Paragraph Embeddings & Attention

Kuan-Yu Chen (陳冠宇)

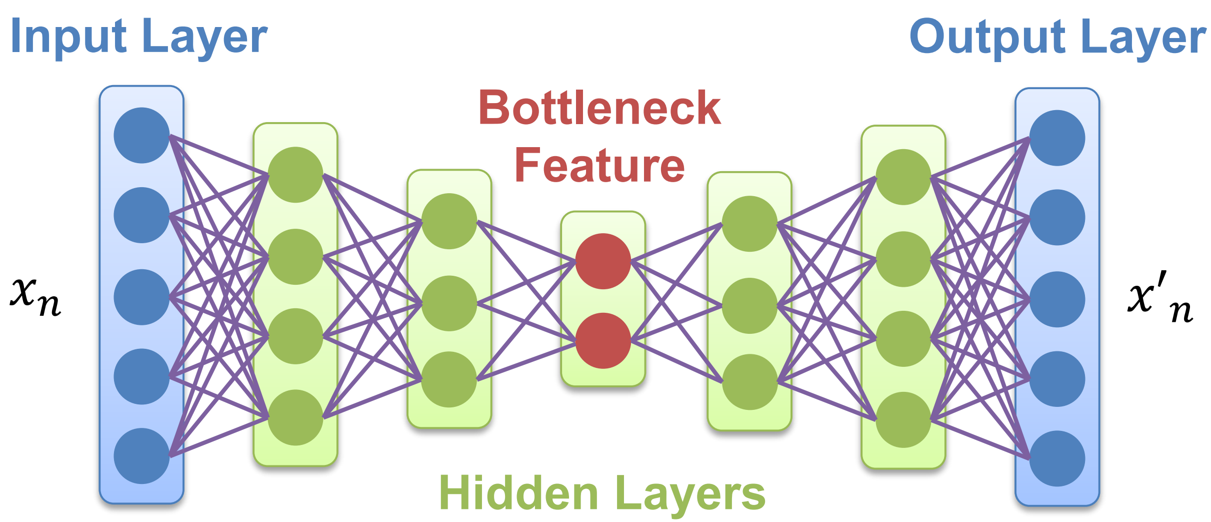
2018/05/03 @ NTUST

Autoencoder.

- An autoencoder is a DNN-based **unsupervised learning** of efficient codings
  - The training objective is to minimize the reconstructed errors

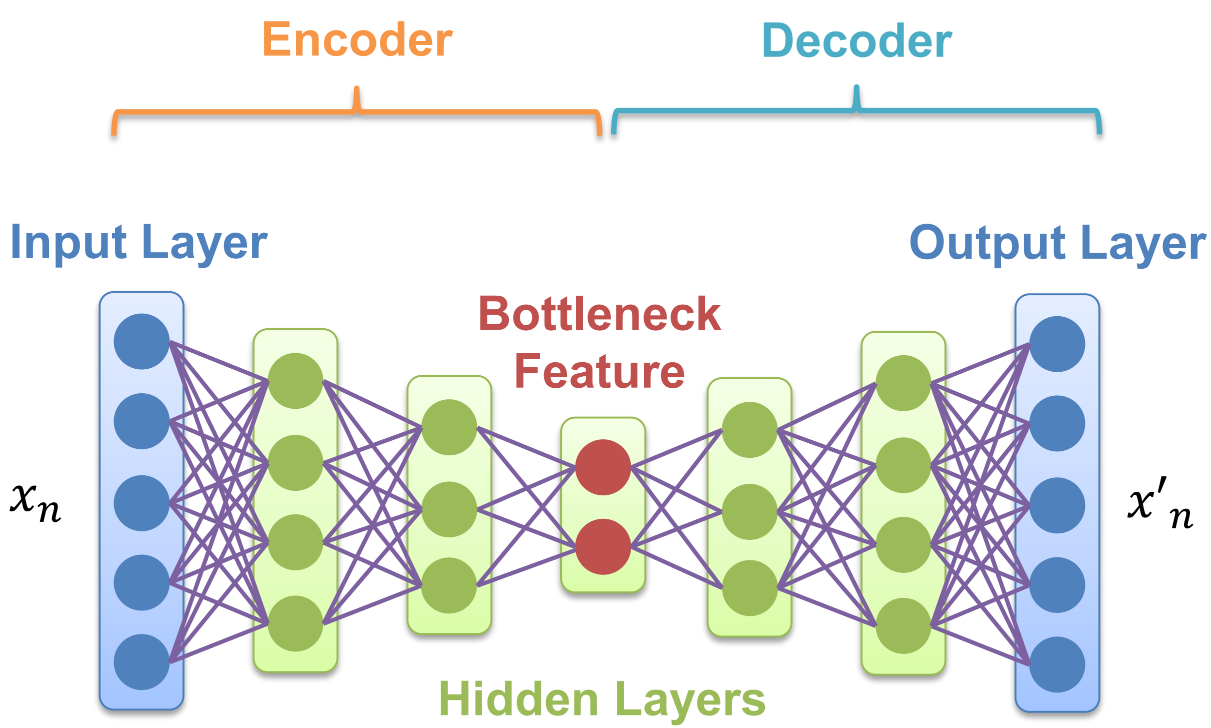
$$\min \frac{1}{N}\sum_{n=1}^{N}(x_n - x'_n)^2$$

$$\min -\sum_{n=1}^{N} x_n log(x'_n)$$



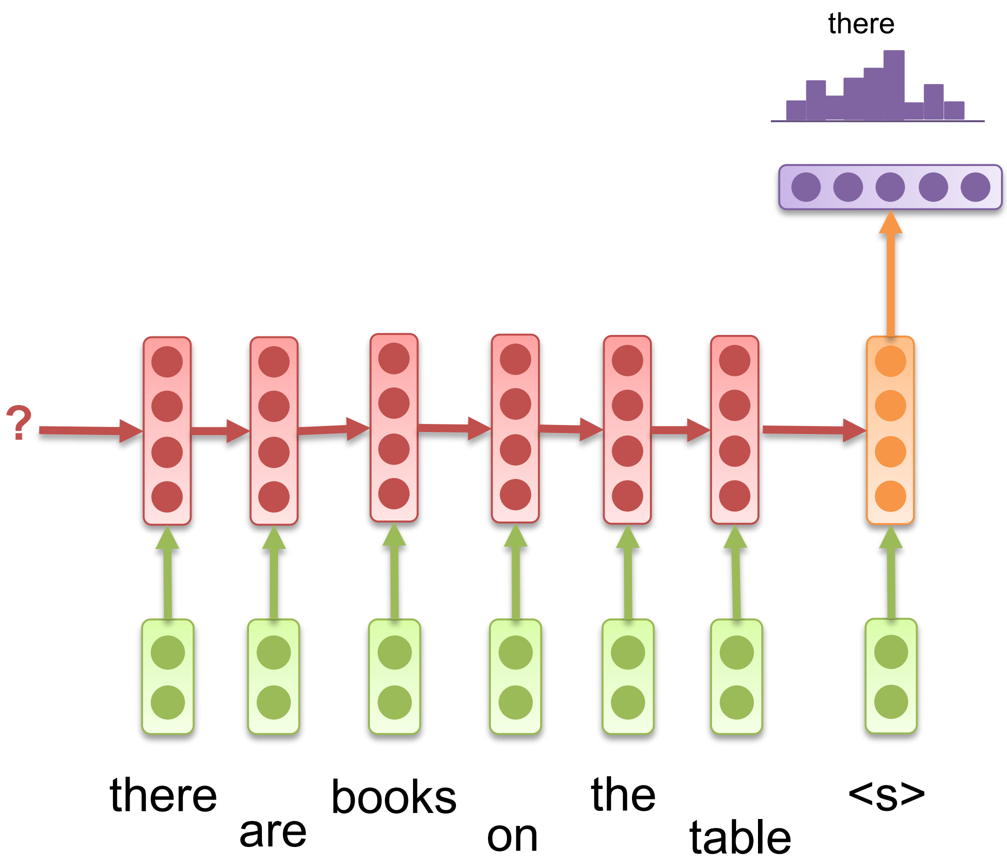
Autoencoder..

