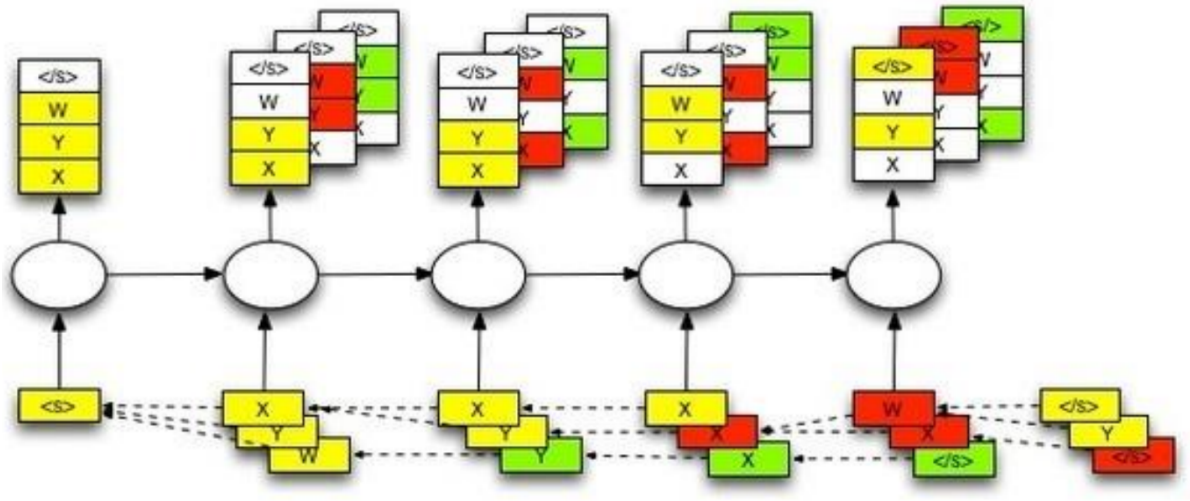
The green path has higher score. However, it is not possible to check all the paths.



Beam Search keep several best path at each step (Beam Size = 2)



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



The size of beam is 3 in this example.

