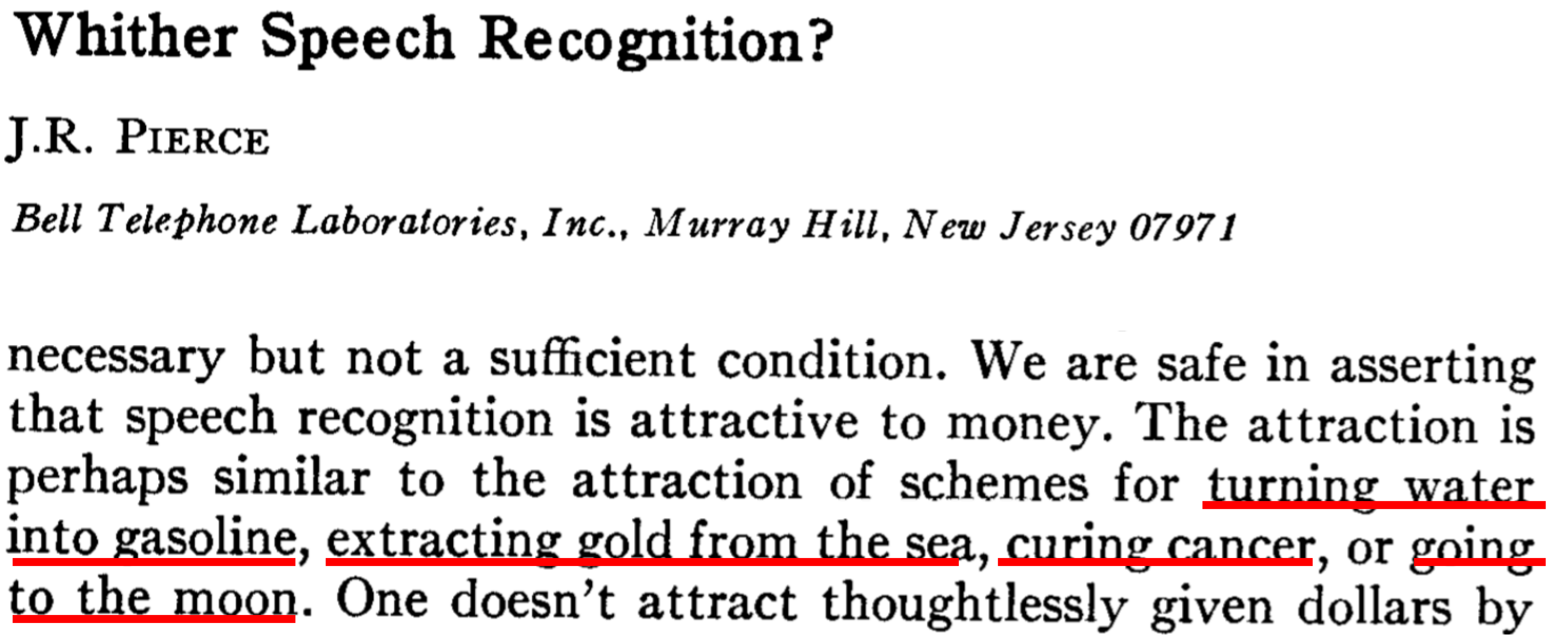
Speech Recognition

HUNG-YI LEE 李宏毅

Speech Recognition is Difficult?



Whither Speech Recognition?

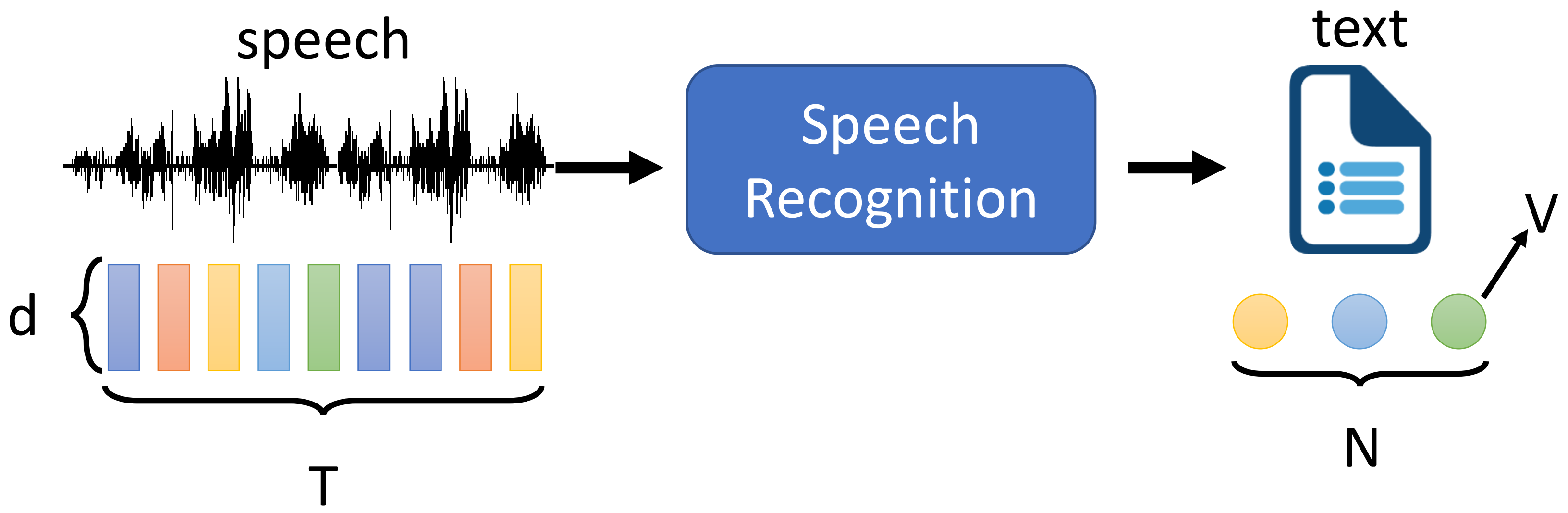
J.R. Pierce

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07971

necessary but not a sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by

I heard the story from Prof Haizhou Li.

Speech Recognition



Speech: a sequence of vector (length T, dimension d)

Text: a sequence of token (length N, V different tokens)

Usually T > N

Token

Phoneme: a unit of sound

W AH N	P AH N CH	M AE N
one	punch	man

Lexicon: word to phonemes

cat ⟶ K AE T

good ⟶ G UH D

man ⟶ M AE N

one ⟶ W AH N

punch ⟶ P AH N CH

Grapheme: smallest unit of a writing system

Lexicon free!

one_punch_man

N=13, V=26+?

“一”，“拳”，“超”，“人”

N=4, V≈4000

26 English alphabet

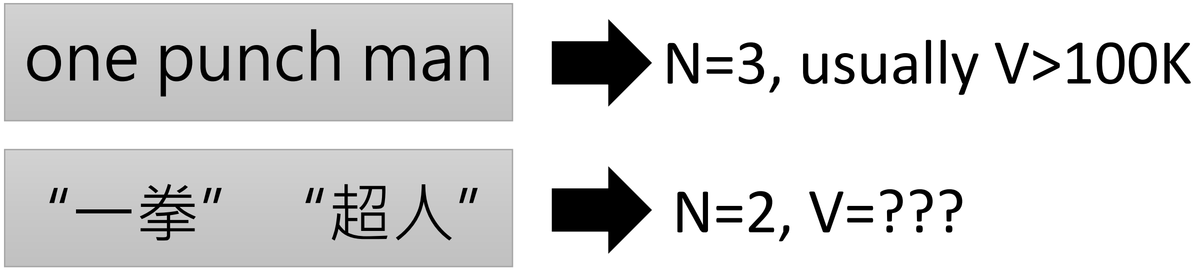
+ { _ } (space)

+ {punctuation marks}

Chinese does not need “space”

Token

Word:



one punch man → N=3, usually V>100K

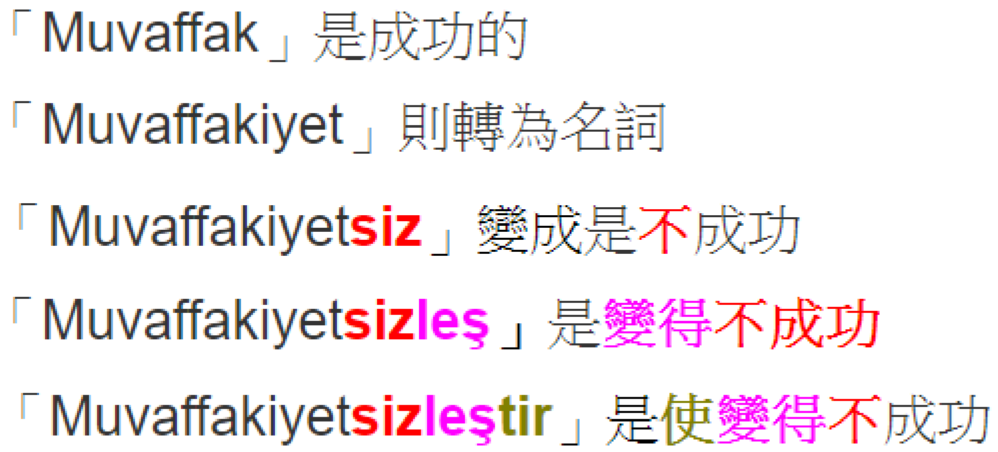
“一拳” “超人” → N=2, V=???

For some languages, V can be too large!

Token

Turkish: Agglutinative language

Source of information: http://tkturkey.com/ (土女時代)



「Muvaffak」是成功的

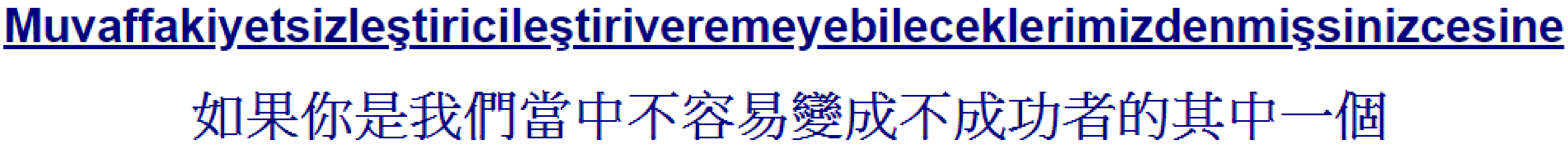
「Muvaffakiyet」則轉為名詞

「Muvaffakiyet**siz**」變成是不成功

「Muvaffakiyet**sizleş**」是變得不成功

「Muvaffakiyet**sizleştir**」是使變得不成功

70 characters?!



Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsinizcesine

如果你是我們當中不容易變成不成功者的其中一個

Token

Word:

one punch man → N=3, usually V>100K

“一拳”　“超人” → N=2, V=???

For some languages, V can be too large!

Morpheme: the smallest meaningful unit (< word, > grapheme)

unbreakable → “un” “break” “able”

rekillable → “re” “kill” “able”

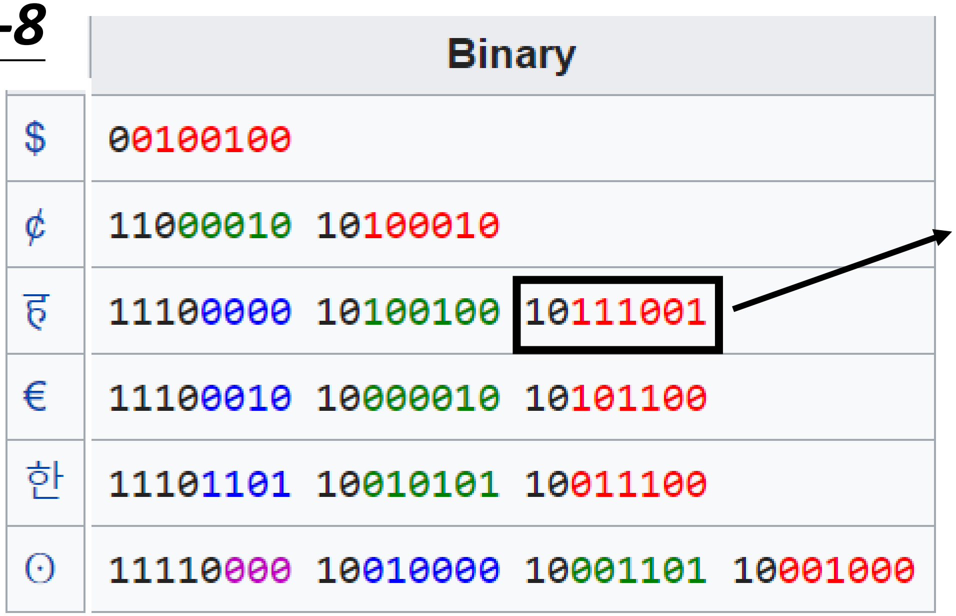
What are the morphemes in a language?

linguistic or statistic

Token

Bytes (!): The system can be **language independent**!

UTF-8



	Binary
$	00100100
¢	11000010 10100010
ह	11100000 10100100 10111001
€	11100010 10000010 10101100
한	11101101 10010101 10011100
𐍈	11110000 10010000 10001101 10001000

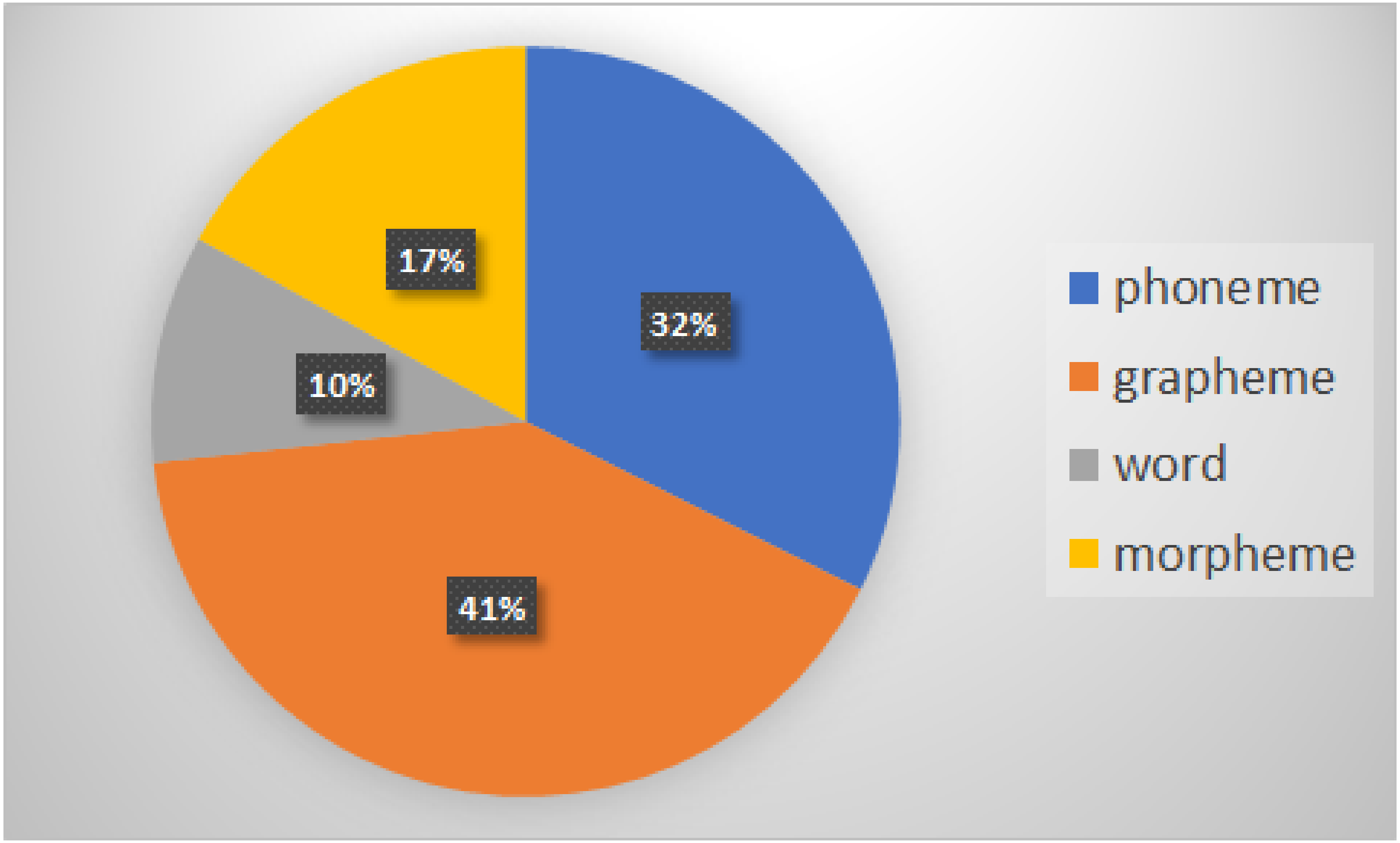
V is always 256

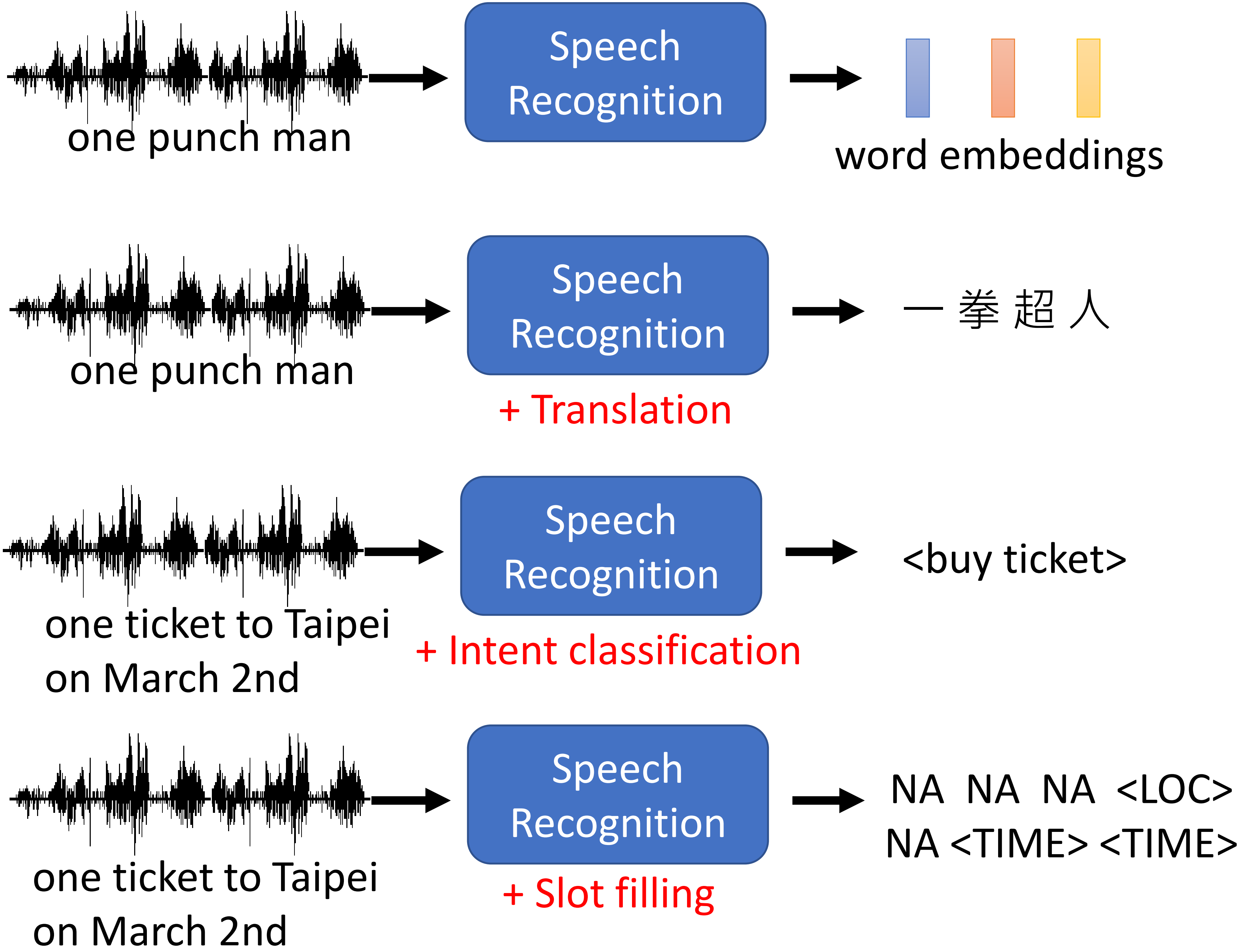
[Li, et al., ICASSP'19]

Token

Go through more than 100 papers in INTERSPEECH’19, ICASSP’19, ASRU’19

感謝助教群的辛勞





one punch man

Speech Recognition

word embeddings

one punch man

Speech Recognition

+ Translation

一 拳 超 人

one ticket to Taipei
on March 2nd

Speech Recognition

+ Intent classification

<buy ticket>

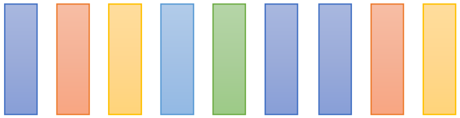
one ticket to Taipei
on March 2nd

Speech Recognition

+ Slot filling

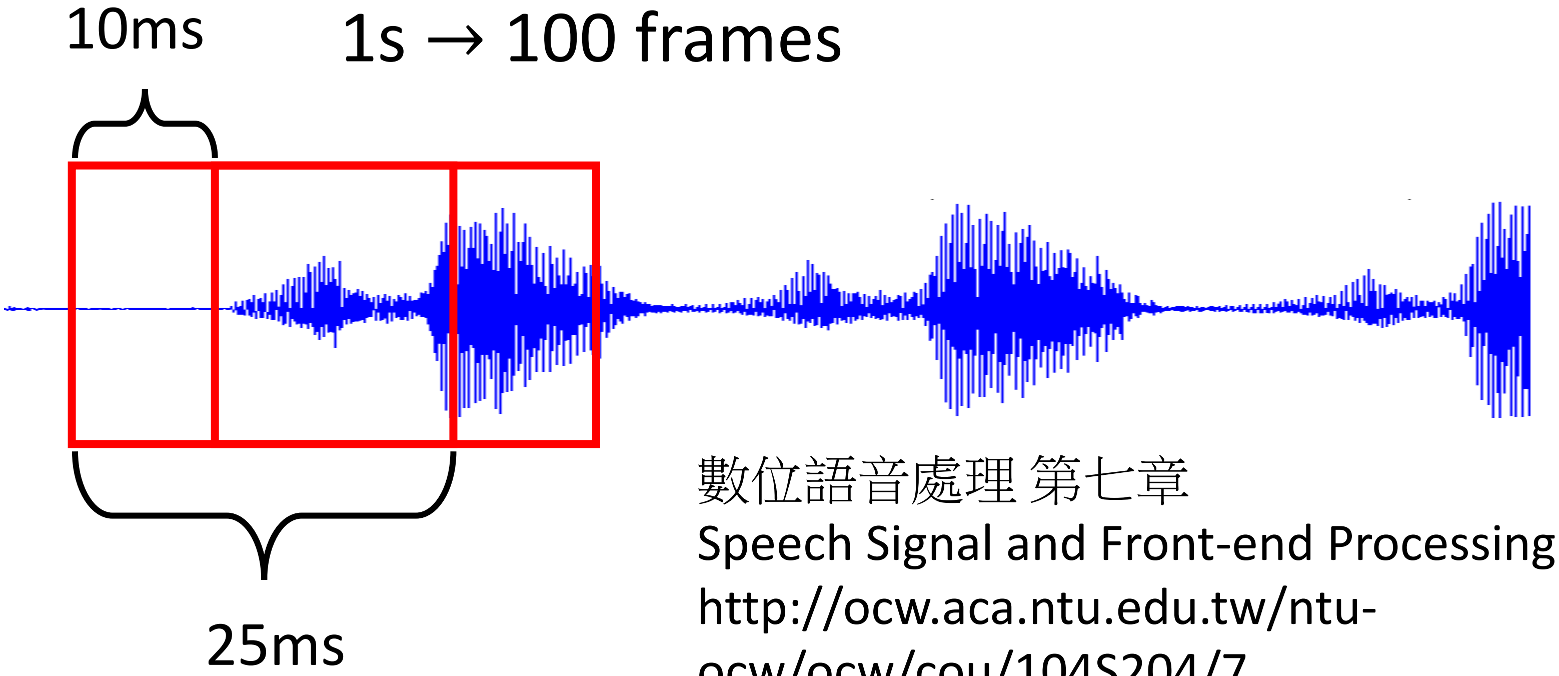
NA NA NA <LOC>
NA <TIME> <TIME>

Acoustic Feature



length T, dimension d

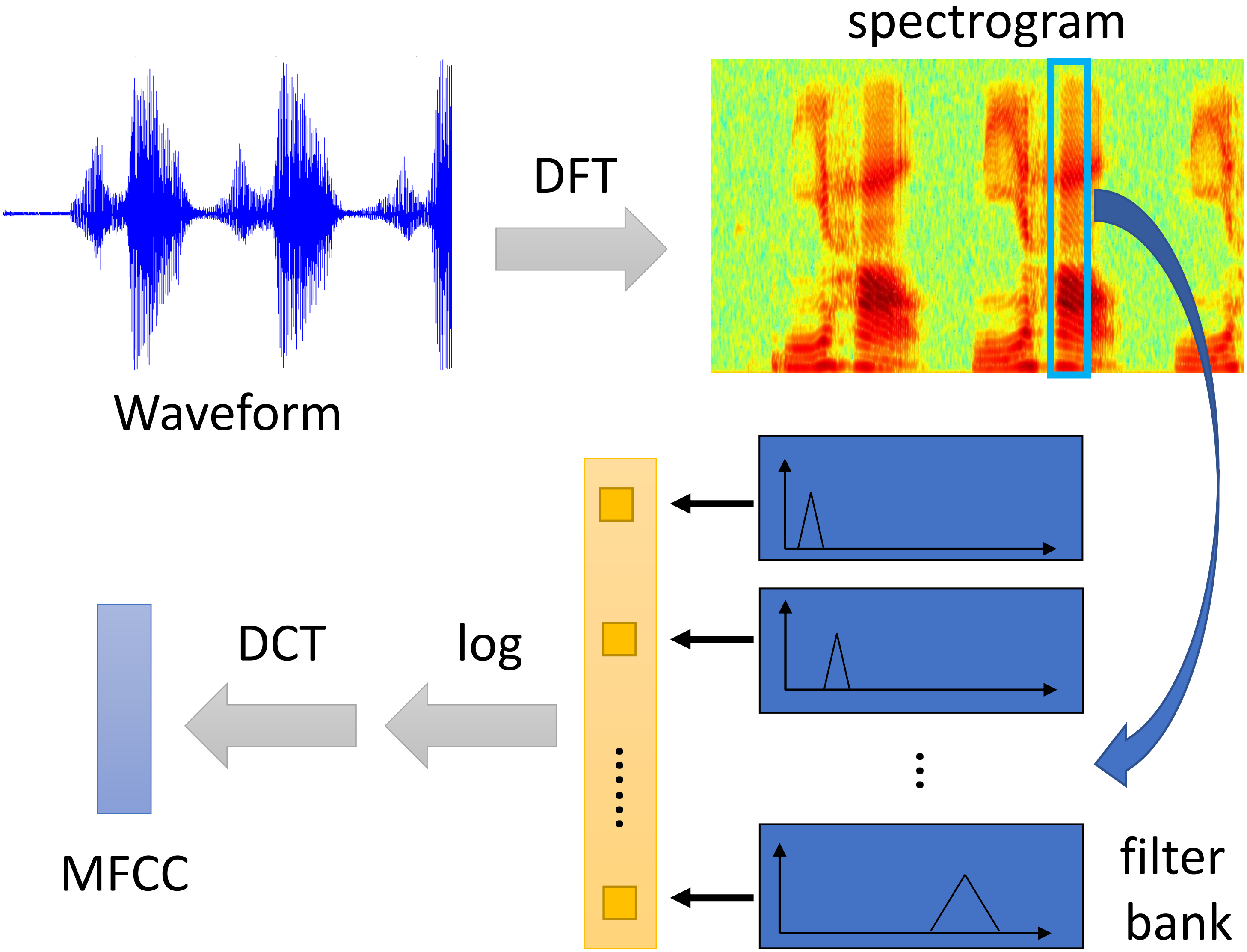
1s → 100 frames



數位語音處理 第七章
Speech Signal and Front-end Processing
http://ocw.aca.ntu.edu.tw/ntu-ocw/ocw/cou/104S204/7

400 sample points (16KHz)
39-dim MFCC
80-dim filter bank output

Acoustic Feature



spectrogram

DFT

Waveform

filter bank

log

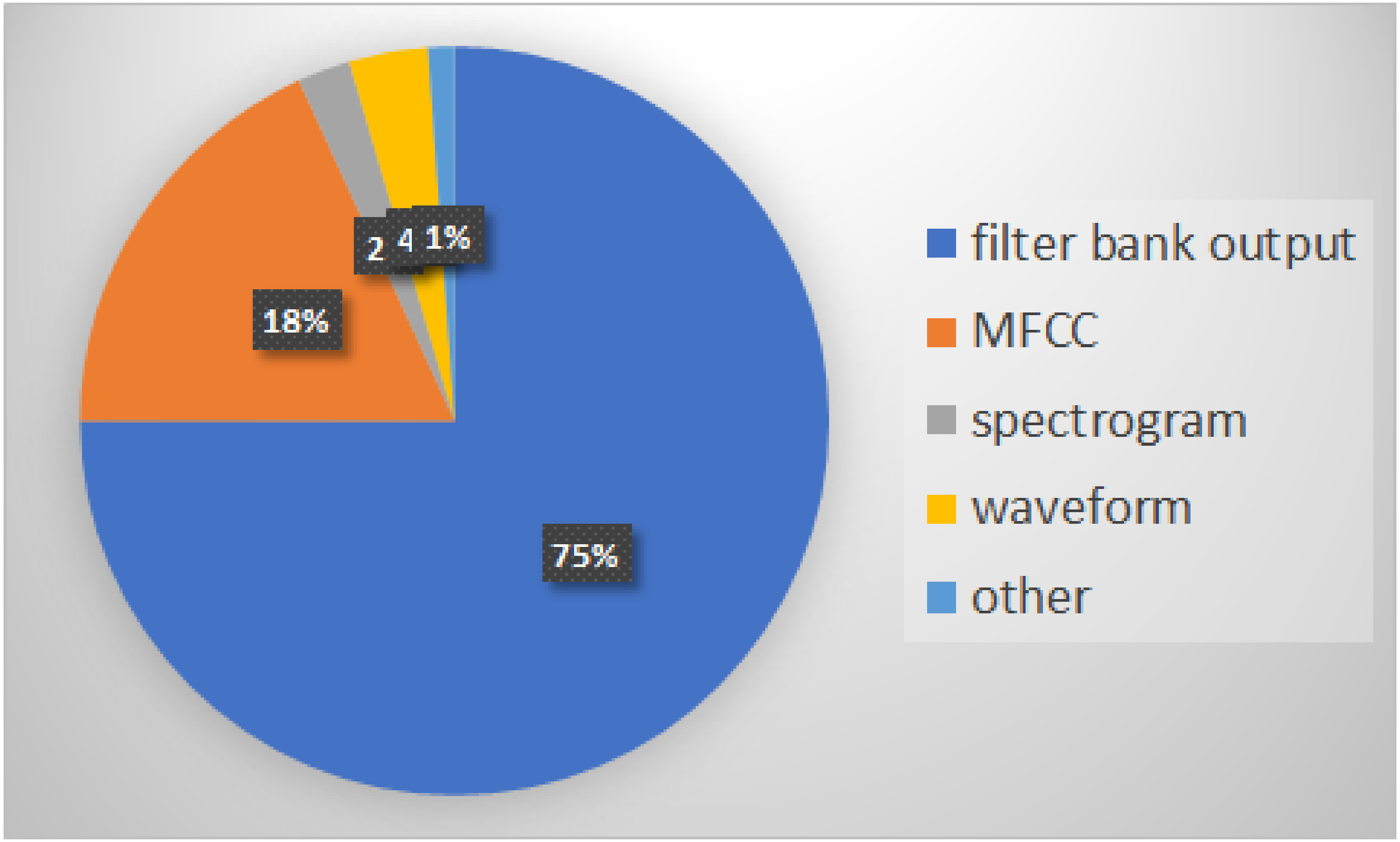
DCT

MFCC

Acoustic Feature

Go through more than 100 papers in INTERSPEECH’19, ICASSP’19, ASRU’19

感謝助教群的辛勞



How much data do we need?