- An autoencoder is a DNN-based **unsupervised learning** of efficient codings
  - The training objective is to minimize the reconstructed errors



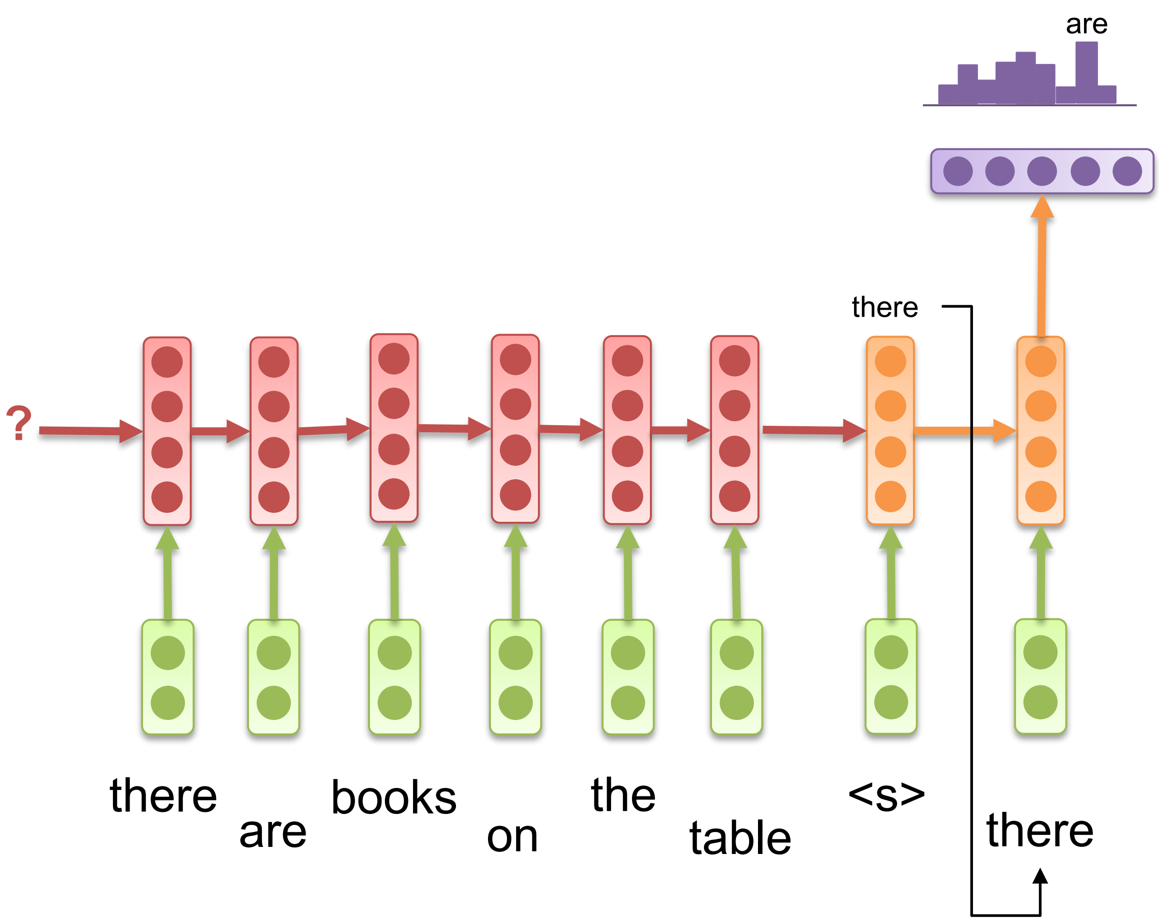
RNN-based Autoencoder.

- RNN can also be used to construct an autoencoder



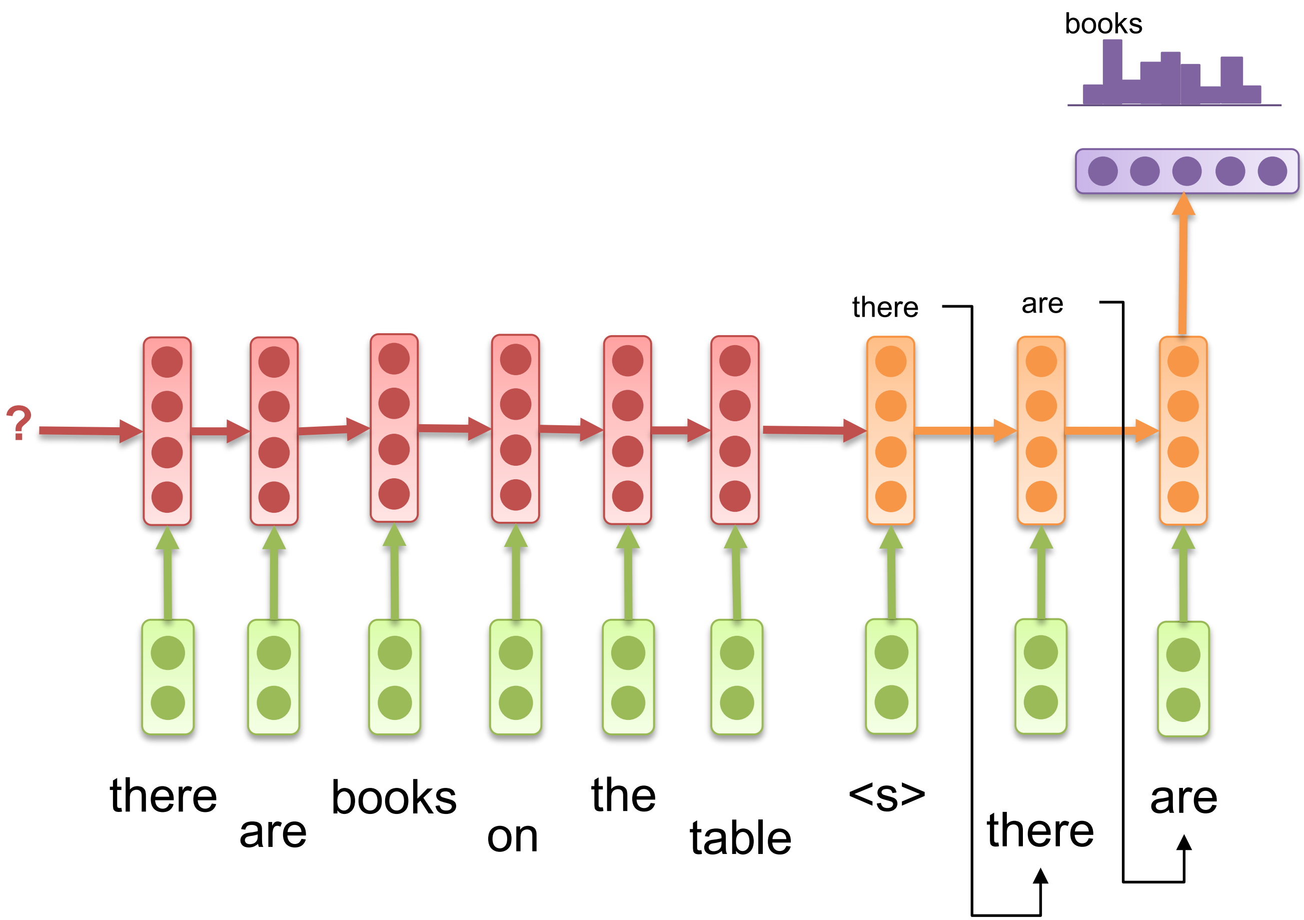
RNN-based Autoencoder..