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

$$C = \sum_t C_t$$

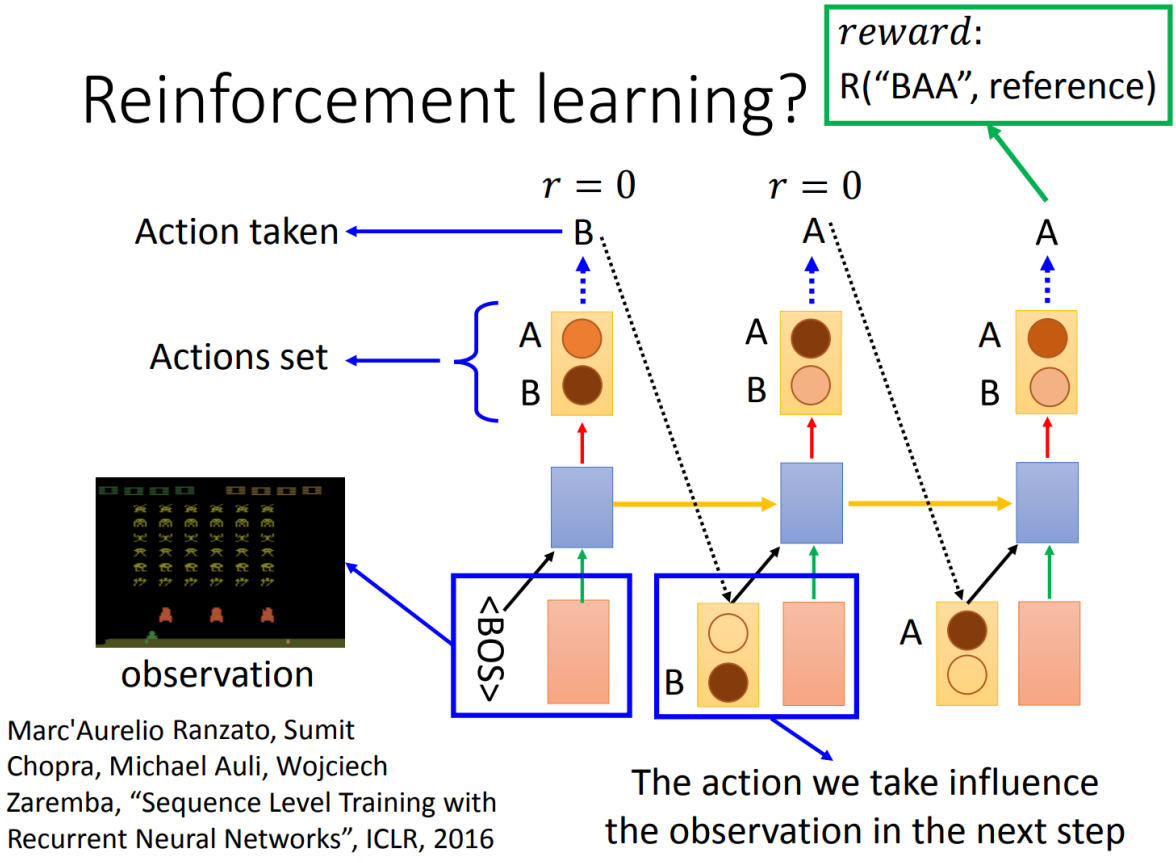
Cross-entropy of each step

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

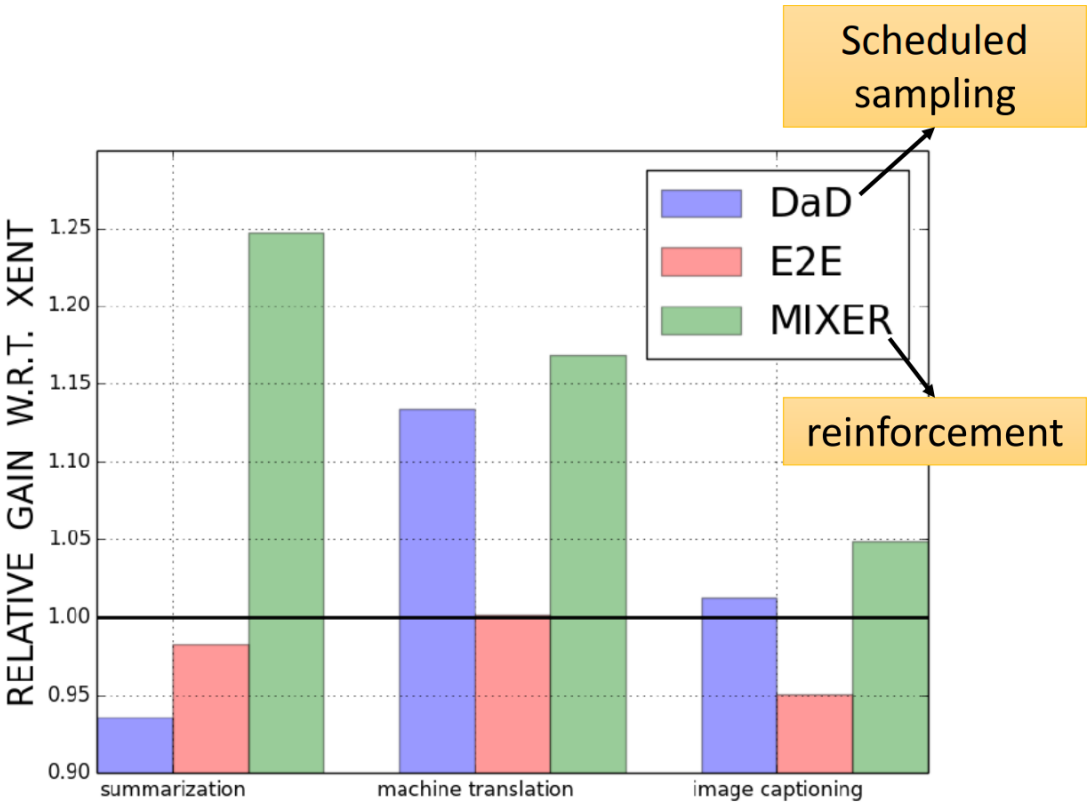
y: generated utterance, $\hat{y}$: ground truth

Reinforcement Learning ?

