NLP项目

Seq2Seq

课程内容

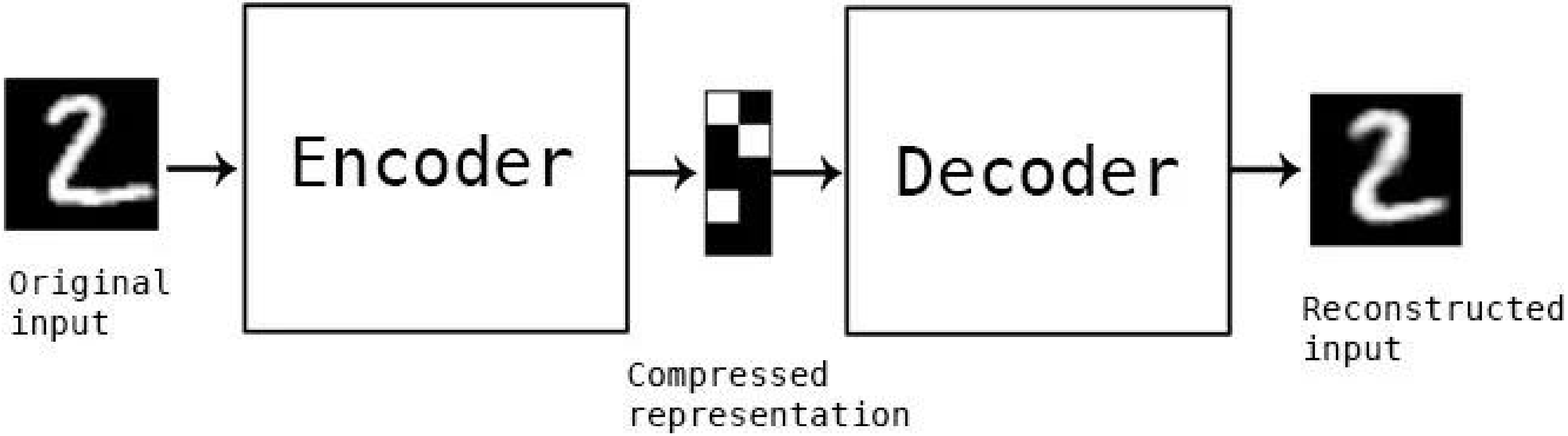
- 自编码神经网络
- RNN、LSTM神经网络
- Seq2Seq网络结构讲解

自编码器

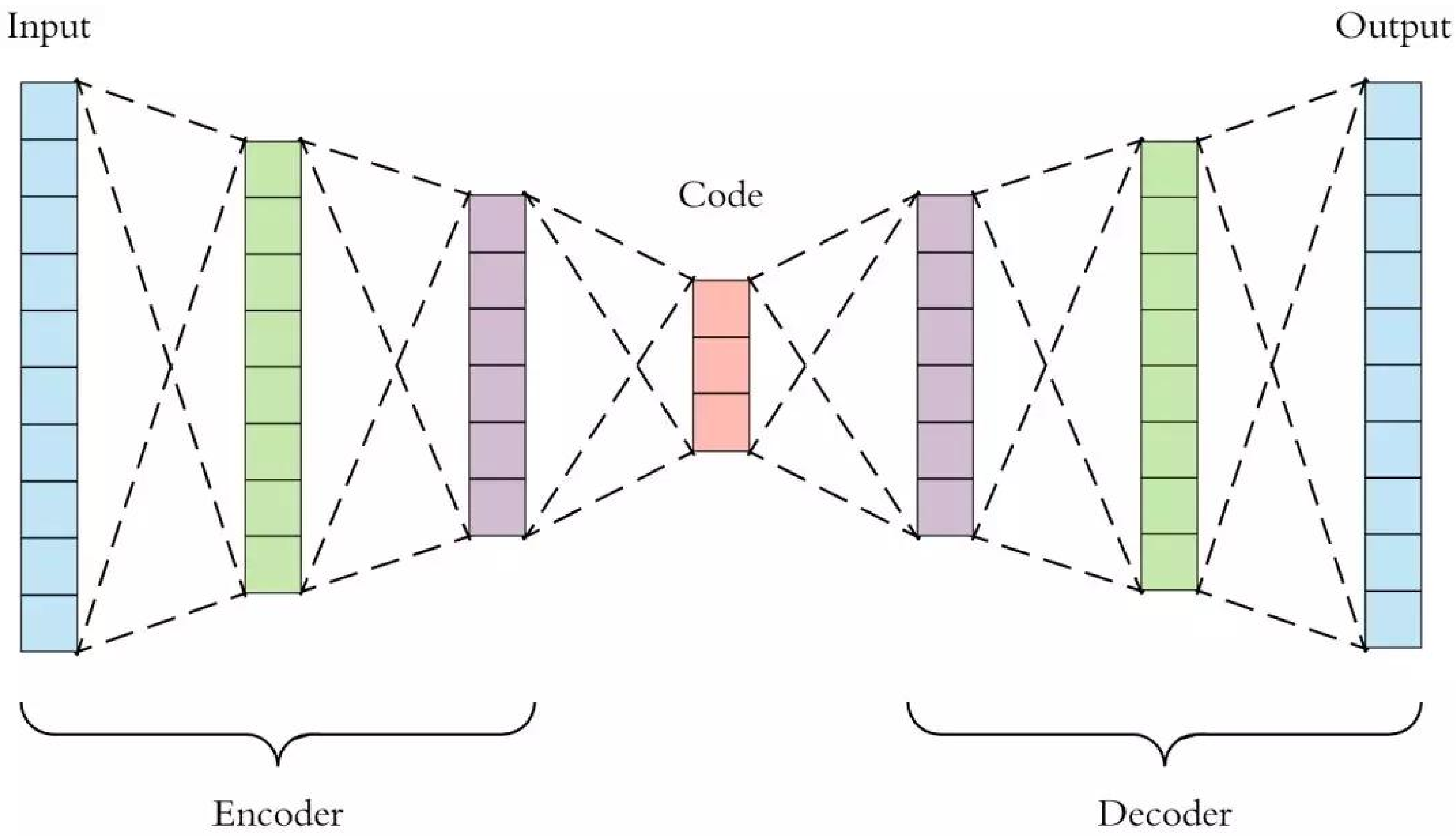
- 自编码器作为一种深度学习领域无监督的算法，本质上是一种数据压缩算法，和生成对抗网络一样，属于生成算法的一种。
- 自编码器(AutoEncoder，AE)就是一种利用反向传播使得输出值等于输入值的神经网络，它将输入压缩成潜在特征/高阶特征，然后将这种表征重构输出。主要包含以下三个特征：
  - 数据相关性。
  - 数据有损性。
  - 自动学习性。

自编码器

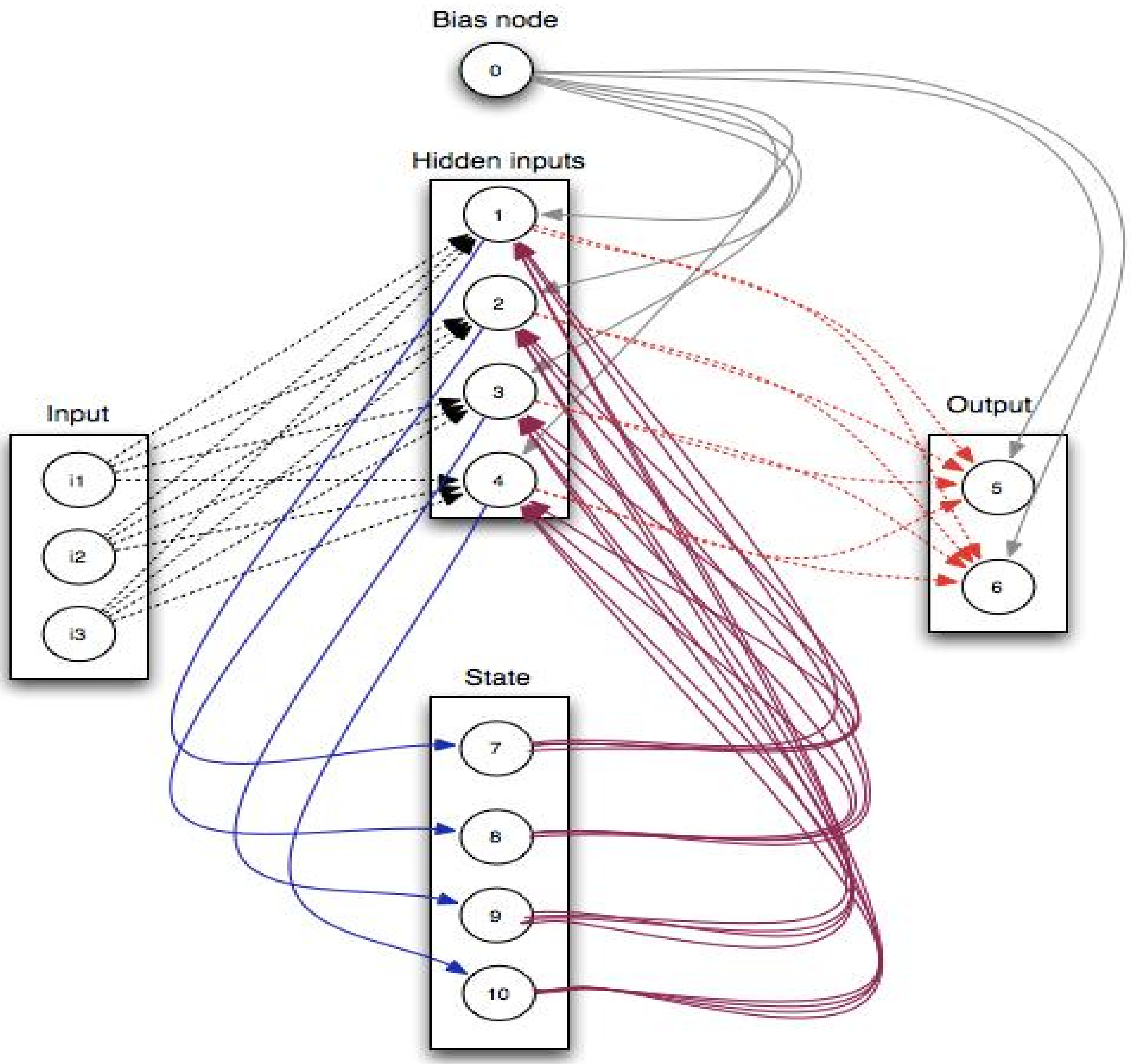
- 构建一个自编码器主要包括两部分：编码器(Encoder)和解码器(Decoder)。编码器将输入压缩为潜在空间特征，解码器将潜在空间特征重构输出。
- **自编码的核心价值是在于提取潜在的高阶空间特征信息。**主要应用是两个方面：**数据去燥**以及进行**可视化降维。**



自编码器

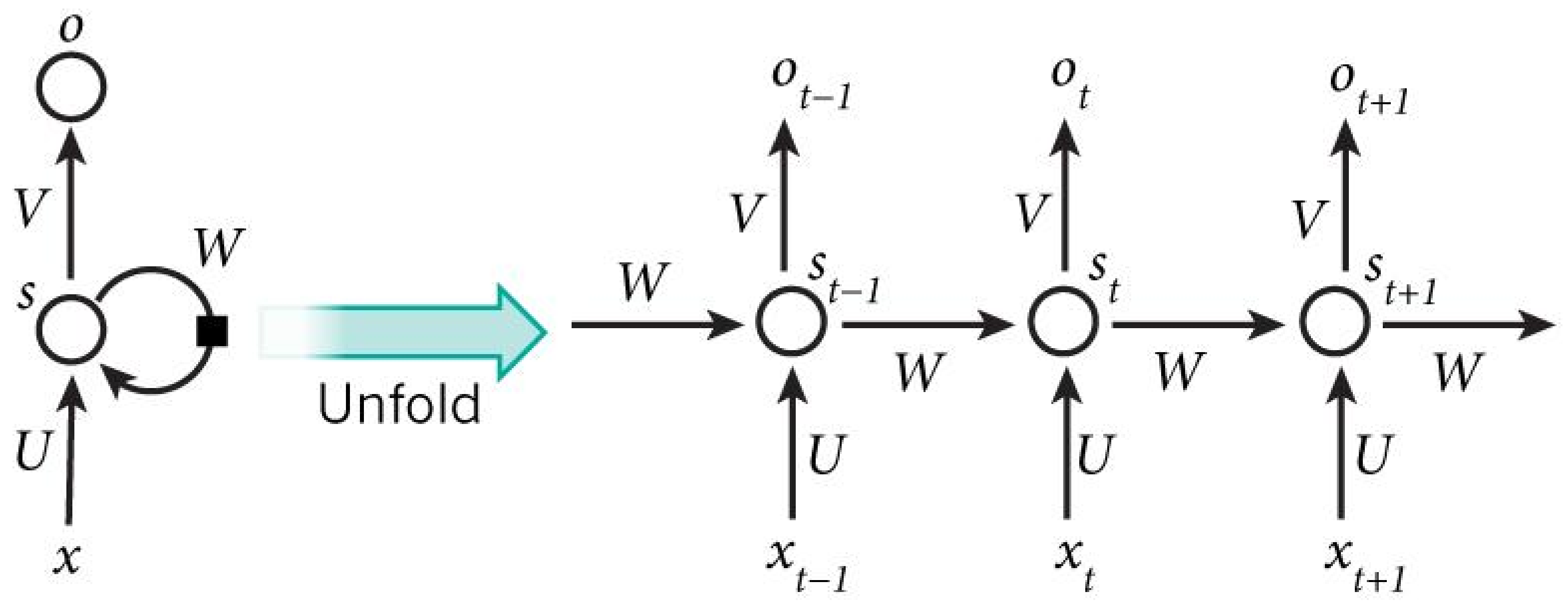


RNN回顾



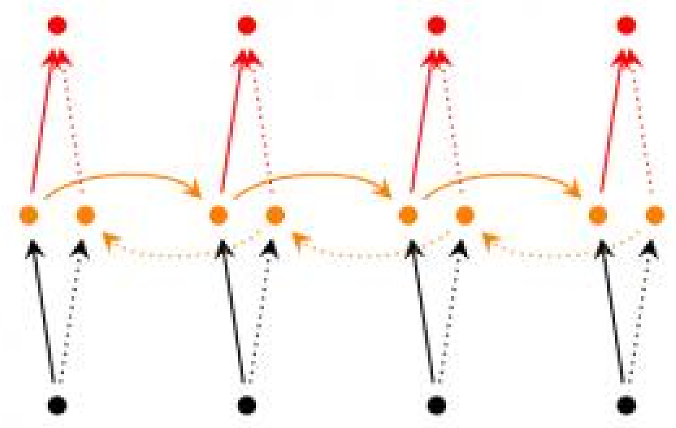
Bias node
0
Hidden inputs
1
2
3
4
Input
i1
i2
i3
Output
5
6
State
7
8
9
10

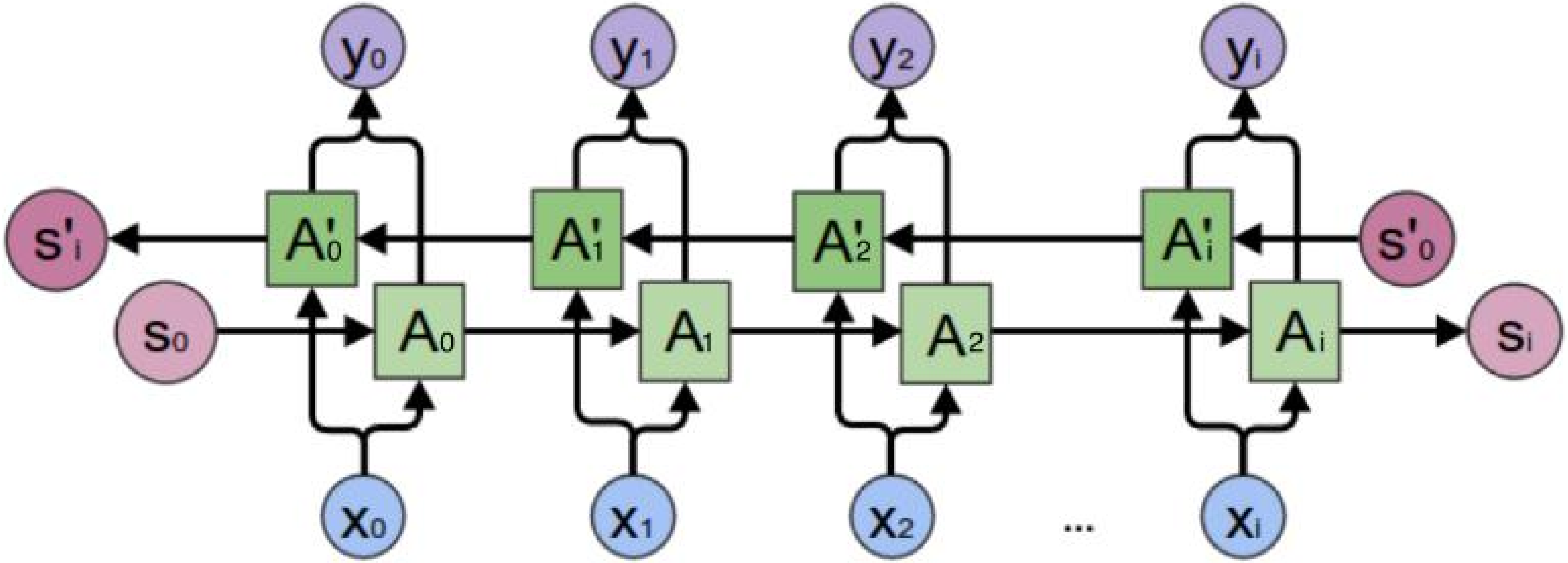
RNN回顾



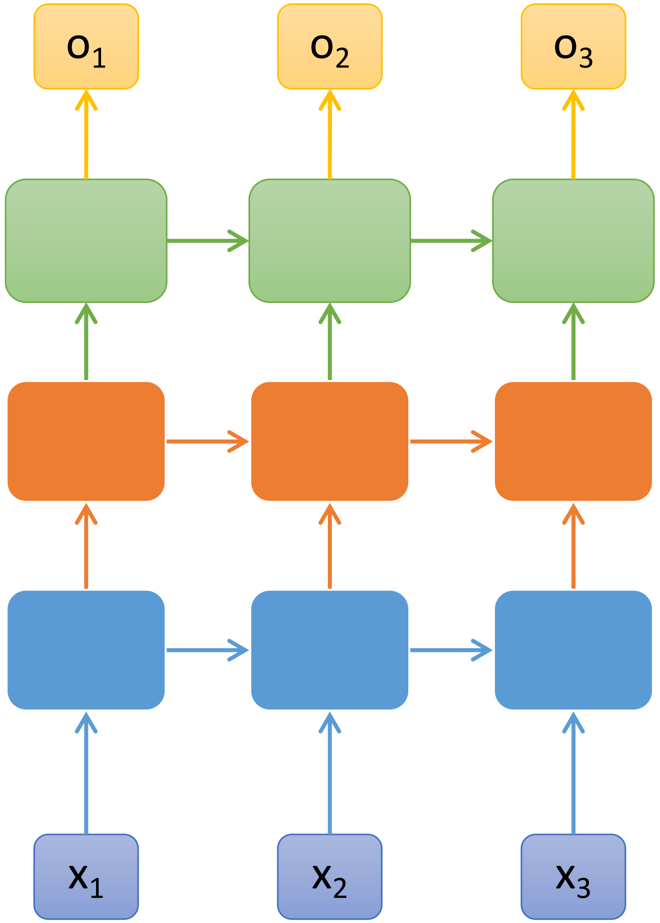
$$s_t = Ux_t + Wh_{t-1}$$
$$h_t = f(Ux_t + Wh_{t-1})$$
$$o_t = g(Vh_t)$$