- RNN can also be used to construct an autoencoder



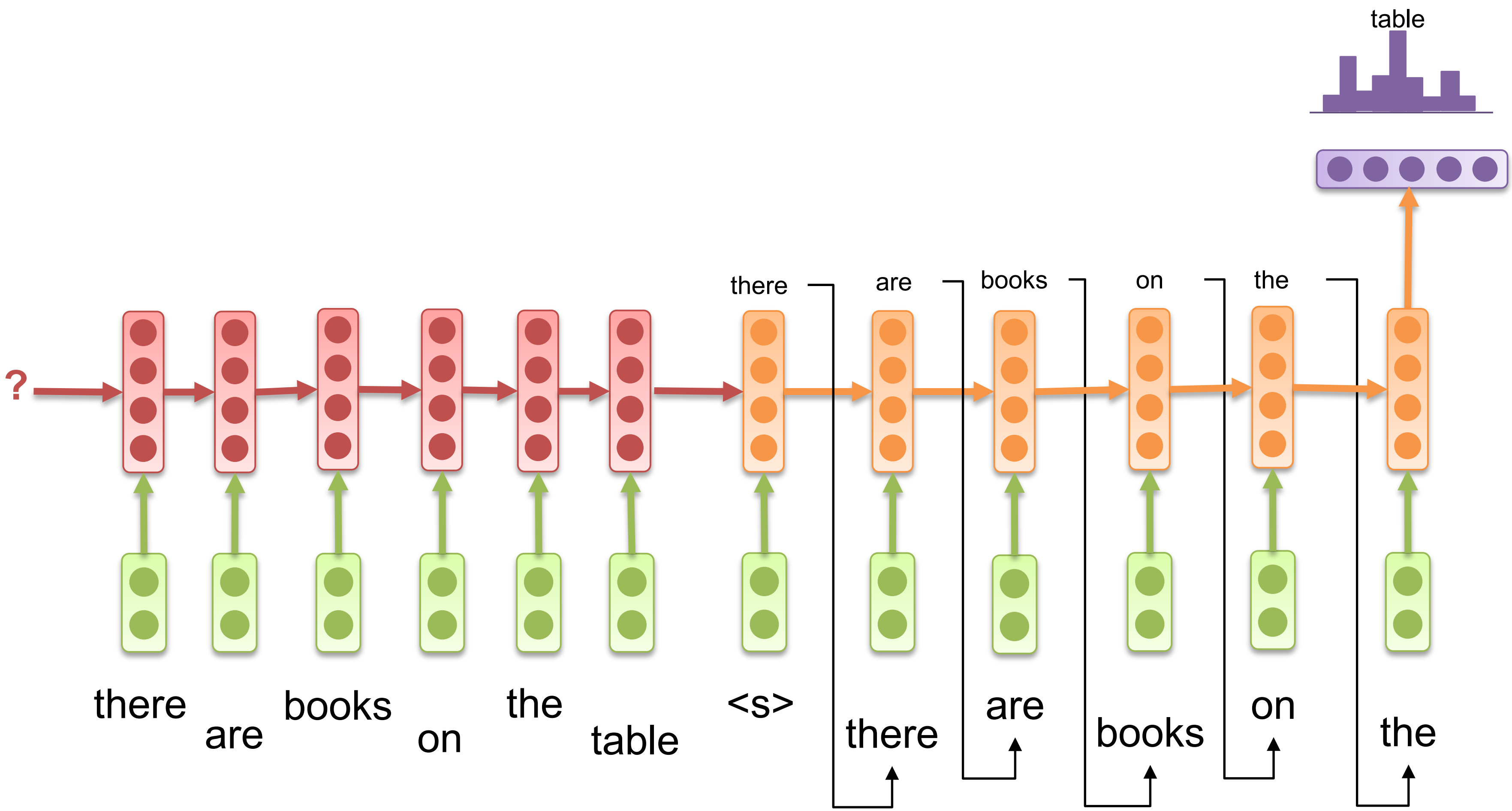
RNN-based Autoencoder...

- RNN can also be used to construct an autoencoder



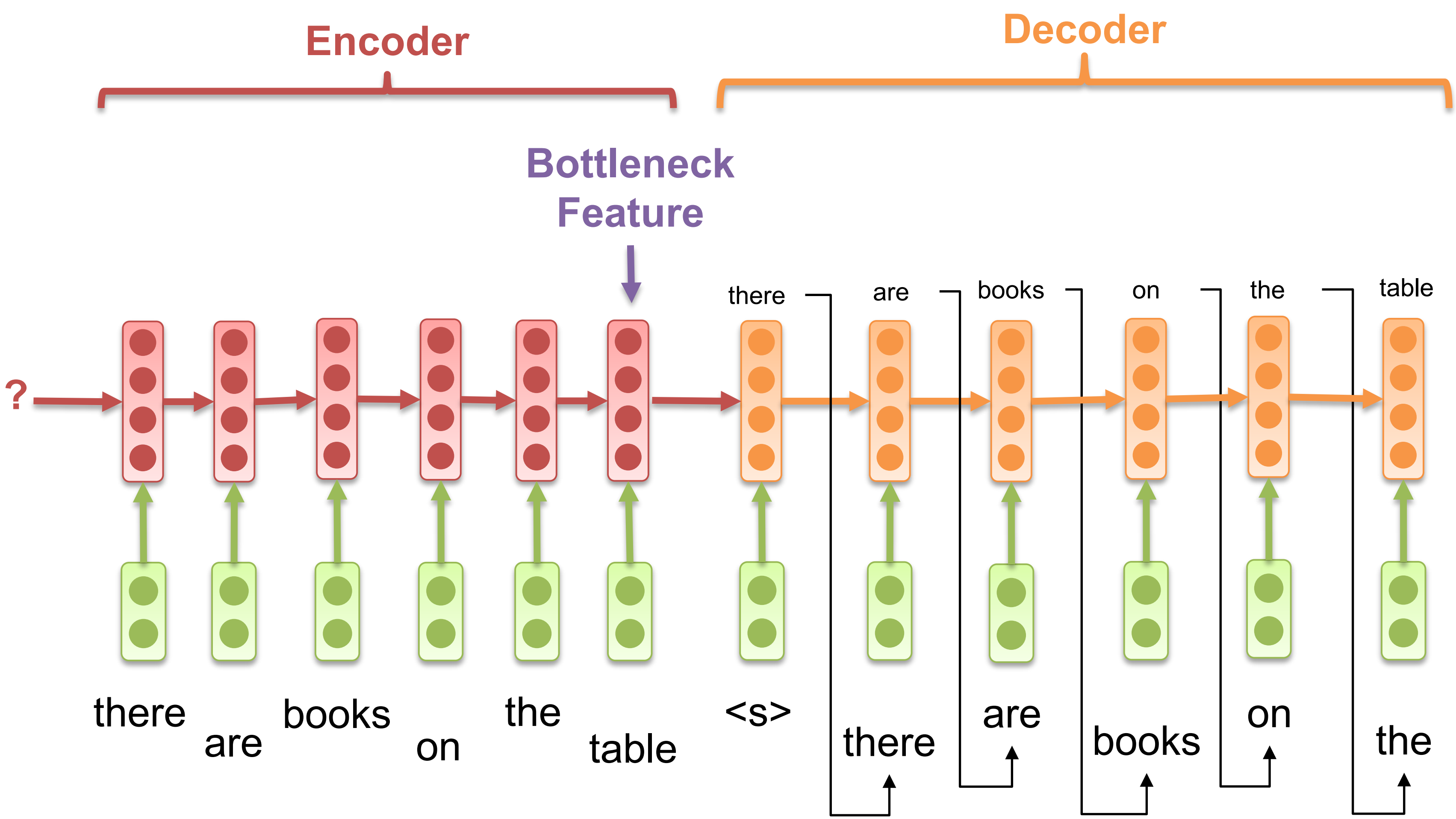
RNN-based Autoencoder….