(English corpora)

MNIST: 28 x 28 x 1 x 60000

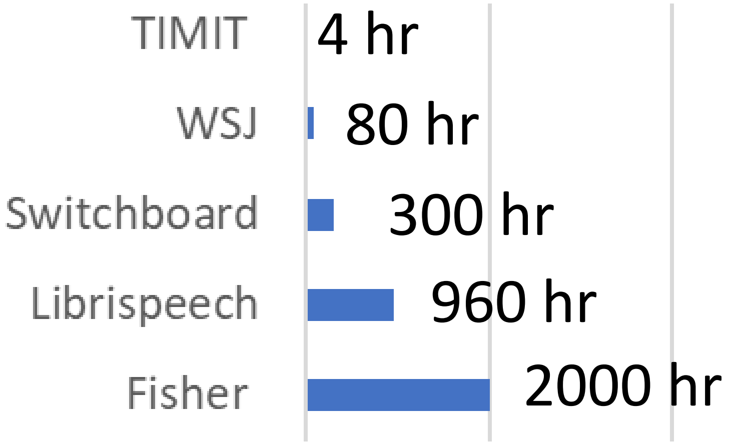
= 47,040,000

= 49 minutes (16kHz)

CIFAR-10: 32 x 32 x 3 x 50000

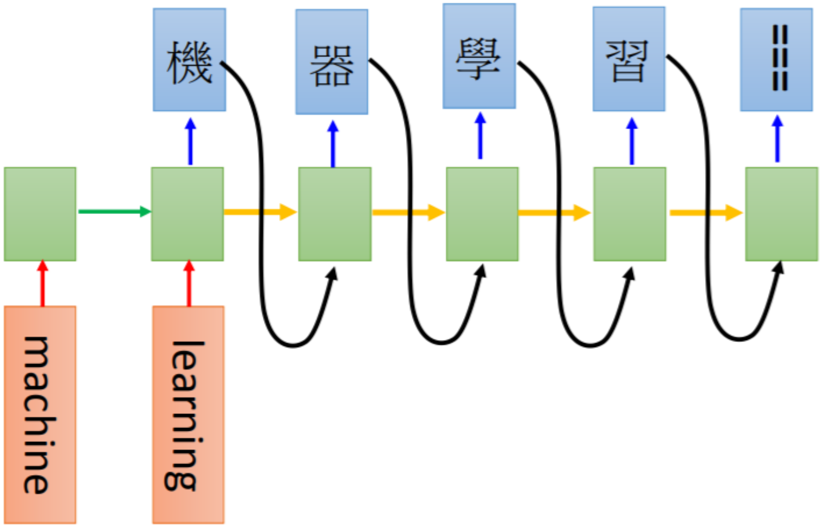
= 153,600,000

= 2 hours 40 minutes

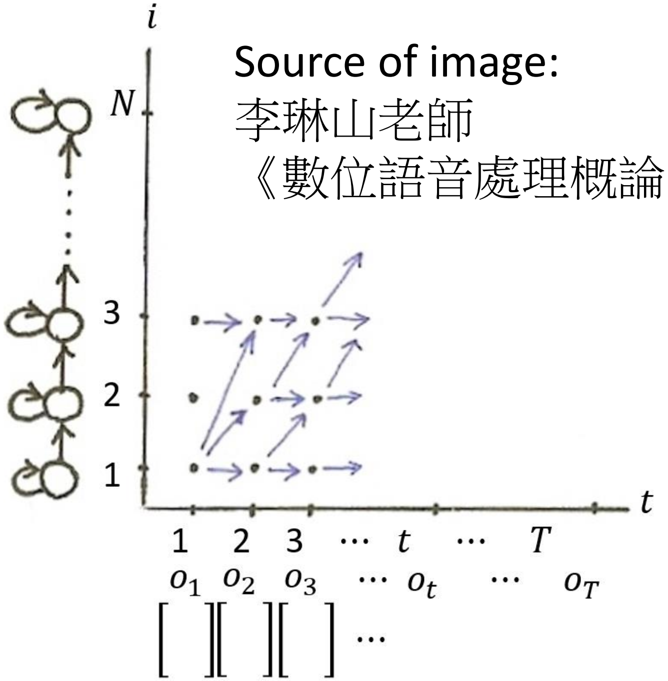


The commercial systems use more than that ……

Two Points of Views



Seq-to-seq



Source of image:
李琳山老師
《數位語音處理概論》

HMM

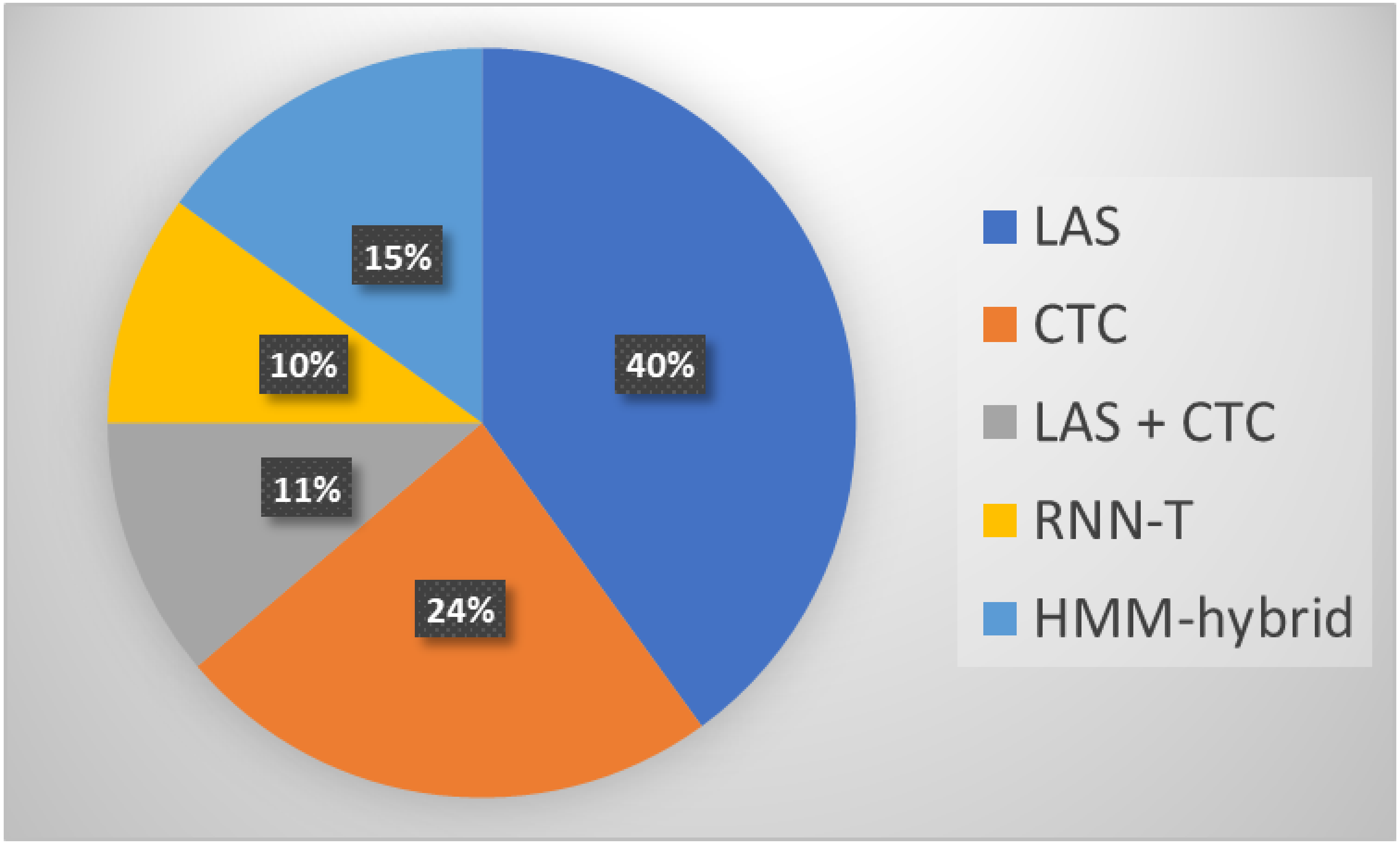
Models to be introduced

- Listen, Attend, and Spell (LAS) [Chorowski. et al., NIPS'15]
- Connectionist Temporal Classification (CTC) [Graves, et al., ICML'06]
- RNN Transducer (RNN-T) [Graves, ICML workshop'12]
- Neural Transducer [Jaitly, et al., NIPS'16]
- Monotonic Chunkwise Attention (MoChA) [Chiu, et al., ICLR'18]

Models

Go through more than 100 papers in INTERSPEECH’19, ICASSP’19, ASRU’19

感謝助教群的辛勞



Models to be introduced

Encoder

Decoder

- Listen, Attend, and Spell (LAS) [Chorowski. et al., NIPS’15]

It is the typical seq2seq with attention.

- Connectionist Temporal Classification (CTC)

[Graves, et al., ICML’06]

- RNN Transducer (RNN-T) [Graves, ICML workshop’12]

- Neural Transducer [Jaitly, et al., NIPS’16]

[Chiu, et al., ICLR’18]

- Monotonic Chunkwise Attention (MoChA)

Listen

- Extract content information
- Remove speaker variance, remove noises

output:

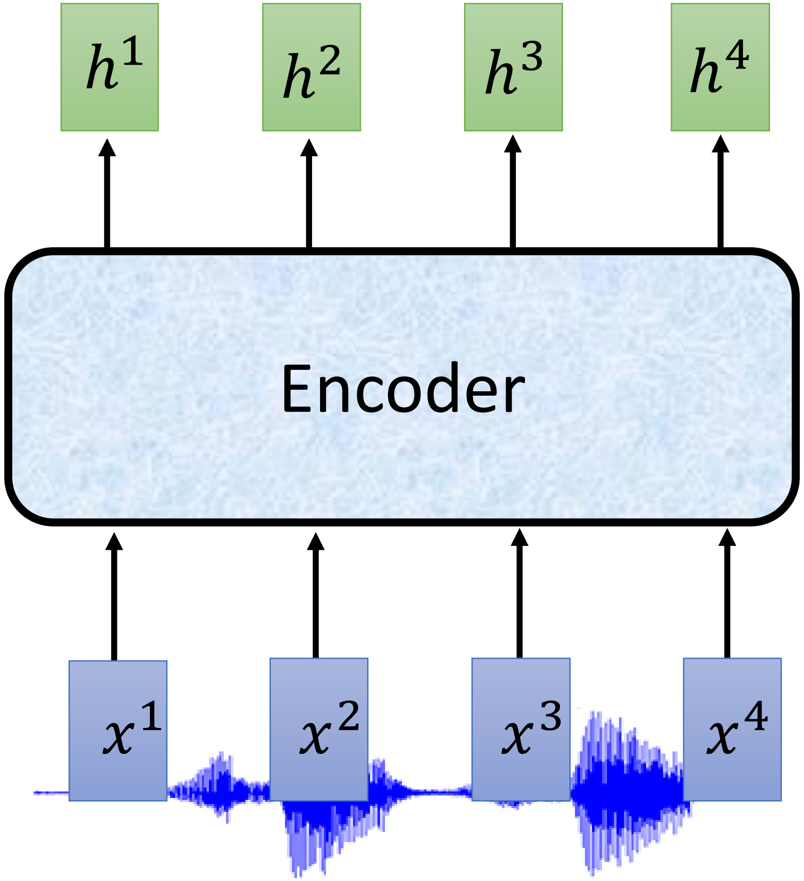
$\{h^1, h^2, \cdots, h^T\}$

high-level representations

Input:

$\{x^1, x^2, \cdots, x^T\}$

acoustic features



Listen

output:

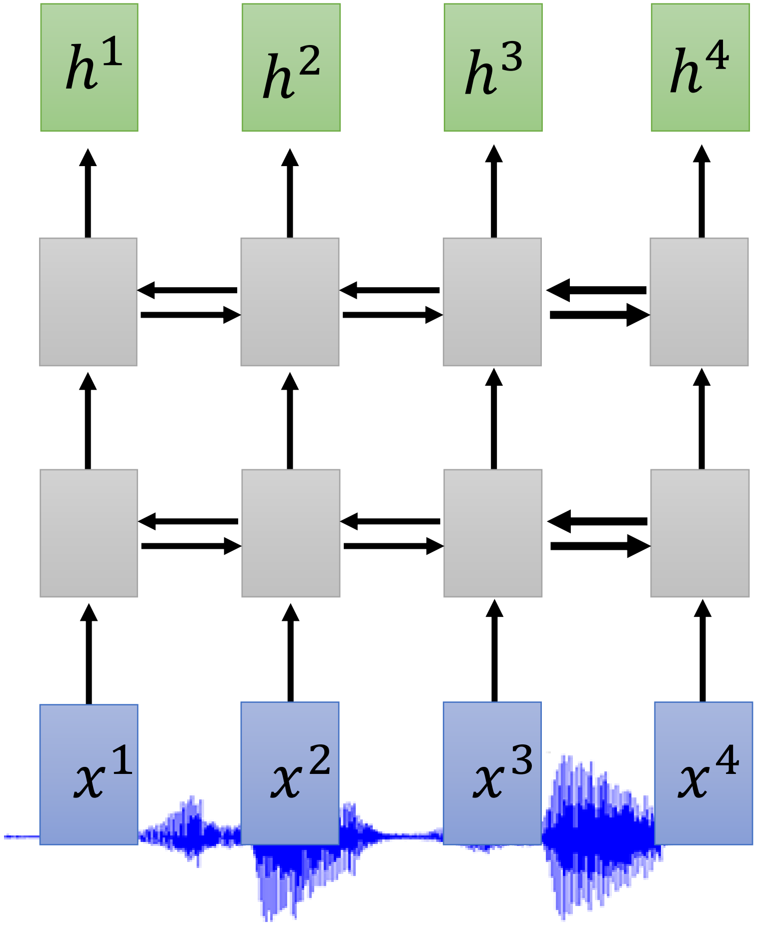
$\{h^1, h^2, \cdots, h^T\}$

high-level representations

Input:

$\{x^1, x^2, \cdots, x^T\}$

acoustic features



RNN

Listen

output:

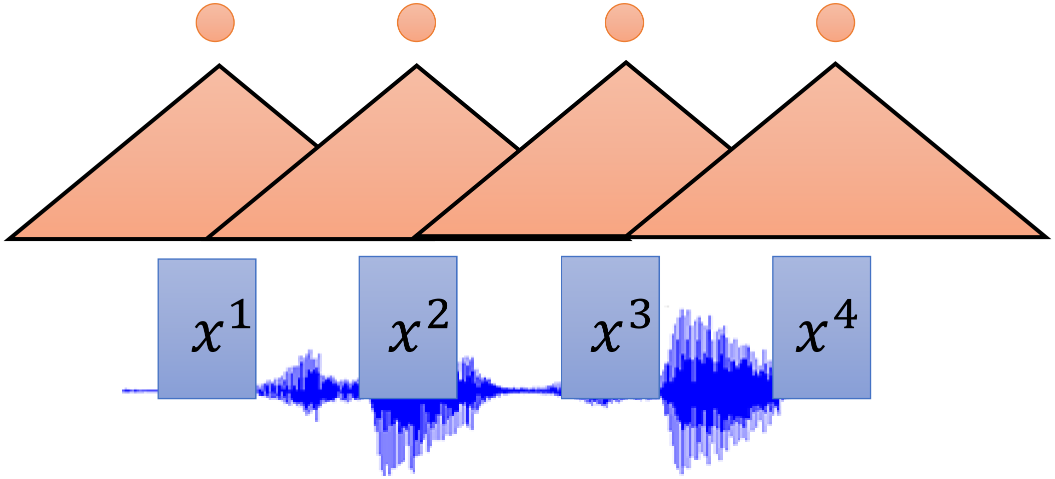
$\{h^1, h^2, \cdots, h^T\}$

high-level representations

Input:

$\{x^1, x^2, \cdots, x^T\}$

acoustic features



1-D CNN

Listen

- Filters in higher layer can consider longer sequence
- CNN+RNN is common

output:

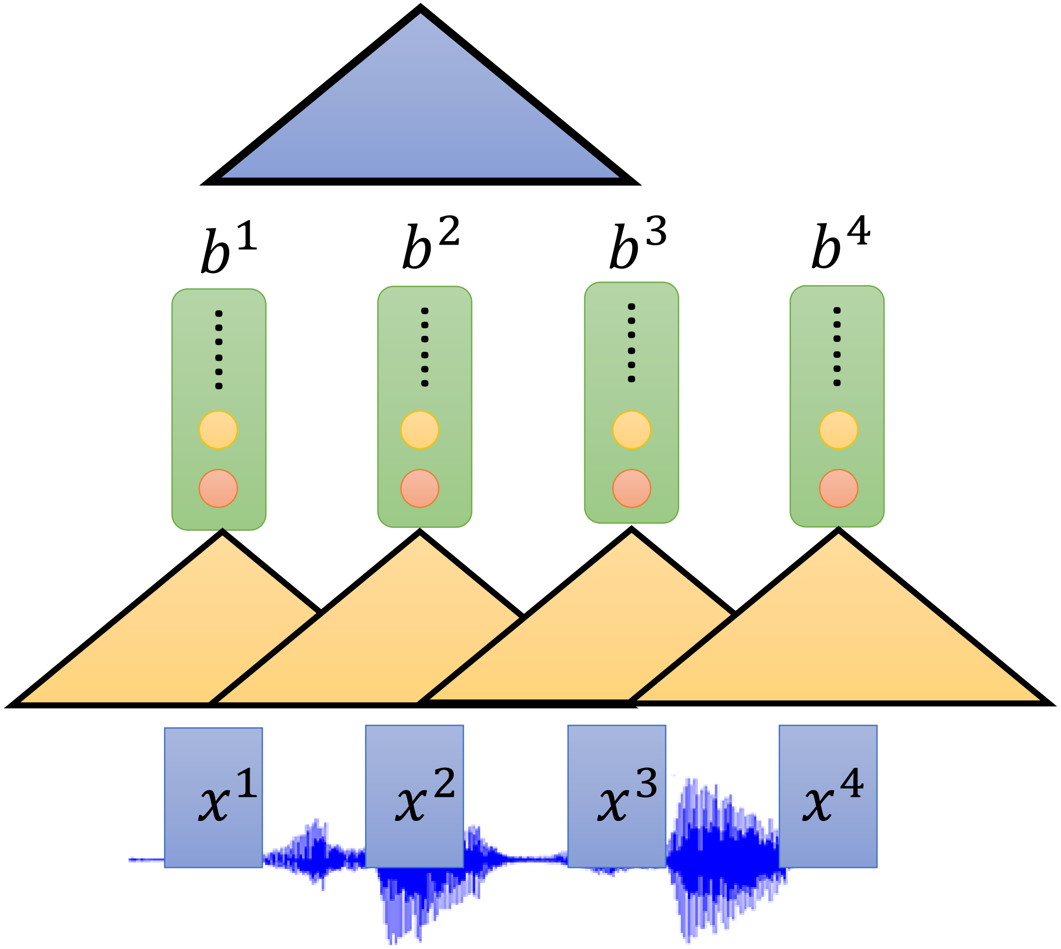
$\{h^1, h^2, \cdots, h^T\}$

high-level representations

Input:

$\{x^1, x^2, \cdots, x^T\}$

acoustic features



1-D CNN

Listen

Please refer to ML video recording: https://www.youtube.com/watch?v=ugWDIIOHtPA

[Zeyer, et al., ASRU'19]

[Karita, et al., ASRU'19]

output:

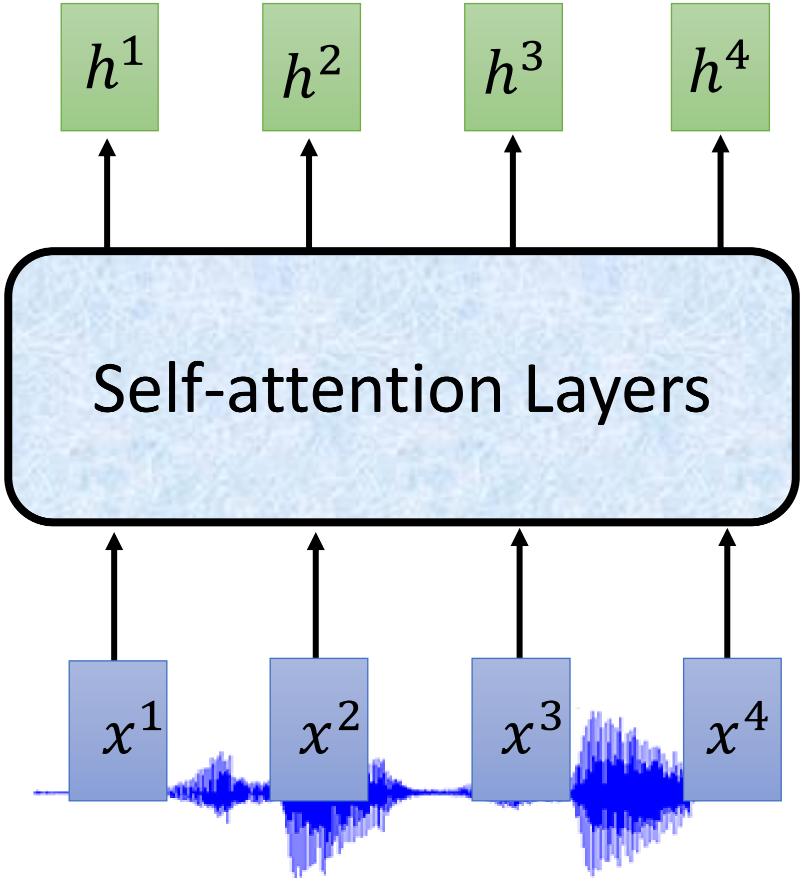
$\{h^1, h^2, \cdots, h^T\}$

high-level representations

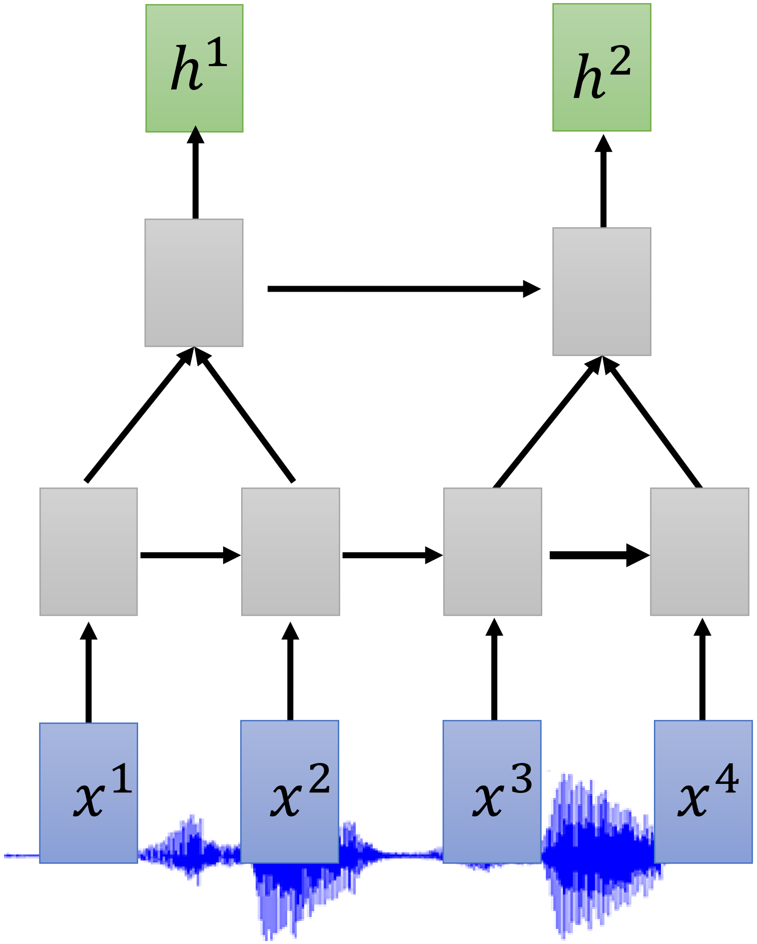
Input:

$\{x^1, x^2, \cdots, x^T\}$

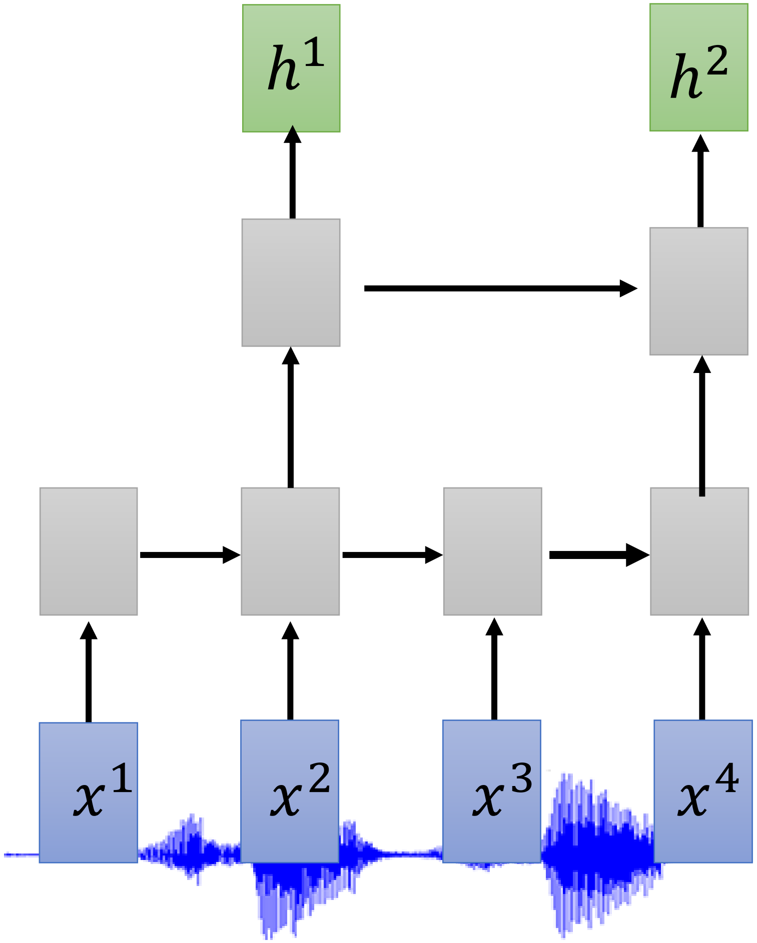
acoustic features



Listen – Down Sampling



Pyramid RNN [Chan, et al., ICASSP'16]