Bidirectional RNN回顾

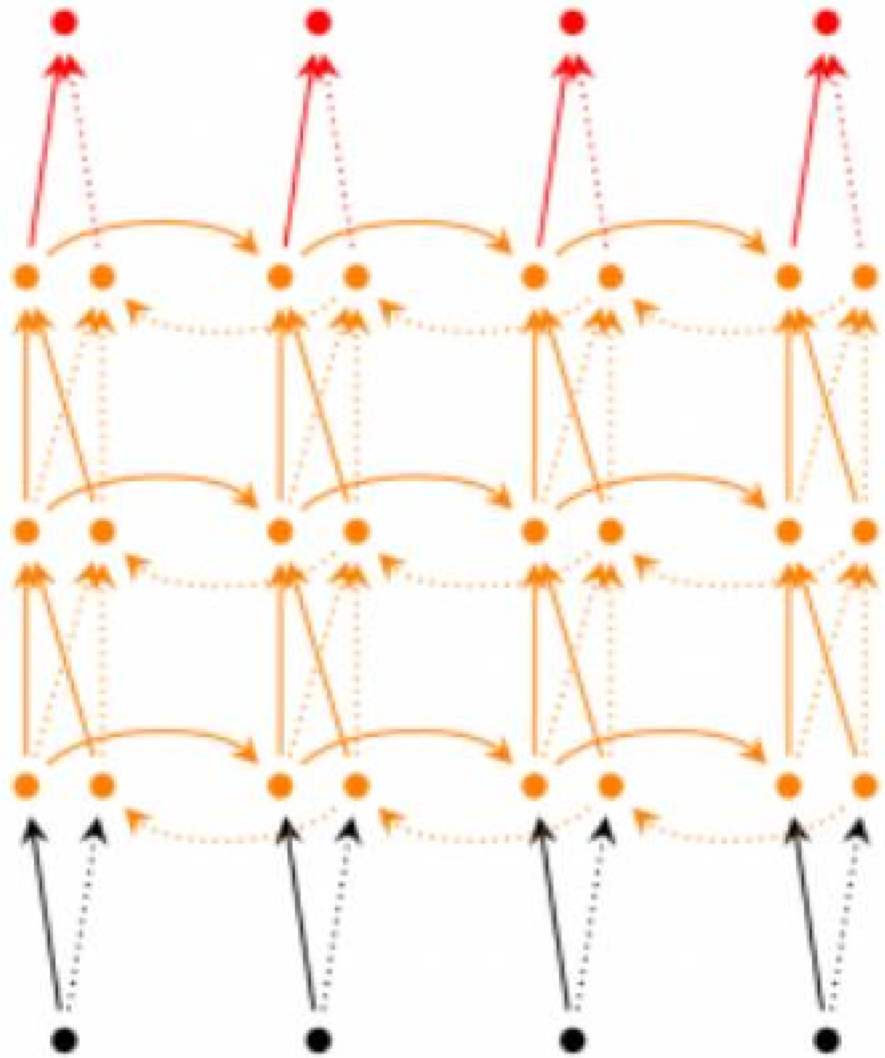




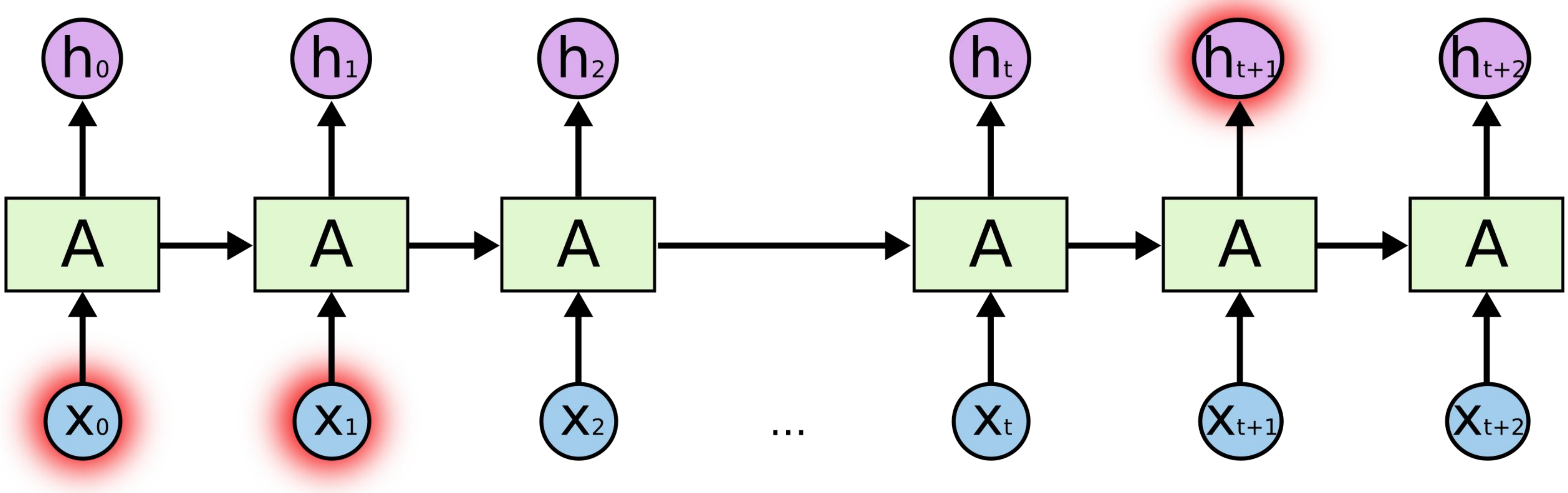
Deep RNN回顾



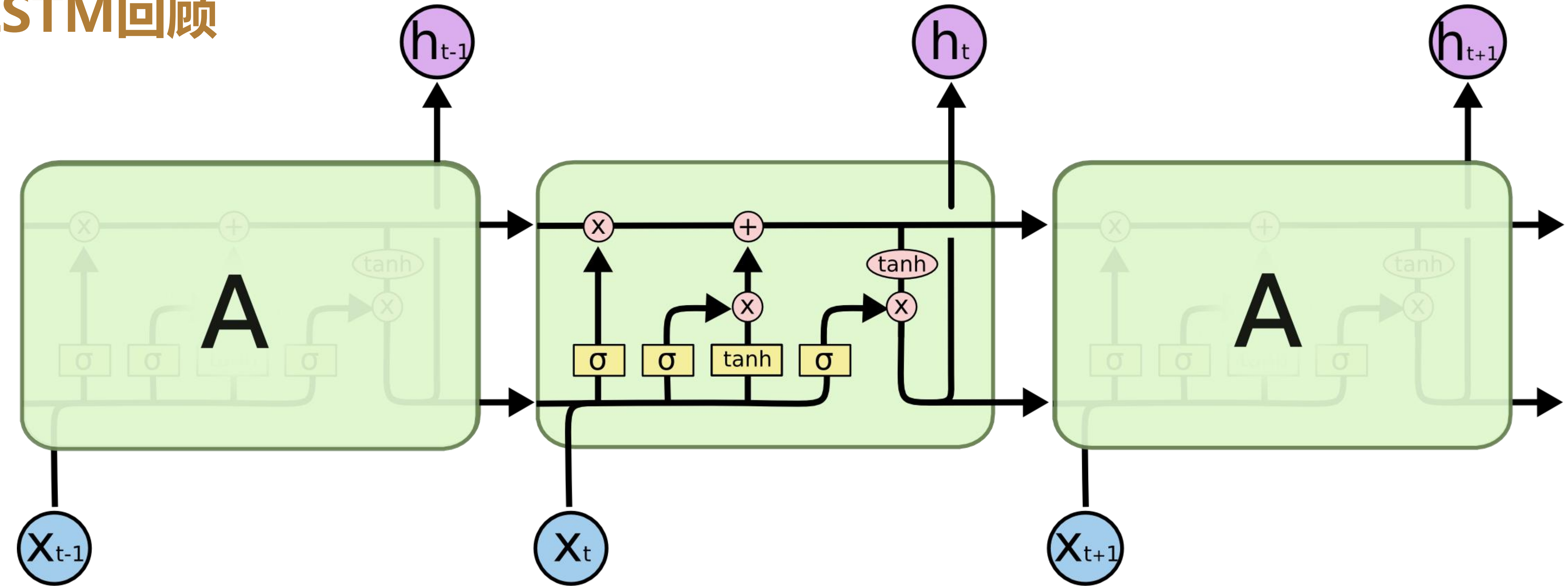
Deep Bidirectional RNN回顾



LSTM回顾



LSTM回顾



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$

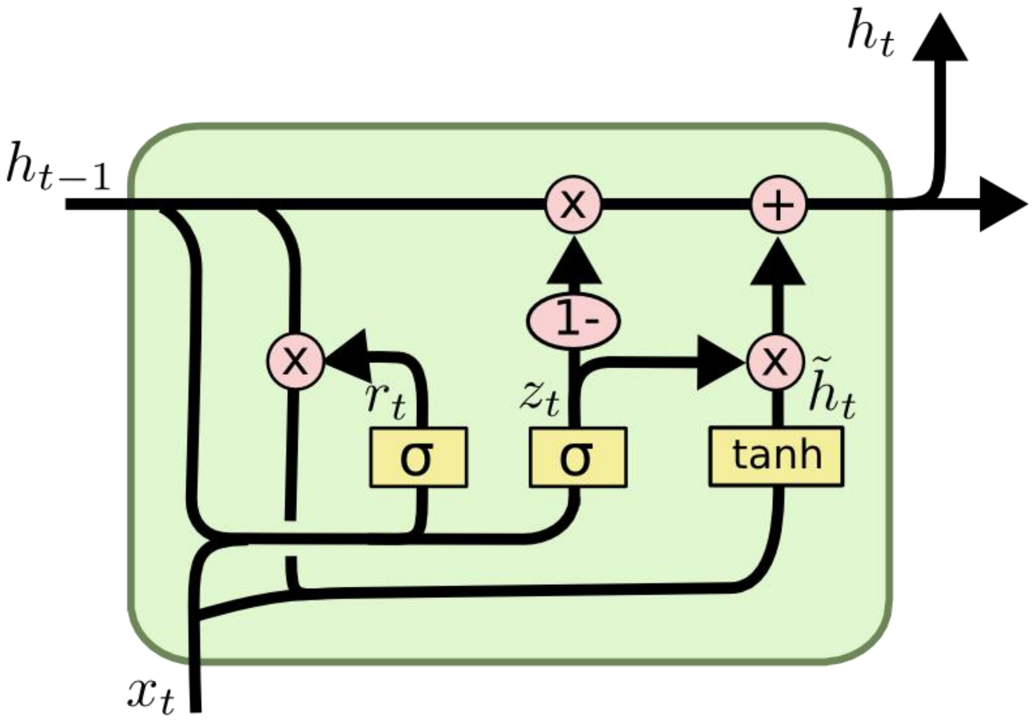
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$o_t = \sigma\left(W_o [h_{t-1}, x_t] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

GRU回顾



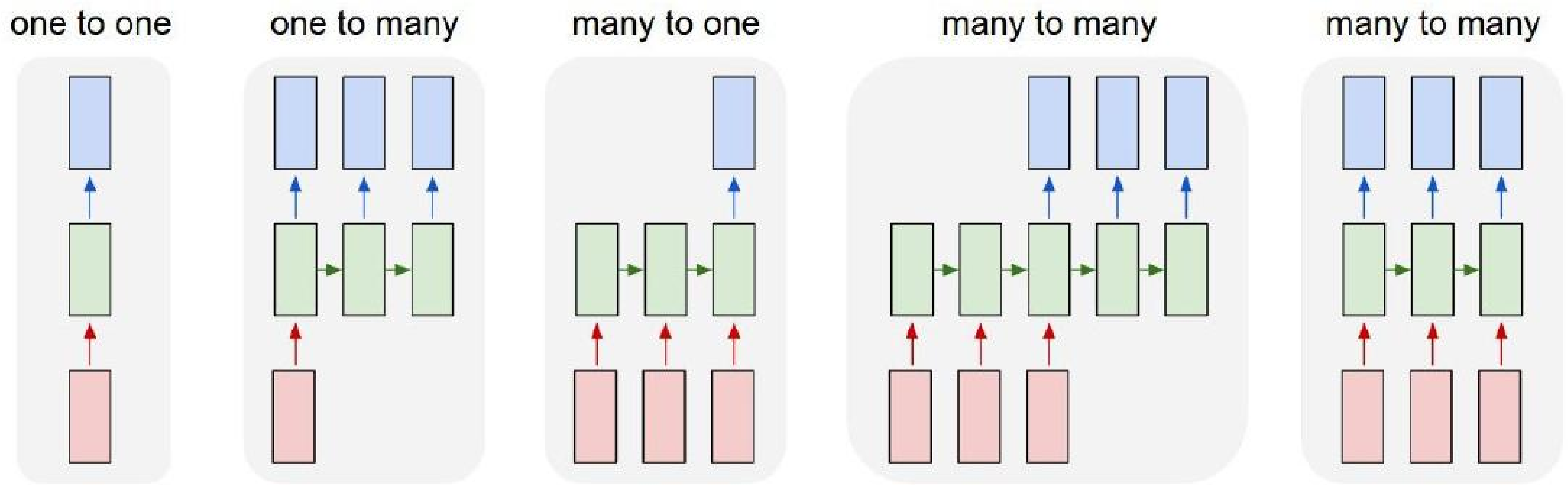
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

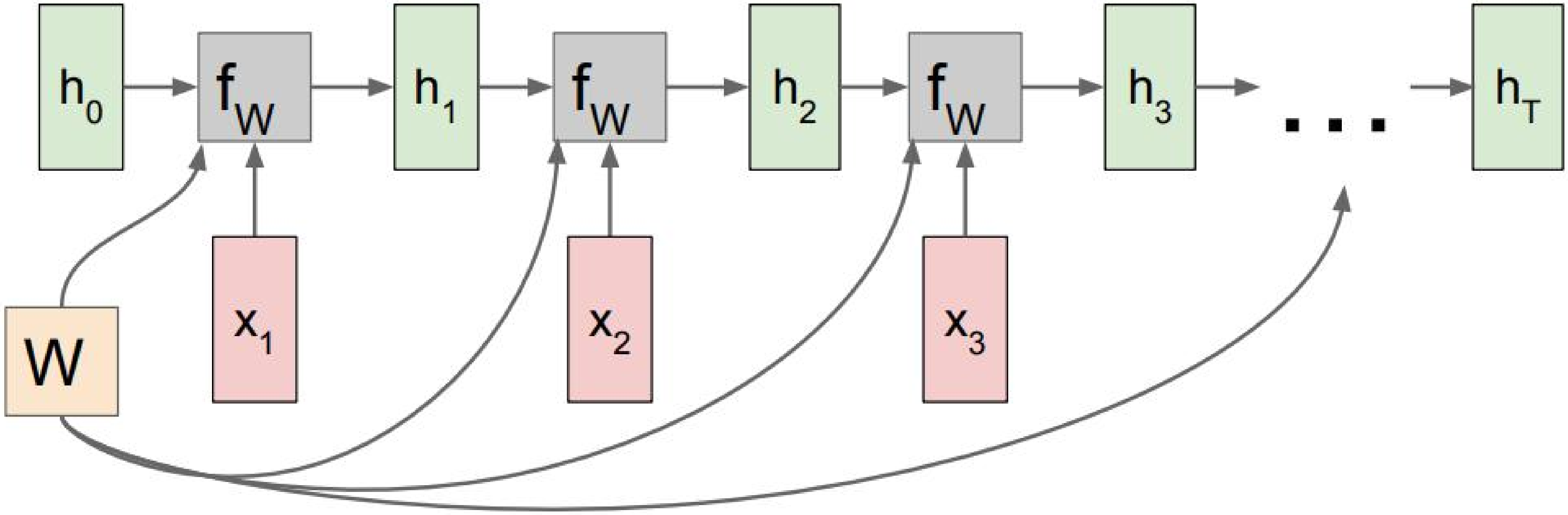
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