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

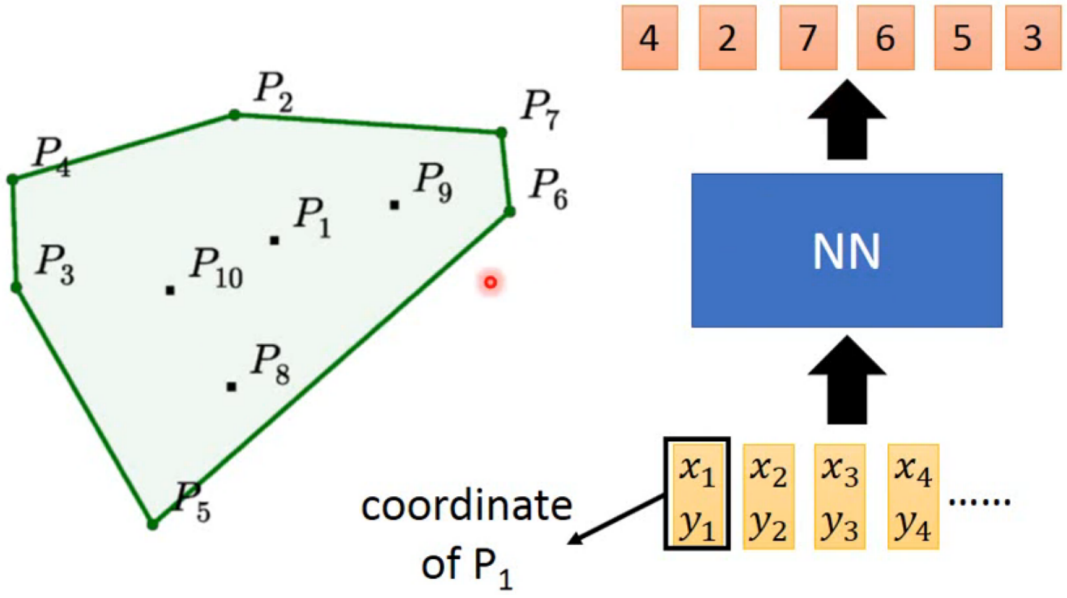


Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, Wojciech Zaremba, "Sequence Level Training with Recurrent Neural Networks", ICLR, 2016