- RNN can also be used to construct an autoencoder



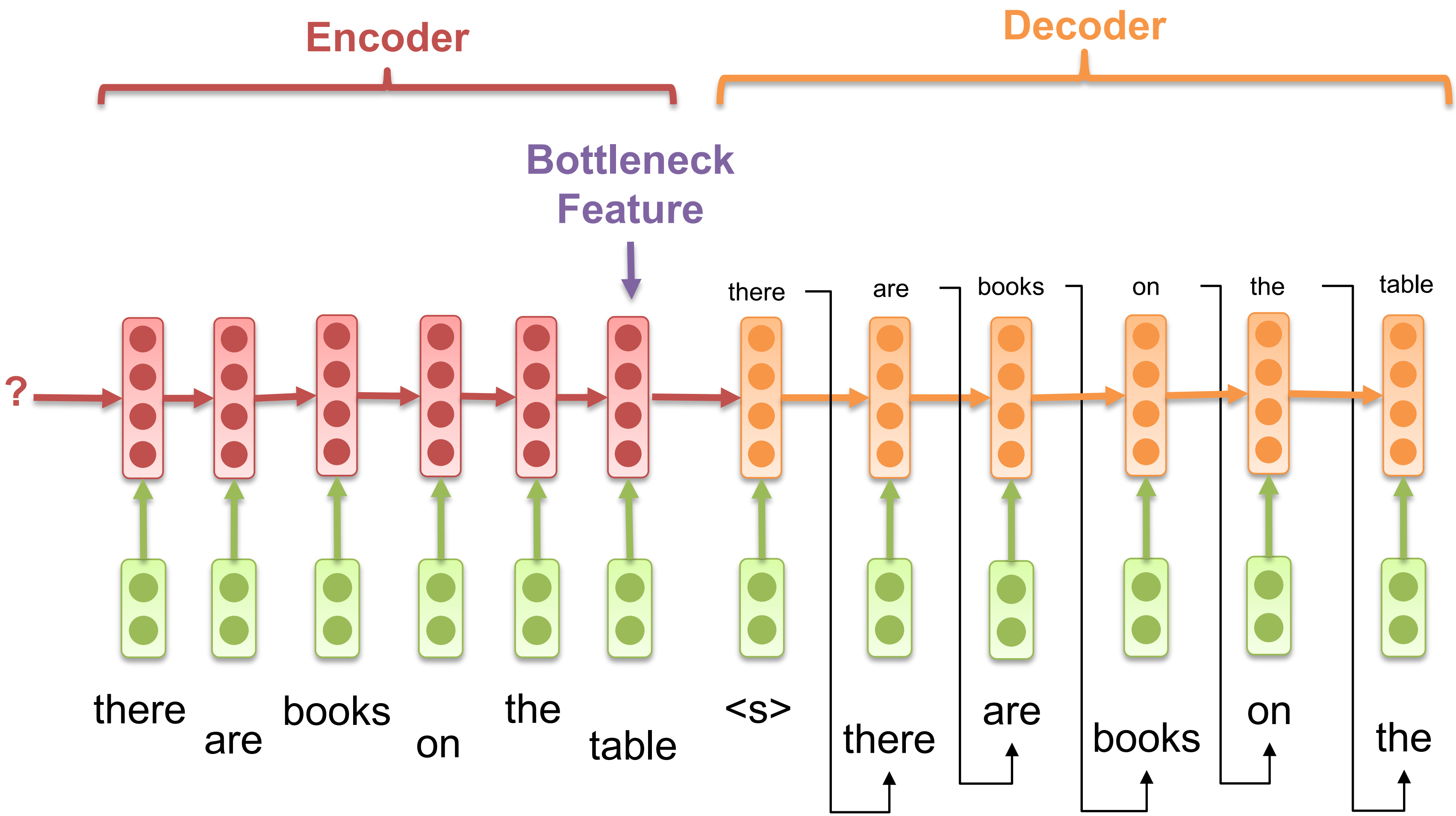
RNN-based Autoencoder.....