RNN结构回顾



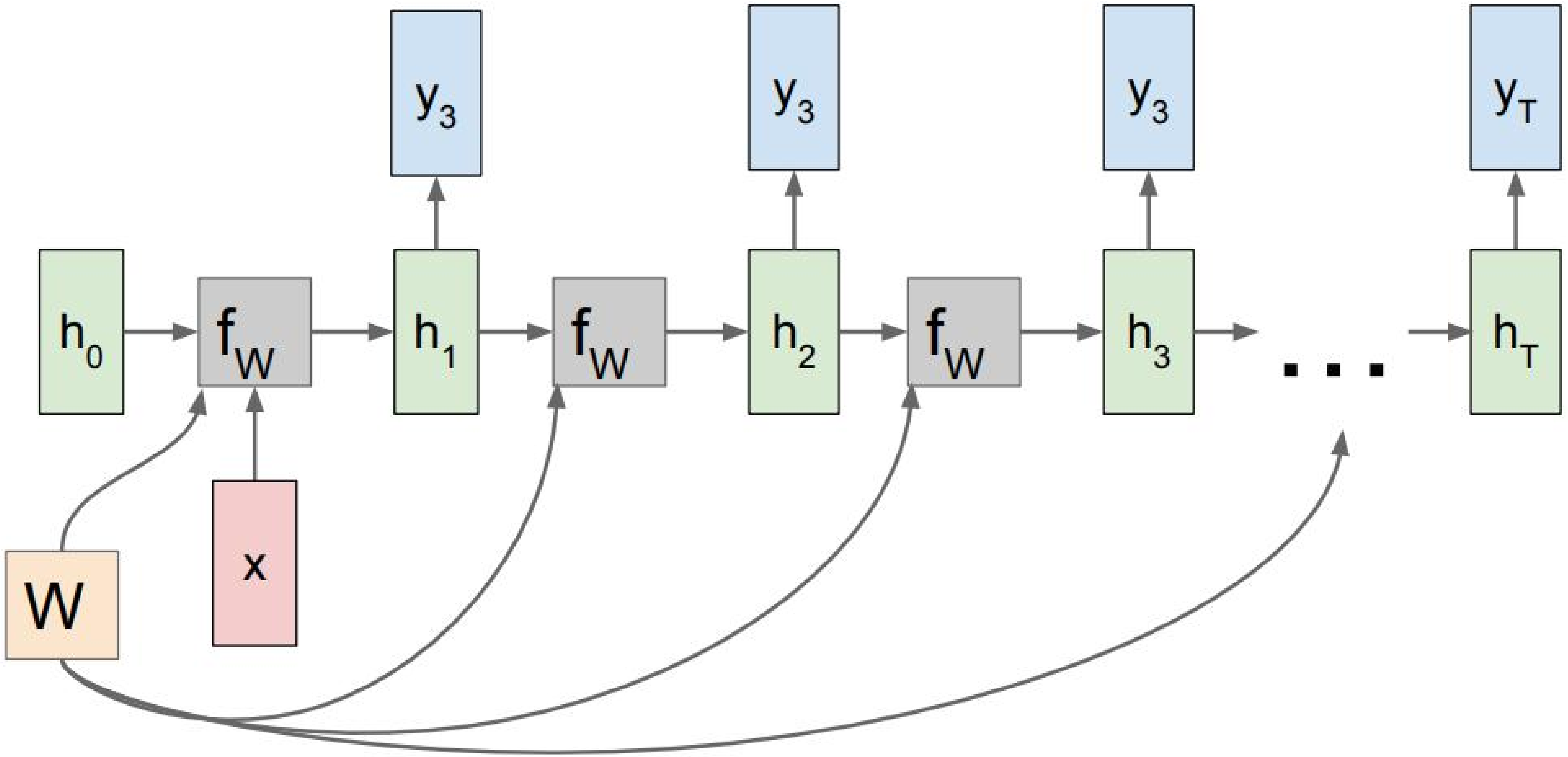
RNN结构回顾

Re-use the same weight matrix at every time-step



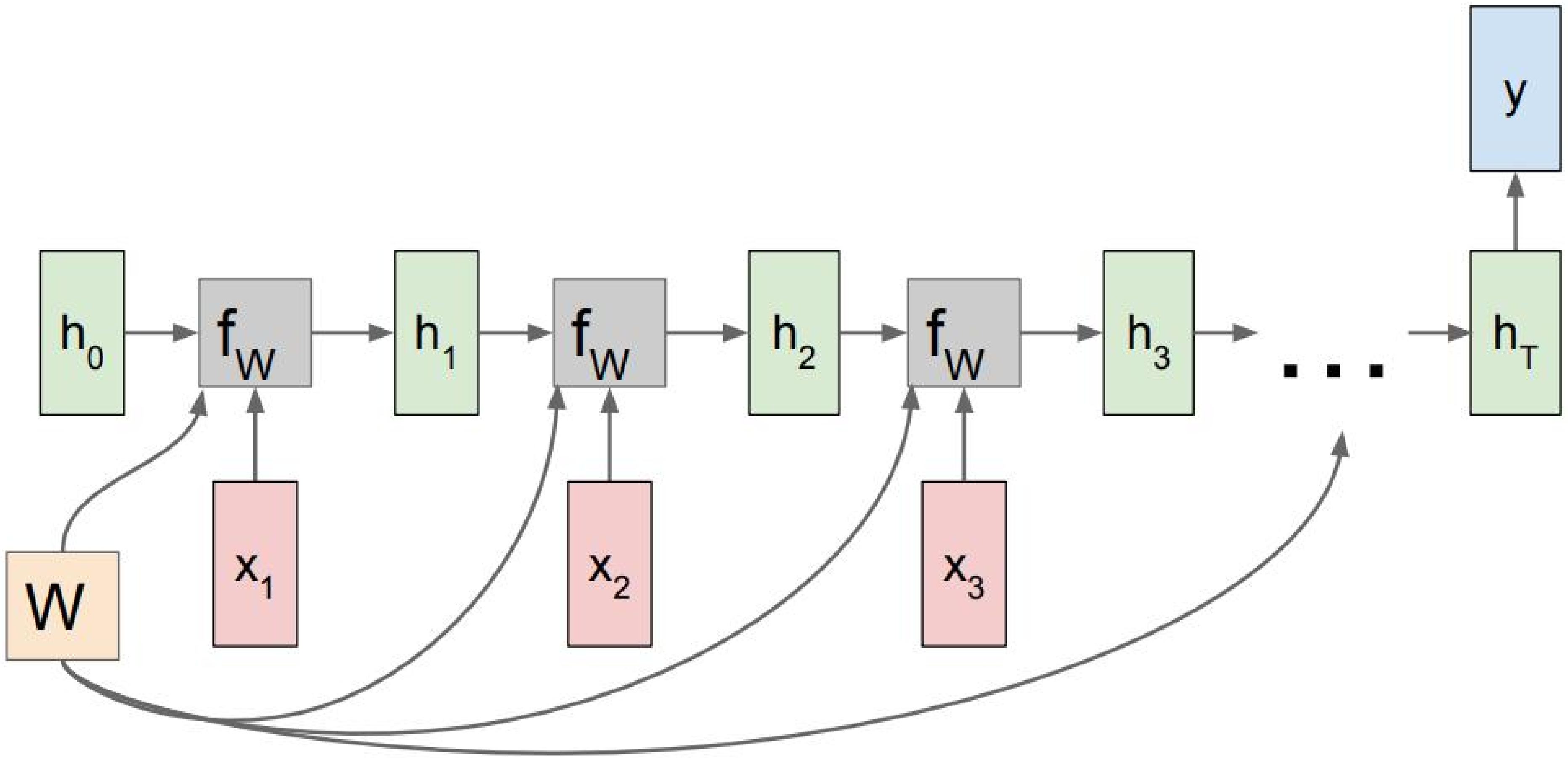
RNN结构回顾

RNN: Computational Graph: One to Many

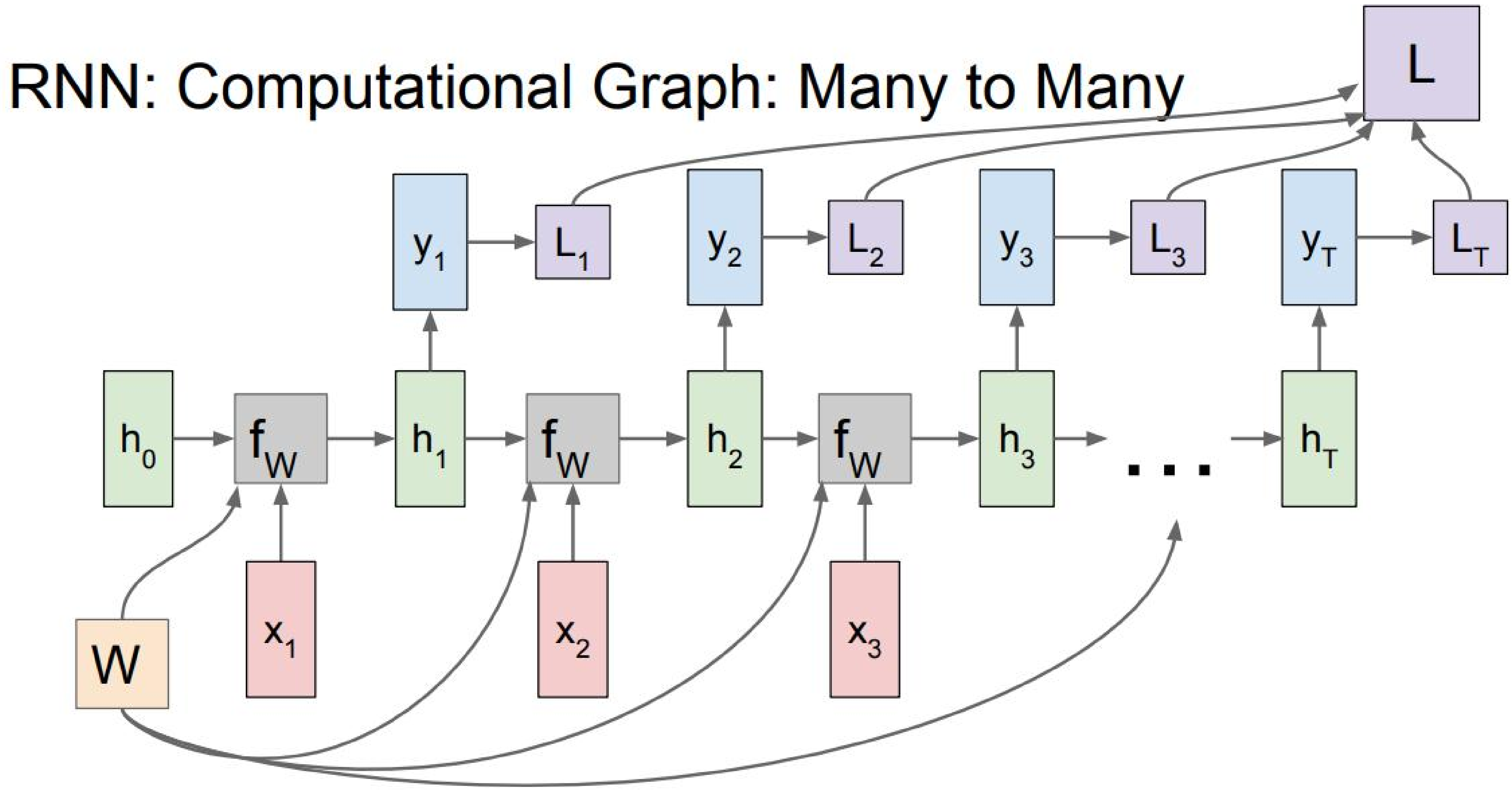


RNN结构回顾

RNN: Computational Graph: Many to One

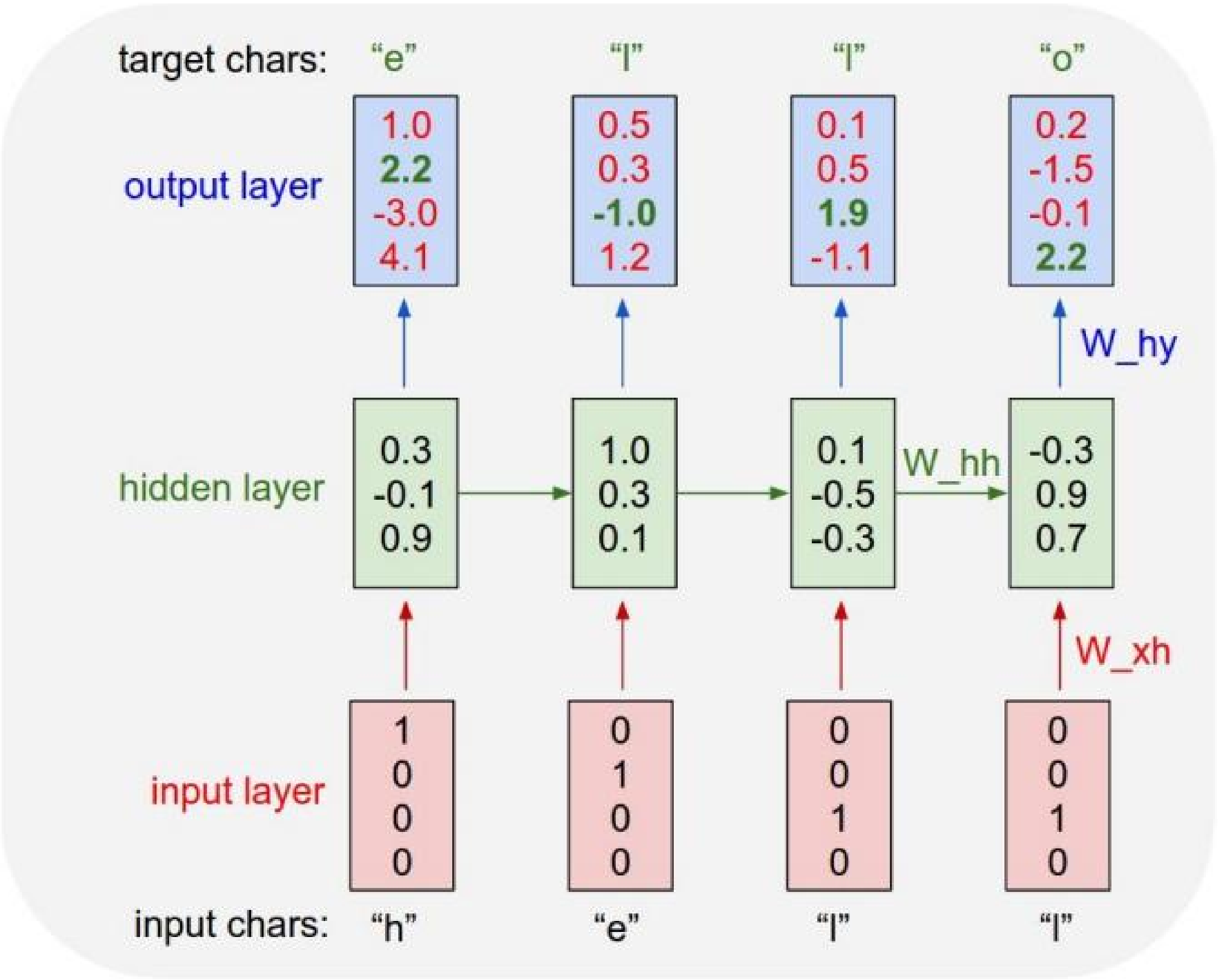


RNN结构回顾

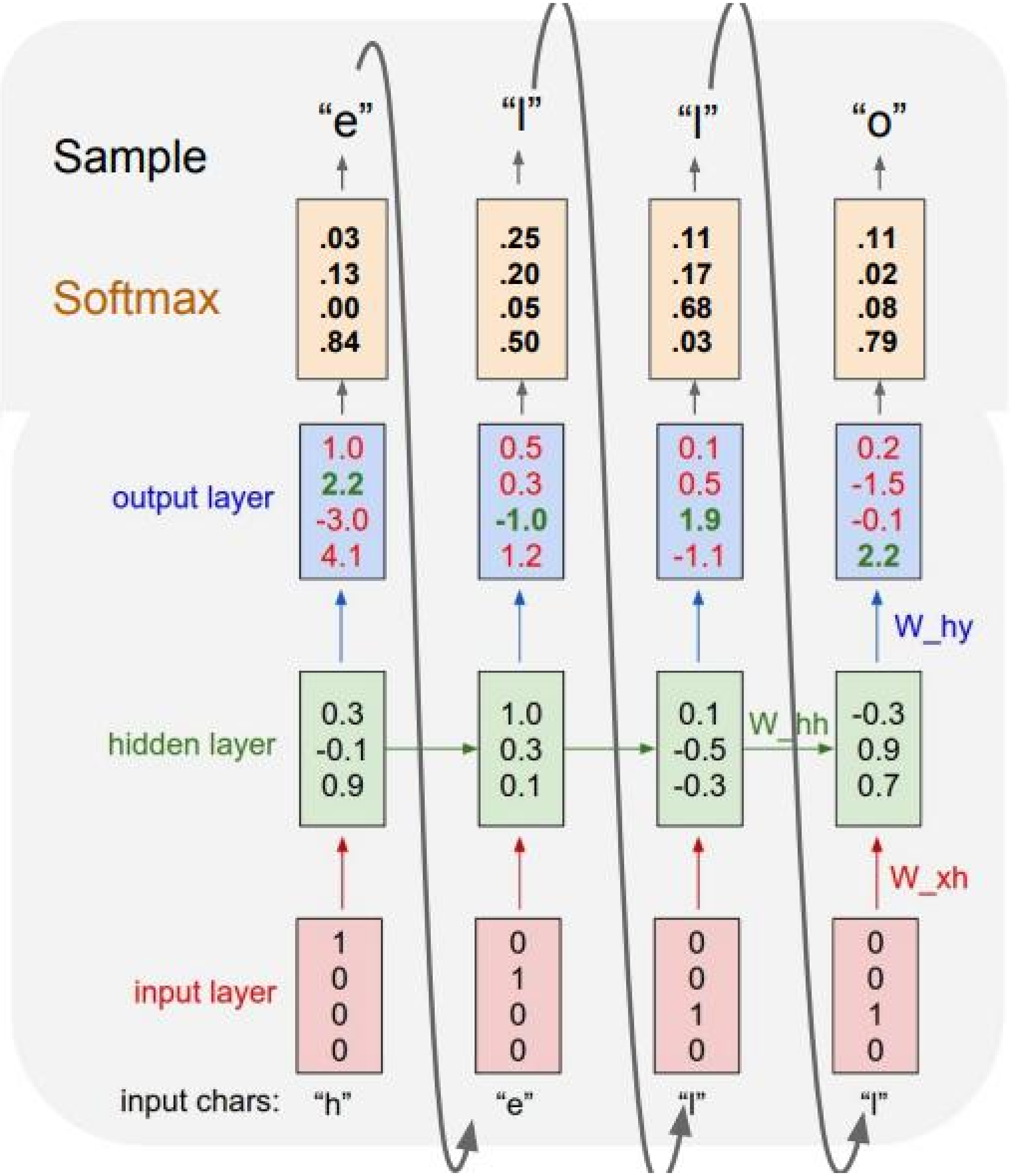


RNN: Computational Graph: Many to Many
L
y1
L1
y2
L2
y3
L3
yT
LT
h0
fW
h1
fW
h2
fW
h3
...
hT
W
x1
x2
x3

RNN结构回顾



训练阶段

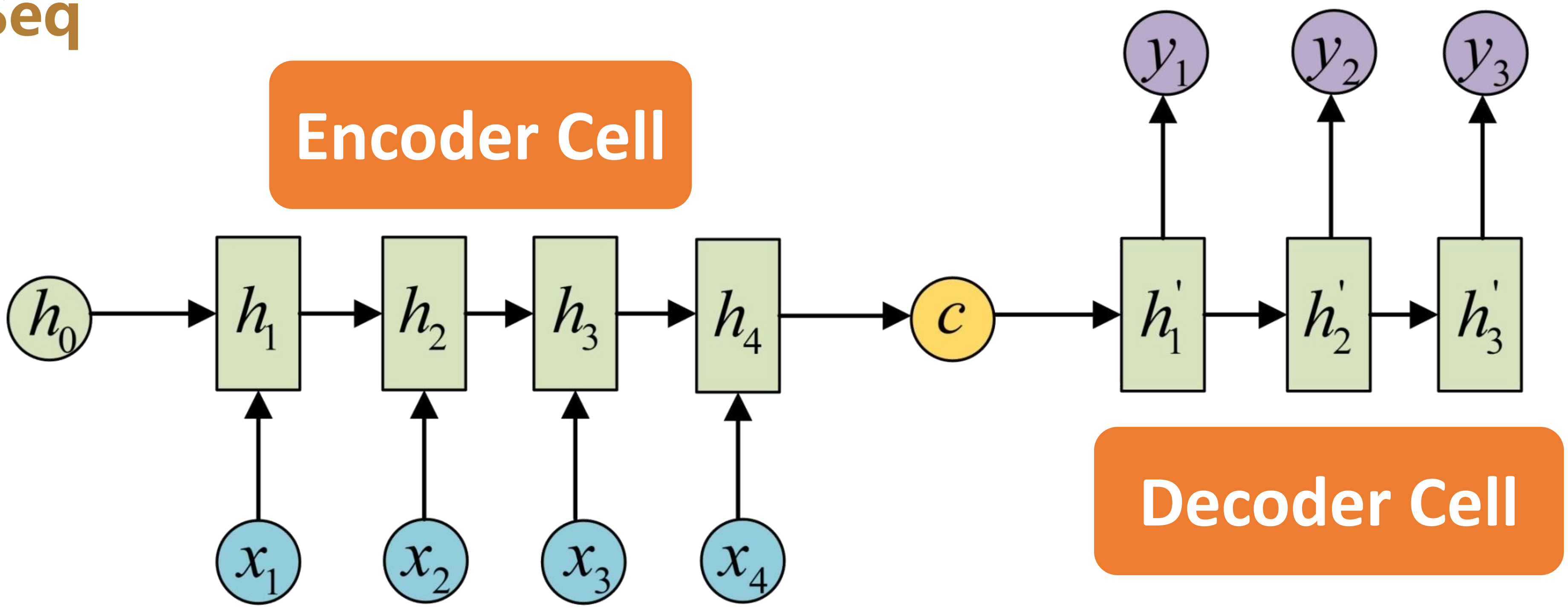


预测阶段

Seq2Seq

- Seq2Seq(Sequence to Sequence)，它被提出于2014年，最早由两篇文章独立地阐述了它主要思想，分别是Google Brain团队的《Sequence to Sequence Learning with Neural Networks》和Yoshua Bengio团队的《Learning Phrase Representation using RNN Encoder-Decoder for Statistical Machine Translation》。