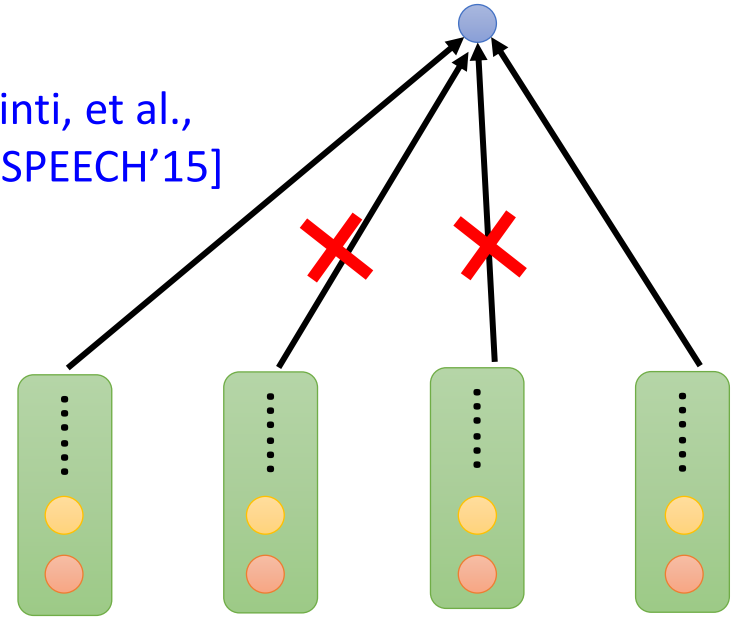
Pooling over time [Bahdanau. et al., ICASSP'16]

Listen – Down Sampling

[Yeh, et al., arXiv'19]

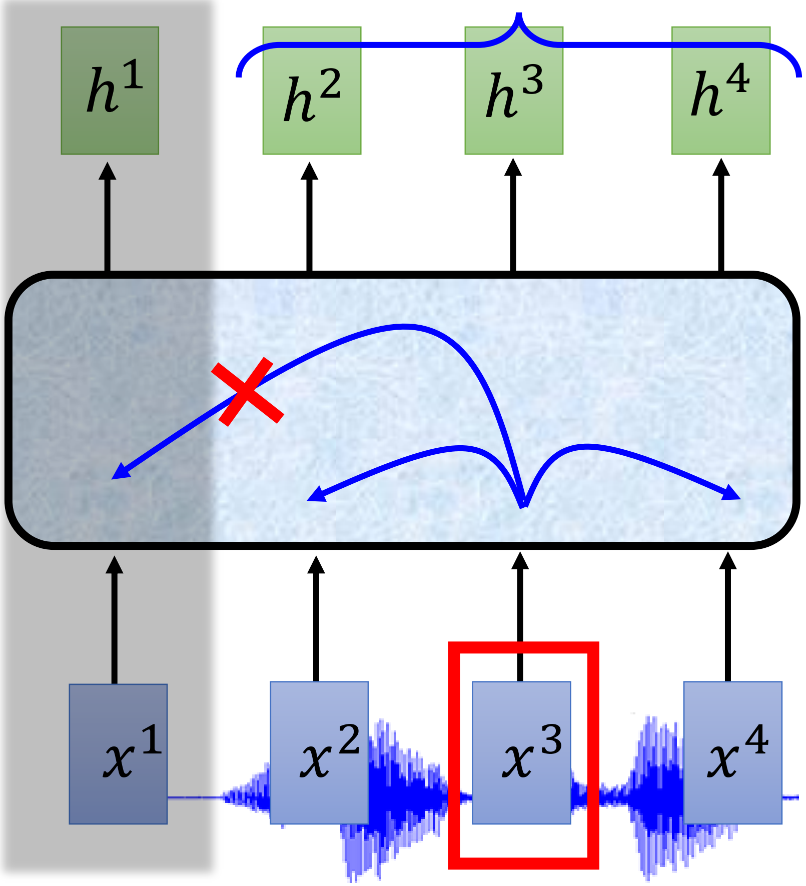
Dilated CNN has the same concept

[Peddinti, et al., INTERSPEECH'15]



Time-delay DNN (TDNN)

Attention in a range

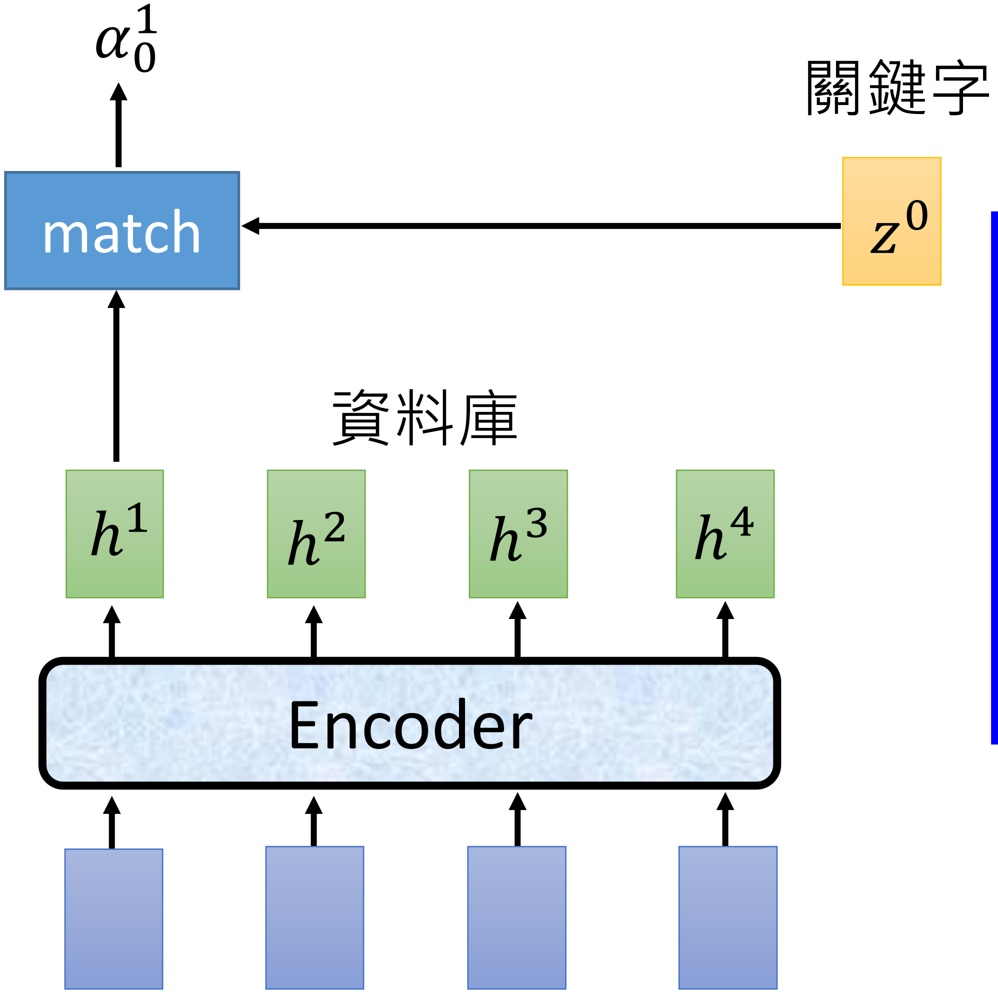


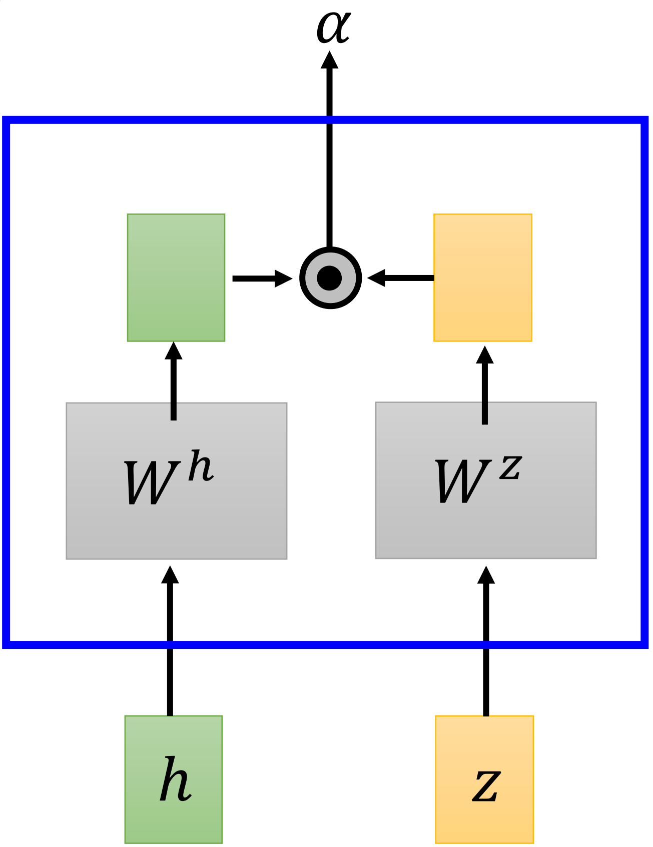
Truncated Self-attention

Attention

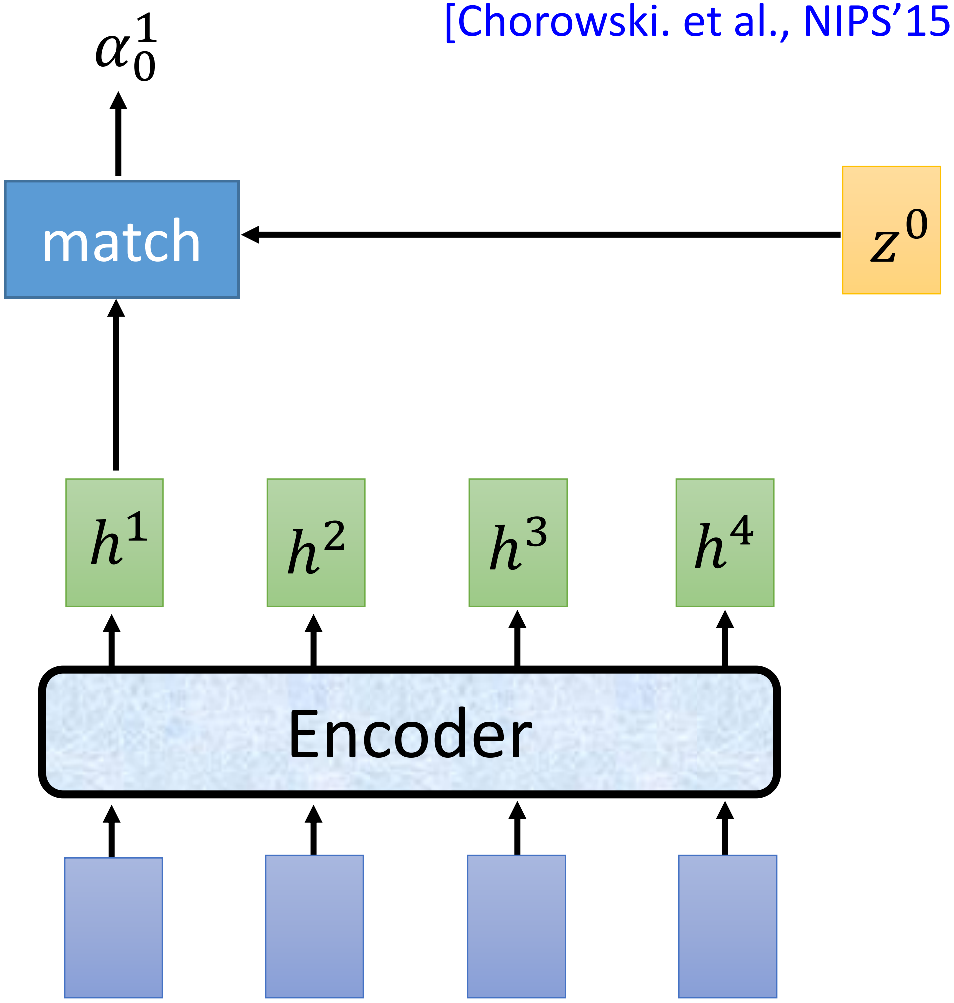
Dot-product Attention

[Chan, et al., ICASSP'16]



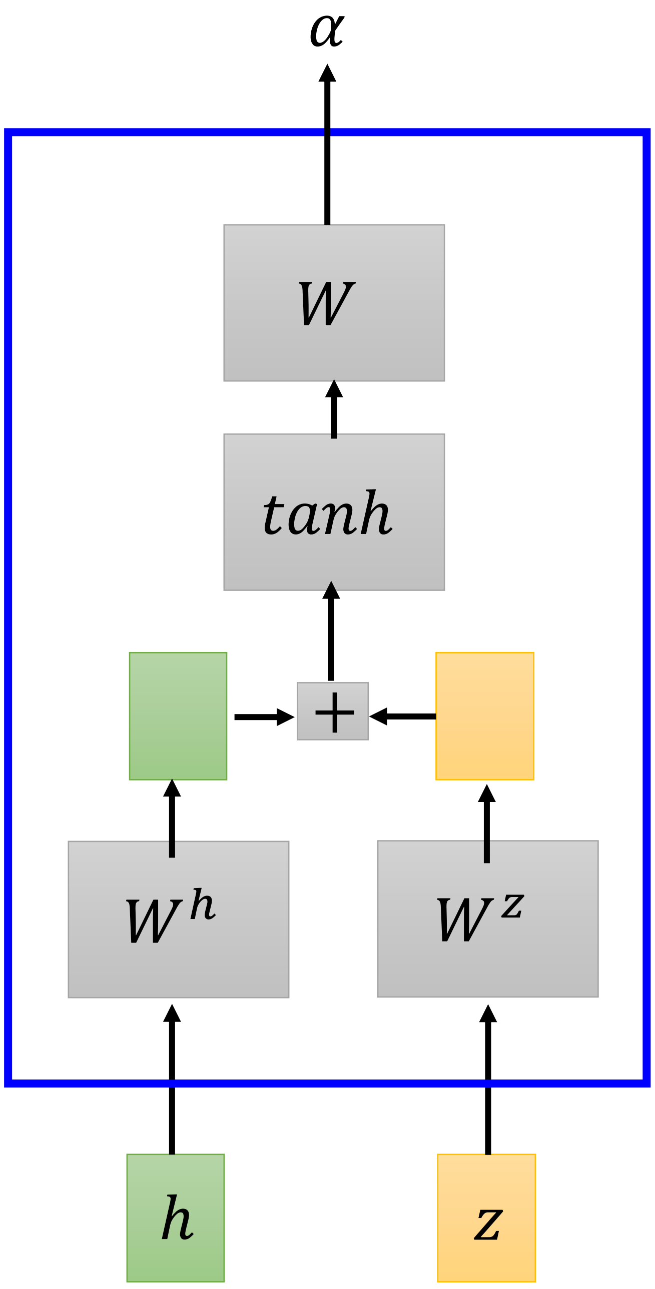


Attention

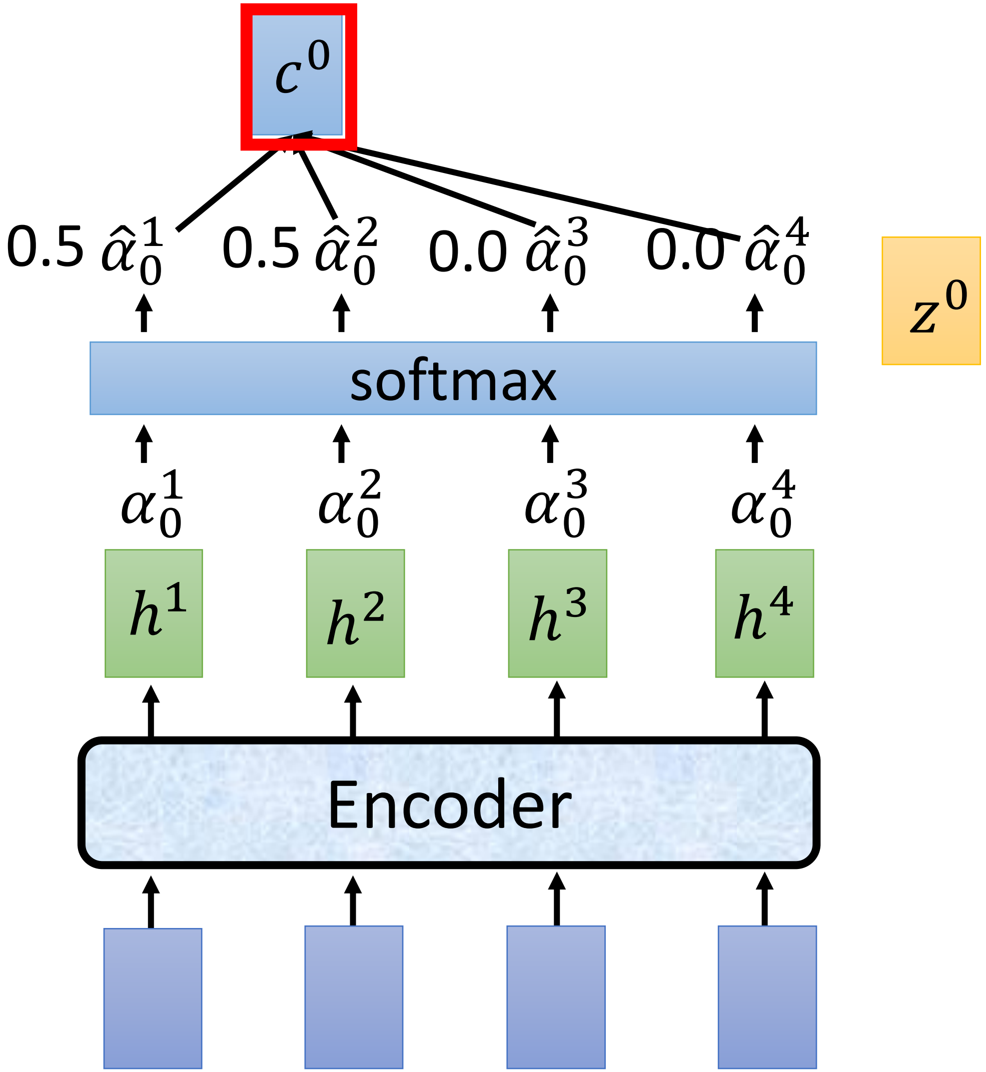


Additive Attention

[Chorowski. et al., NIPS'15]



Attention

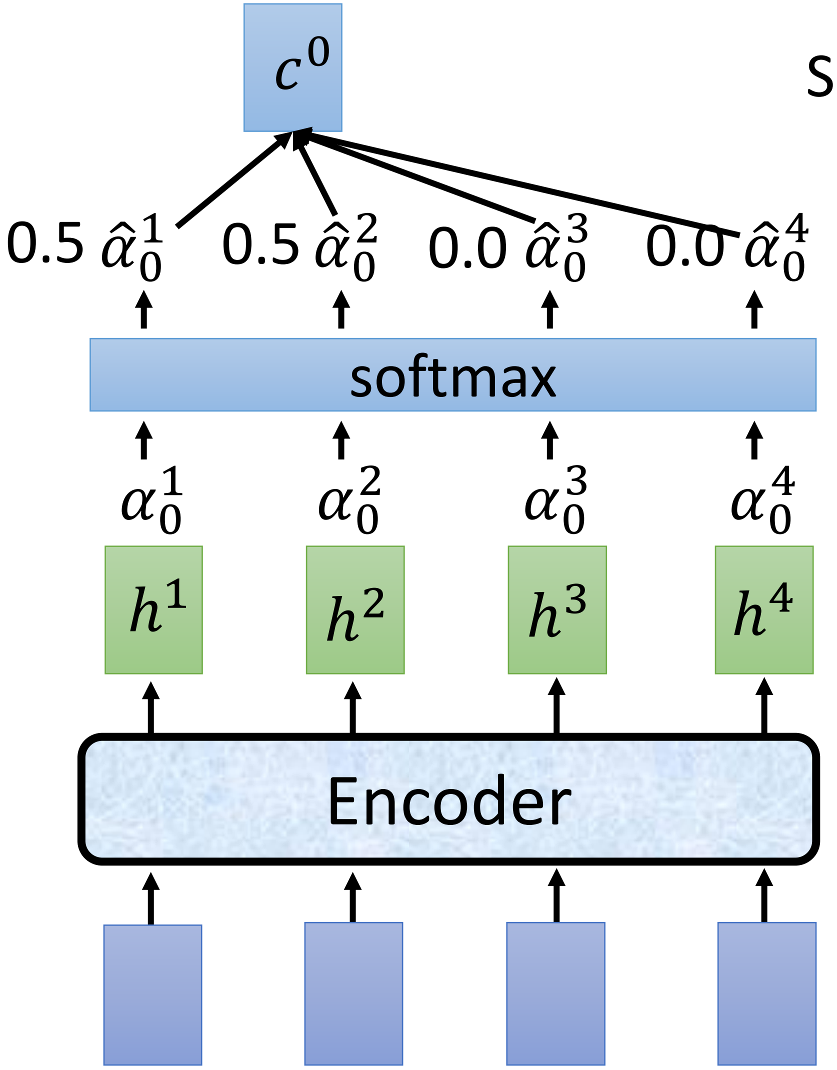


c^0 As RNN input

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

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

Spell



c^0

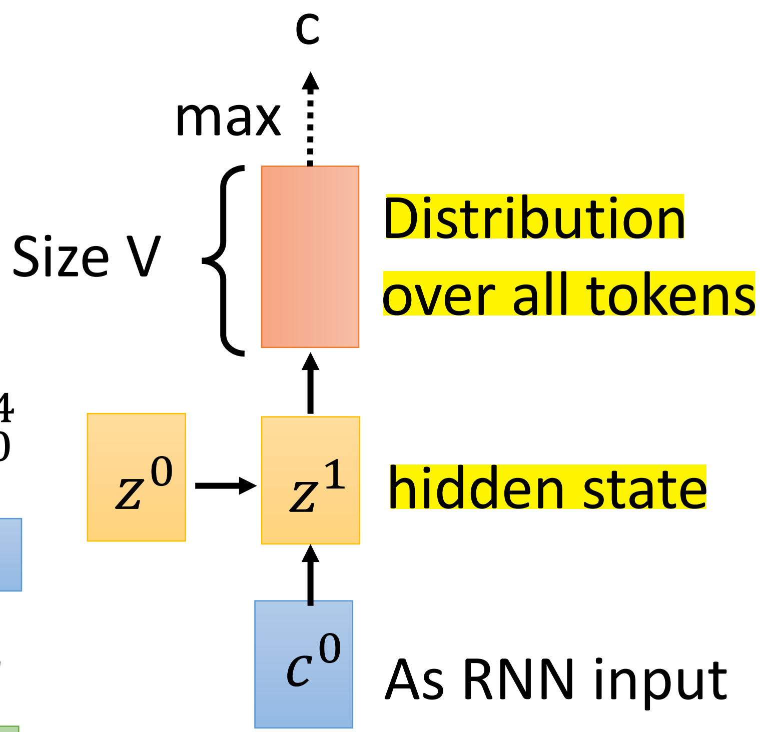
0.5 $\hat{\alpha}_0^1$ 0.5 $\hat{\alpha}_0^2$ 0.0 $\hat{\alpha}_0^3$ 0.0 $\hat{\alpha}_0^4$

softmax

α_0^1 α_0^2 α_0^3 α_0^4

h^1 h^2 h^3 h^4

Encoder



c

max

Size V

Distribution over all tokens

z^0 z^1

hidden state

c^0

As RNN input

a: 0.0

b: 0.1

c: 0.6

d: 0.1

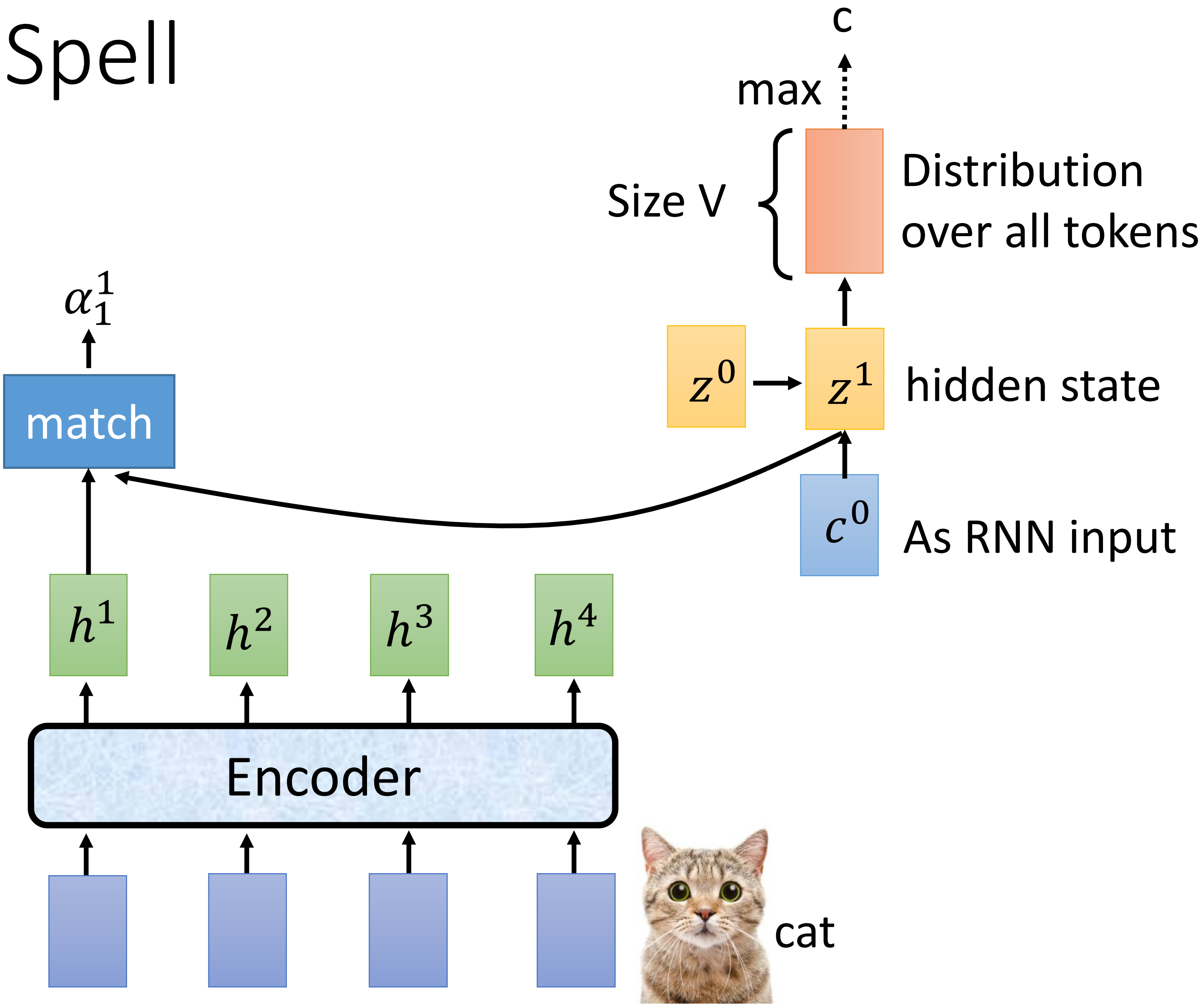
……

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

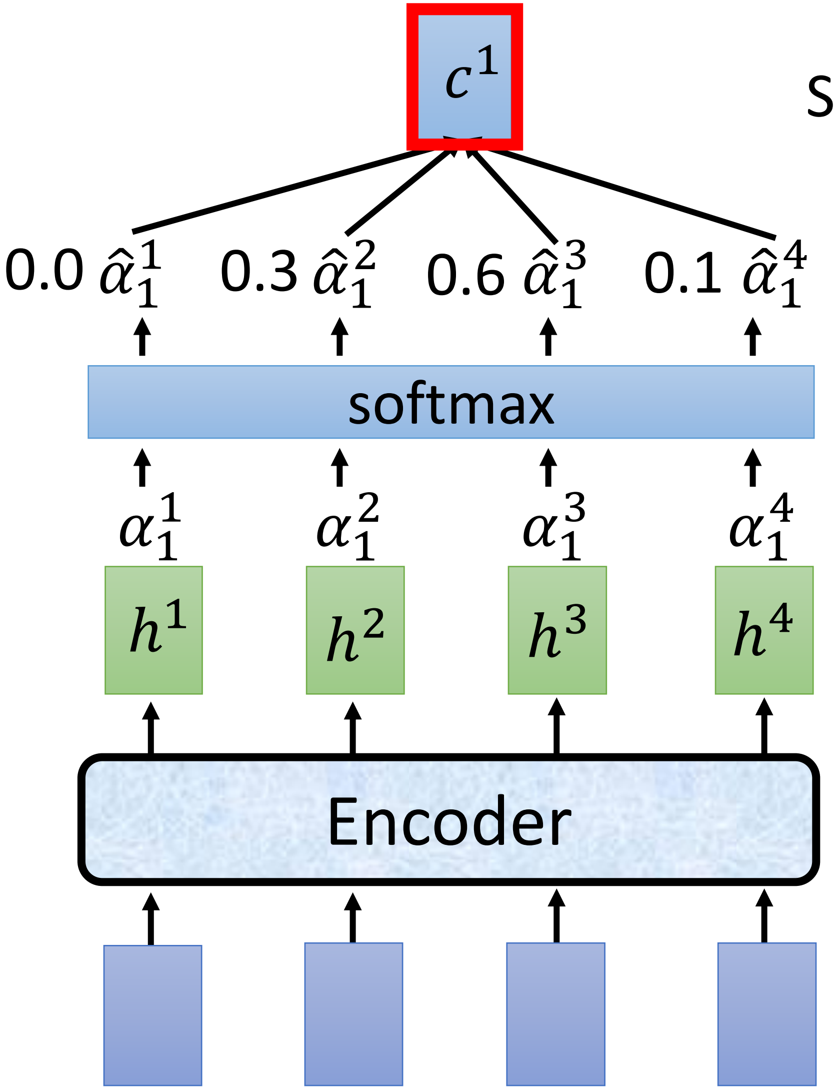
$$= 0.5h^1 + 0.5h^2$$

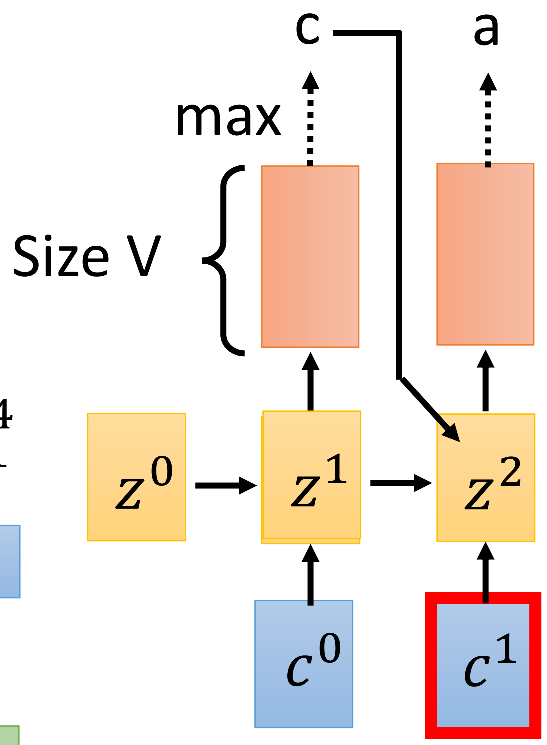
cat

Spell


c
max
Size V
Distribution over all tokens
α_1^1
match
z^0
z^1
hidden state
c^0
As RNN input
h^1
h^2
h^3
h^4
Encoder
cat