- RNN can also be used to construct an autoencoder



Sequence-to-sequence Learning

- Such a methodology also calls sequence-to-sequence (seq2seq) learning

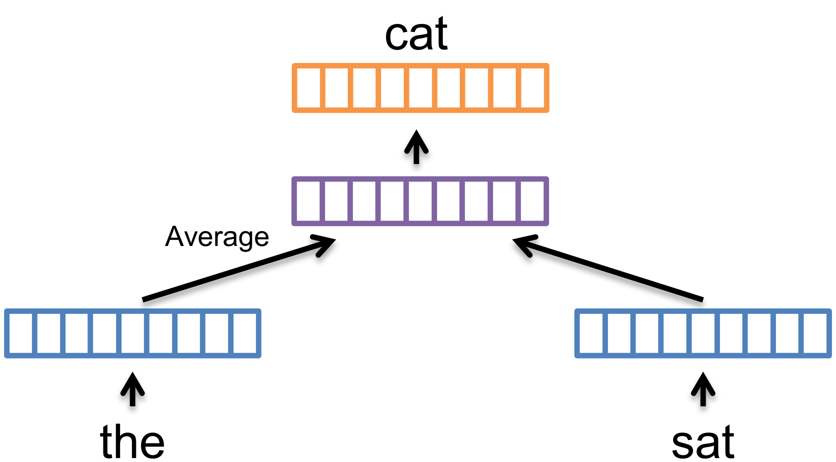


Revisiting Classic Word Embeddings

- CBOW and Skip-gram models are two representative word embedding methods

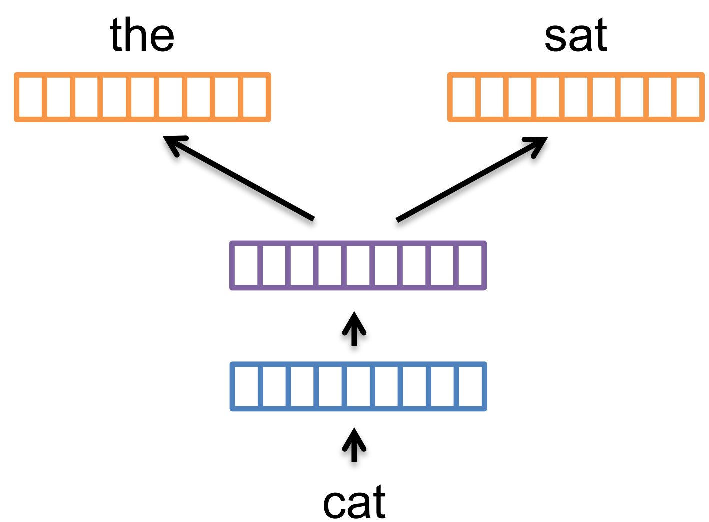
CBOW

$$\prod_{t=1}^{T} P(w_t | w_{t-c}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+c})$$



Skip-gram

$$\prod_{t=1}^{T} P(w_{t-c}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+c} | w_t)$$



Paragraph Embeddings

- Learning of paragraph representations is more reasonable and suitable for some tasks
  - Summarization, Retrieval, and Sentiment Analysis

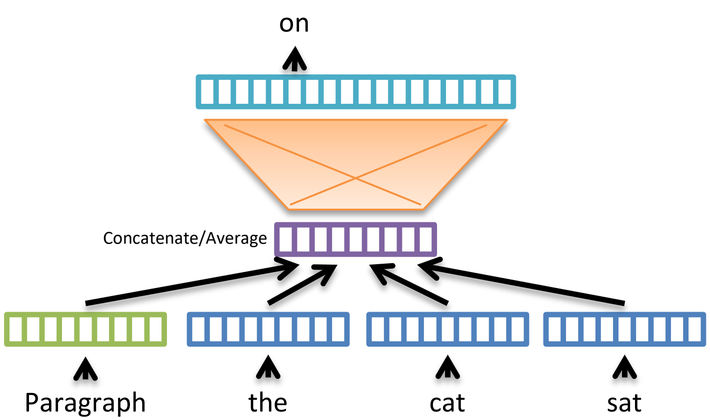
- A straightforward method is to represent a paragraph by averaging the vector representations of words occurring in the paragraph

$$\vec{d} = \sum_{w \in d} \frac{c(w,d)}{|d|} v_w$$

Distributed Memory (DM) Model

- Learning of paragraph representations is more reasonable and suitable for some tasks
  - The distributed memory model, the distributed bag-of-words model, and the thought vector model

- The DM model is inspired from the CBOW model

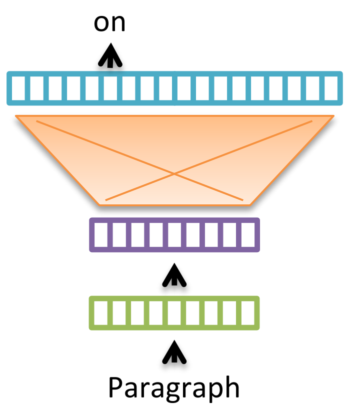


$$\prod_{t=1}^{T} P(w_t | w_{t-c}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+c}, d)$$

  - The idea is that a given paragraph also contributes to the prediction of a next word

Distributed Bag-of-words (DBOW) Model

- Opposite to the DM model, a simplified version is to only leverage the paragraph representation to predict all of the words occurring in the paragraph

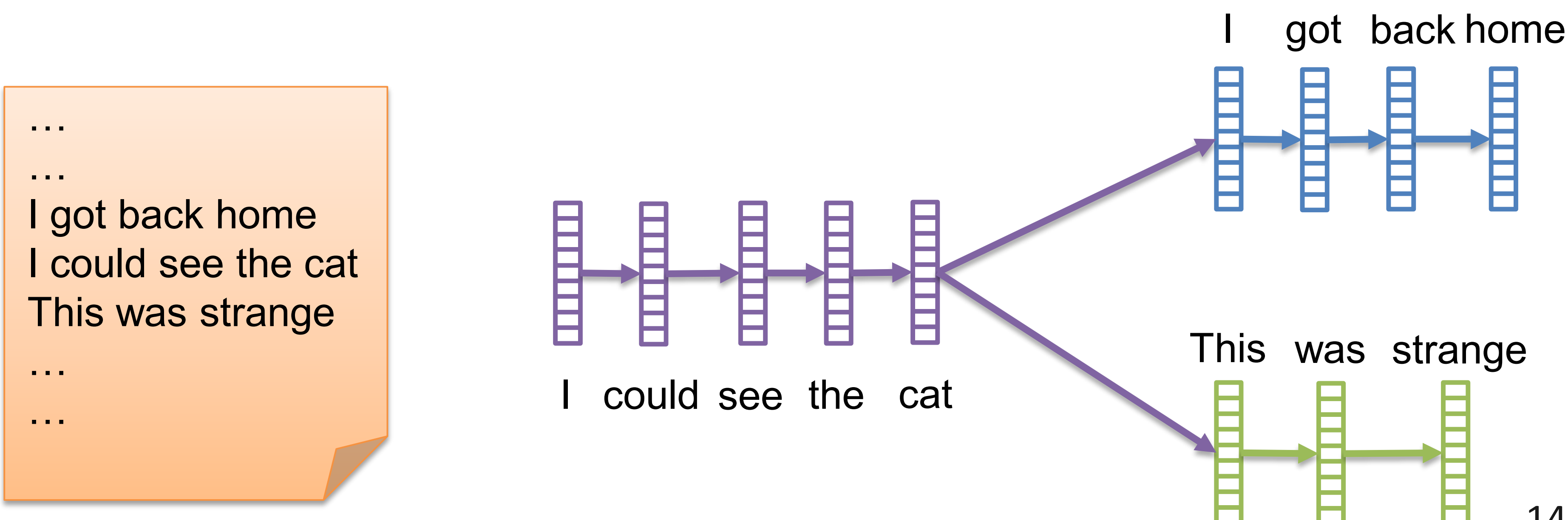


$$\prod_{t=1}^{T} P(w_t|d)$$

- Since the model ignores the contextual words at the input layer, it is named the distributed bag-of-words (DBOW) model

Skip-Thought Vector Model

- The skip-thought vector model presents an objective function that abstracts the **skip-gram** model to the sentence level
  - Instead of using a word to predict its surrounding context, thought vector encodes a sentence to predict the sentences around it