- Seq2Seq属于一种Encoder-Decoder结构。

Seq2Seq



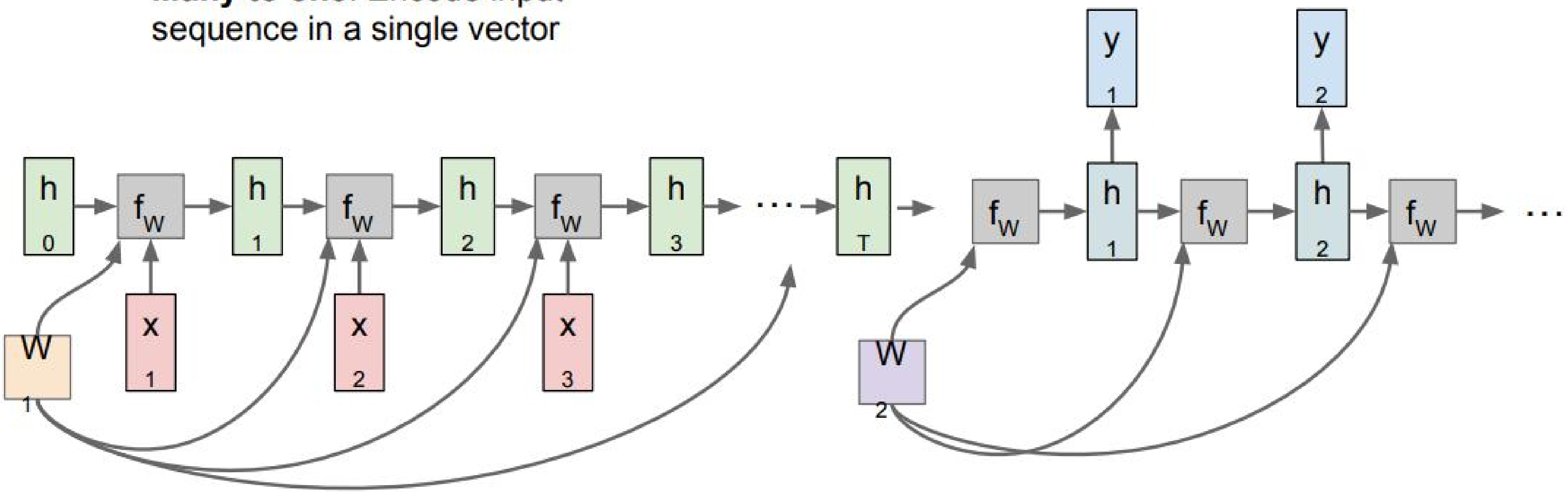
- Encoder-Decoder 这种结构的，其中 Encoder 是一个RNNCell（RNN ， GRU，LSTM 等） 结构。每个 timestep，我们向 Encoder 中输入一个字/词（一般是表示这个字/词的一个实数向量），直到我们输入这个句子的最后一个字/词 X_T，然后输出整个句子的语义向量 c（一般情况下， $c=h_T=F([X_T;h_{T-1}]W)$, X^T 是最后一个timestep输入）。因为 RNN 的特点就是把前面每一步的输入信息都考虑进来了，所以理论上这个 c 就能够把整个句子的信息都包含了，我们可以把 c 当成这个句子的一个语义表示，也就是一个句向量。在 Decoder 中，我们根据 Encoder 得到的句向量 c， 一步一步地把蕴含在其中的信息分析出来。

Seq2Seq

Sequence to Sequence: Many-to-one + one-to-many

One to many: Produce output sequence from single input vector

Many to one: Encode input sequence in a single vector



Seq2Seq

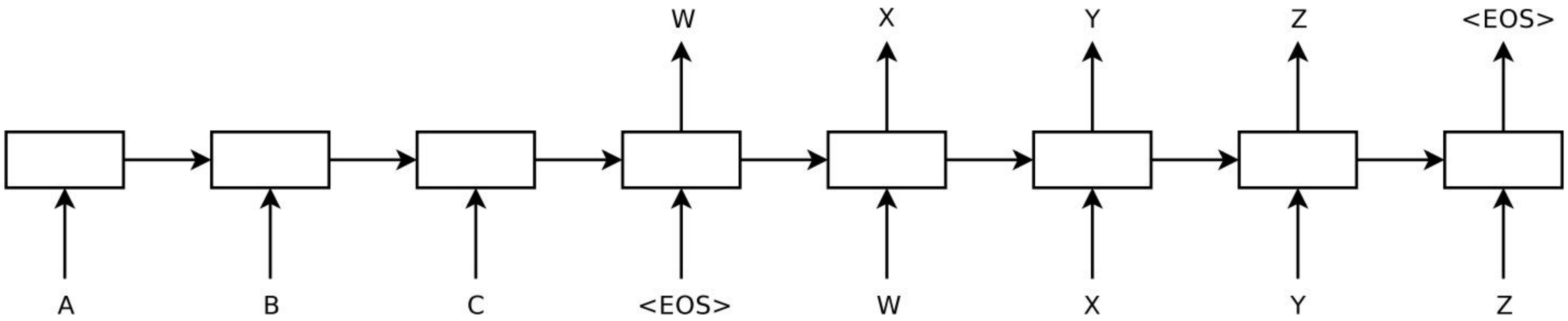
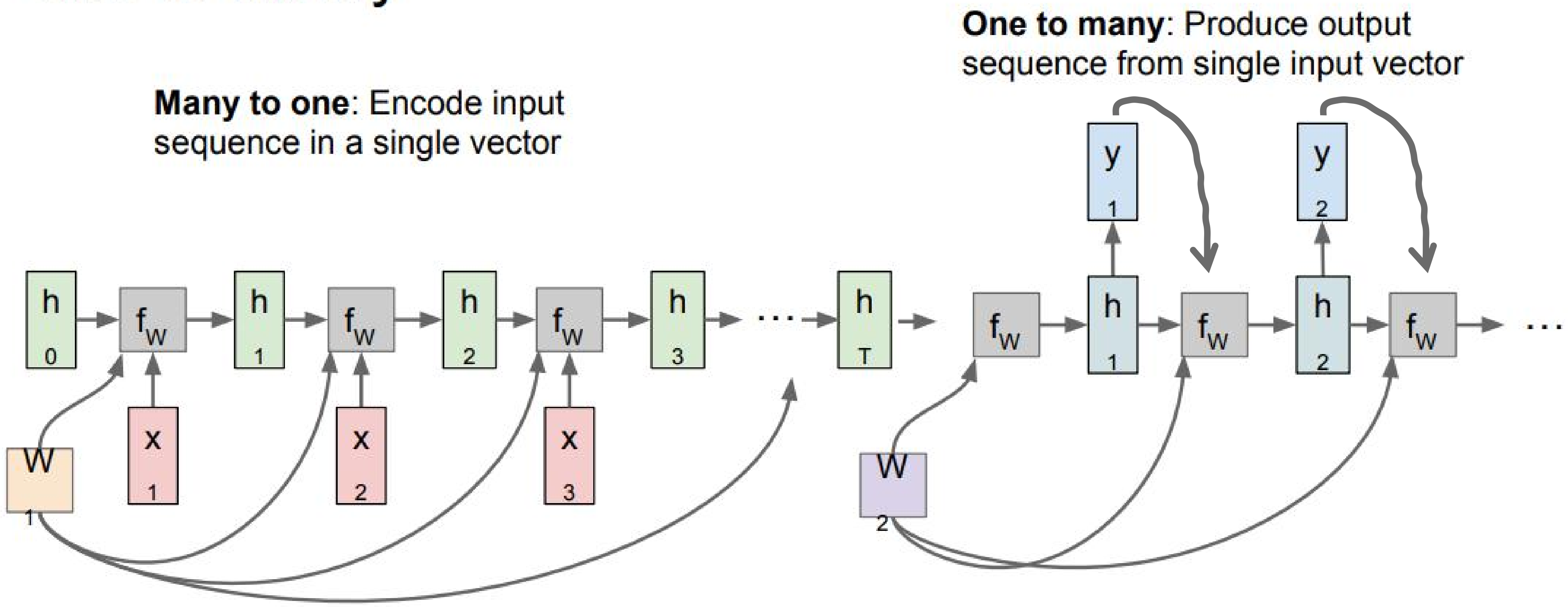