Spell





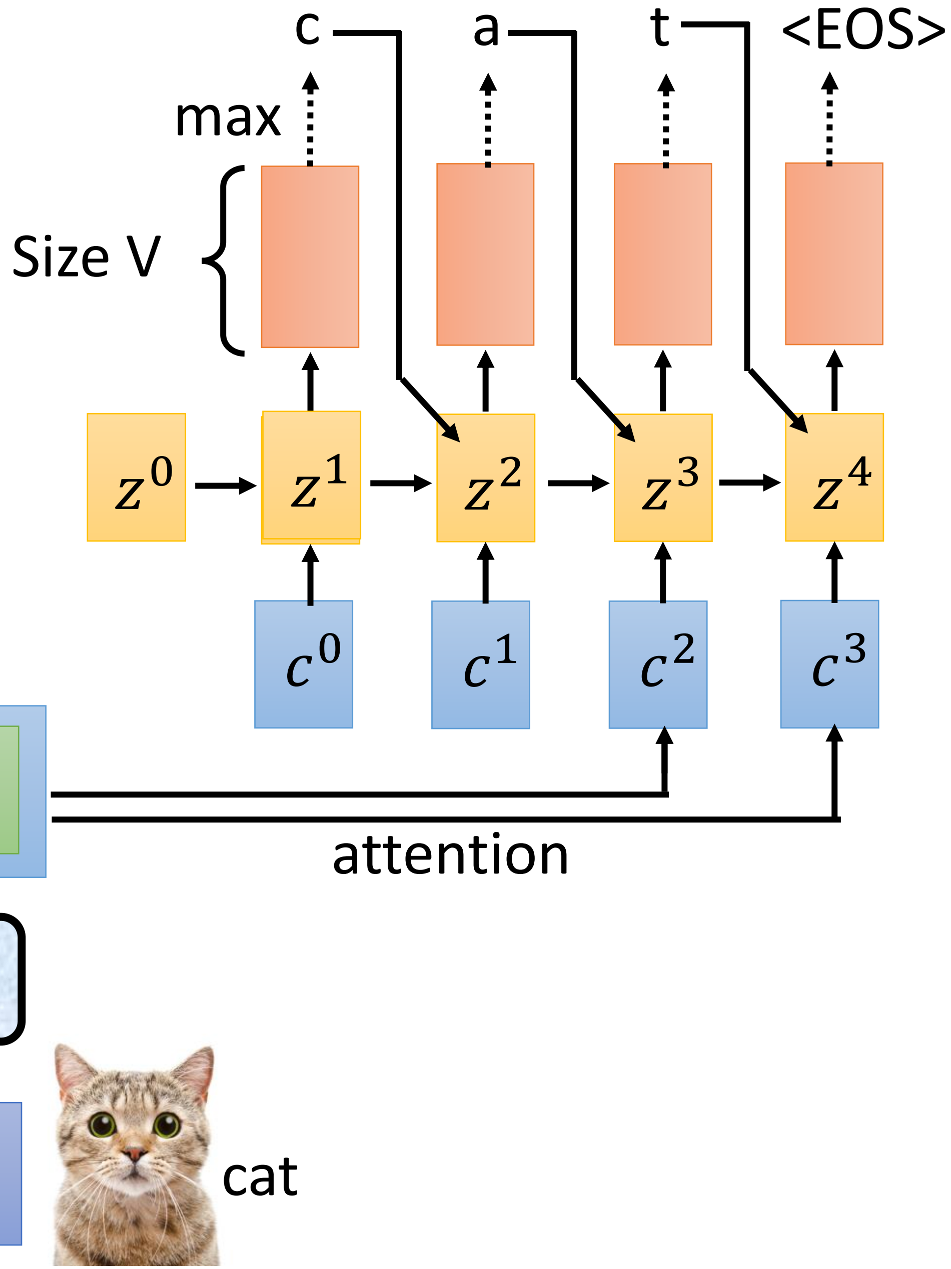
cat

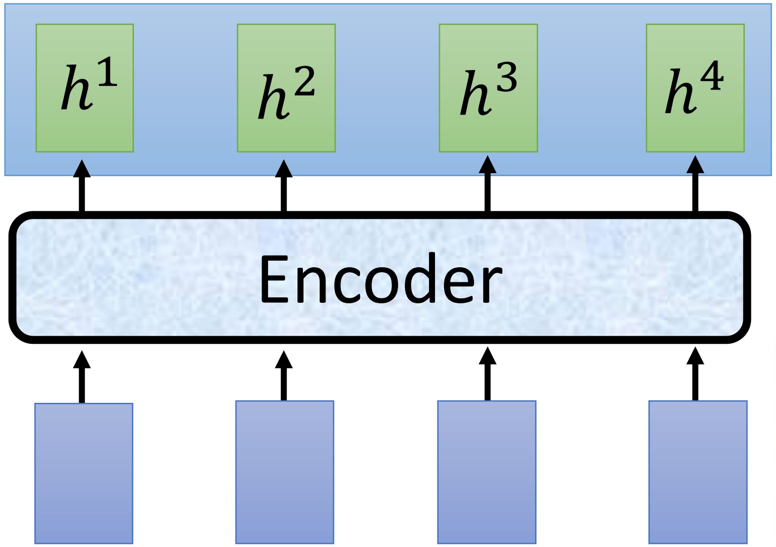
$$c^1 = \sum \hat{\alpha}_1^t h^t$$

$$= 0.3h^2 + 0.6h^3 + 0.1h^4$$

Spell

Beam Search is usually used





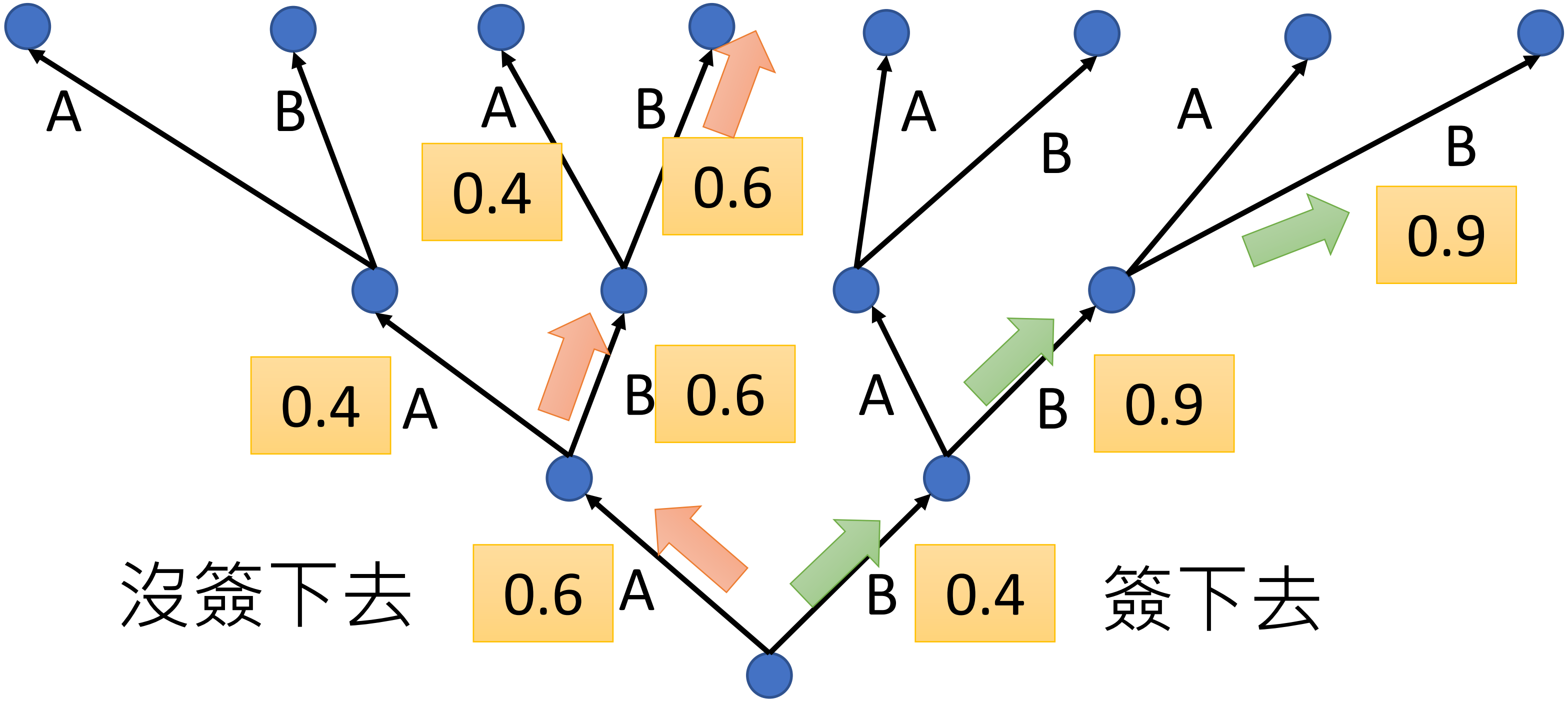
Beam Search

Assume there are only two tokens (V=2).

The red path is ***Greedy Decoding***.

The green path is the best one.

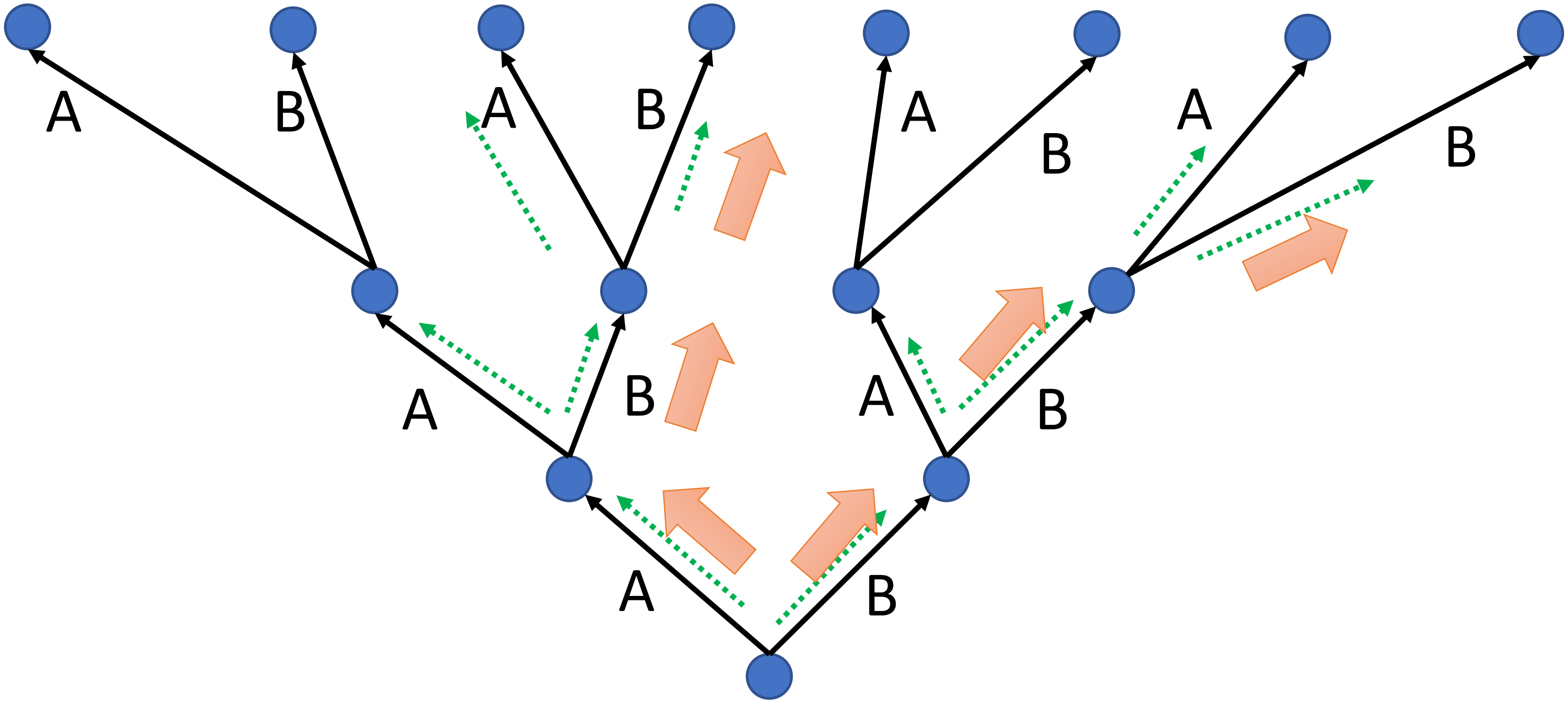
Not possible to check all the paths …



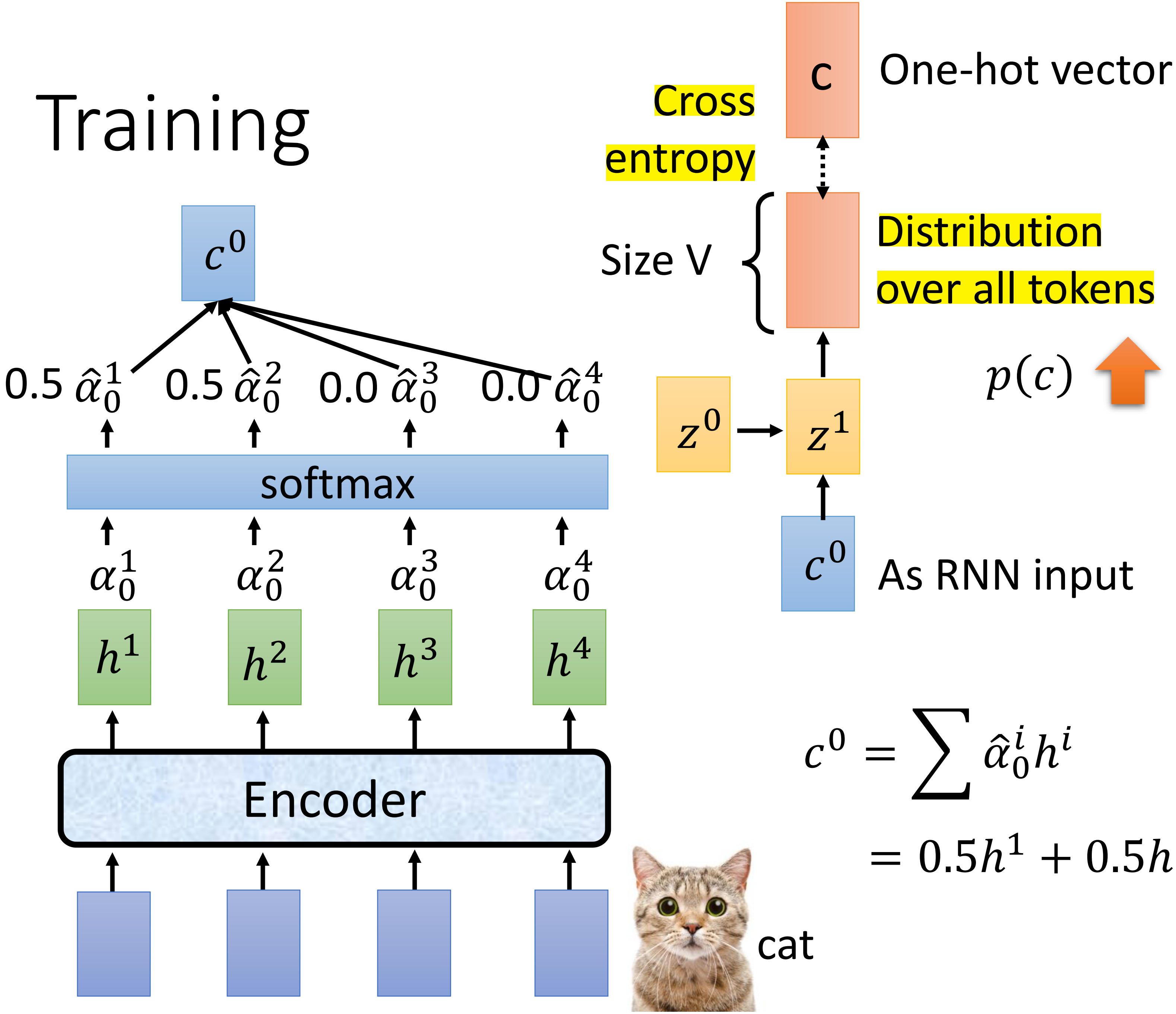
Beam Search

Keep **B** best pathes at each step

B (Beam size) = 2



Training



c^0

0.5 $\hat{\alpha}_0^1$ 0.5 $\hat{\alpha}_0^2$ 0.0 $\hat{\alpha}_0^3$ 0.0 $\hat{\alpha}_0^4$

softmax

α_0^1 α_0^2 α_0^3 α_0^4

h^1 h^2 h^3 h^4

Encoder

cat

c

One-hot vector

Cross entropy

Size V

Distribution over all tokens

$p(c)$

z^0 z^1

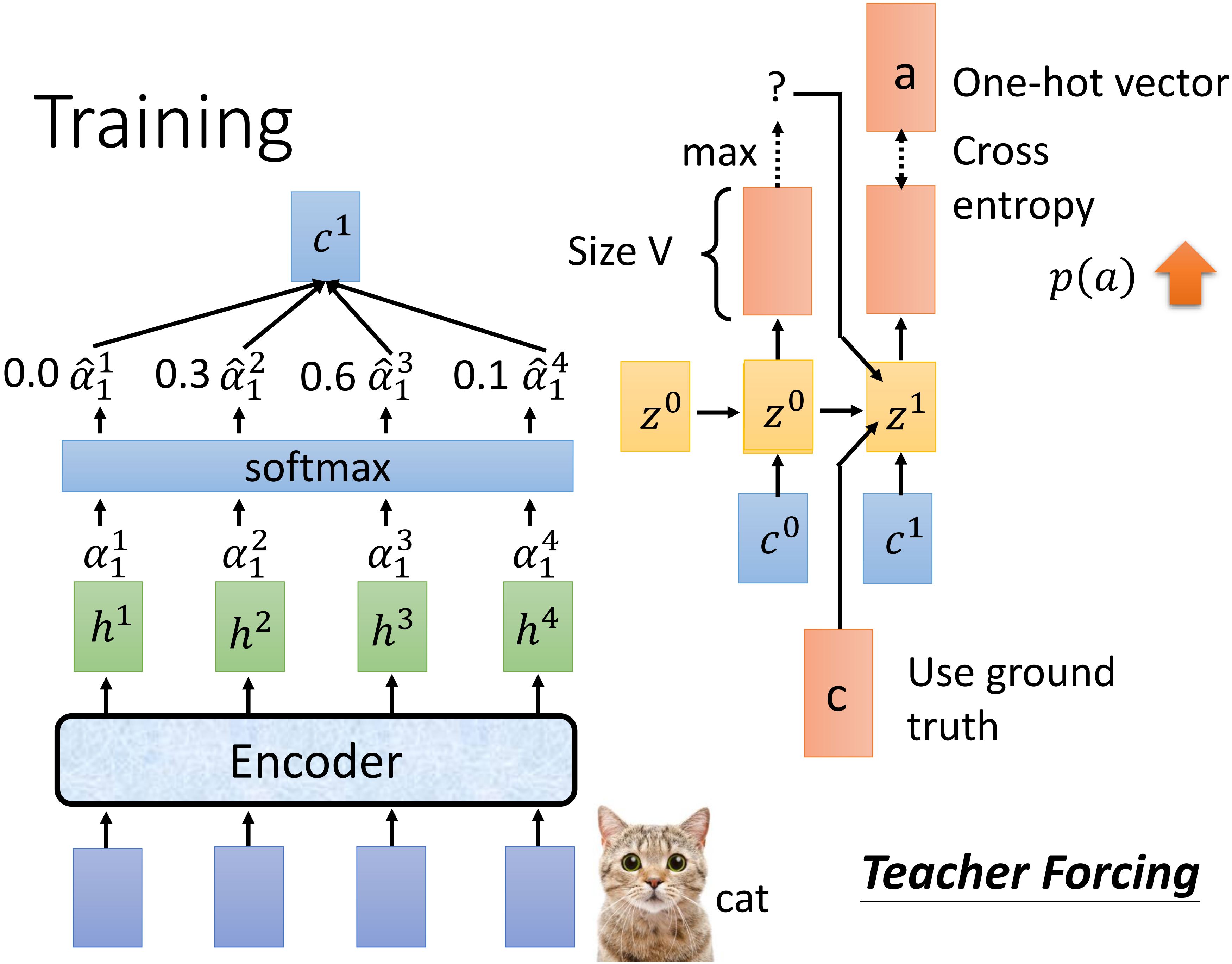
c^0

As RNN input

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

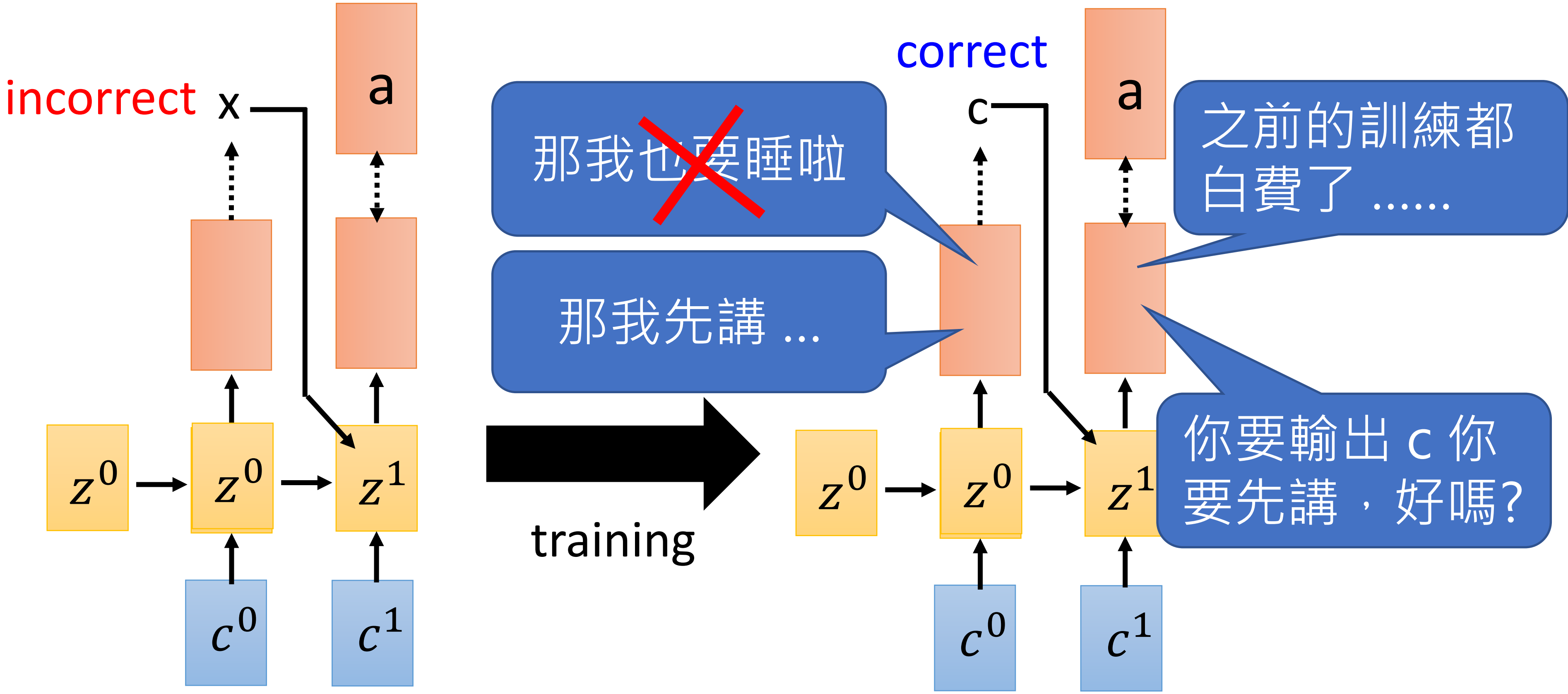
$$= 0.5h^1 + 0.5h^2$$

Training



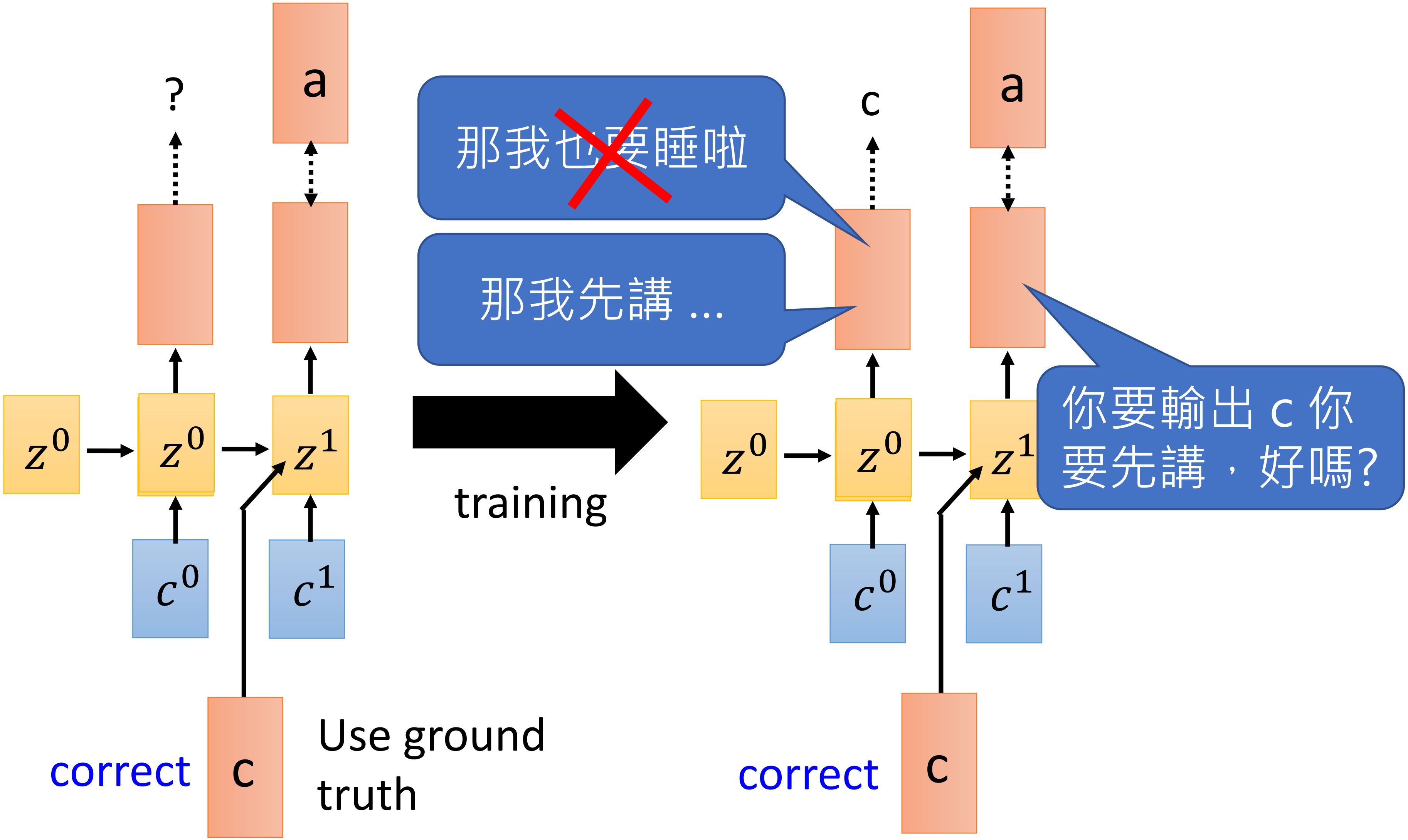
Teacher Forcing

Why Teacher Forcing?