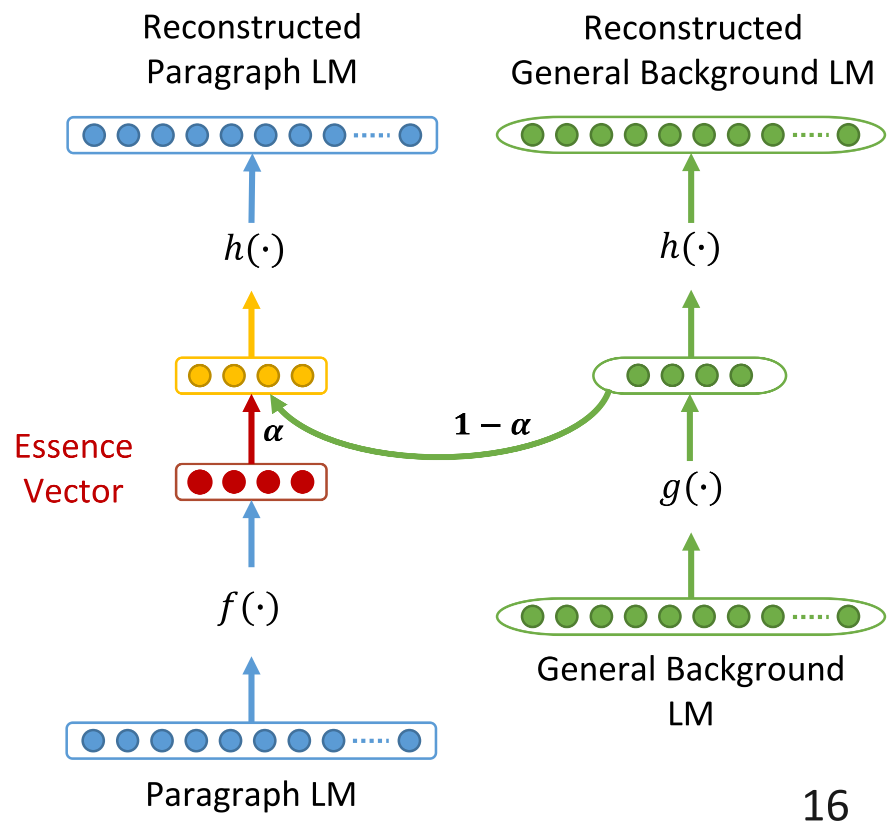
Classic Paragraph Embedding Methods

- Classic paragraph embedding methods infer the representation of a given paragraph by **considering all of the words** occurring in the paragraph
  - Such as the Distributed Memory model, the Distributed Bag-of-words model, and the skip-though vector model

- The **stop** or **function words** that occur frequently may mislead the embedding learning process
  - The learned representation for the paragraph might be undesired
  - The performance is limited
  - Our goal is to
    - Distill the most representative information from a given paragraph
    - Get rid of the general background information

Learning to Distill

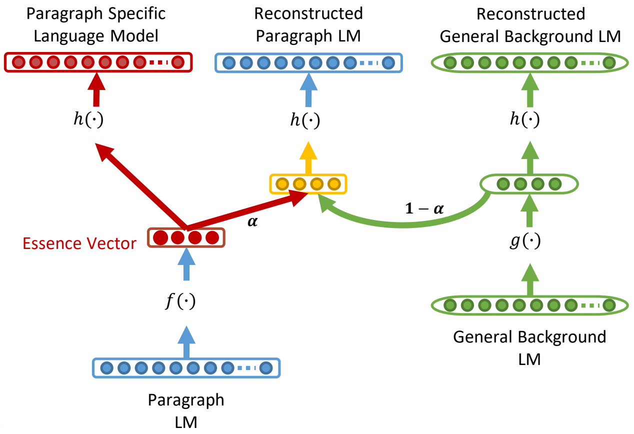
- We assume that each paragraph can be assembled by the **paragraph specific information** and the **general background information**
  - This assumption also holds in the low-dimensional representation space
  - Three modules
    - Paragraph encoder $f(\cdot)$
    - Background encoder $g(\cdot)$
    - Decoder $h(\cdot)$



Essence Vector-based Language Model

- A brilliant property inherits in the EV model is that it can be readily inferred a "paragraph" specific language model

$$\hat{P}(w) \equiv h\left(f\left(P_{D_t}\right)\right)$$



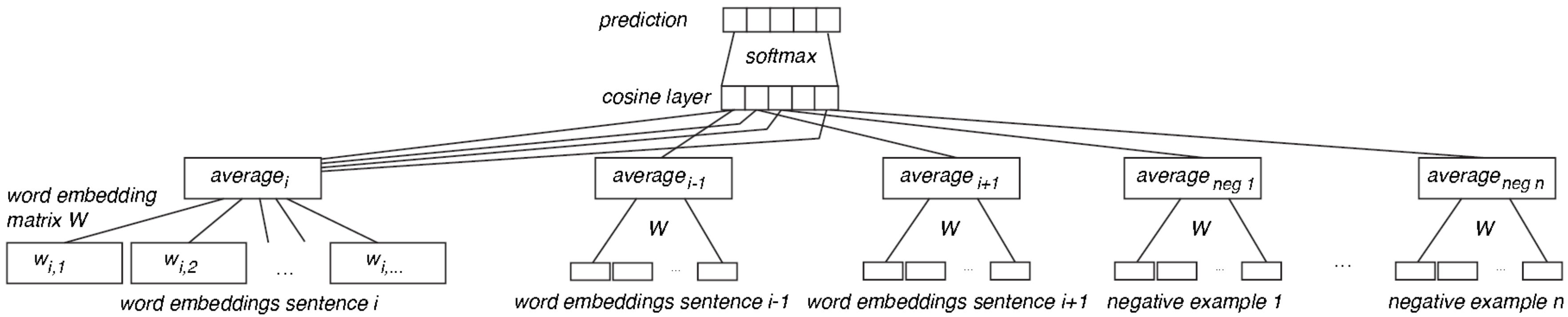
Paragraph Specific Language Model

Original Paragraph Language Model

Background Language Model

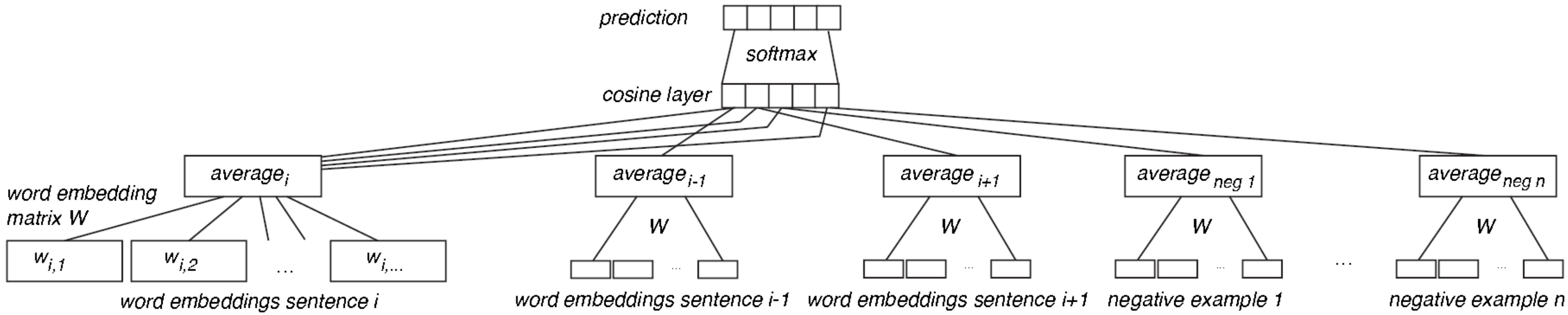
Siamese CBOW.