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

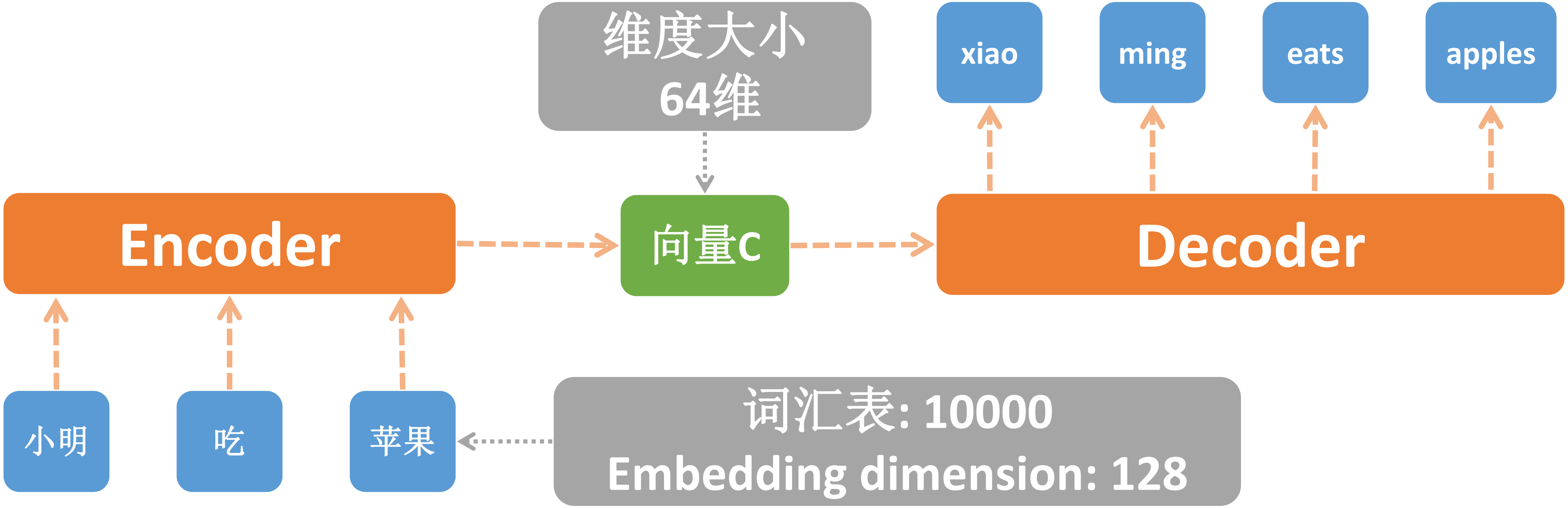
Seq2Seq

Sequence to Sequence: Many-to-one + one-to-many

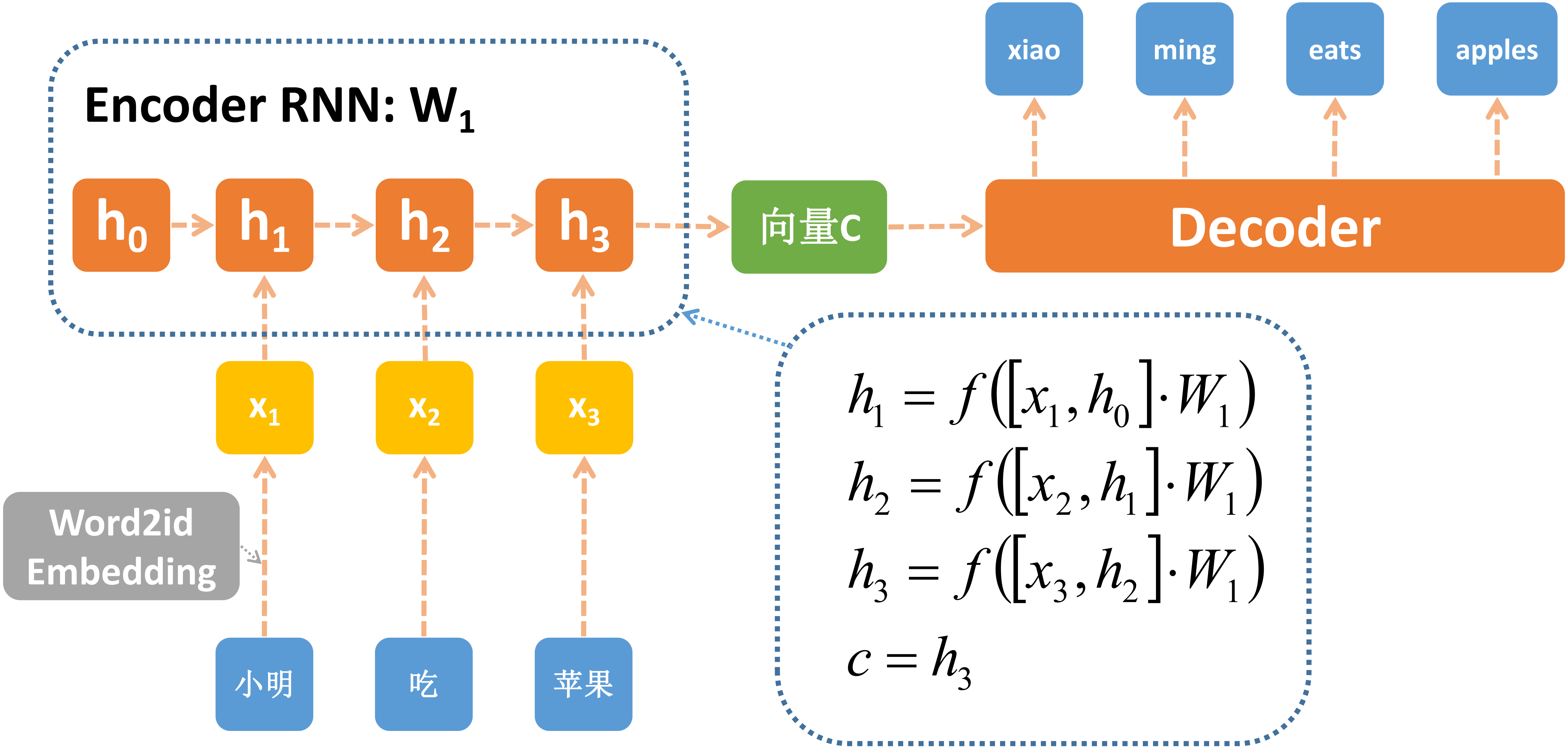


Seq2Seq理解

- 输入：小明　吃　苹果
- 希望输出：xiao ming　eats　apples



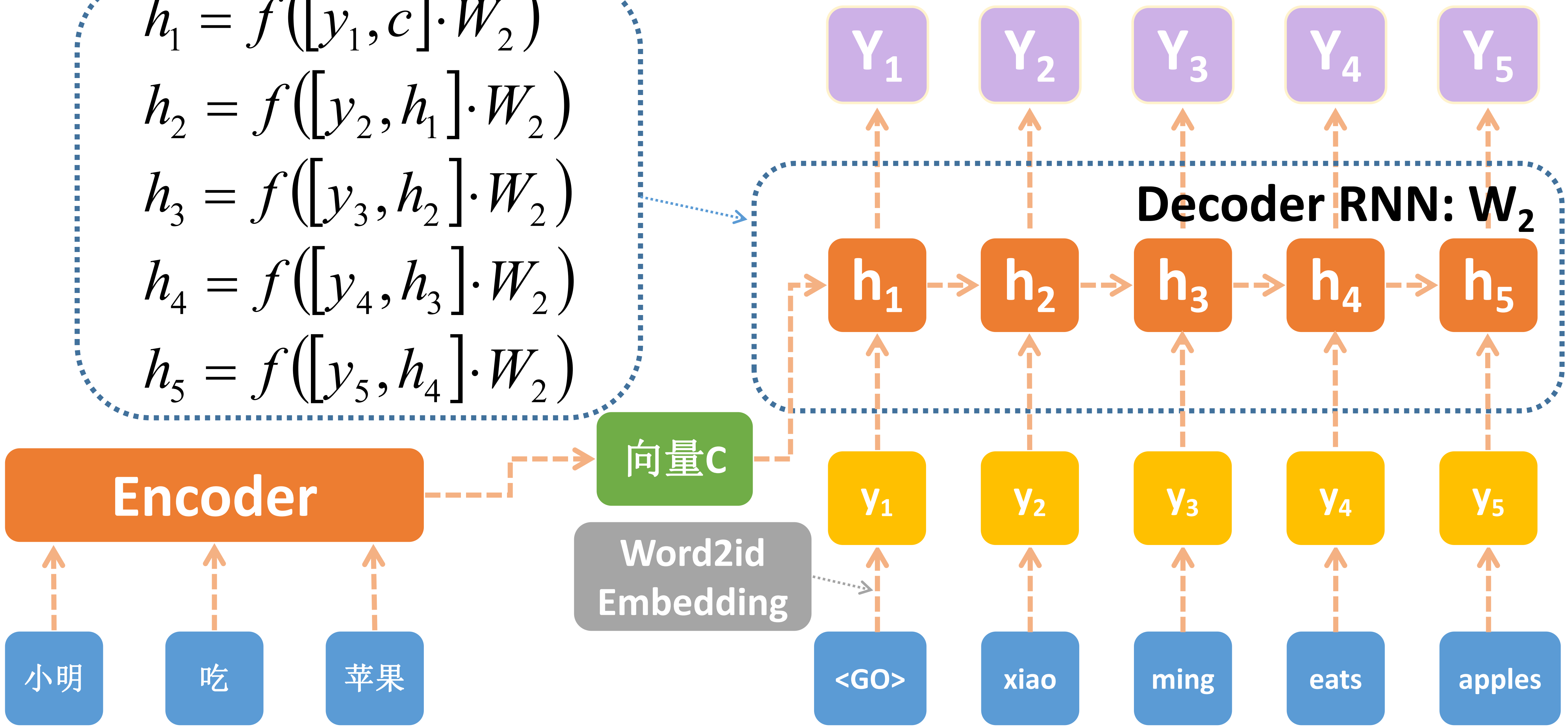
Seq2Seq理解



$$h_1 = f\left(\left[x_1, h_0\right] \cdot W_1\right)$$
$$h_2 = f\left(\left[x_2, h_1\right] \cdot W_1\right)$$
$$h_3 = f\left(\left[x_3, h_2\right] \cdot W_1\right)$$
$$c = h_3$$

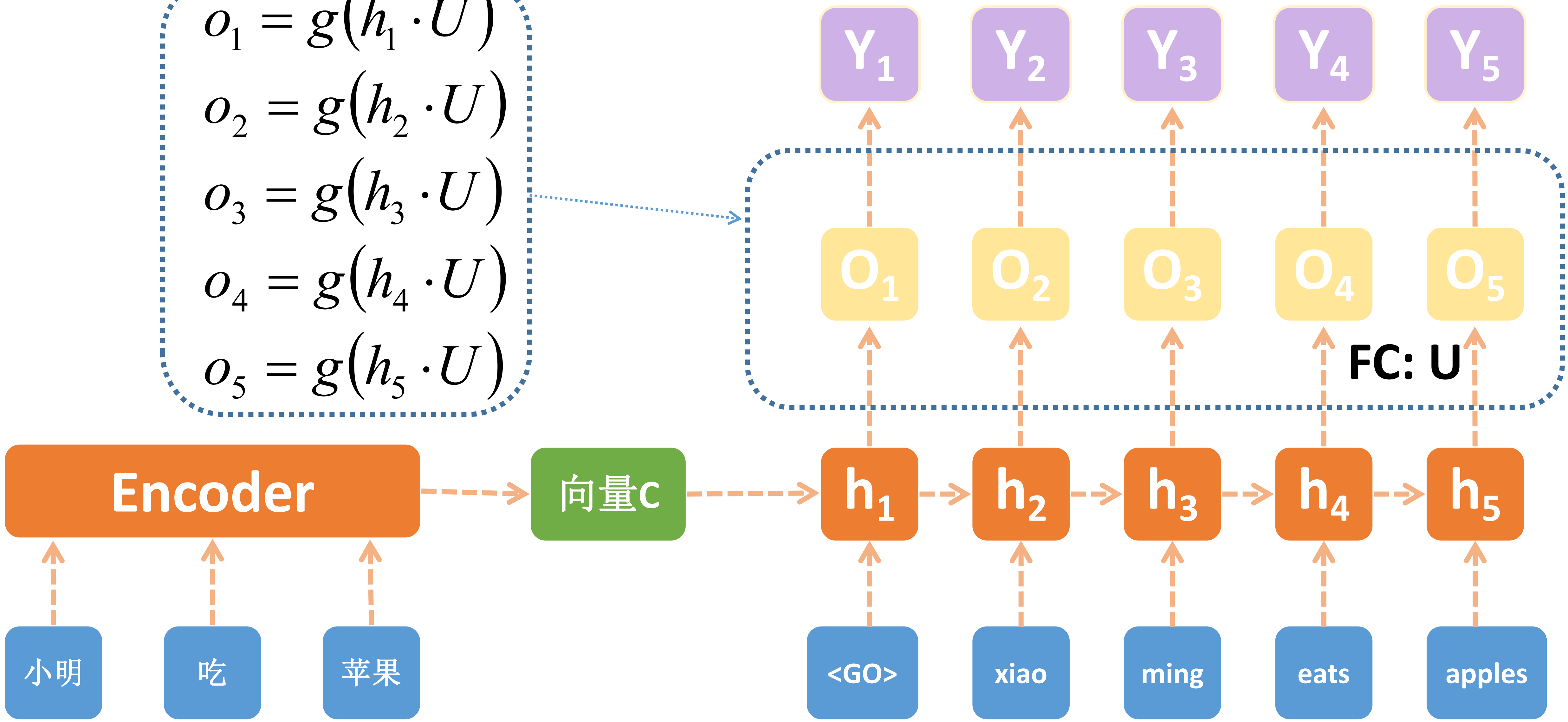
Seq2Seq理解

$$h_1 = f([y_1, c] \cdot W_2)$$
$$h_2 = f([y_2, h_1] \cdot W_2)$$
$$h_3 = f([y_3, h_2] \cdot W_2)$$
$$h_4 = f([y_4, h_3] \cdot W_2)$$
$$h_5 = f([y_5, h_4] \cdot W_2)$$

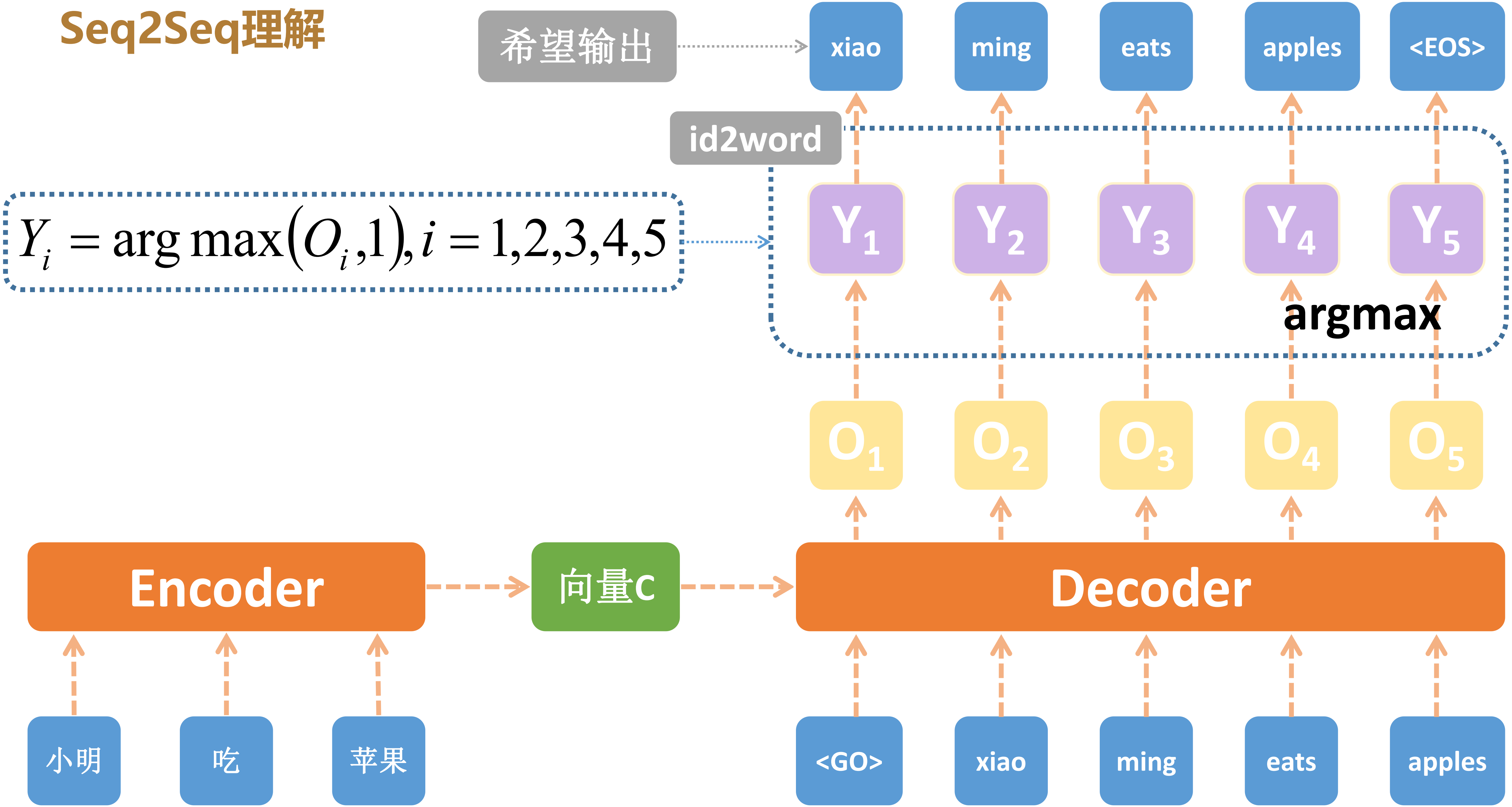


Seq2Seq理解

$$o_1 = g(h_1 \cdot U)$$
$$o_2 = g(h_2 \cdot U)$$
$$o_3 = g(h_3 \cdot U)$$
$$o_4 = g(h_4 \cdot U)$$
$$o_5 = g(h_5 \cdot U)$$