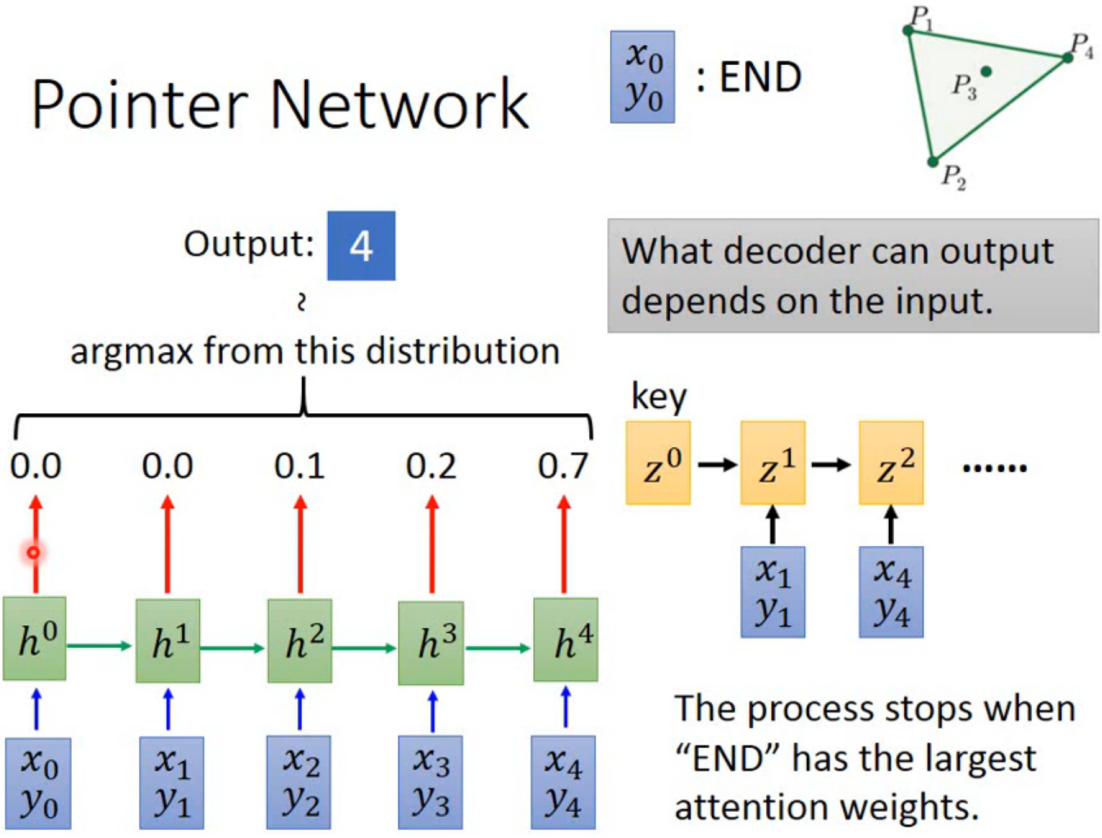
Pointer Network

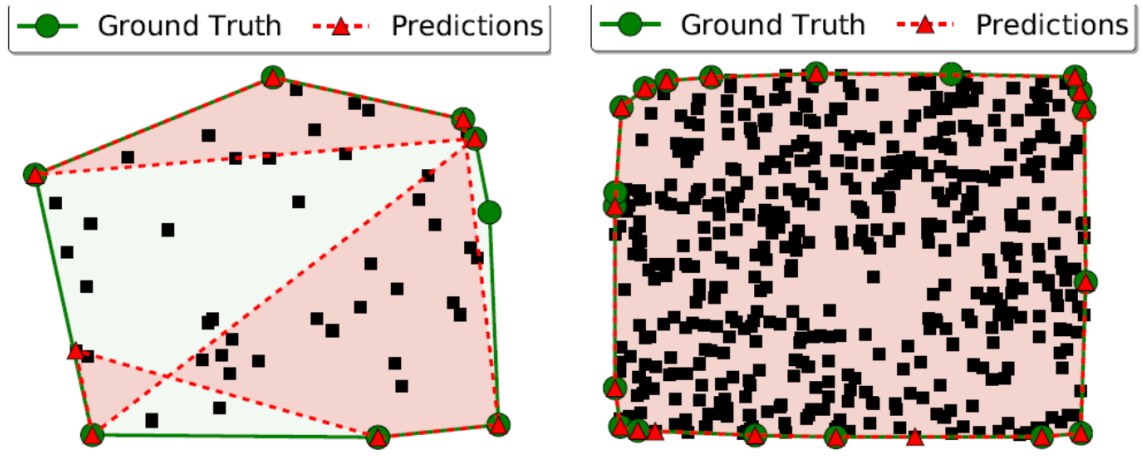
> Oriol Vinyals, Meire Fortunato, Navdeep Jaitly, Pointer Network, NIPS, 2015



Pointer Network can be seen as a **seq2seq** task, however, it can not be trained perfectly because of the size of the output limited to the training set when doing inference.