- Siamese CBOW model aims at learning a set of word embeddings which can be directly used for the purpose of being averaged



K.Y. Chen, *et al.*, “Learning to distill: the essence vector modeling framework,” in *Proc. of Coling*, 2016.

Siamese CBOW..



$$L = -\sum_{s_j \in \{S^+, S^-\}} P(s_i, s_j) \log P'(s_i, s_j)$$

sentences that occur next to the target sentence

$$P(s_i, s_j) = \begin{cases} \dfrac{1}{|S^+|}, & if\ s_j \in S^+ \\ 0, & if\ s_j \in S^- \end{cases}$$

randomly chosen sentences that do not occur next to the target sentence

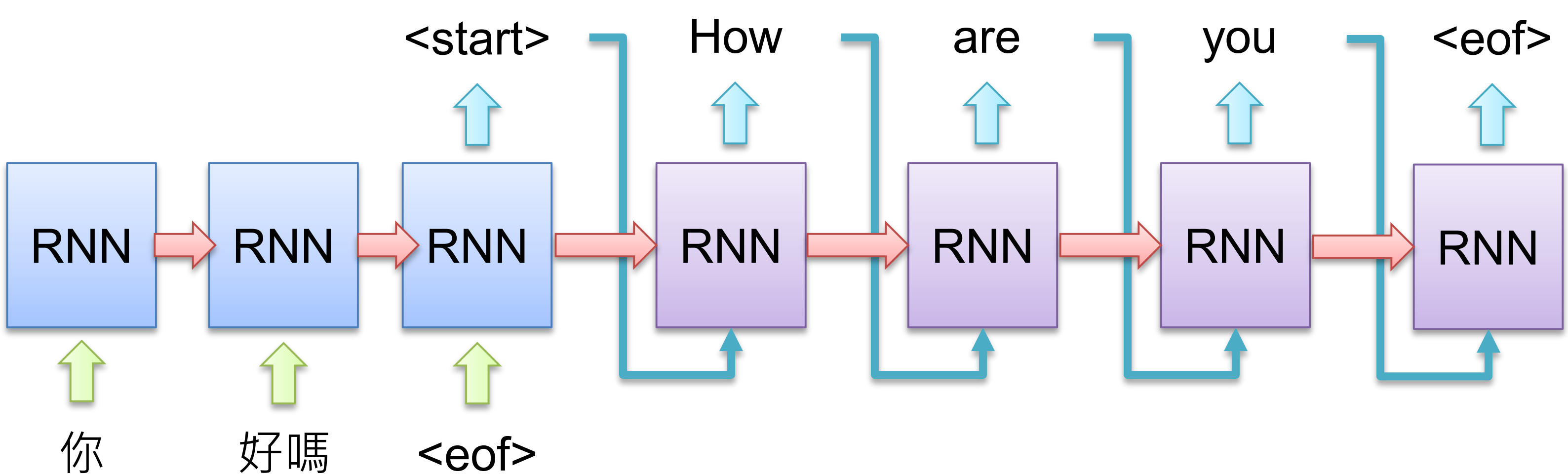
…
…
I got back home
I could see the cat
This was strange
…
…

$$P'(s_i, s_j) = \frac{e^{\cos(\vec{s_i}, \vec{s_j})}}{\sum_{s_k \in \{S^+, S^-\}} e^{\cos(\vec{s_i}, \vec{s_k})}}$$

sentence representations

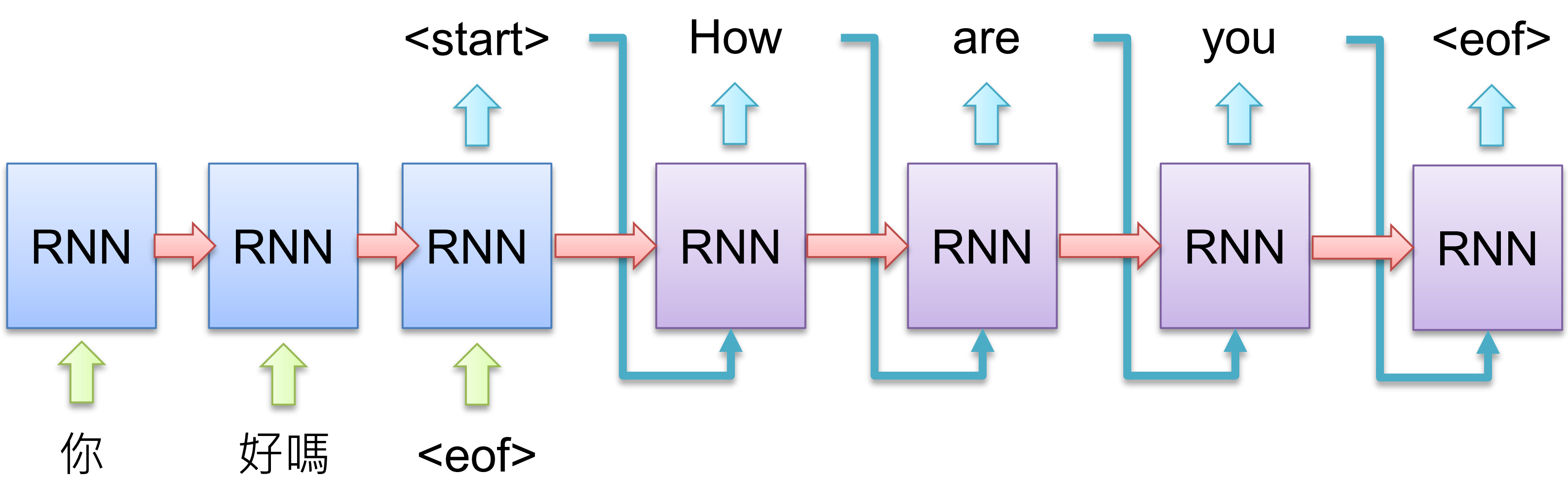
Machine Translation.

- RNN can be used to encode a variable-length source sentence, and then a variable-length target sentence will be generated by considering the encoded information
  - RNN Encoder-Decoder
  - Seq2seq
  - It is suitable for machine translation task



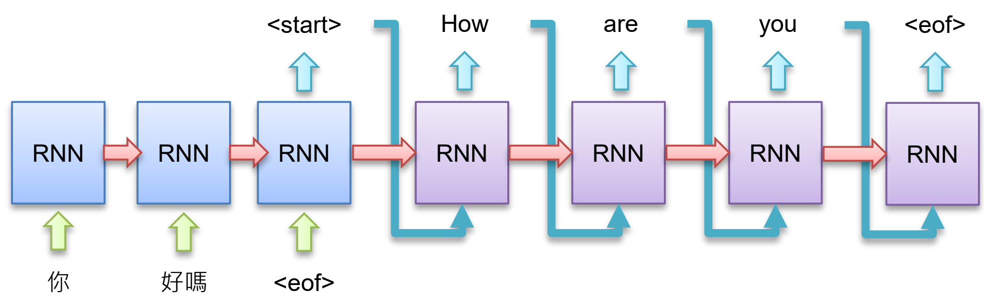
Machine Translation..