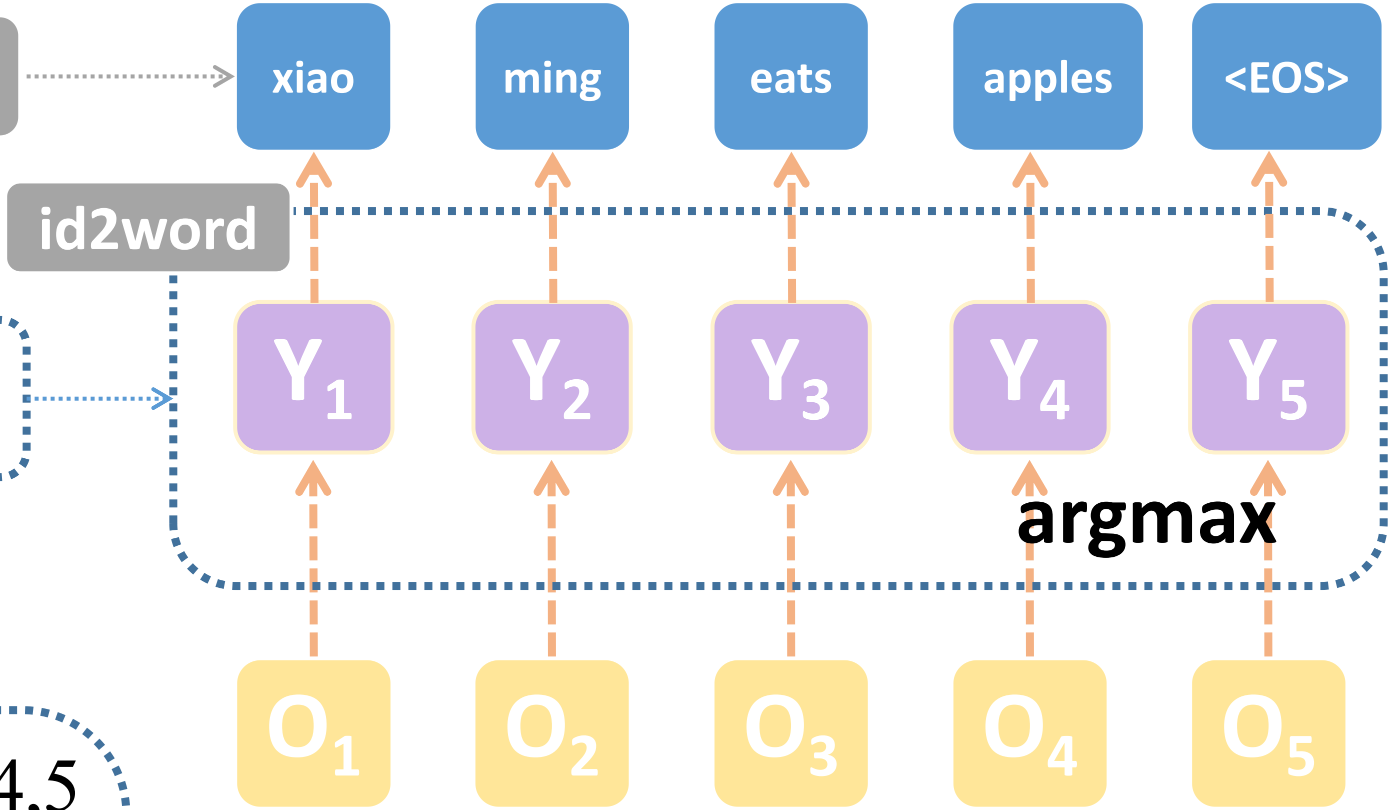
Seq2Seq理解



希望输出
xiao
ming
eats
apples
<EOS>
id2word
$Y_i = \arg\max(O_i, 1), i = 1,2,3,4,5$
Y_1 Y_2 Y_3 Y_4 Y_5
argmax
O_1 O_2 O_3 O_4 O_5
Encoder
向量C
Decoder
小明
吃
苹果
<GO>
xiao
ming
eats
apples

Seq2Seq理解

希望输出



$$Y_i = \arg\max\left(O_i, 1\right), i = 1,2,3,4,5$$

$$k_i = word2id\left(word_i\right), i = 1,2,3,4,5$$

$$p_i = O_i$$

$$E = \sum_{i=1}^{5} E_i$$

$$E_i = -\ln p_{i,k_i}$$

LOSS

Seq2Seq理解预测

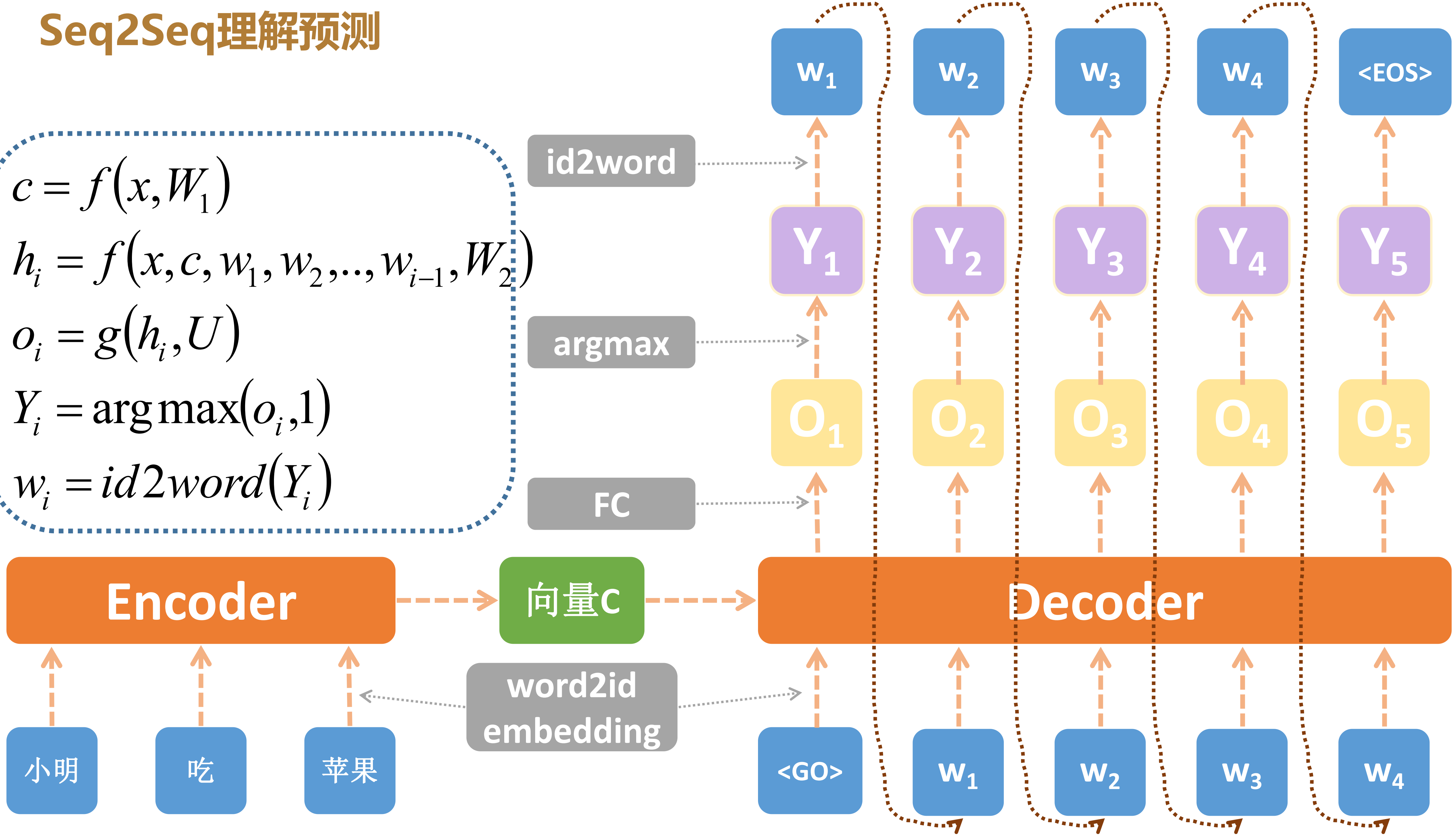
$$c = f(x, W_1)$$

$$h_i = f(x, c, w_1, w_2, \ldots, w_{i-1}, W_2)$$

$$o_i = g(h_i, U)$$

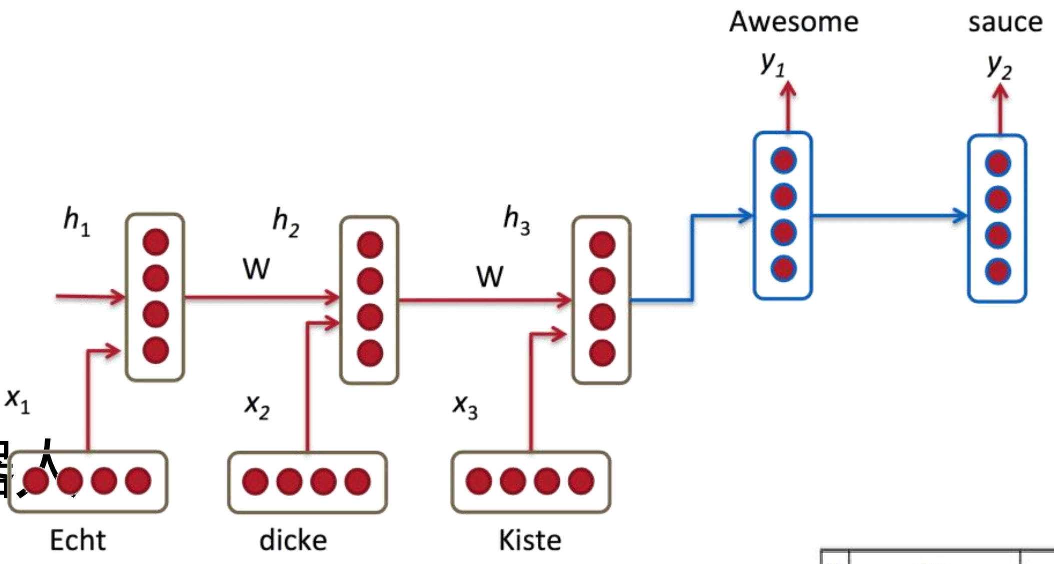
$$Y_i = \arg\max(o_i, 1)$$

$$w_i = id2word(Y_i)$$



Seq2Seq应用

- 机器翻译



- 自动对话机器人

Post	Model		Response(Original)	Response(Translated)
比情人节更庸俗的事就是，对情人节的接待规格待遇提出要求。What is more vulgar than Valentine's day is to make specific demands on the reception of the Valentine's day.	Seq2Seq		这才是真正的情人节。	This is a real Valentine's day.
	ECM	*Like*	情人节快乐！	Happy Valentine's Day!
		Happiness	哈哈，这也太浪漫了吧！	Aha, this is too romantic!
		Sadness	我也想要这样的情人节。	I also want this kind of Valentine's day, sigh.
		Disgust	这就是所谓的情人节！	This is the so-called Valentine's day!
		Anger	这是赤裸裸的炫耀啊！	This is shameless show-off!
知道真相的我眼泪笑下来。Knowing the truth, I laughed myself to tears.	Seq2Seq		知道真相的我眼泪掉下来。	After knowing the truth, I burst into tears.
	ECM	*Like*	太可爱了，太可爱了！	Too lovely, it's too lovely.
		Happiness	笑得我眼泪都出来了。	I can't help laugh myself to tears.
		Sadness	知道真相的我眼泪掉下来。	After knowing the truth, I burst into tears.
		Disgust	这是真的还是假的？	Is this true?
		Anger	什么意思？没看懂。	What do you mean? I don't understand.

ID	Post	NRM-glo	NRM-loc	NRM-hyb	Rtr.-based
1	每年元旦都来高烧。 High fever attacks me every New Year's Day	多休息，增强免疫力！ Rest more to boost your immune system.	每年元旦都去医院看病 Go to see a doctor every New Year's Day	祝早日康复，身体健康！ Get well soon and stay healthy!	高烧… High fever…
2	又老了一岁，感谢我们组，我很高兴 I gain one more year. Grateful to my group, so happy	生日快乐！永远支持你！ Happy birthday! Will stand by you forever!	生日快乐！永远年轻漂亮！ Happy birthday! Wish you always stay young and pretty!	老了老了，岁月不饶人啊！ Getting old now. Time has no mercy.	我只是老了一岁 I just gain only one more year.
3	吃素第一天，坚持住，崔朵拉。 First day of being a vegetarian. Hold on, **CuiDuoLa**	必须来支持！加油！ I come to offer my support. Keep it up!	坚持住，坚持下去。 Hold on, keep it up.	每天都是素食主义者 I have been being a vegetarian everyday	坚持住！ Hold on!
4	这种天气真应该带个帅锅出去踏青赏花 We should go out with some cute guys to enjoy a great outing in such a nice weather.	这是哪儿啊？真美！ Where is it? It is so beautiful!	阳光明媚，心情舒畅。 Such a nice sunny day! I am in a great mood.	这种天气真不是盖的。 It is indeed a very nice weather.	文山啊 出去踏青寻找灵感 哈哈 **WenShan**, let's go out to get some inspiration. Ha! Ha!

Seq2Seq应用

- 文档摘要自动生成

- 文本生成

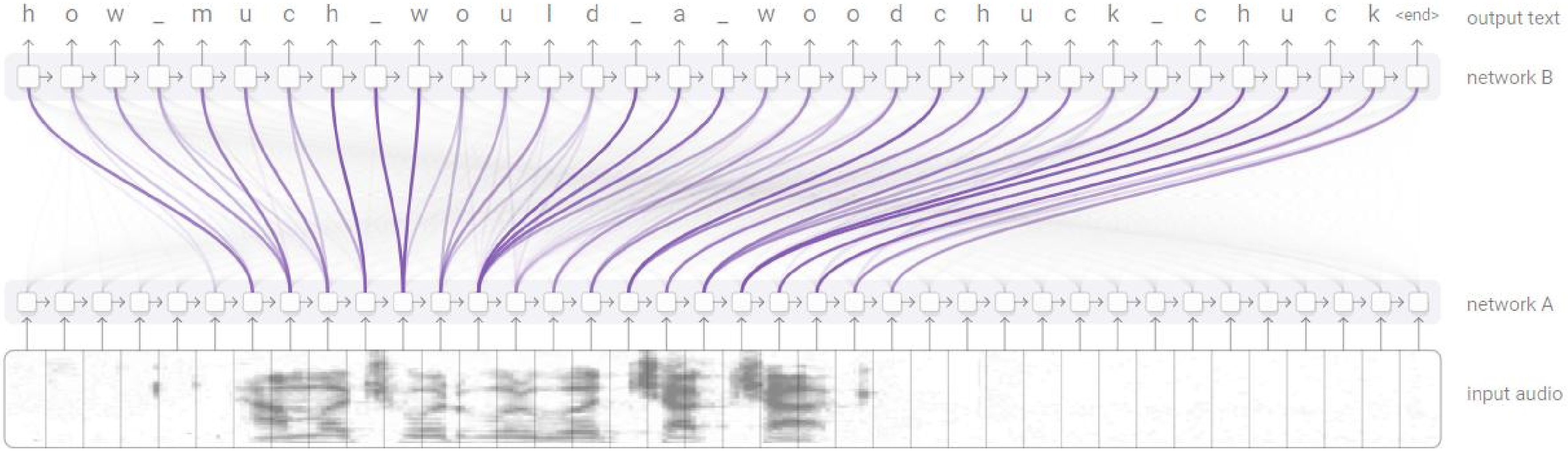
Good quality summary output
S: a man charged with the murder last year of a british backpacker confessed to the slaying on the night he was charged with her killing , according to police evidence presented at a court hearing tuesday . ian douglas previte , ## , is charged with murdering caroline stuttle , ## , of yorkshire , england
T: man charged with british backpacker 's death confessed to crime police officer claims
O: man charged with murdering british backpacker confessed to murder

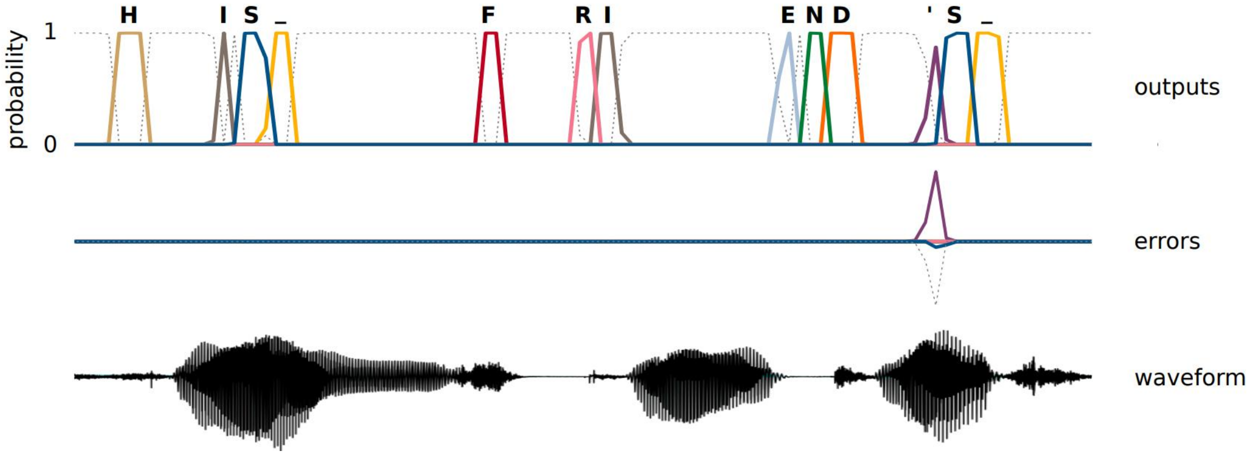
秋夕湖上	秋夕湖上
By a Lake at Autumn Sunset	By a Lake at Autumn Sunset
一夜秋凉雨湿衣，	荻花风里桂花浮，
A cold autumn rain wetted my clothes last night,	The wind blows reeds with osmanthus flying,
西窗独坐对夕晖。	恨竹生云翠欲流。
And I sit alone by the window and enjoy the sunset.	And the bamboos under clouds are so green as if to flow down.
湖波荡漾千山色，	谁拂半湖新镜面，
With mountain scenery mirrored on the rippling lake,	The misty rain ripples the smooth surface of lake,
山鸟徘徊万籁微。	飞来烟雨暮天愁。
A silence prevails over all except the hovering birds.	And I feel blue at sunset .