Add attention to pointer network

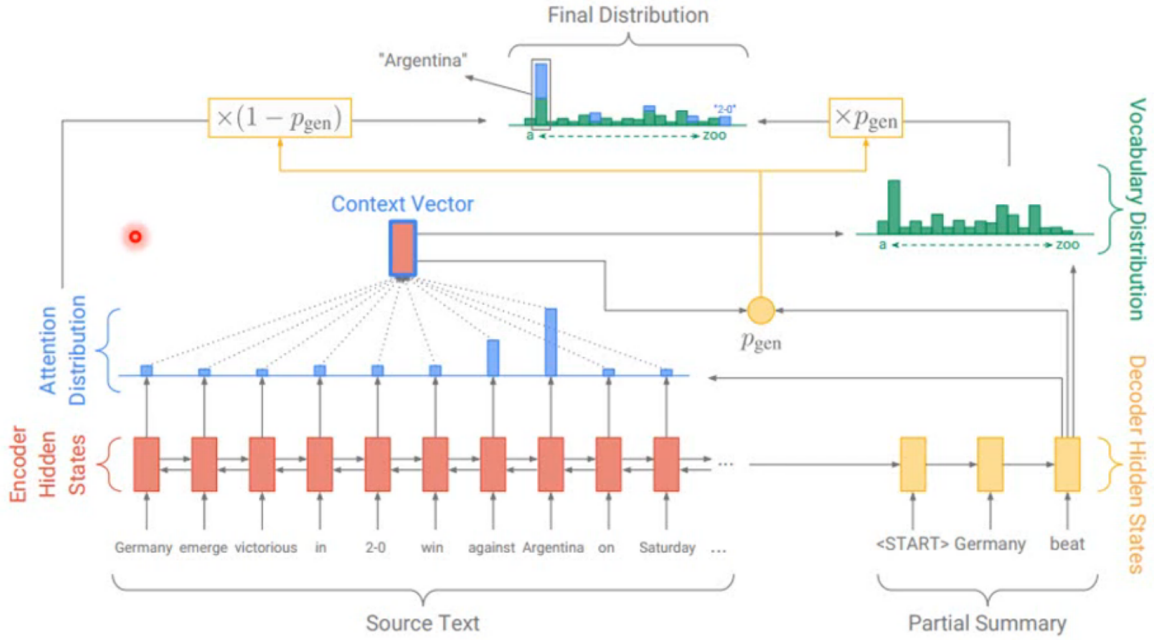




(a) LSTM, m=50, n=50 (d) Ptr-Net, m=5-50, n=500

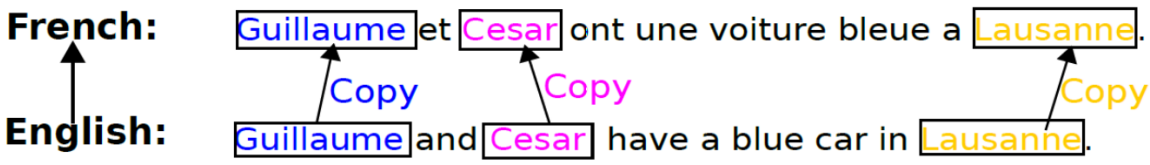
Applications

Summarization



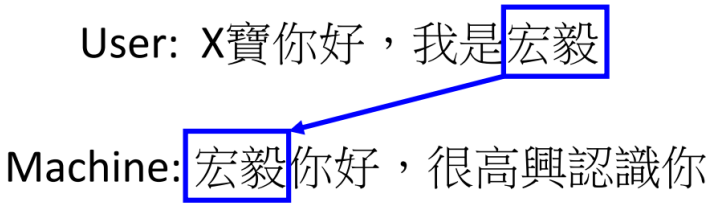
Machine Translation

> Jiatao Gu, Zhengdong Lu, Hang Li, Victor O.K. Li, "Incorporating Copying Mechanism in Sequence-toSequence Learning", ACL, 2016



Chat-bot

> Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, Yoshua Bengio, "Pointing the Unknown Words", ACL, 2016


User: X寶你好，我是宏毅
Machine: 宏毅你好，很高興認識你