Use previous output

Why Teacher Forcing?

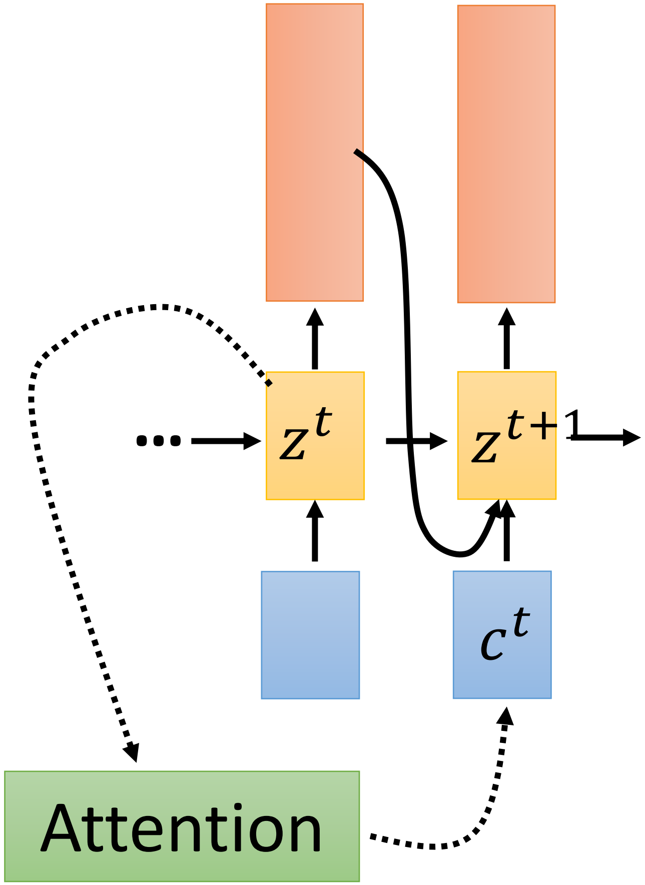


那我~~也要~~睡啦

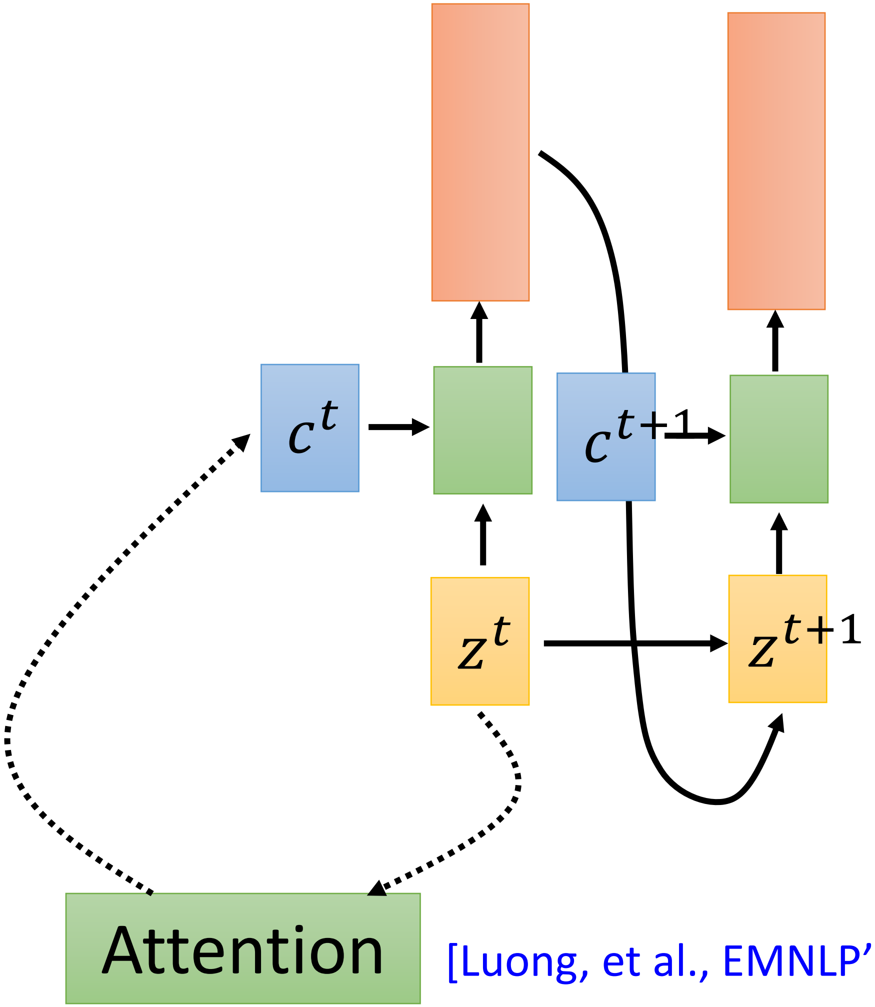
那我先講 ...

你要輸出 c 你要先講，好嗎?

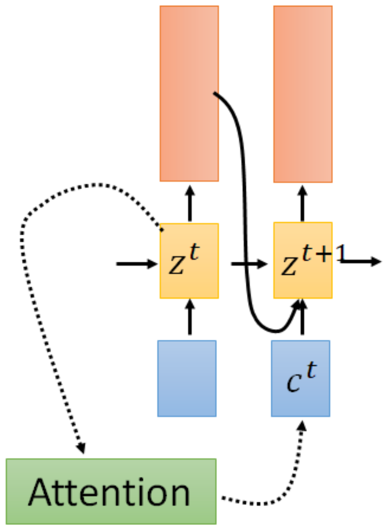
Back to Attention

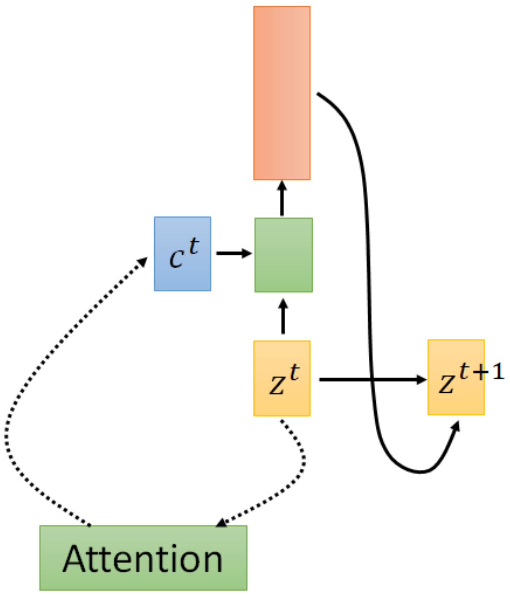


[Bahdanau. et al., ICLR’15]



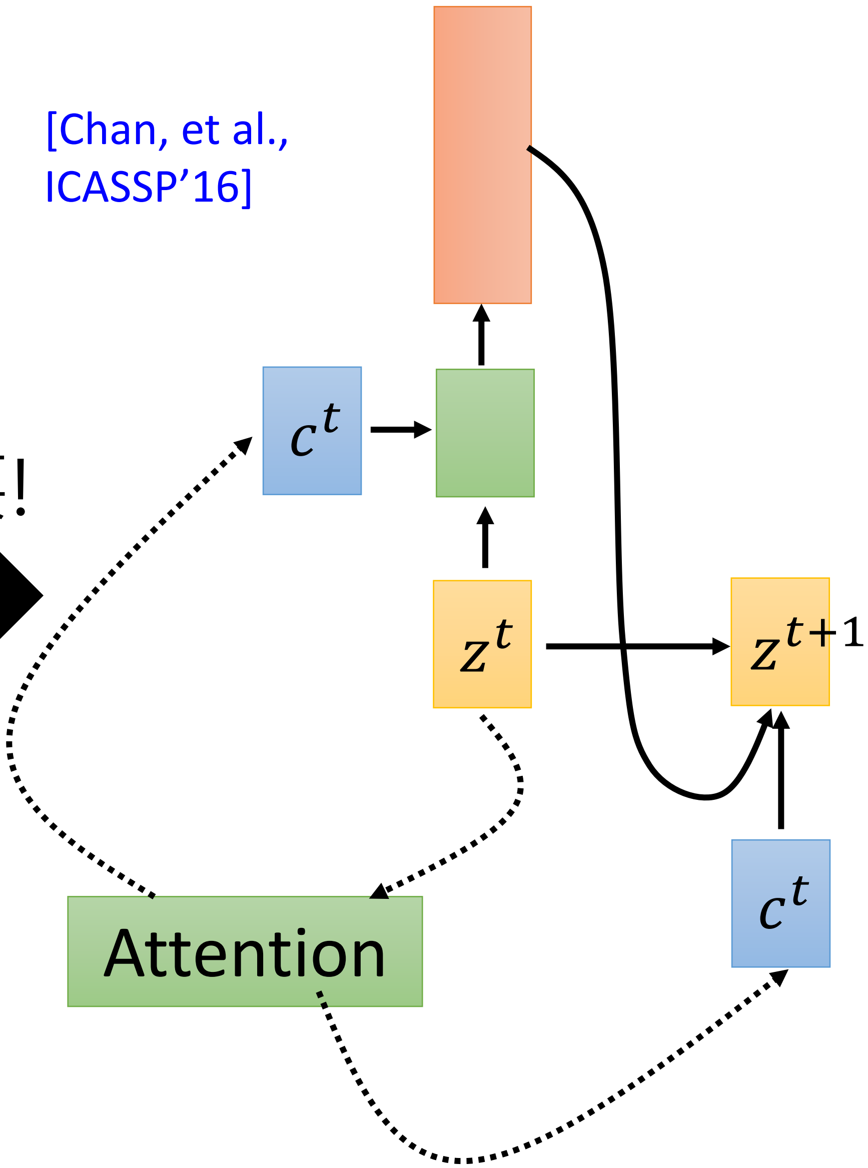
[Luong, et al., EMNLP’15]



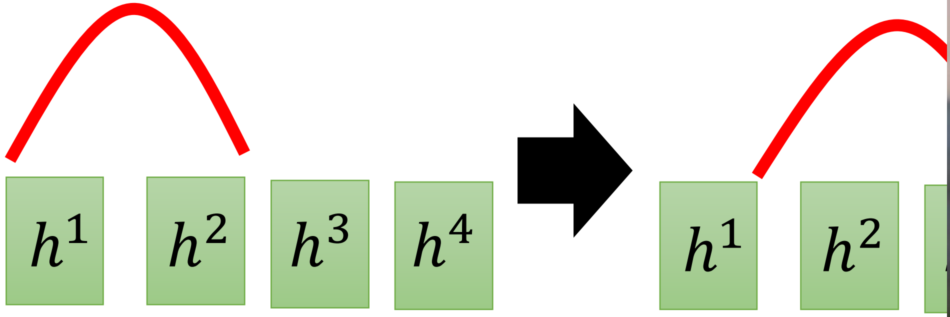


我全都要!

[Chan, et al., ICASSP’16]



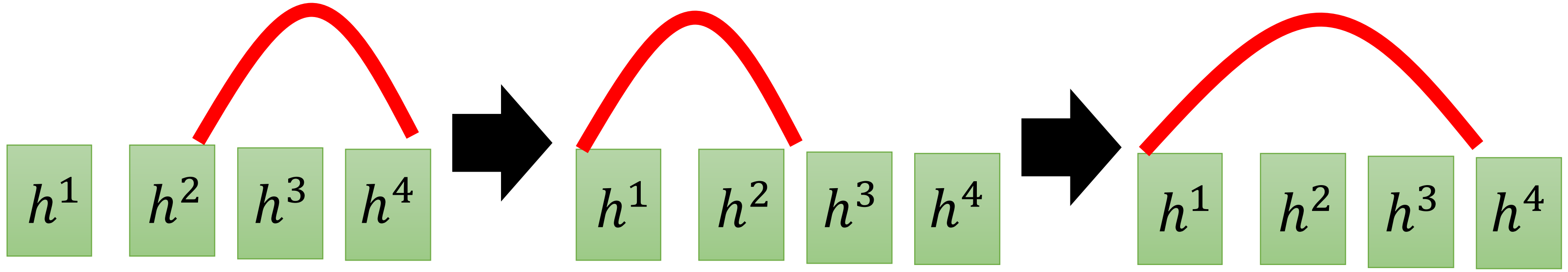
Back to Attention



generate 1st token

generate 2

n



generate 1st token

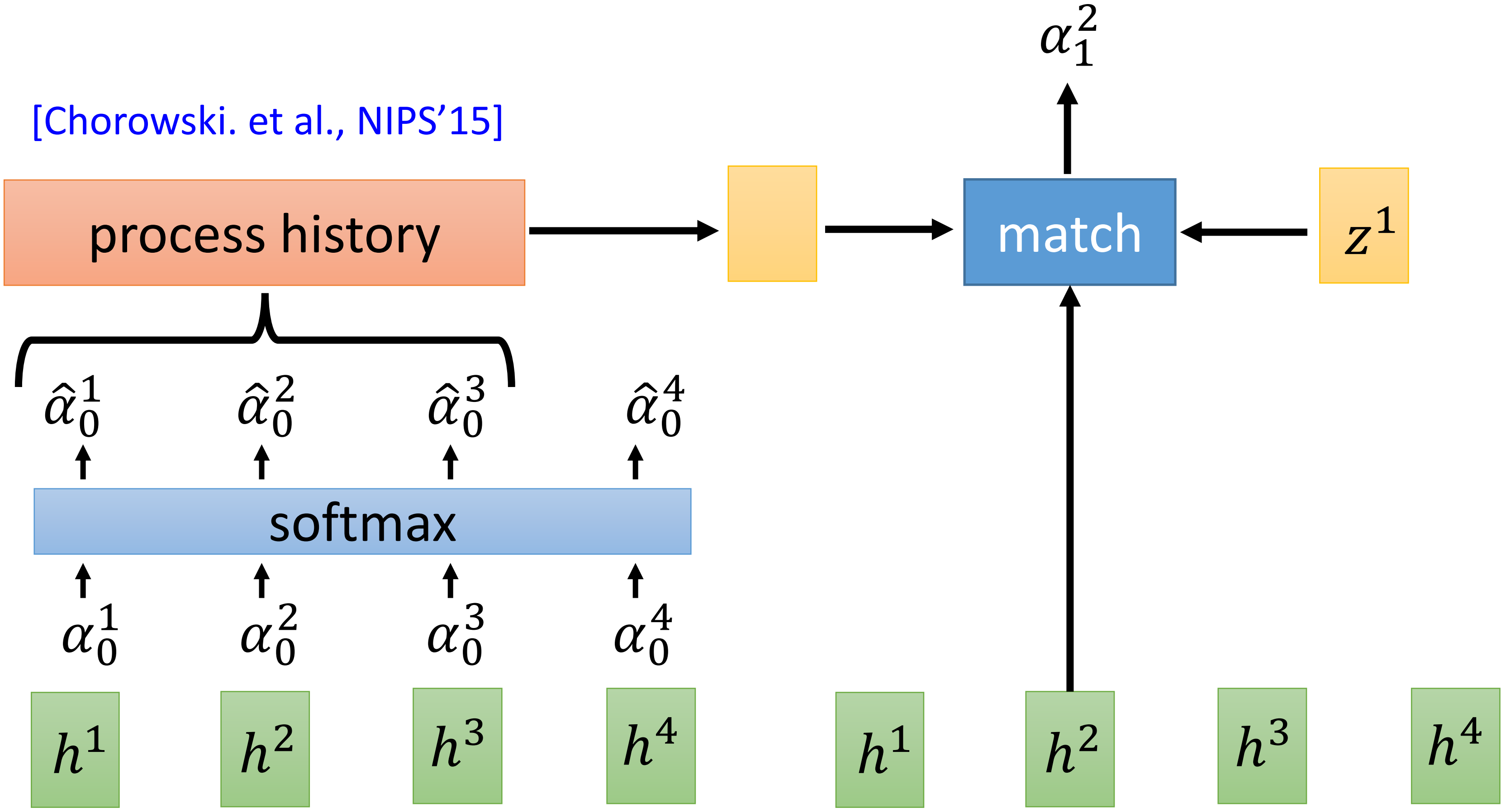
generate 2nd token

generate 3rd token

原始論文難以理解，感謝劉浩然同學指點

Location-aware attention

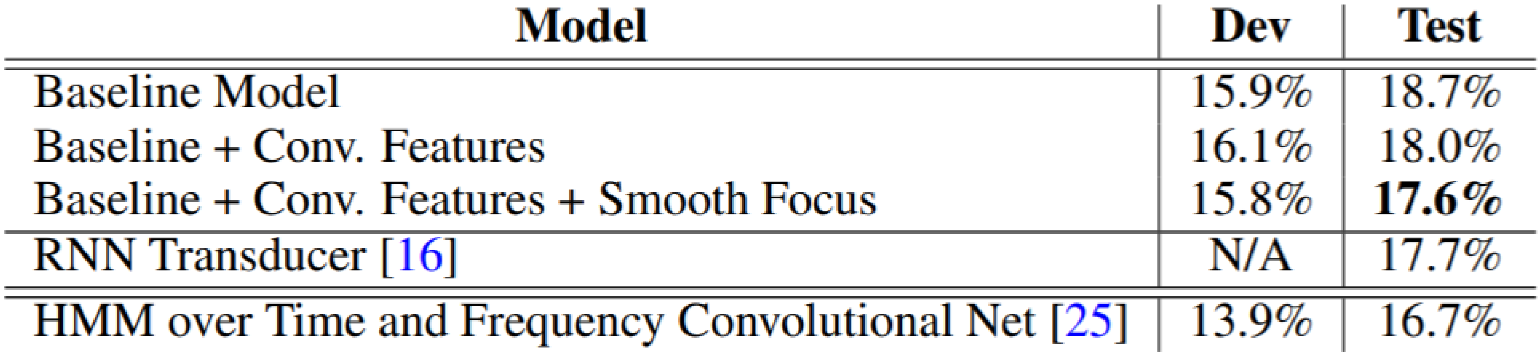
[Chorowski. et al., NIPS’15]



generate the 1st token

generate the 2nd token

LAS – Does it work?



Model	Dev	Test
Baseline Model	15.9%	18.7%
Baseline + Conv. Features	16.1%	18.0%
Baseline + Conv. Features + Smooth Focus	15.8%	**17.6%**
RNN Transducer [16]	N/A	17.7%
HMM over Time and Frequency Convolutional Net [25]	13.9%	16.7%

TIMIT

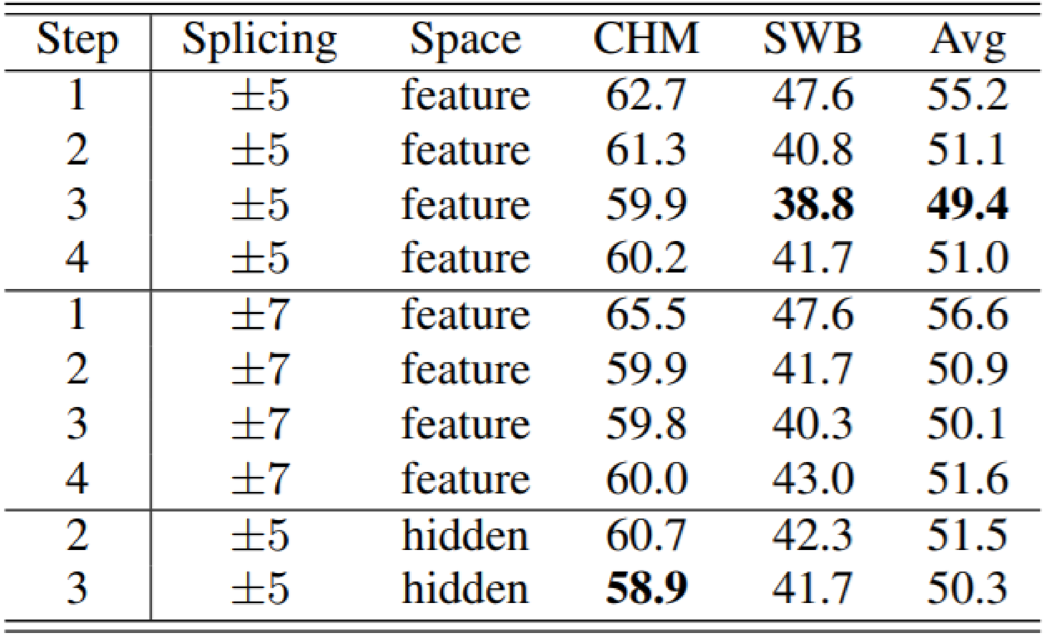
[Chorowski. Et al., NIPS'15]

10.4% on SWB …

[Soltau, et al., ICASSP'14]

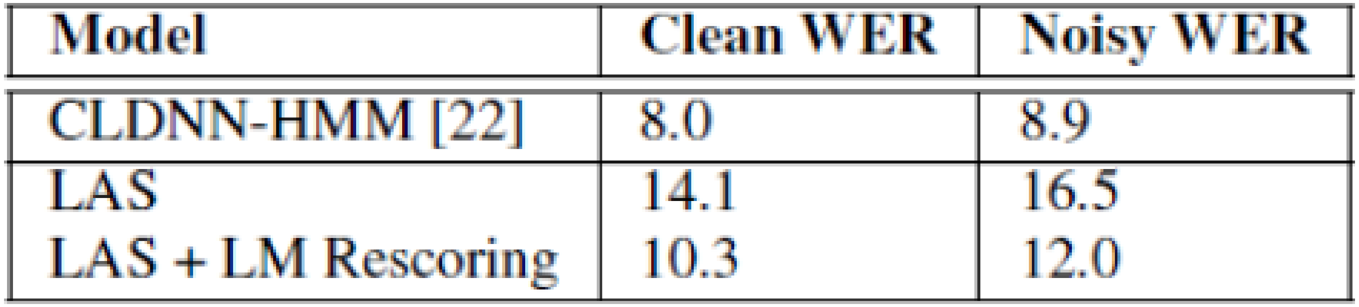
300 hours

[Lu, et al., INTERSPEECH'15]



Step	Splicing	Space	CHM	SWB	Avg
1	±5	feature	62.7	47.6	55.2
2	±5	feature	61.3	40.8	51.1
3	±5	feature	59.9	**38.8**	**49.4**
4	±5	feature	60.2	41.7	51.0
1	±7	feature	65.5	47.6	56.6
2	±7	feature	59.9	41.7	50.9
3	±7	feature	59.8	40.3	50.1
4	±7	feature	60.0	43.0	51.6
2	±5	hidden	60.7	42.3	51.5
3	±5	hidden	**58.9**	41.7	50.3

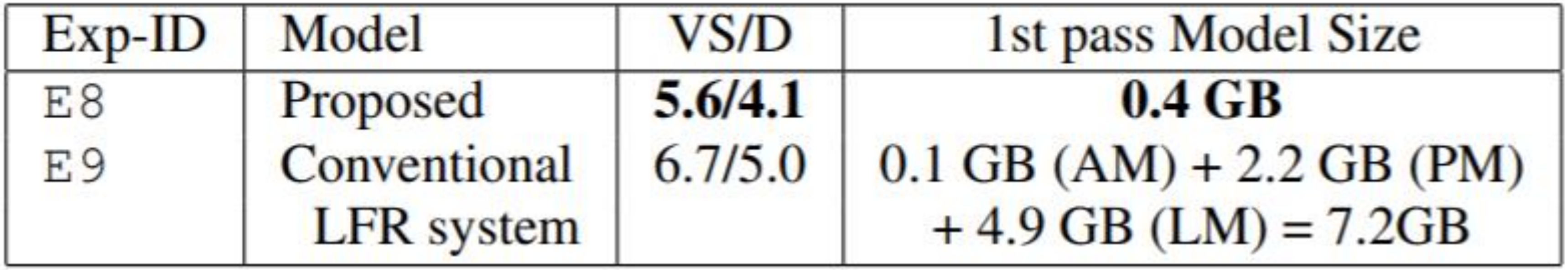
LAS – Yes, it works!