▲ 图5：左边是机器生成的诗词，右边是一首宋代诗词

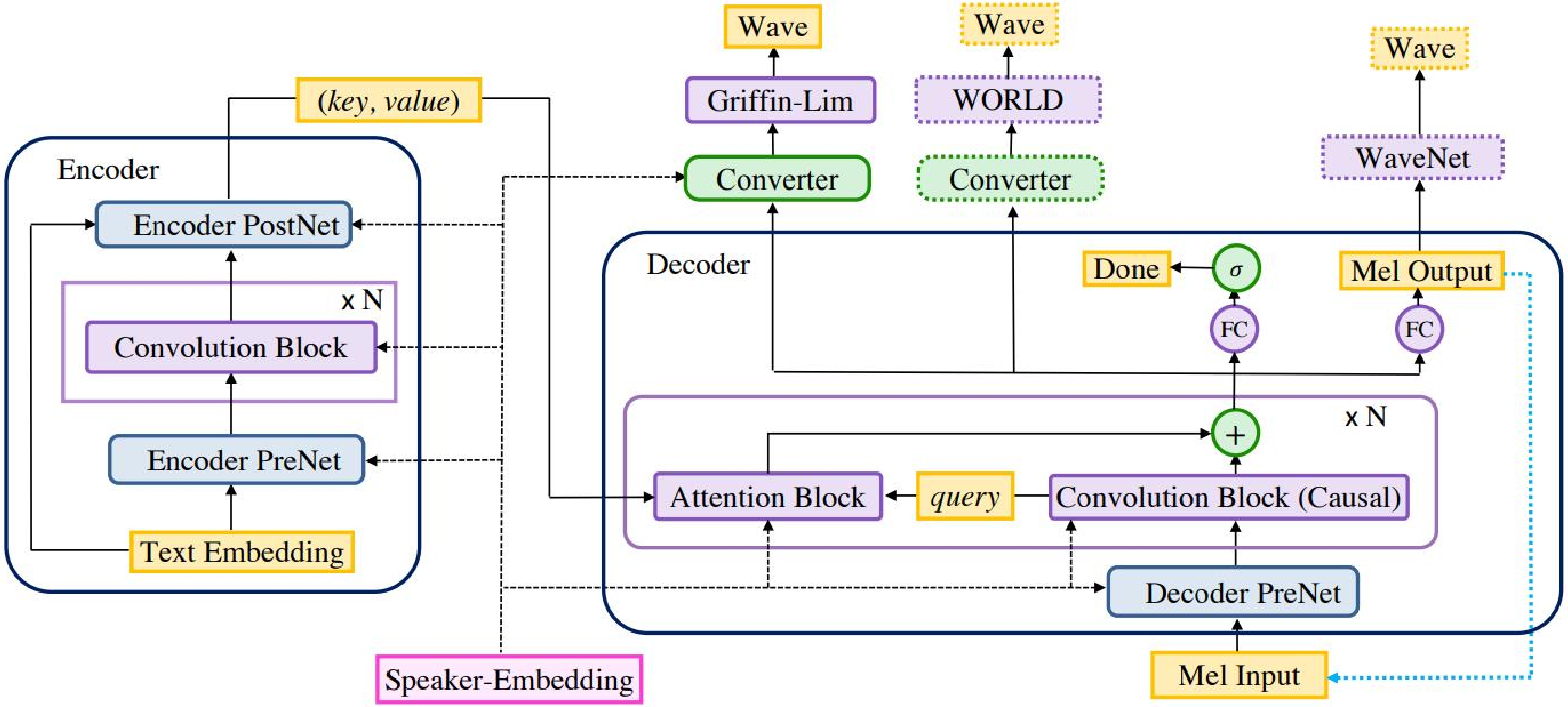
Seq2Seq应用

- 语音识别/合成/语音-文本转换





Seq2Seq应用



百度Deep Voice v3

https://github.com/Kyubyong/deepvoice3

Seq2Seq应用

- 图片描述自动生成

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

Seq2Seq应用

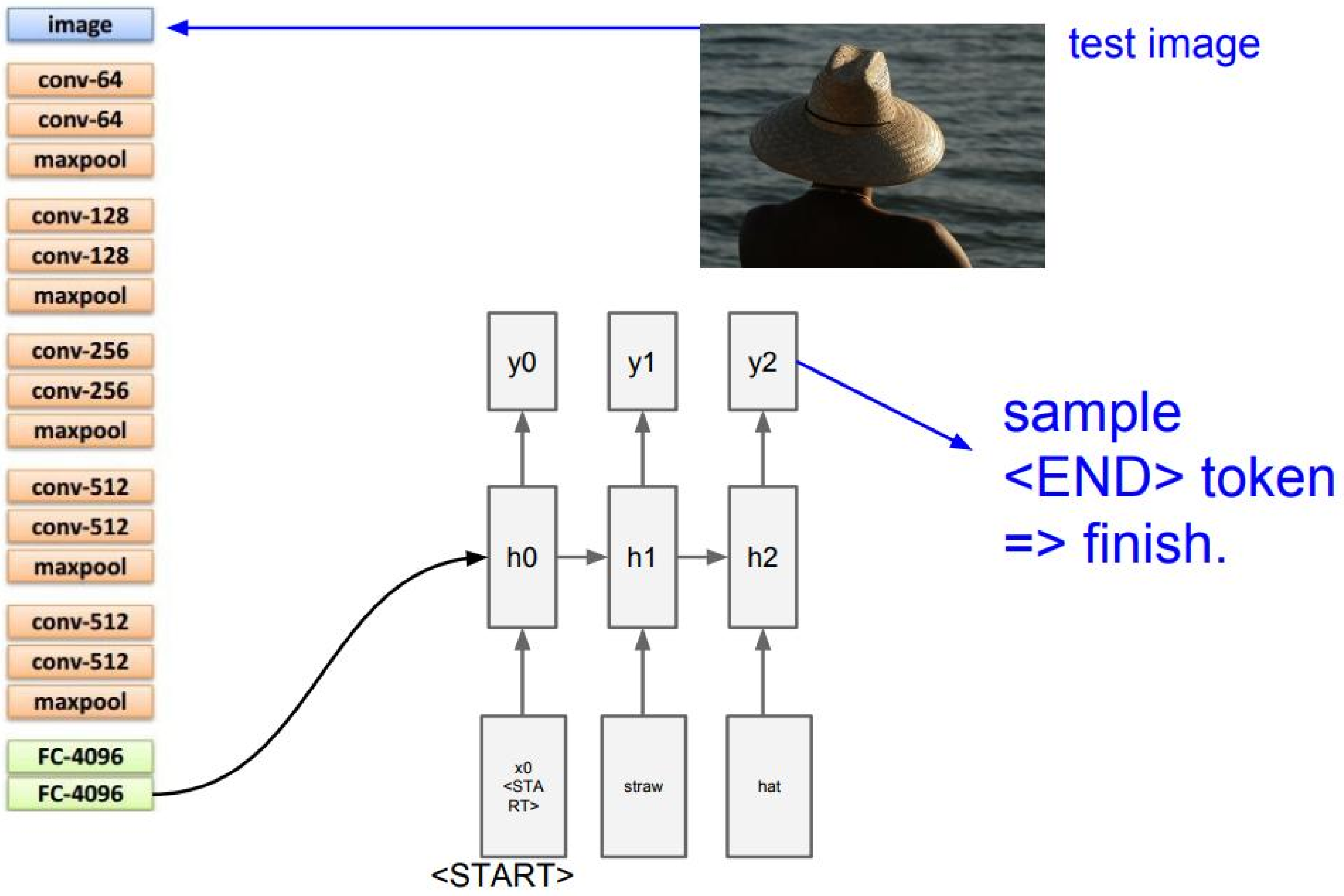
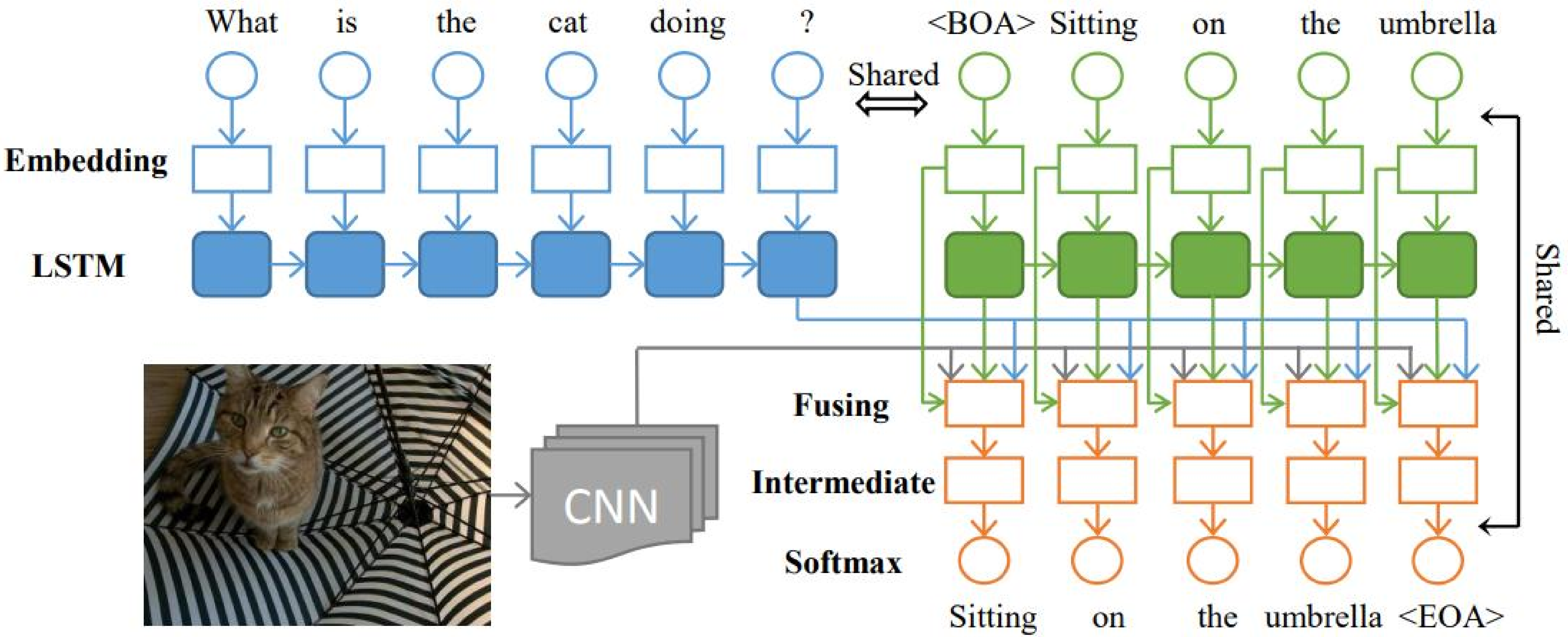


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
y2
h0
h1
h2
x0 <STA RT>
straw
hat
<START>
sample <END> token => finish.

Seq2Seq应用

- Visual Question Answering(VQA, 视觉问答系统)
  - https://visualqa.org/index.html

Seq2Seq应用

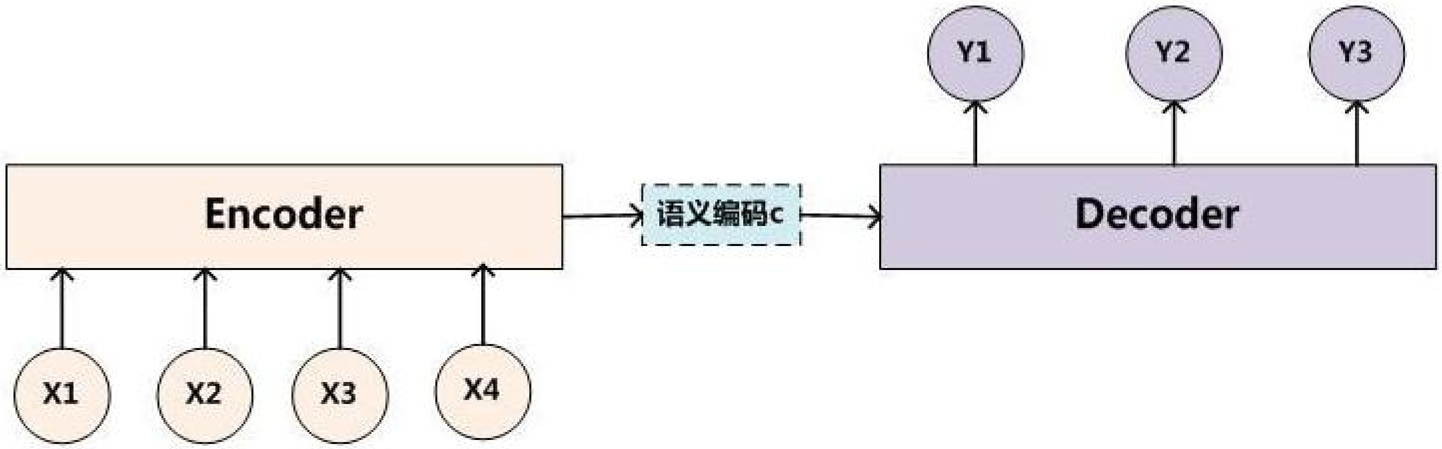

What
is
the
cat
doing
?
Shared
<BOA>
Sitting
on
the
umbrella
Embedding
LSTM
Shared
CNN
Fusing
Intermediate
Softmax
Sitting
on
the
umbrella
<EOA>

Seq2Seq应用总结

- 总而言之，Seq2Seq应用场景，包括了经典的机器翻译、文本摘要和对话生成等，也包括了一些非常有趣的应用，比如：根据公式图片生成 latex 代码，生成 commit message 等。自然语言生成（NLG）是一个非常有意思，也非常有前途的研究领域，简单地说，就是解决一个条件概率 p(output| context)的建模问题，即根据 context 来生成 output，这里的 context 可以非常零活多样，大家都是利用深度学习模型对这个条件概率进行建模，同时加上大量的训练数据和丰富的想象力，可以实现很多有趣的工作。Seq2Seq是一个简单易用的框架，开源的实现也非常多，但并不意味着直接生搬硬套就可以了，需要具体问题具体分析。此外，对于生成内容的控制，即 decoding 部分的研究也是一个非常有意思的方向，比如：如何控制生成文本的长度，控制生成文本的多样性，控制生成文本的信息量大小，控制生成文本的情感等等。

Seq2Seq原理

- 最基础的Seq2Seq模型包含了三个部分，即Encoder、Decoder以及连接两者的中间状态向量，Encoder通过学习输入，将其编码成一个固定大小的状态向量c，继而将c传给Decoder，Decoder再通过对状态向量c的学习来进行输出。下图中，图中每一个box代表了一个**RNN Cell**单元，通常是**LSTM**或者**GRU**。



Seq2Seq

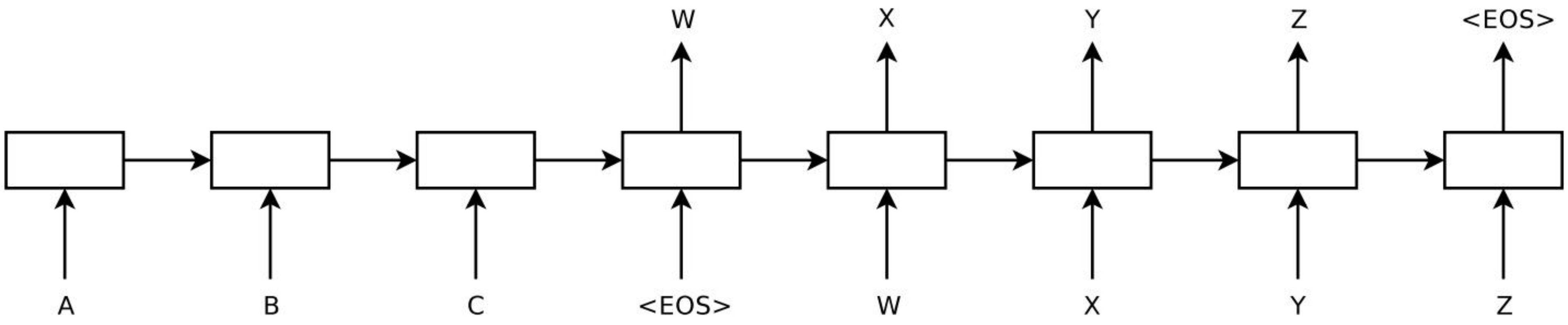


Figure 1: Our model reads an input sentence “ABC” and produces “WXYZ” as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