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

2000 hours

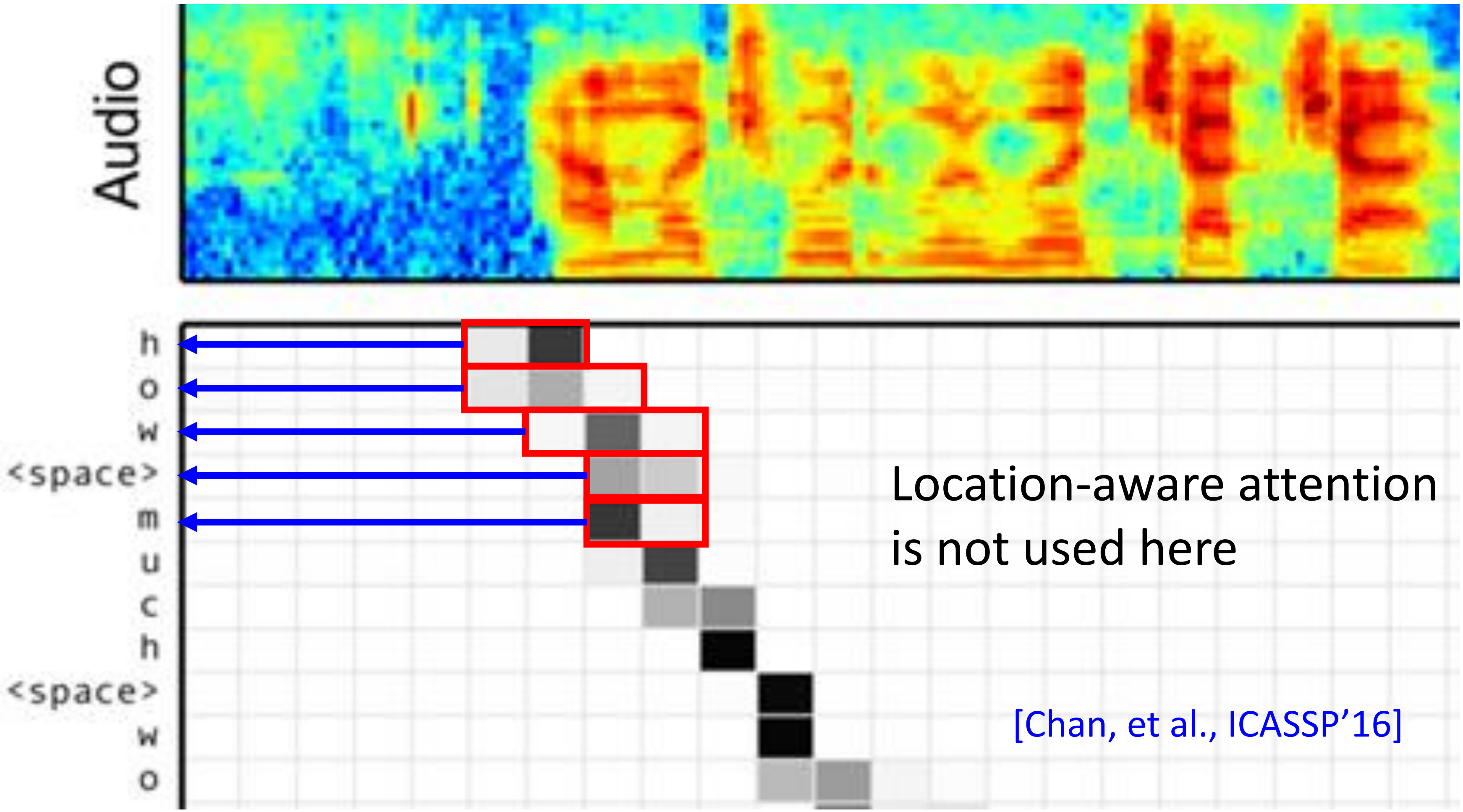
[Chan, et al., ICASSP’16]



Exp-ID	Model	VS/D	1st pass Model Size
E8	Proposed	**5.6/4.1**	**0.4 GB**
E9	Conventional LFR system	6.7/5.0	0.1 GB (AM) + 2.2 GB (PM) + 4.9 GB (LM) = 7.2GB

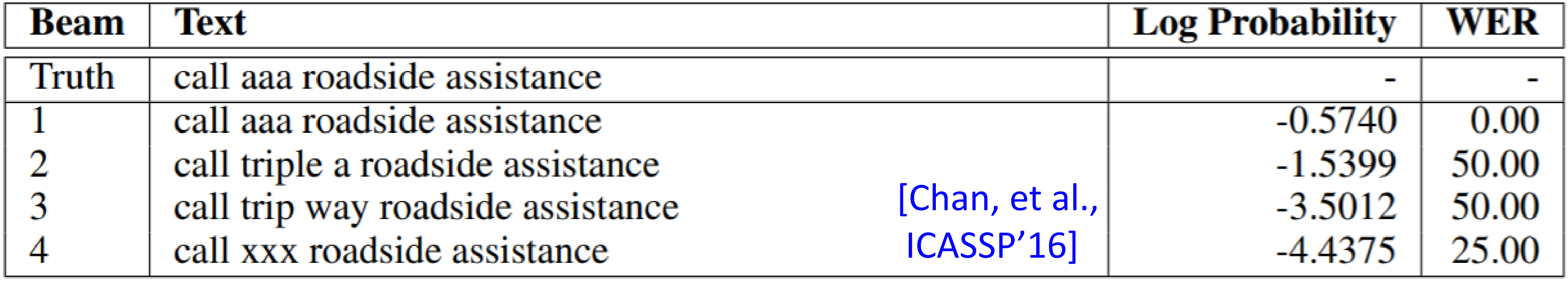
12500 hours

[Chiu, et al., ICASSP, 2018]



Location-aware attention is not used here

[Chan, et al., ICASSP’16]

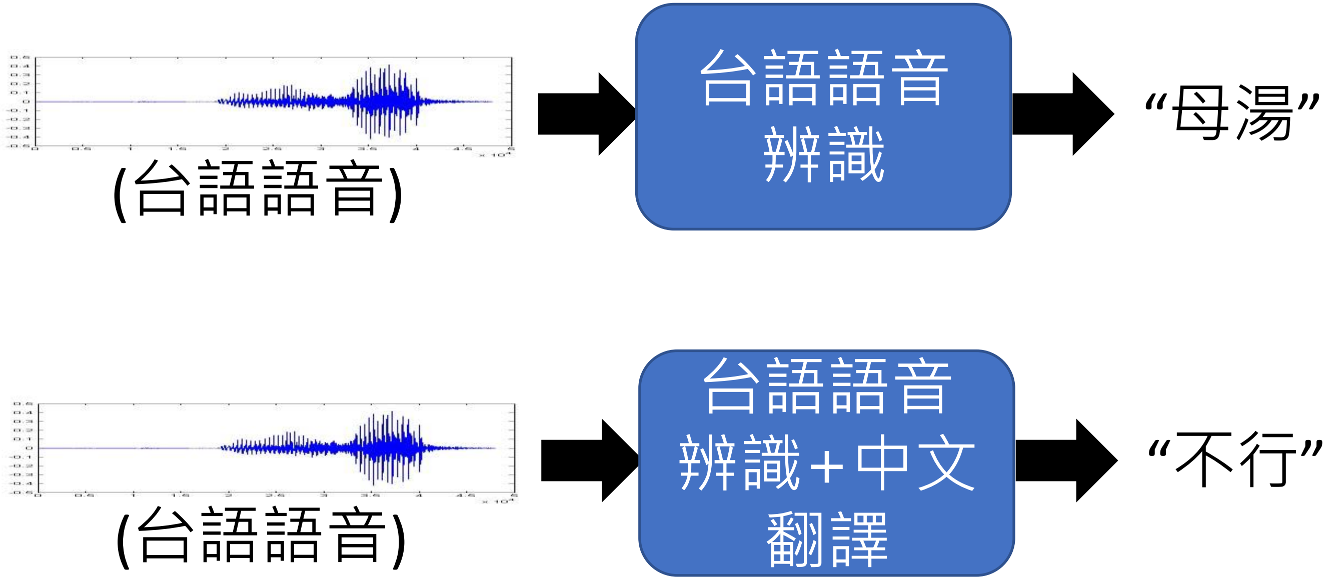


Beam	Text	Log Probability	WER
Truth	call aaa roadside assistance	-	-
1	call aaa roadside assistance	-0.5740	0.00
2	call triple a roadside assistance	-1.5399	50.00
3	call trip way roadside assistance	-3.5012	50.00
4	call xxx roadside assistance	-4.4375	25.00

[Chan, et al., ICASSP’16]

感謝 李仲翊 同學提供實驗結果

Hokkien (閩南語、台語)



訓練資料: YouTube 上的鄉土劇
(台語語音、中文字幕)，約 1500 小時

然後就直接用 LAS 訓練下去

Hokkien (閩南語、台語)

- 有背景音樂、音效？
- 語音和字幕沒有對齊？
- 台羅拼音？

不管 ...

不管 ...

不會 QQ ...

只有用深度學習
“硬train一發”

Results

Accuracy = 62.1%

你 的 身 體 撐 不 住

沒 事 你 為 什 麼 要 請 假

要 生 了 嗎

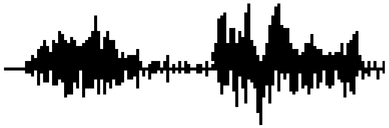
正解:不 會 膩 嗎

我 有 幫 廠 長 拜 託

正解: 我 拜 託 廠 長 了

Limitation of LAS

- LAS outputs the first token after listening the whole input.
- Users expect on-line speech recognition.



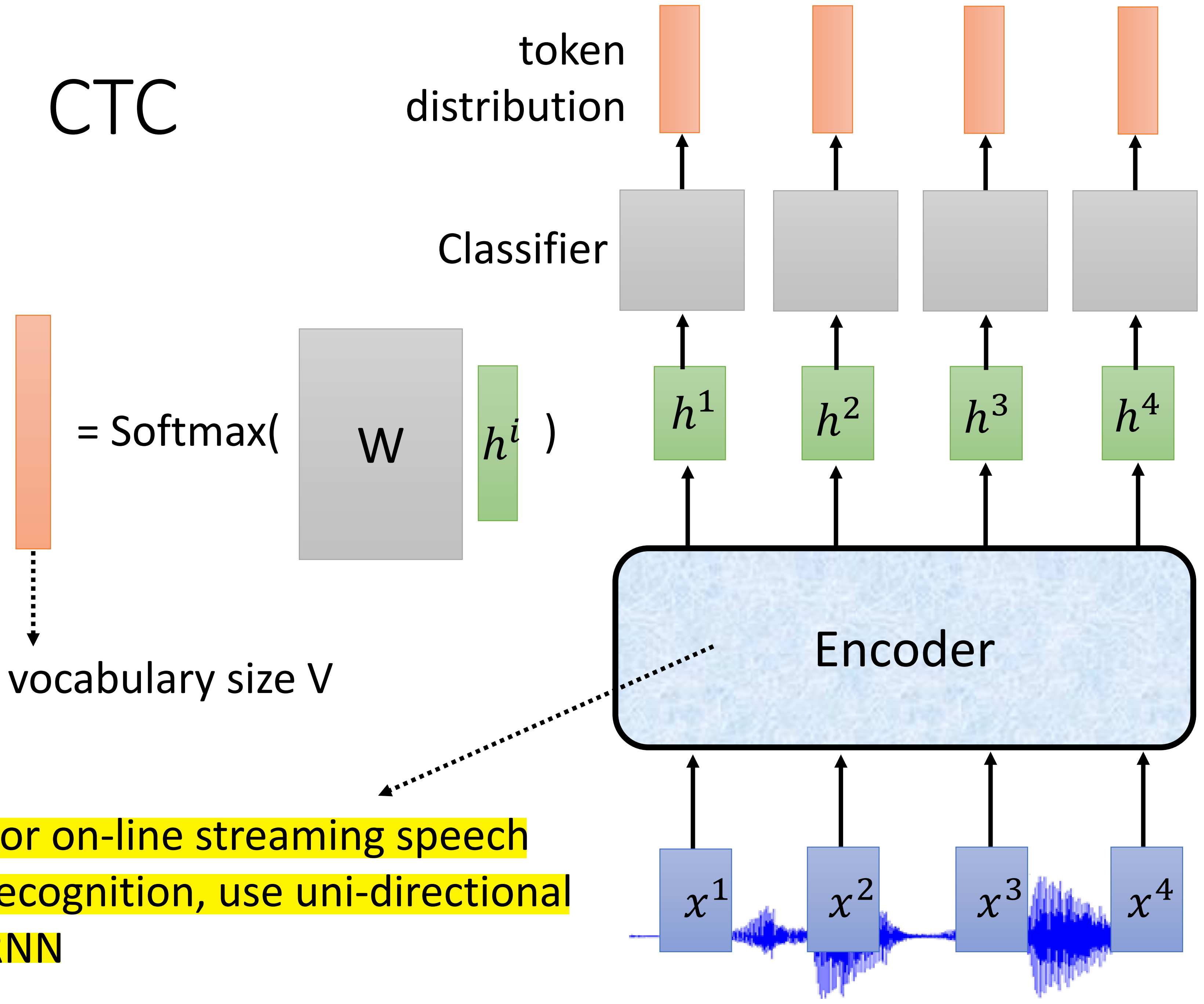
今　天　的　天　氣　非　常　好

LAS is not the final solution of ASR!

Models to be introduced

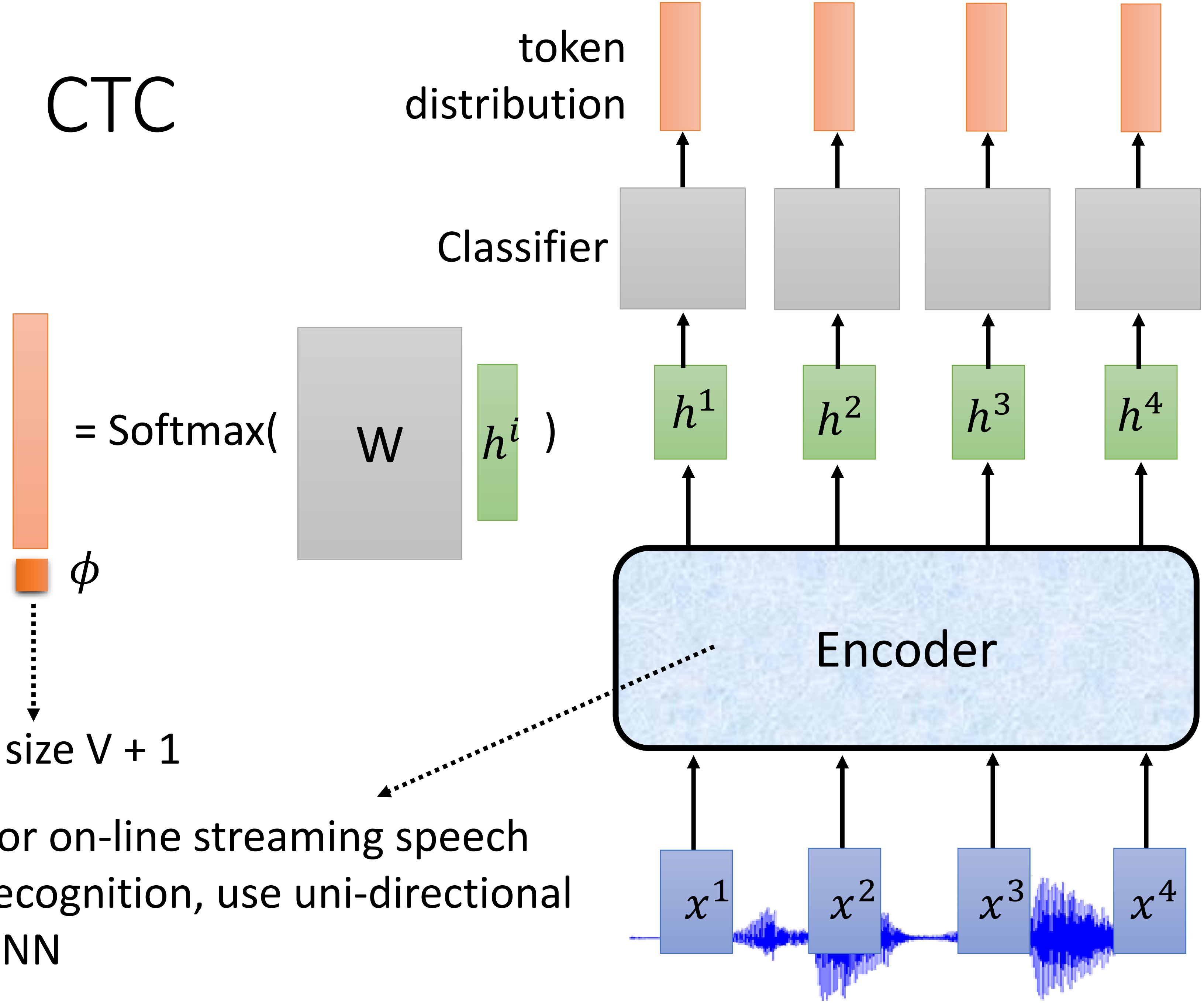
- Listen, Attend, and Spell (LAS) [Chorowski. et al., NIPS’15]
- Connectionist Temporal Classification (CTC) [Graves, et al., ICML’06]
- RNN Transducer (RNN-T) [Graves, ICML workshop’12]
- Neural Transducer [Jaitly, et al., NIPS’16]
- Monotonic Chunkwise Attention (MoChA) [Chiu, et al., ICLR’18]

CTC



For on-line streaming speech recognition, use uni-directional RNN

CTC



For on-line streaming speech recognition, use uni-directional RNN

CTC

- Input T acoustic features, output T tokens (ignoring down sampling)
- Output tokens including ϕ, merging duplicate tokens, removing ϕ

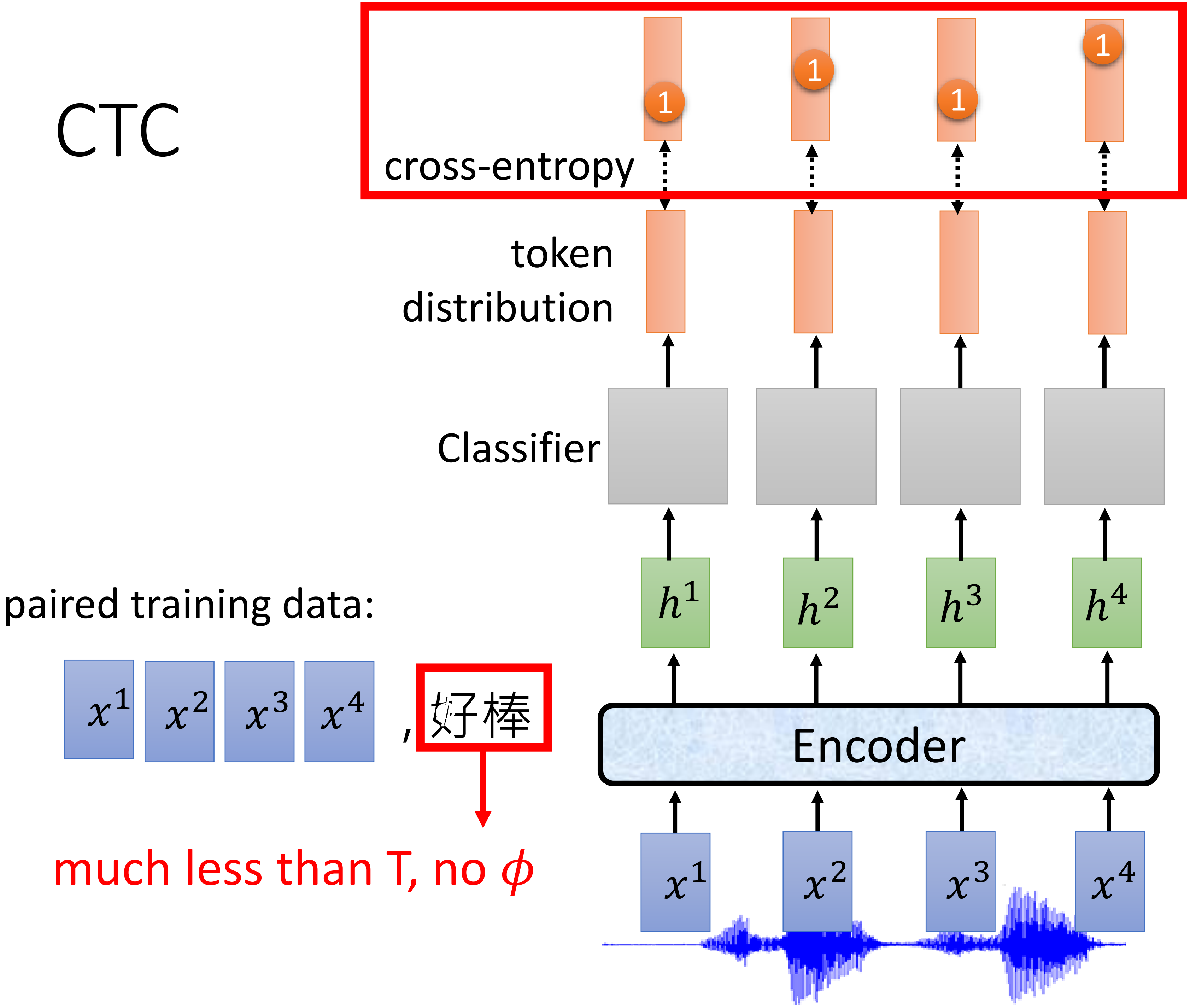
ϕ ϕ d d ϕ e ϕ e ϕ p p → d e e p

ϕ ϕ d d ϕ e e e ϕ p p → d e p

好 好 棒 棒 棒 棒 棒 → 好 棒

好 ϕ 棒 ϕ 棒 ϕ ϕ → 好 棒 棒

CTC



paired training data:

x^1 x^2 x^3 x^4 , 好棒

much less than T, no ϕ

CTC – Training

paired training data:

x^1 x^2 x^3 x^4 , 好棒

All of them are used in training! (How?!)

x^1 x^2 x^3 x^4 ,好好棒ϕ

x^1 x^2 x^3 x^4 ,ϕ好棒棒

x^1 x^2 x^3 x^4 ,好棒棒棒

x^1 x^2 x^3 x^4 , 好棒$\phi\phi$

x^1 x^2 x^3 x^4 , 好ϕ棒ϕ

x^1 x^2 x^3 x^4 , 好$\phi\phi$棒

alignment

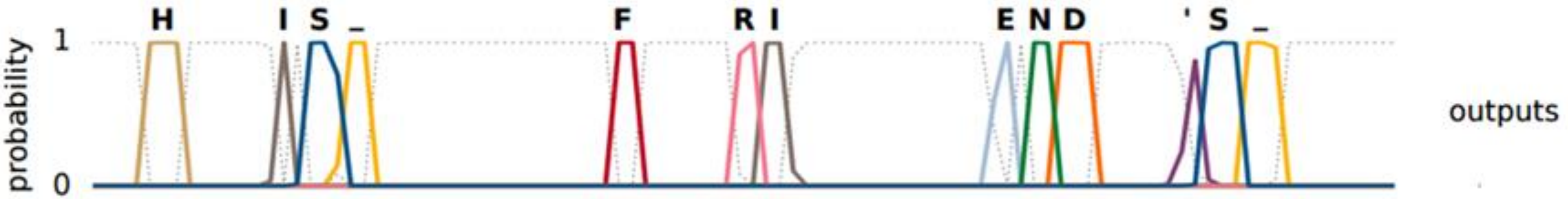
x^1 x^2 x^3 x^4 ,ϕ好棒ϕ

x^1 x^2 x^3 x^4 ,ϕ好ϕ棒

x^1 x^2 x^3 x^4 ,$\phi\phi$好棒

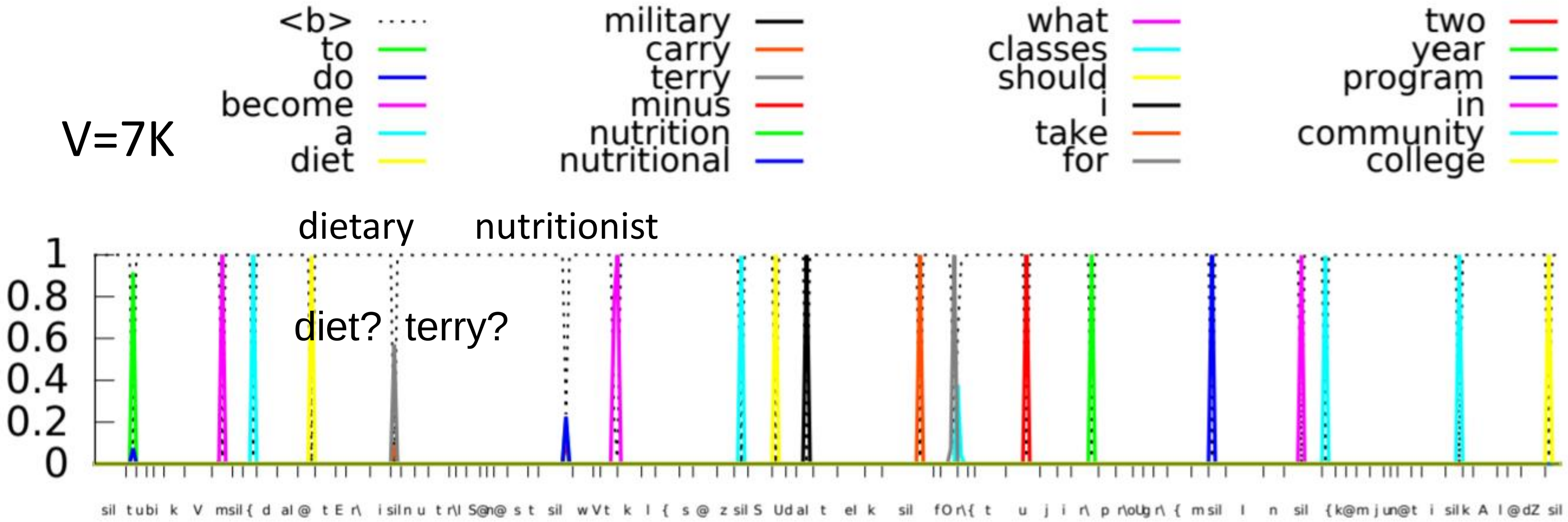
x^1 x^2 x^3 x^4 , 好棒ϕ棒

Does CTC work?



[Graves, et al., ICML'14]

V=7K



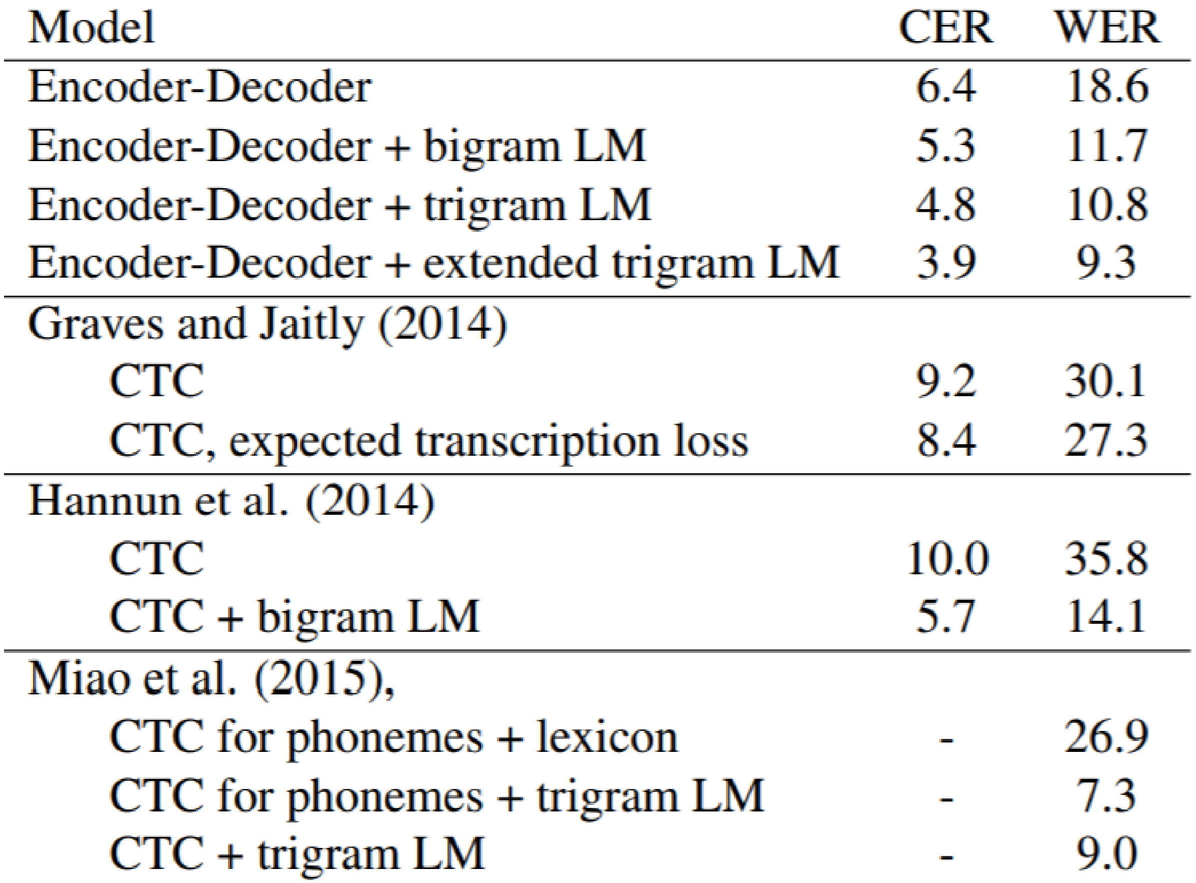
dietary nutritionist

diet? terry?

One can increase V to obtain better performance

[Sak, et al., INTERSPEECH'15]

Does CTC work?



Model	CER	WER
Encoder-Decoder	6.4	18.6
Encoder-Decoder + bigram LM	5.3	11.7
Encoder-Decoder + trigram LM	4.8	10.8
Encoder-Decoder + extended trigram LM	3.9	9.3
Graves and Jaitly (2014)		
CTC	9.2	30.1
CTC, expected transcription loss	8.4	27.3
Hannun et al. (2014)		
CTC	10.0	35.8
CTC + bigram LM	5.7	14.1
Miao et al. (2015),		
CTC for phonemes + lexicon	-	26.9
CTC for phonemes + trigram LM	-	7.3
CTC + trigram LM	-	9.0

80 hours

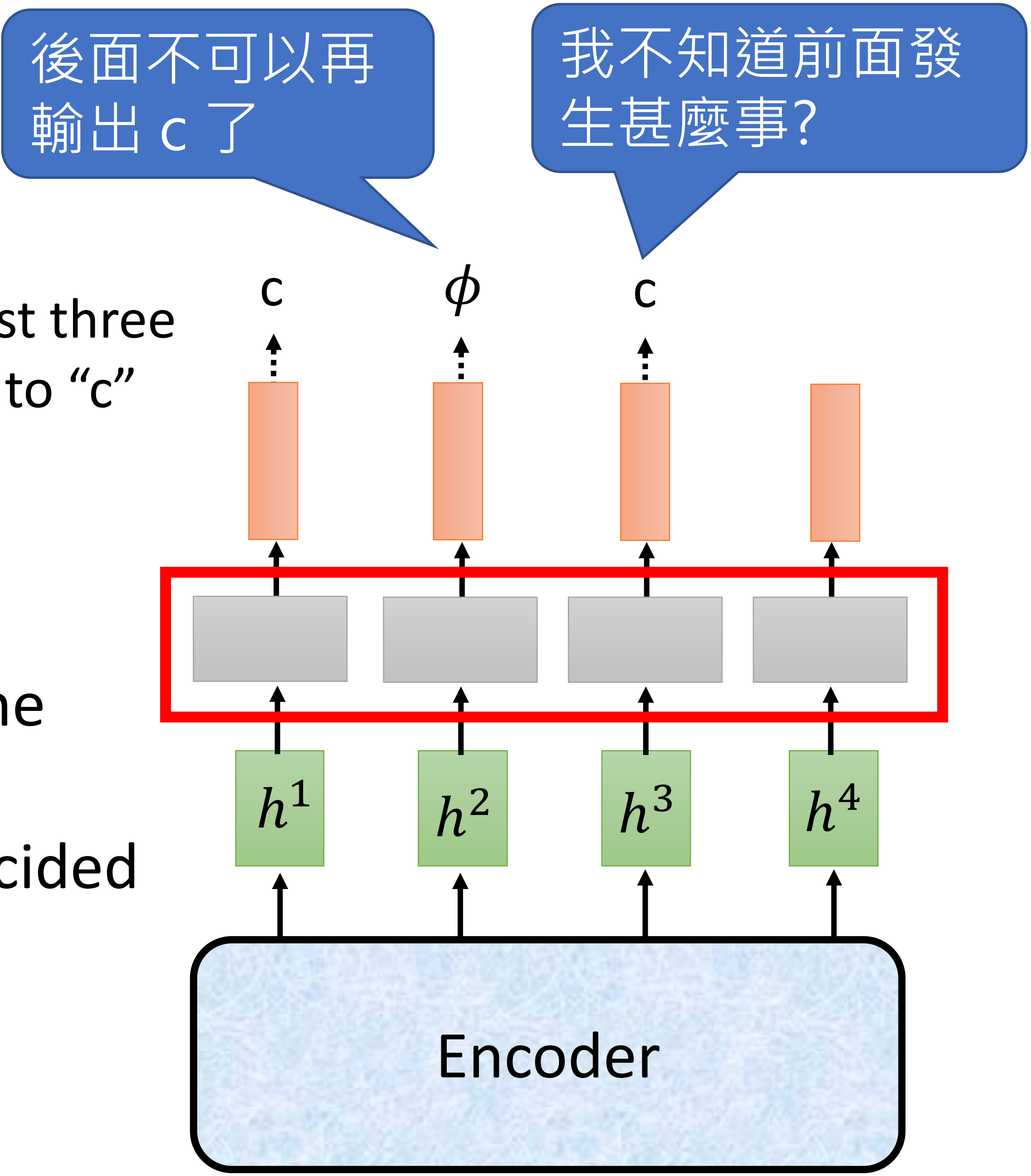
[Bahdanau. et al., ICASSP’16]

Issue

Assume the first three frames belong to “c”

“Decoder”:

- Only attend on one vector
- Each output is decided independently



Models to be introduced

- Listen, Attend, and Spell (LAS) [Chorowski. et al., NIPS'15]
- Connectionist Temporal Classification (CTC) [Graves, et al., ICML'06]
- RNN Transducer (RNN-T) [Graves, ICML workshop'12]
- Neural Transducer [Jaitly, et al., NIPS'16]
- Monotonic Chunkwise Attention (MoChA) [Chiu, et al., ICLR'18]

RNA

Recurrent Neural Aligner

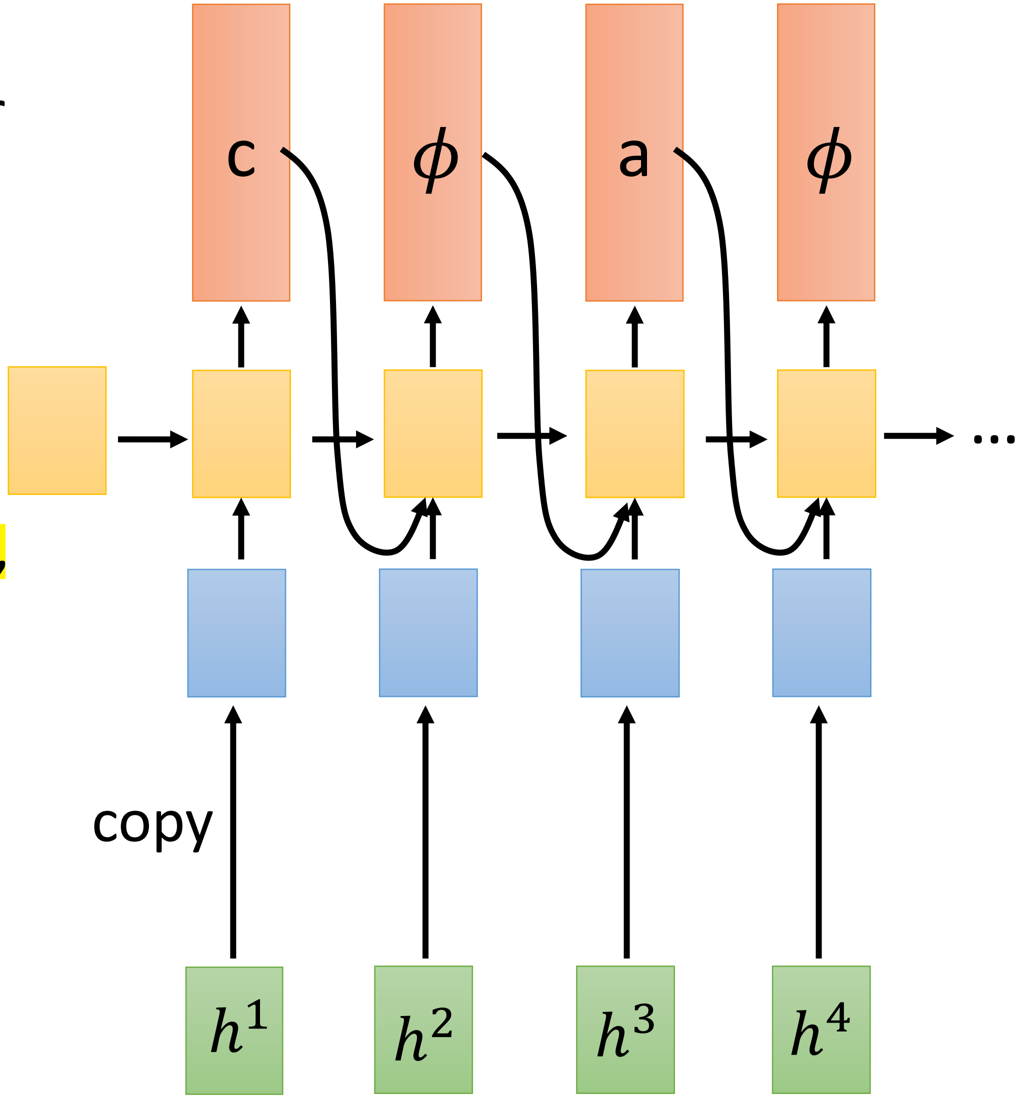
[Sak, et al., INTERSPEECH'17]

CTC Decoder:
take one vector as input,
output one token

RNA adds dependency

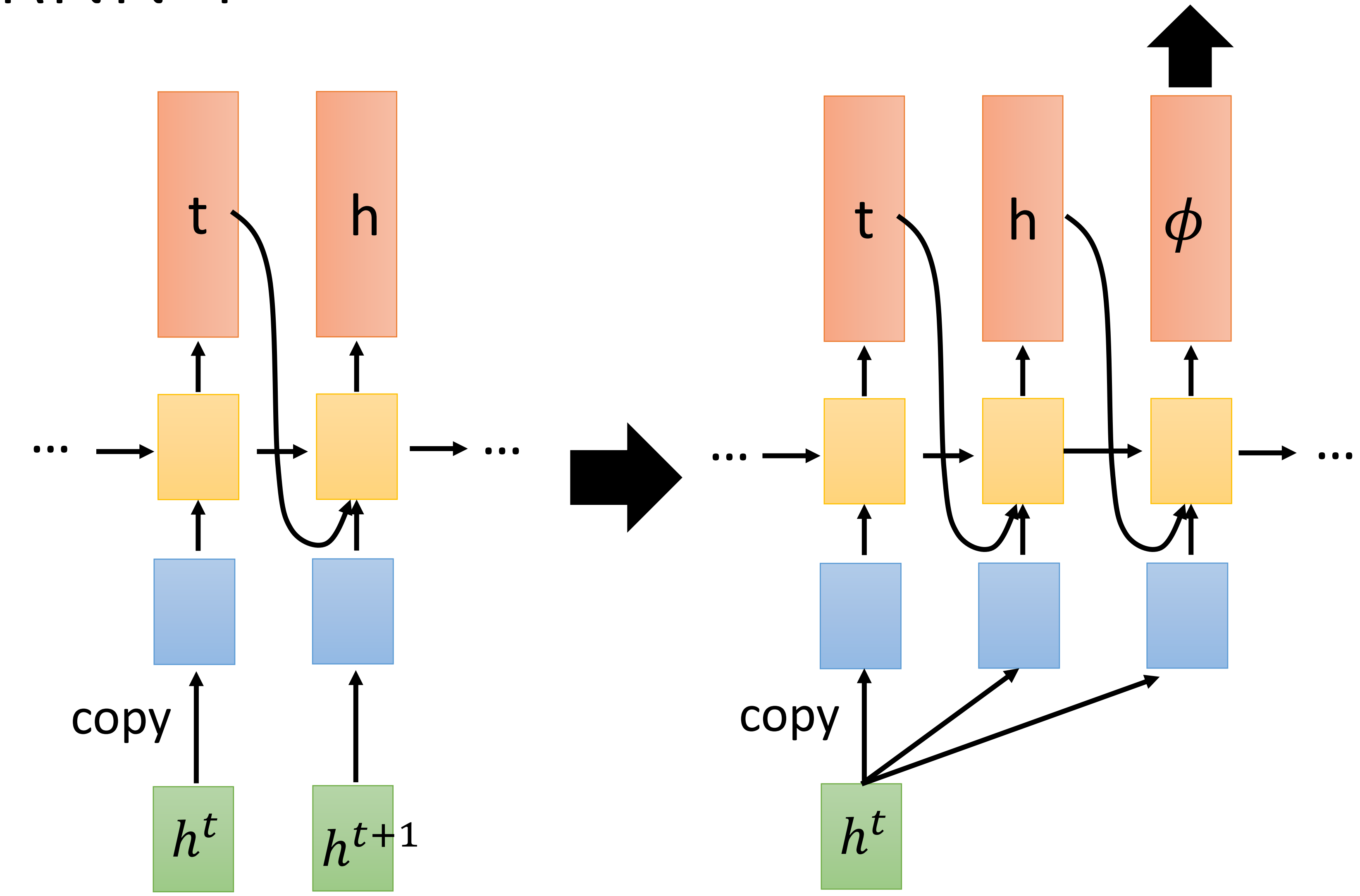
Can one vector map to multiple tokens?

for example, "*th*"

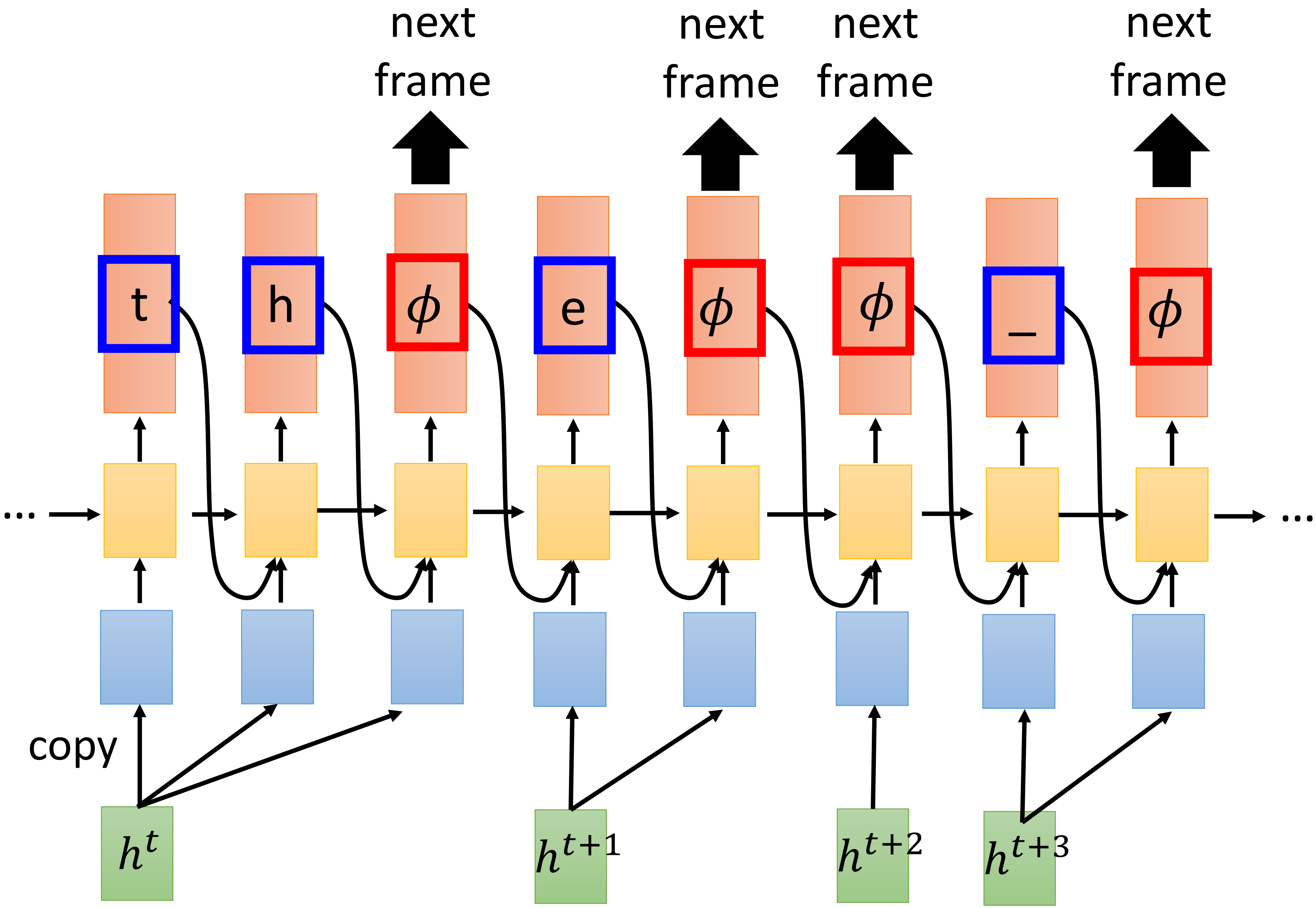


RNN-T

Give me the next frame



There are T “ϕ” in the output.



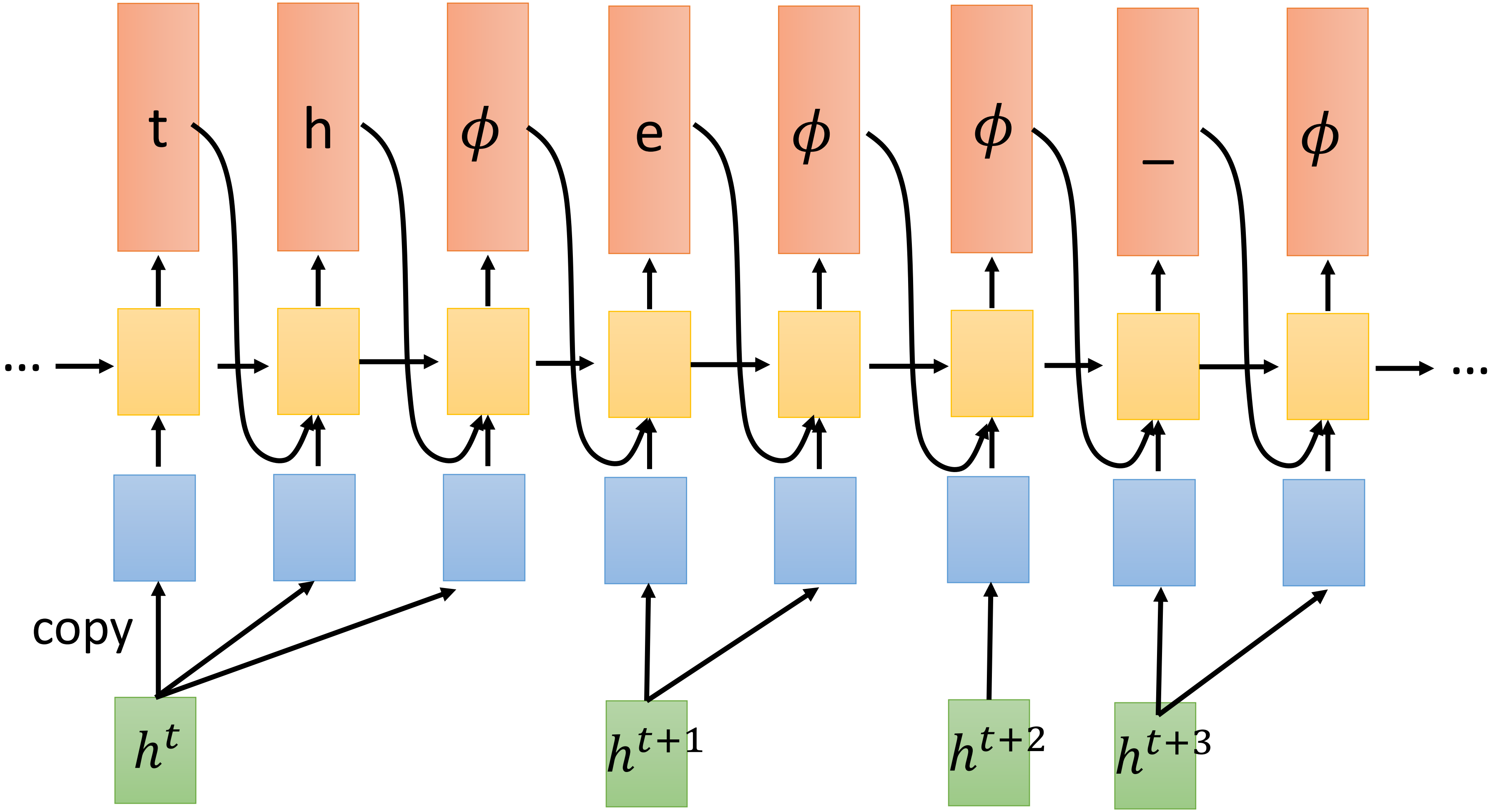
There are T “ϕ” in the output.

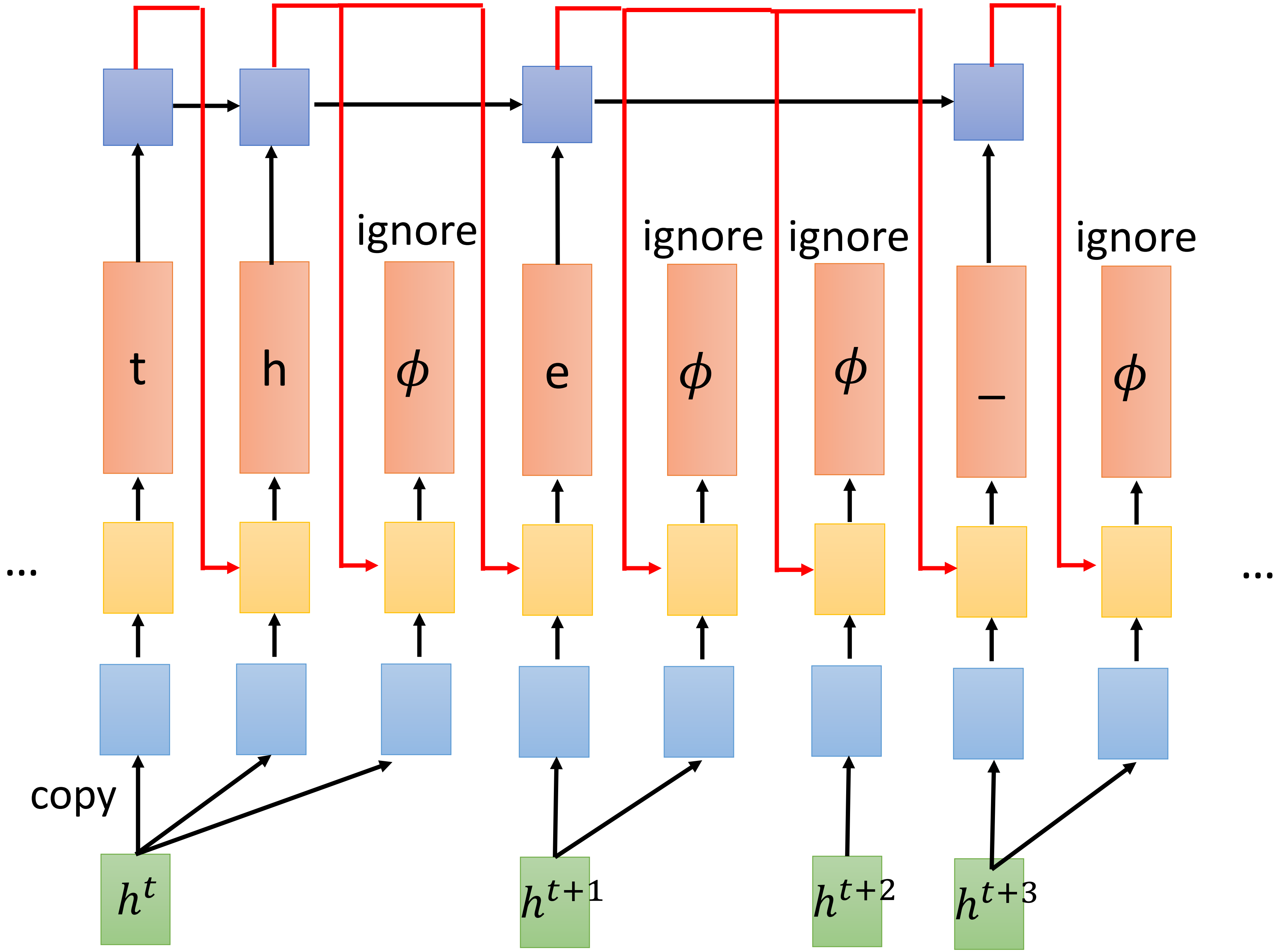
ϕ_1 好 ϕ_2 ϕ_3 ϕ_4 ϕ_5 棒 ϕ_6

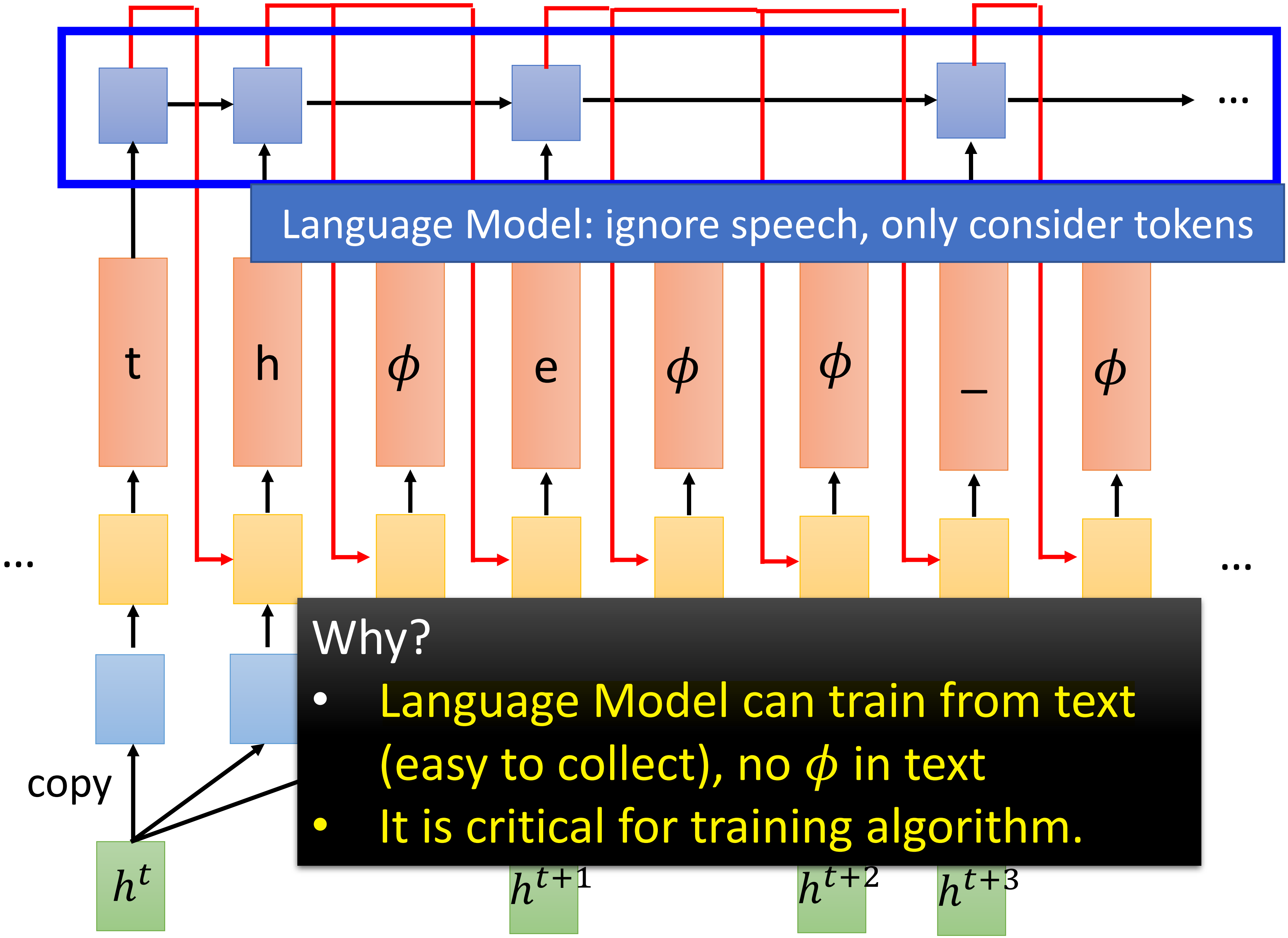
x^1 x^2 x^3 x^4 , 好棒

ϕ_1 ϕ_2 ϕ_3 ϕ_4 ϕ_5 好 棒 ϕ_6

All of them are used in training! (How?!)




ignore
ignore
ignore
ignore
t
h
ϕ
e
ϕ
ϕ
_
ϕ
…
…
copy
h^t
h^{t+1}
h^{t+2}
h^{t+3}



Language Model: ignore speech, only consider tokens

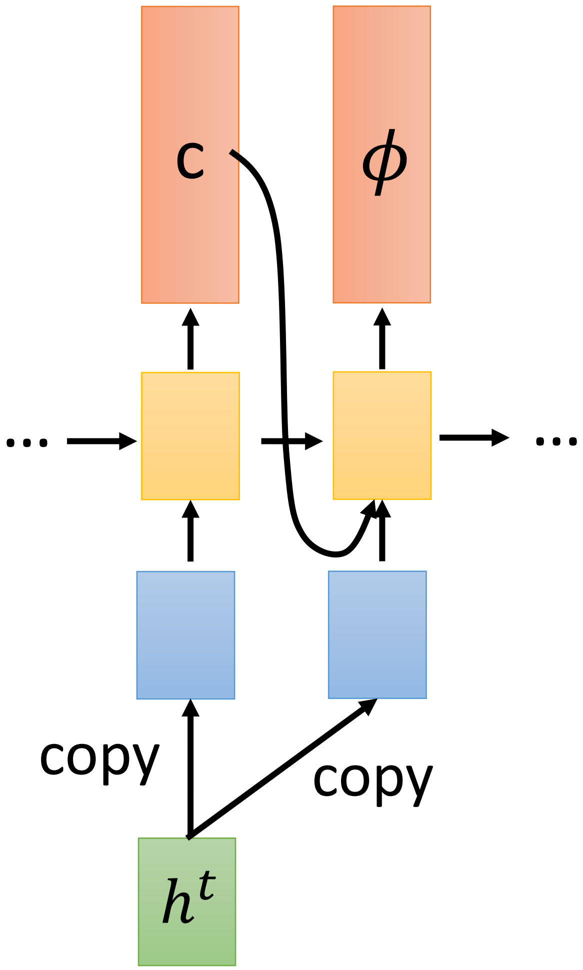
Why?

- Language Model can train from text (easy to collect), no ϕ in text
- It is critical for training algorithm.