Seq2Seq

- Encoder-Decoder框架可以这么直观地去理解：可以把它看作适合处理由一个句子（或篇章）生成另外一个句子（或篇章）的通用处理模型。对于句子对<X,Y>，我们的目标是给定输入句子X，期待通过Encoder-Decoder框架来生成目标句子Y。X和Y可以是同一种语言，也可以是两种不同的语言。而X和Y分别由各自的单词序列构成：

$$X = (x_1, x_2, \ldots, x_m)$$

$$Y = (y_1, y_2, \ldots, y_n)$$

Seq2Seq

- Encoder顾名思义就是对输入句子X进行编码，将输入句子通过非线性变换转化为中间语义表示C：

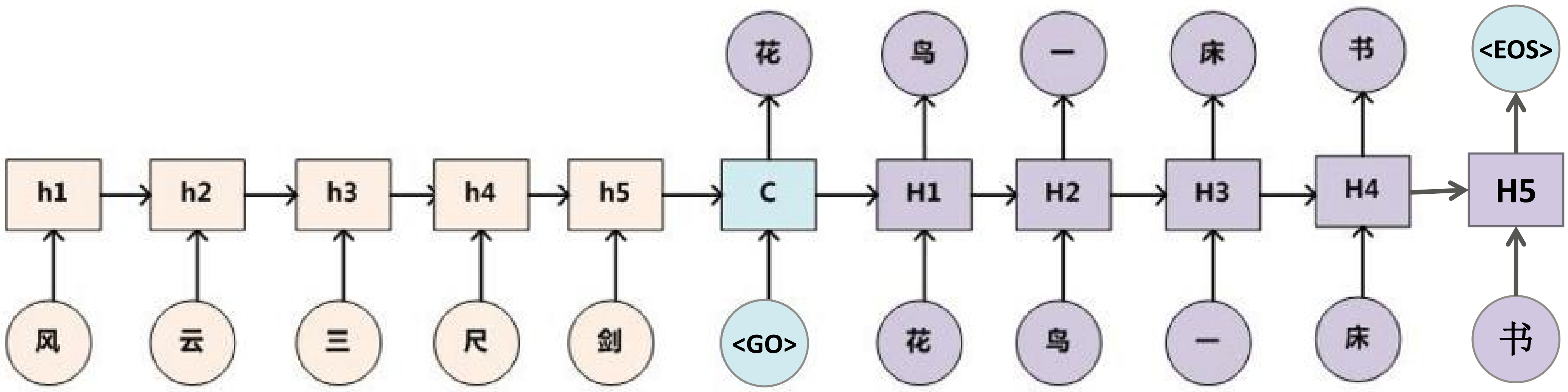
$$C = F(x_1, x_2, \cdots, x_m)$$

- 对于解码器Decoder来说，其任务是根据句子X的中间语义表示C和之前已经生成的历史信息y1,y2....yi-1来生成i时刻要生成的单词yi:每个yi都依次这么产生，那么看起来就是整个系统根据输入句子X生成了目标句子Y。

$$y_i = G(C, y_1, y_2, \cdots, y_{i-1})$$

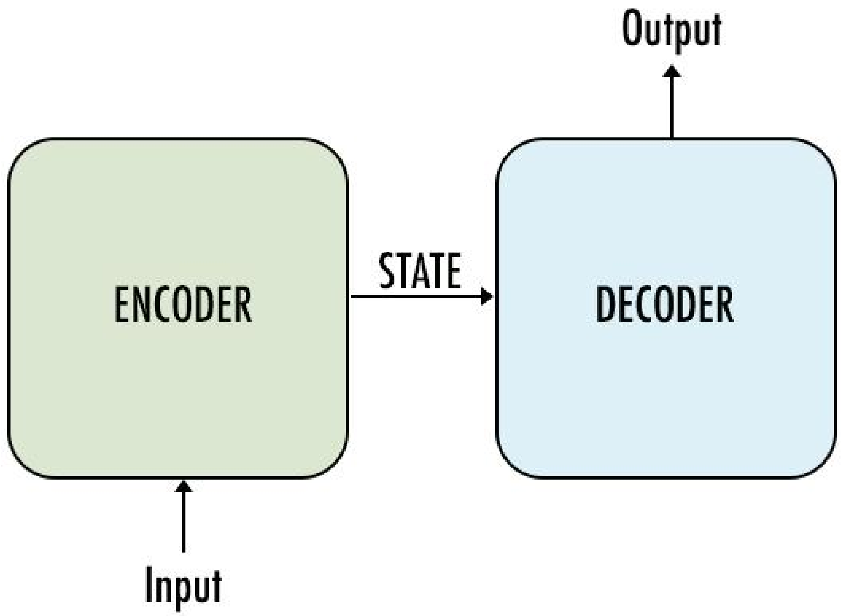
Seq2Seq案例

- 只需要找到大量的对联数据对这个模型进行训练，那么即可利用这个模型，输入上联，机器自动产生下联了。



Seq2Seq案例

Inputs	Target
How are you?	I am good
Can you fly that thing?	Not yet



编码器：tf.nn.dynamic_rnn
解码器：tf.contrib.seq2seq.dynamic_rnn_decoder
(NOTE：但是我们一般不用这个API，这个比较麻烦)
举例：tf.nn.dynamic_rnn(cell, inputs, sequence_length=None, initial_state=None, dtype=None, parallel_iterations=None, swap_memory=False, time_major=False, scope=None)

Cell为前面构建的RNN

cell(tf.contrib.rnn.BasicLSTMCell);

Inputs,为输入的文本数据，通常是嵌入层的输出。以及initial_state

Seq2Seq

<PAD>	在训练中，我们将数据按批次输入。但同一批次中必须有相同的Sequence Length(序列长度/time_steps)。所以我们会用<PAD>填充较短的输入。
<EOS>	它能告诉解码器句子在哪里结束，并且它允许解码器在其输出中表明句子结束的位置
<UNK>	忽视词汇表中出现频率不够高而不足以考虑在内的文字,将这些单词替换为 <UNK>
<GO>	解码器的第一个时间步骤的输入，以使解码器知道何时开始产生输出

Seq2Seq案例

0	<PAD>	11	can
1	<EOS>	12	you
2	<UNK>	13	fly
3	<GO>	14	that
4	how	15	thing
5	are	16	not
6	you	17	yet
7	?		
8	i		
9	am		
10	good		

Seq2Seq案例

Inputs	Target
How are you?	I am good
Can you fly that thing?	Not yet

填充

how	are	you	?	<PAD>	<PAD>
can	you	fly	that	thing	?

Seq2Seq案例

Inputs	Target
How are you?	I am good
Can you fly that thing?	Not yet

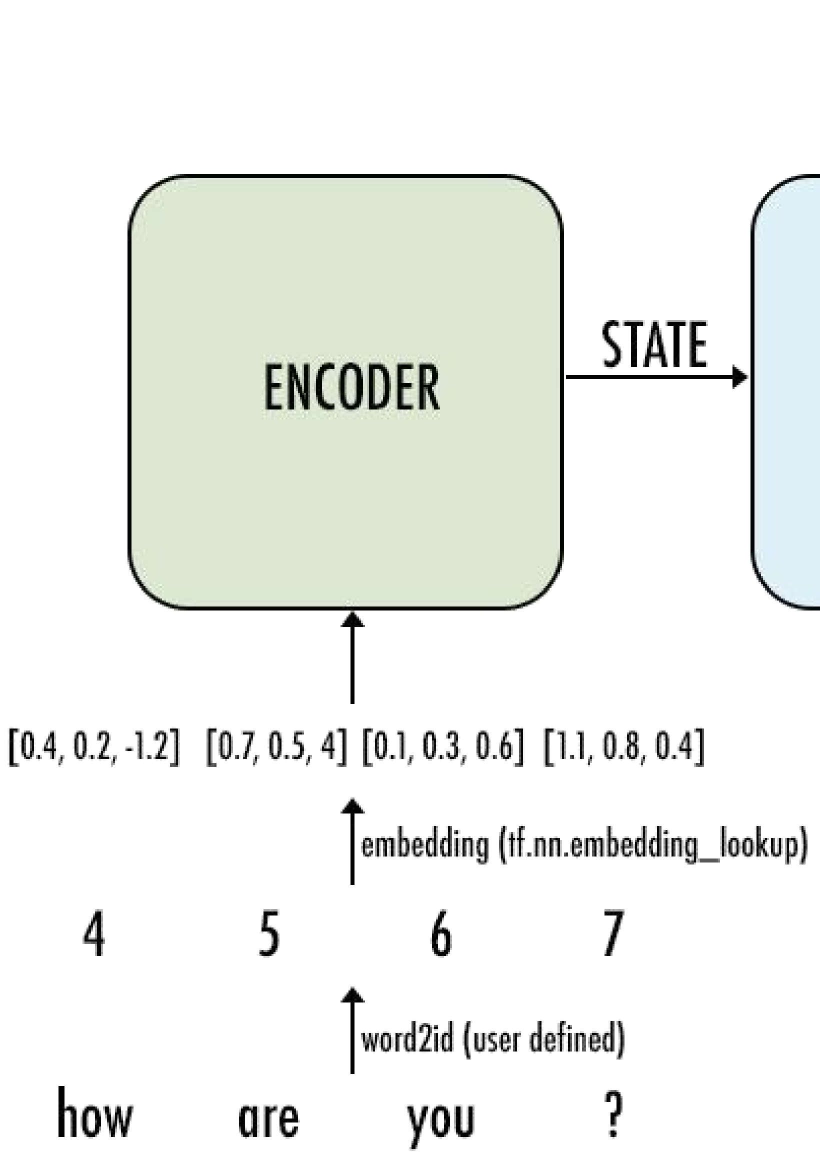
填充

how	are	you	?	<PAD>	<PAD>
can	you	fly	that	thing	?

Word2id

4	5	6	7	0	0
11	12	13	14	15	7

Seq2Seq案例



Output
ENCODER
STATE
DECODER
[0.4, 0.2, -1.2] [0.7, 0.5, 4] [0.1, 0.3, 0.6] [1.1, 0.8, 0.4]
embedding (tf.nn.embedding_lookup)
4 5 6 7
word2id (user defined)
how are you ?

Seq2Seq案例

Inputs

How are you?

Can you fly that thing?

Target

I am good

Not yet

填充

<GO>	i	am	good	<EOS>
<GO>	not	yet	<EOS>	<PAD>

Seq2Seq案例

Inputs	**Target**
How are you?	I am good
Can you fly that thing?	Not yet

填充

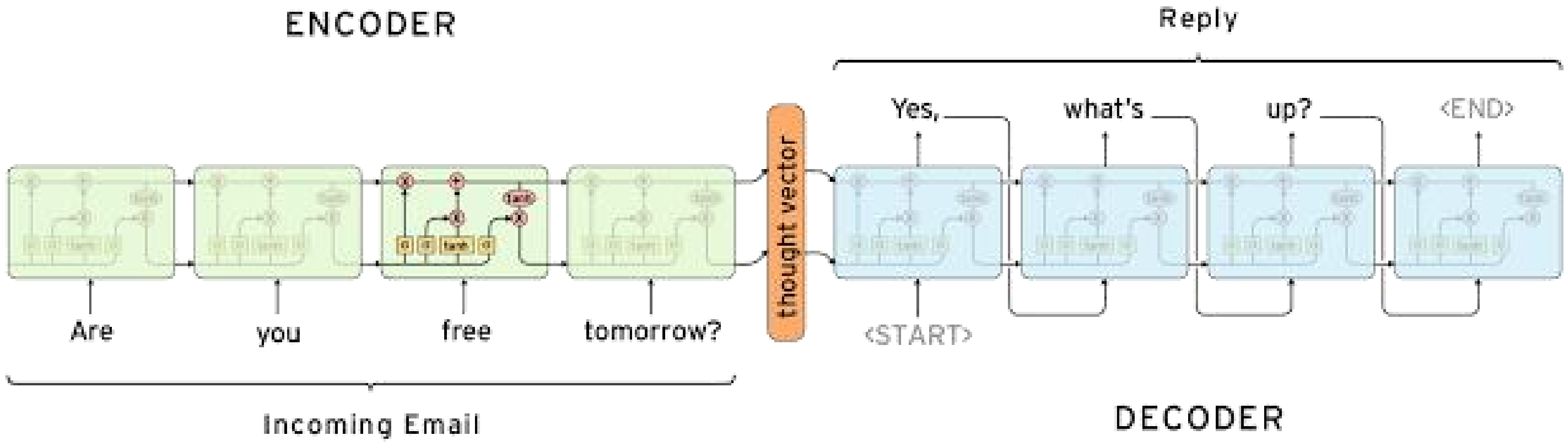
<GO>	i	am	good	<EOS>
<GO>	not	yet	<EOS>	<PAD>

Word2id

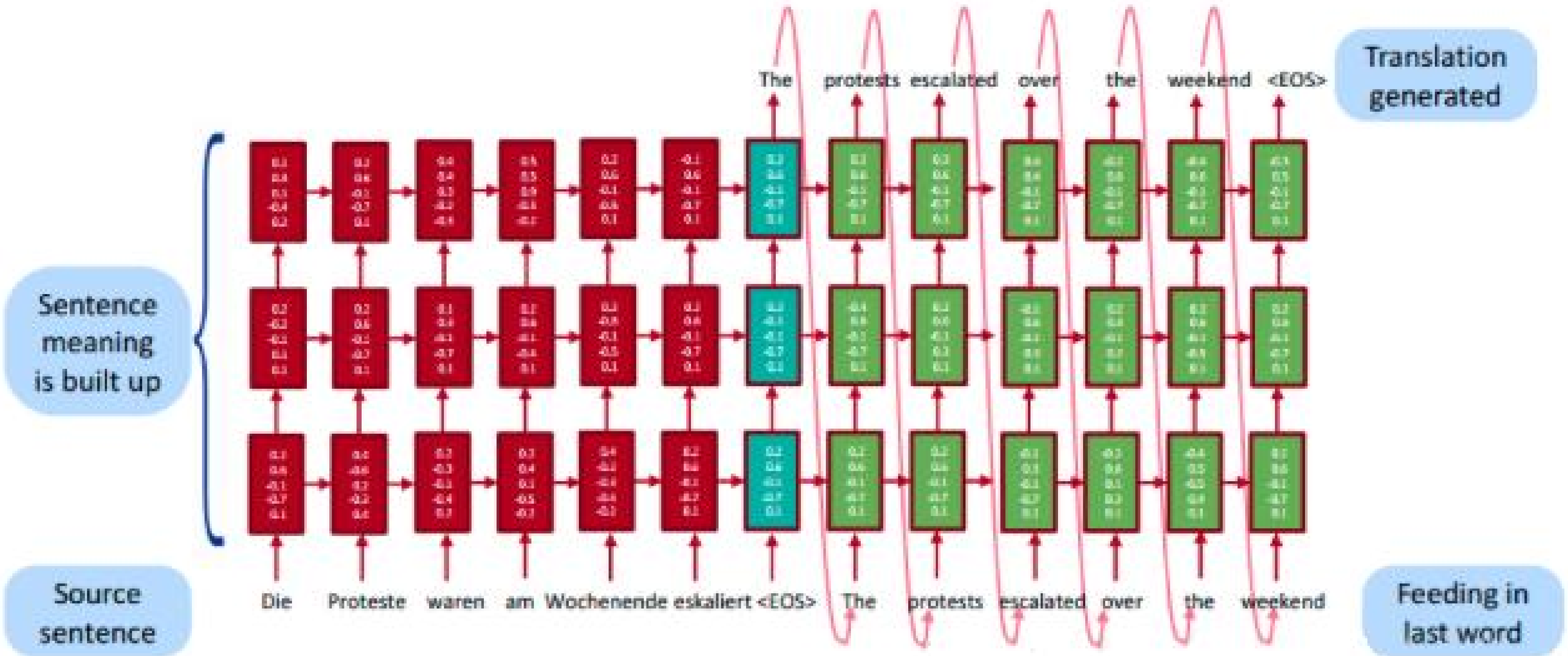
3	8	9	10	1
3	16	17	1	0

Seq2Seq

- 将RNN模块换成LSTM，则效果如下图。Encoder 和 Decoder 都是 4 个时间步长的 LSTM(但是只有两个RNN Cell)。小技巧：将源句子顺序颠倒后再输入 Encoder 中，比如源句子为“A B C”，那么输入 Encoder 的顺序为 “C B A”，经过这样的处理后，取得了很大的提升，而且这样的处理使得模型能够很好地处理长句子。



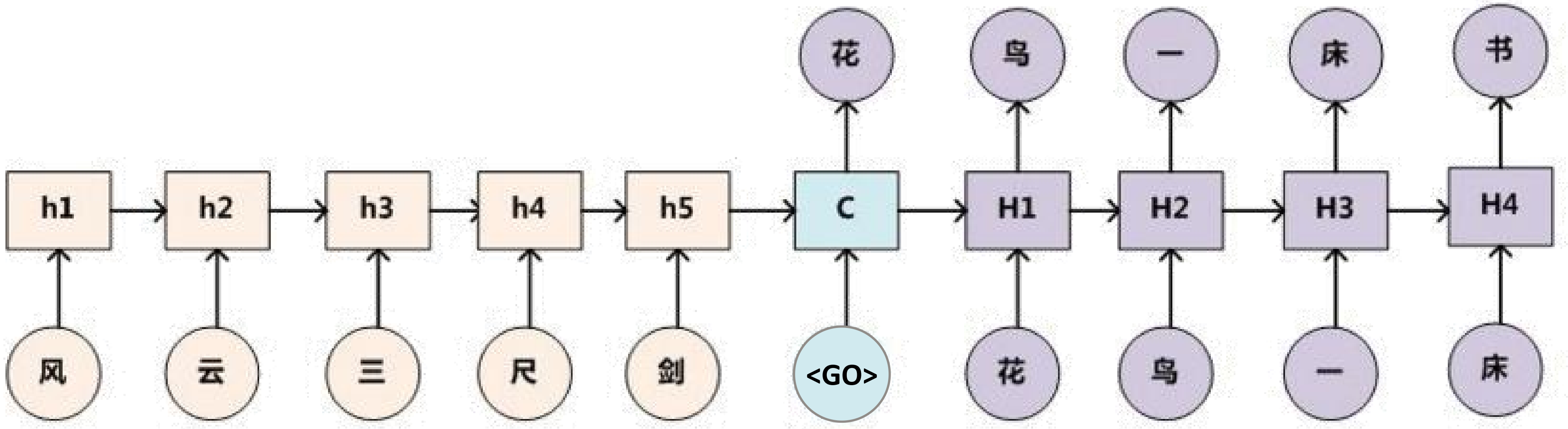
Seq2Seq



Translation generated
The
protests
escalated
over
the
weekend
<EOS>
Sentence meaning is built up
Source sentence
Die
Proteste
waren
am
Wochenende
eskaliert
<EOS>
The
protests
escalated
over
the
weekend
Feeding in last word

Seq2Seq问题

- 问题描述：" 风"对应的特征对于下联的影响是最弱的。



1. 字句对等；
2. 词性对品；
3. 结构对应；
4. 节律对拍；
5. 平仄对立；
6. 形对意联；

THANKS!