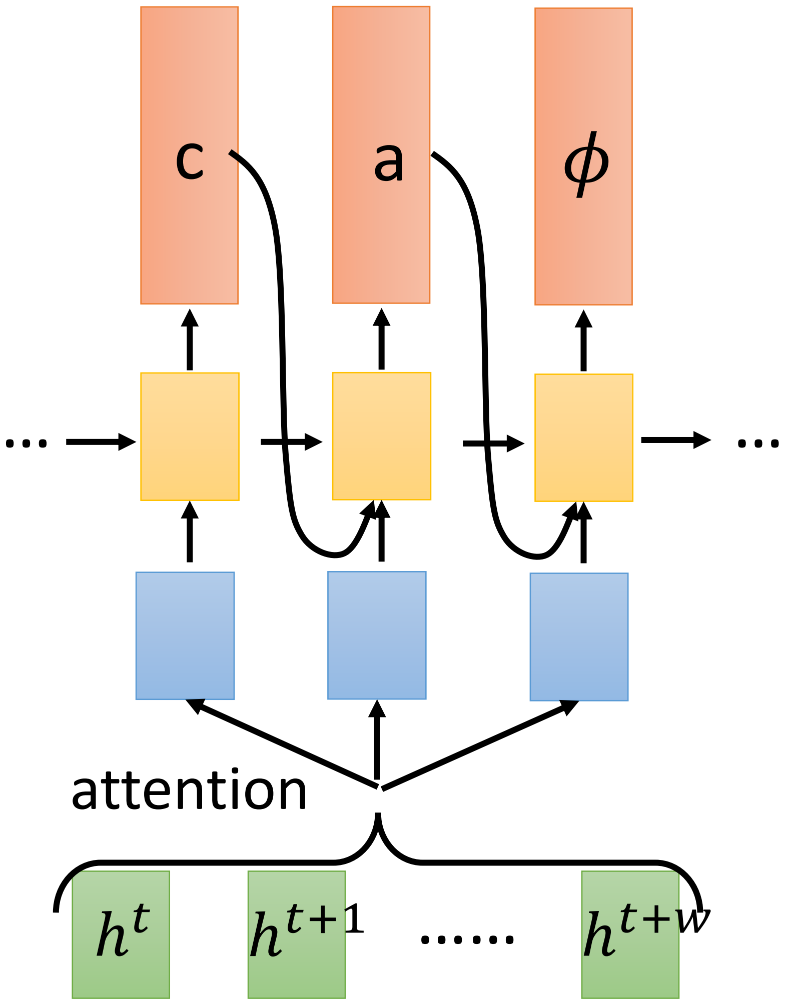
Models to be introduced

- Listen, Attend, and Spell (LAS) [Chorowski. et al., NIPS’15]
- Connectionist Temporal Classification (CTC) [Graves, et al., ICML’06]
- RNN Transducer (RNN-T) [Graves, ICML workshop’12]
- Neural Transducer [Jaitly, et al., NIPS’16]
- Monotonic Chunkwise Attention (MoChA) [Chiu, et al., ICLR’18]

Neural Transducer

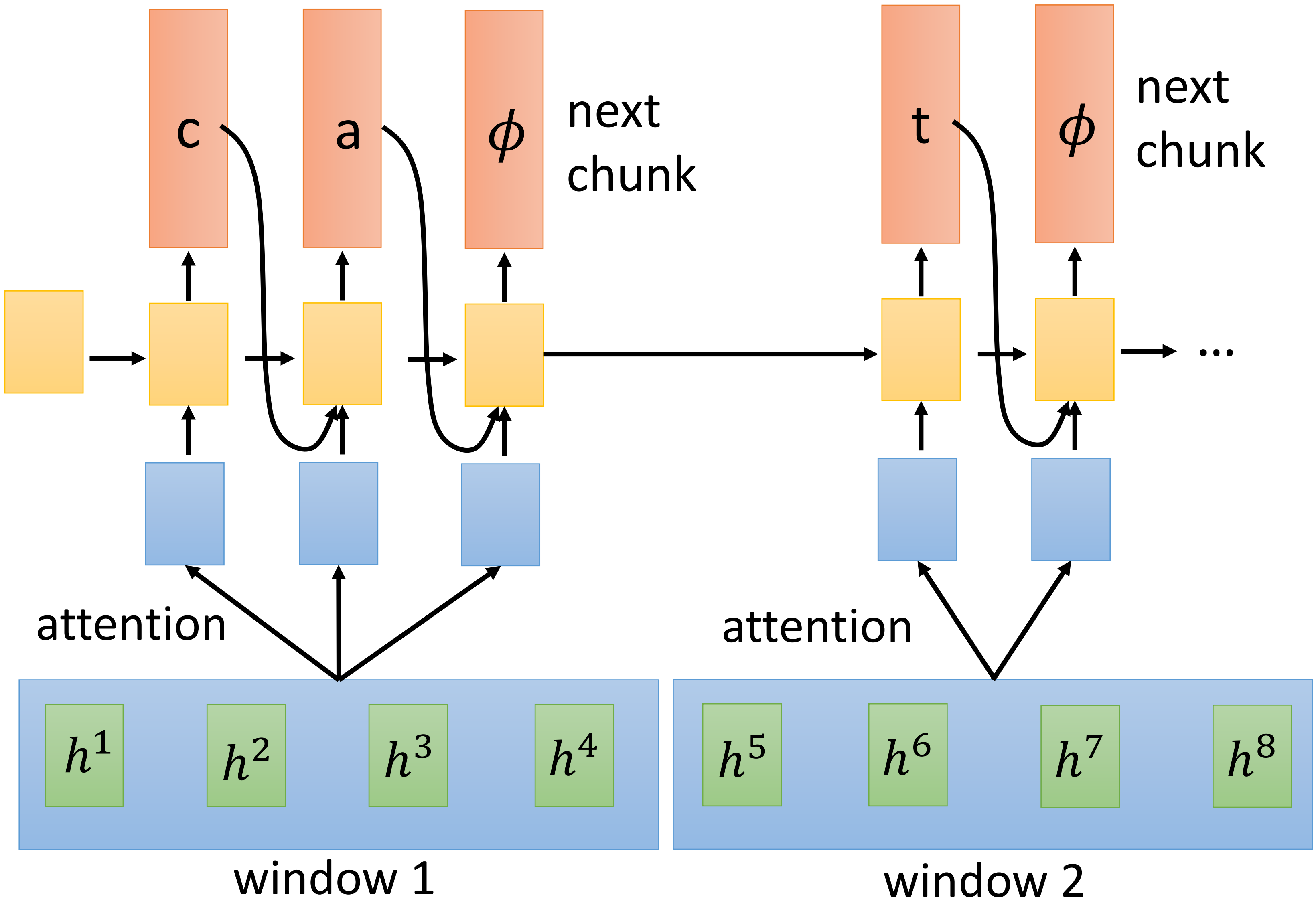


CTC, RNA, RNN-T

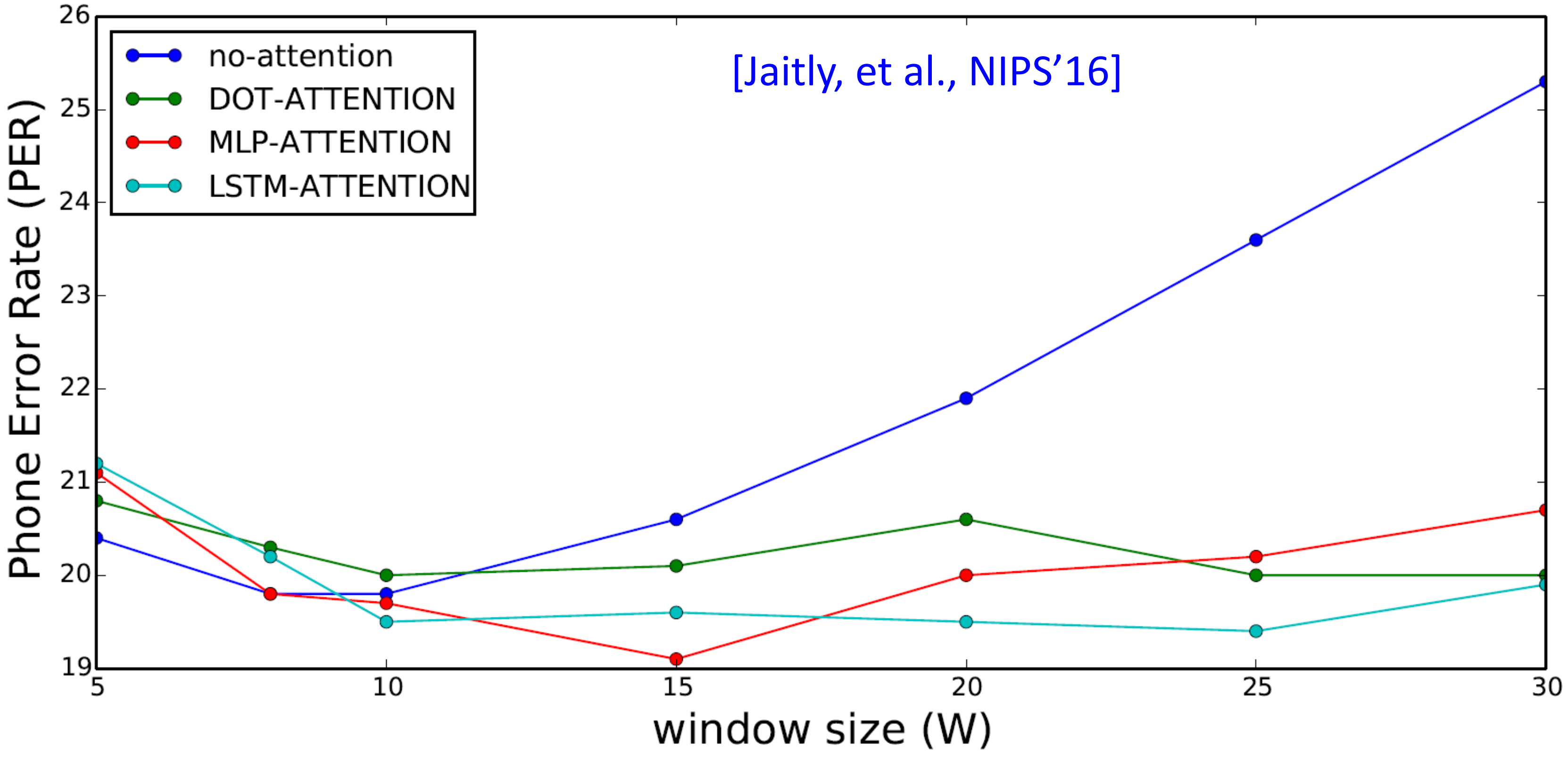


Neural Transducer

Neural Transducer



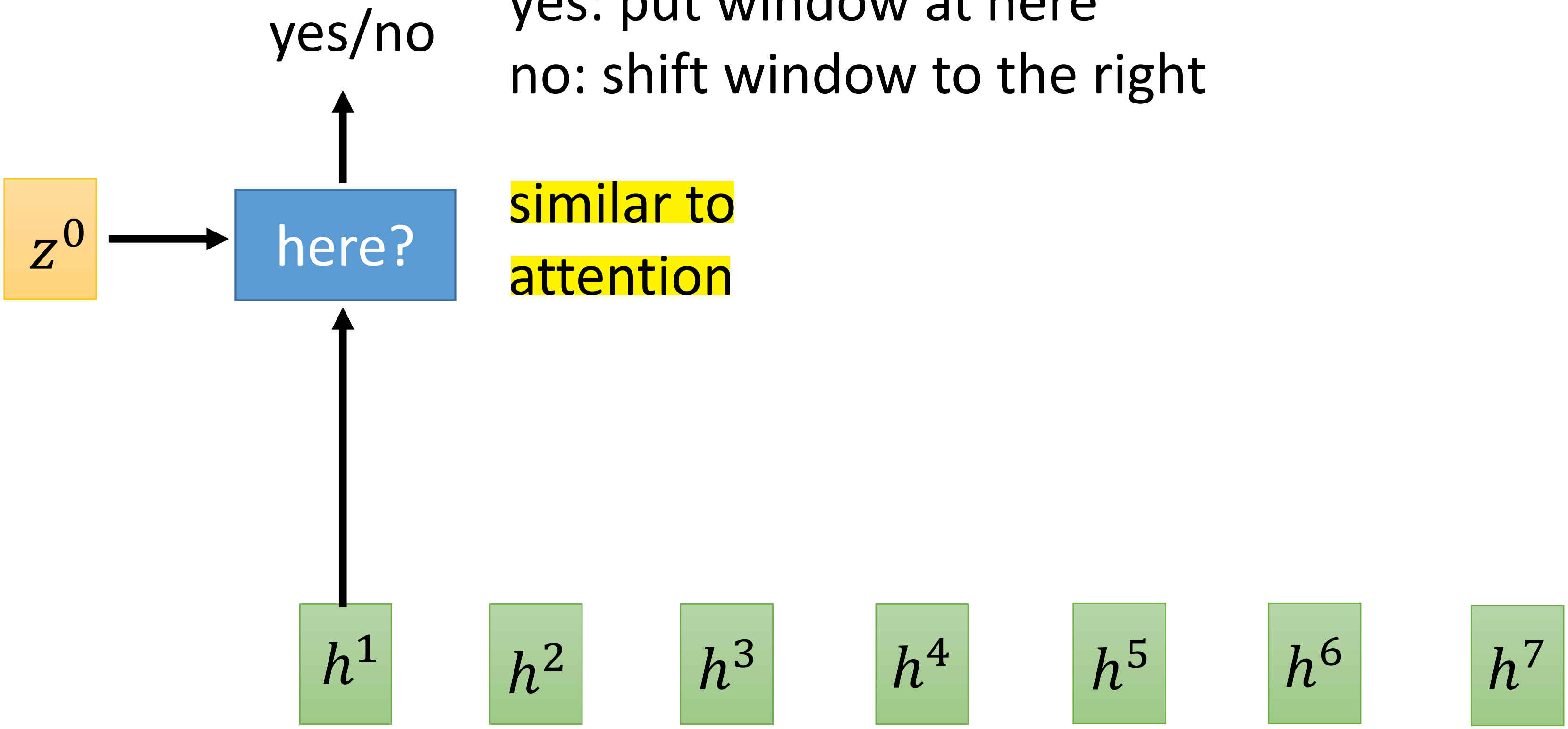
Neural Transducer



Models to be introduced

- Listen, Attend, and Spell (LAS) [Chorowski. et al., NIPS'15]
- Connectionist Temporal Classification (CTC) [Graves, et al., ICML'06]
- RNN Transducer (RNN-T) [Graves, ICML workshop'12]
- Neural Transducer [Jaitly, et al., NIPS'16]
- Monotonic Chunkwise Attention (MoChA) [Chiu, et al., ICLR'18]

MoChA: Monotonic Chunkwise Attention



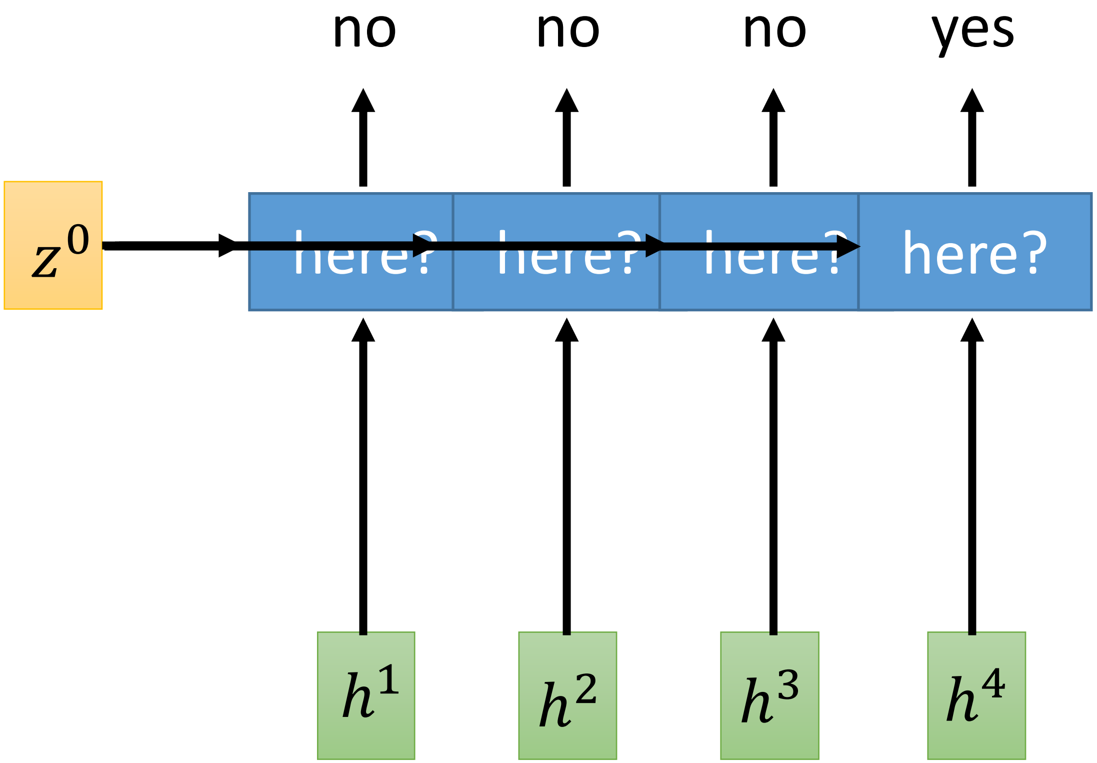
yes: put window at here

no: shift window to the right

similar to attention

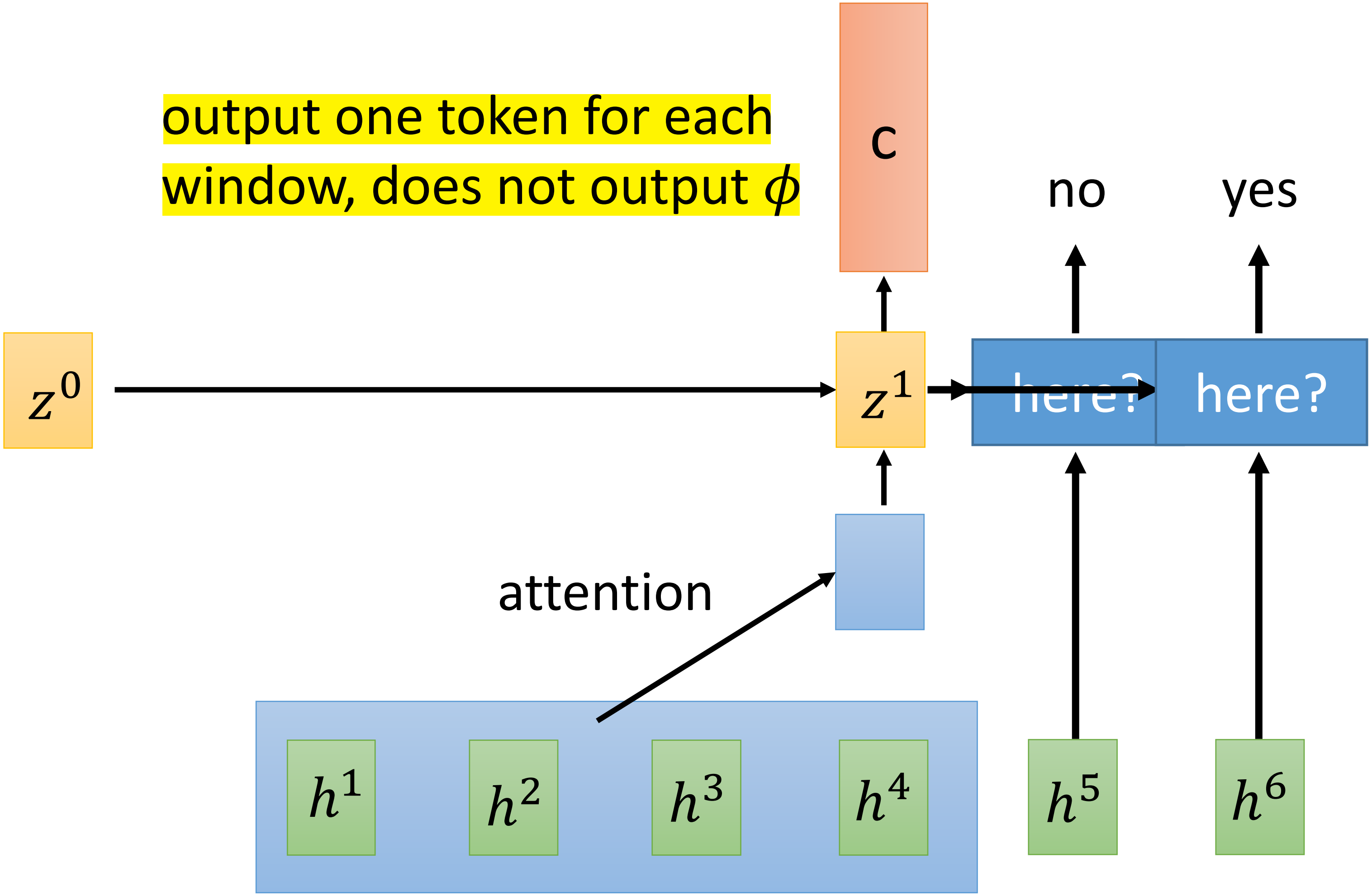
dynamically shift the window

MoChA



MoChA

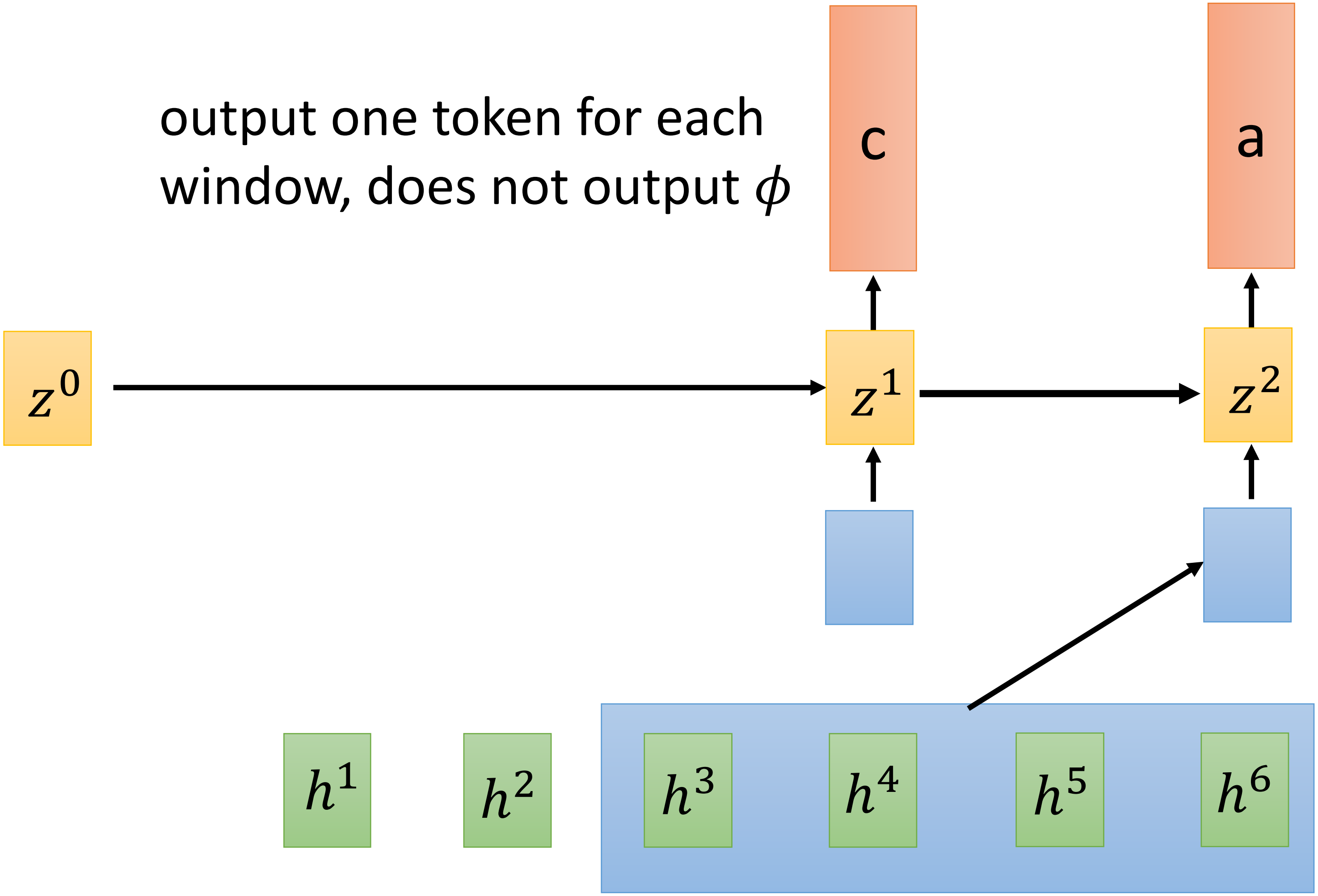
output one token for each window, does not output ϕ



MoChA

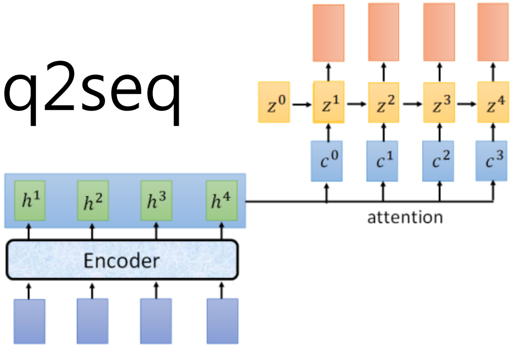
Please refer to the original paper for model training [Chiu, et al., ICLR'18]

output one token for each window, does not output ϕ

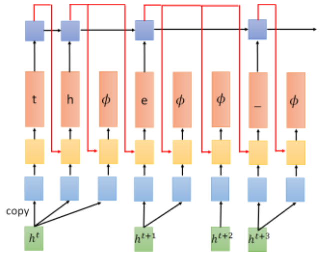


Summary

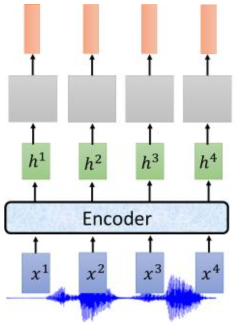
LAS: 就是 seq2seq



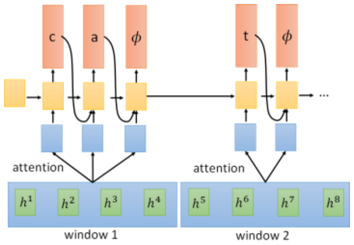
RNN-T: 輸入一個東西可以輸出多個東西的 seq2seq



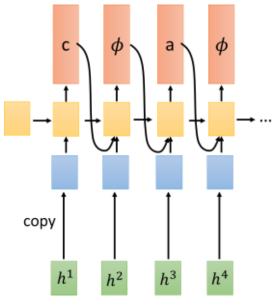
CTC: decoder 是 linear classifier 的 seq2seq



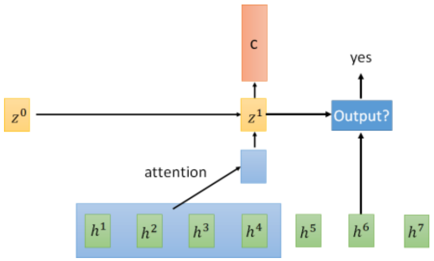
Neural Transducer: 每次輸入一個 window 的 RNN-T



RNA: 輸入一個東西就要輸出一個東西的 seq2seq



MoCha: window 移動伸縮自如的 Neural Transducer



Reference

- [Li, et al., ICASSP’19] Bo Li, Yu Zhang, Tara Sainath, Yonghui Wu, William Chan, Bytes are All You Need: End-to-End Multilingual Speech Recognition and Synthesis with Bytes, ICASSP 2019
- [Bahdanau. et al., ICLR’15] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, ICLR, 2015
- [Bahdanau. et al., ICASSP’16] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, Yoshua Bengio, End-to-End Attention-based Large Vocabulary Speech Recognition, ICASSP, 2016
- [Chan, et al., ICASSP’16] William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, Listen, Attend and Spell, ICASSP, 2016
- [Chiu, et al., ICLR’18] Chung-Cheng Chiu, Colin Raffel, Monotonic Chunkwise Attention, ICLR, 2018
- [Chiu, et al., ICASSP’18] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, Michiel Bacchiani, State-of-the-art Speech Recognition With Sequence-to-Sequence Models, ICASSP, 2018

Reference

- [Chorowski. et al., NIPS’15] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio, Attention-Based Models for Speech Recognition, NIPS, 15
- [Huang, et al., arXiv’19] Hongzhao Huang, Fuchun Peng, An Empirical Study of Efficient ASR Rescoring with Transformers, arXiv, 2019
- [Graves, et al., ICML’06] Alex Graves, Santiago Fernández, Faustino Gomez, Jurgen Schmidhuber, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". In Proceedings of the International Conference on Machine Learning, ICML, 2006
- [Graves, ICML workshop’12] Alex Graves, Sequence Transduction with Recurrent Neural Networks, ICML workshop, 2012
- [Graves, et al., ICML’14] Alex Graves, Navdeep Jaitly, Towards end-to-end speech recognition with recurrent neural networks, ICML, 2014
- [Lu, et al., INTERSPEECH’15] Liang Lu, Xingxing Zhang, Kyunghyun Cho, Steve Renals, A Study of the Recurrent Neural Network Encoder-Decoder for Large Vocabulary Speech Recognition, INTERSPEECH, 2015
- [Luong, et al., EMNLP’15] Minh-Thang Luong, Hieu Pham, Christopher D. Manning, Effective Approaches to Attention-based Neural Machine Translation, EMNLP, 2015

Reference

- [Karita, et al., ASRU'19] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, Wangyou Zhang, A Comparative Study on Transformer vs RNN in Speech Applications, ASRU, 2019
- [Soltau, et al., ICASSP'14] Hagen Soltau, George Saon, Tara N. Sainath, Joint training of convolutional and non-convolutional neural networks, ICASSP, 2014
- [Sak, et al., INTERSPEECH'15] Haşim Sak, Andrew Senior, Kanishka Rao, Françoise Beaufays, Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition, INTERSPEECH, 2015
- [Sak, et al., INTERSPEECH'17] Haşim Sak, Matt Shannon, Kanishka Rao, Françoise Beaufays, Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping, INTERSPEECH, 2017
- [Jaitly, et al., NIPS'16] Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, Ilya Sutskever, David Sussillo, Samy Bengio, An Online Sequence-to-Sequence Model Using Partial Conditioning, NIPS, 2016

Reference

- [Rao, et al., ASRU’17] Kanishka Rao, Haşim Sak, Rohit Prabhavalkar, Exploring Architectures, Data and Units For Streaming End-to-End Speech Recognition with RNN-Transducer, ASRU. 2017
- [Peddinti, et al., INTERSPEECH’15] Vijayaditya Peddinti, Daniel Povey, Sanjeev Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts, INTERSPEECH, 2015
- [Yeh, et al., arXiv’19] Ching-Feng Yeh, Jay Mahadeokar, Kaustubh Kalgaonkar, Yongqiang Wang, Duc Le, Mahaveer Jain, Kjell Schubert, Christian Fuegen, Michael L. Seltzer, Transformer-Transducer: End-to-End Speech Recognition with Self-Attention, arXiv, 2019
- [Zeyer, et al., ASRU’19] Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, Hermann Ney, A Comparison of Transformer and LSTM Encoder Decoder Models for ASR, ASRU, 2019