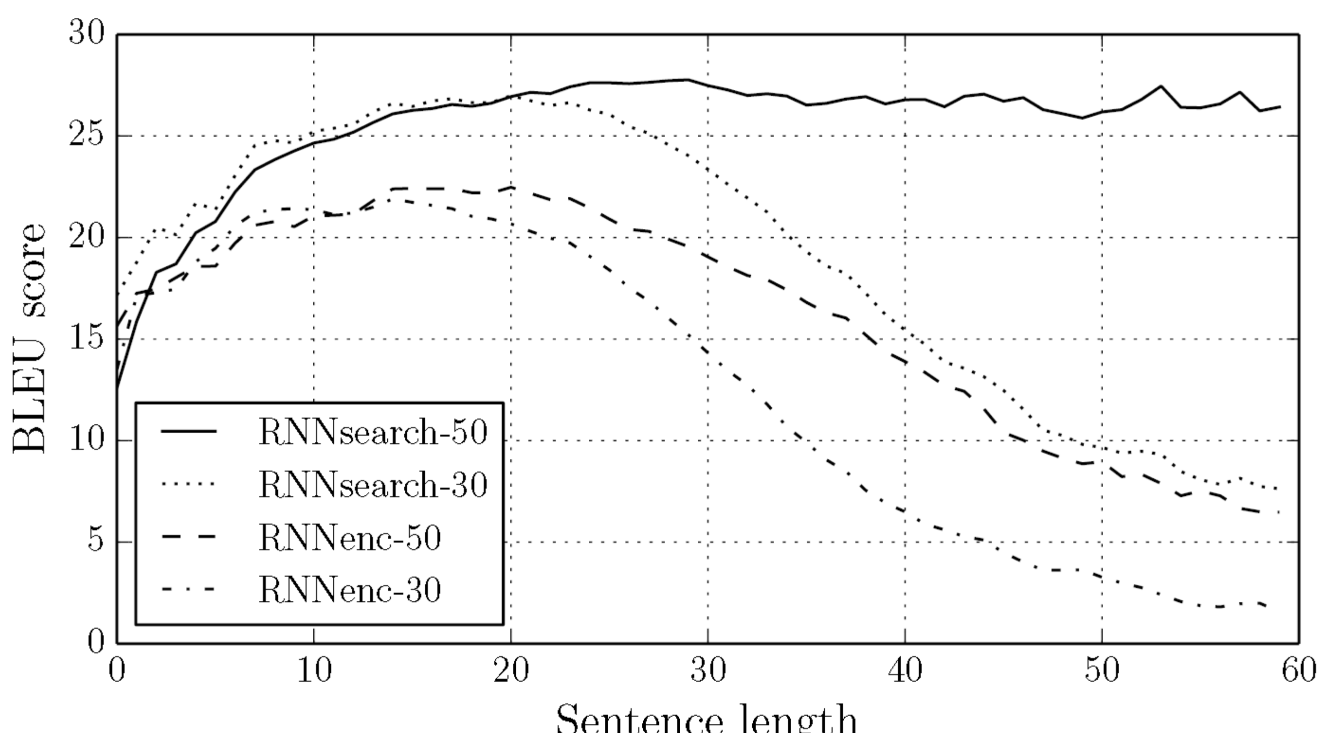
- A potential issue with this encoder–decoder approach is that a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector



Machine Translation…

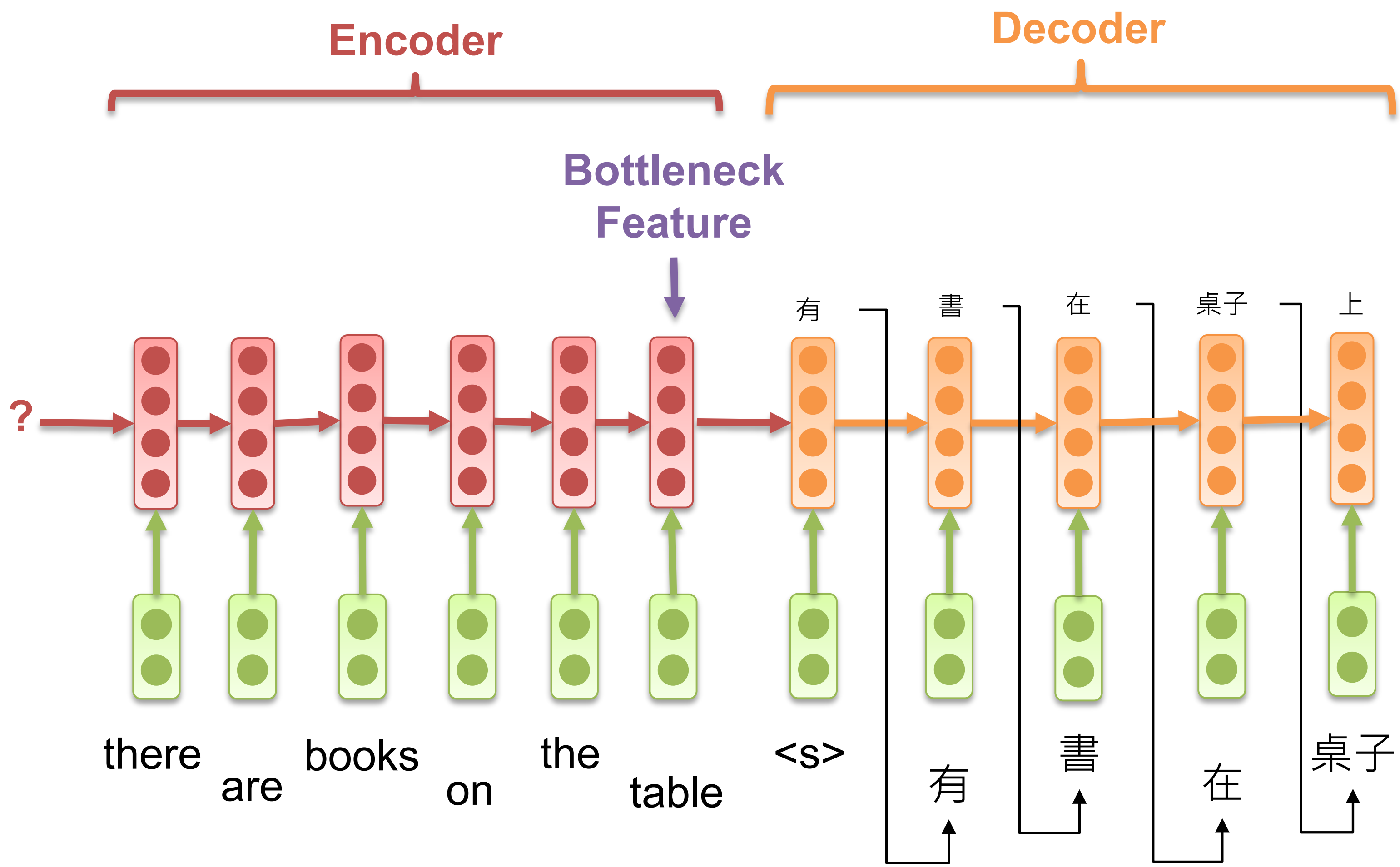
- The performance will drop when the sentence being longer!



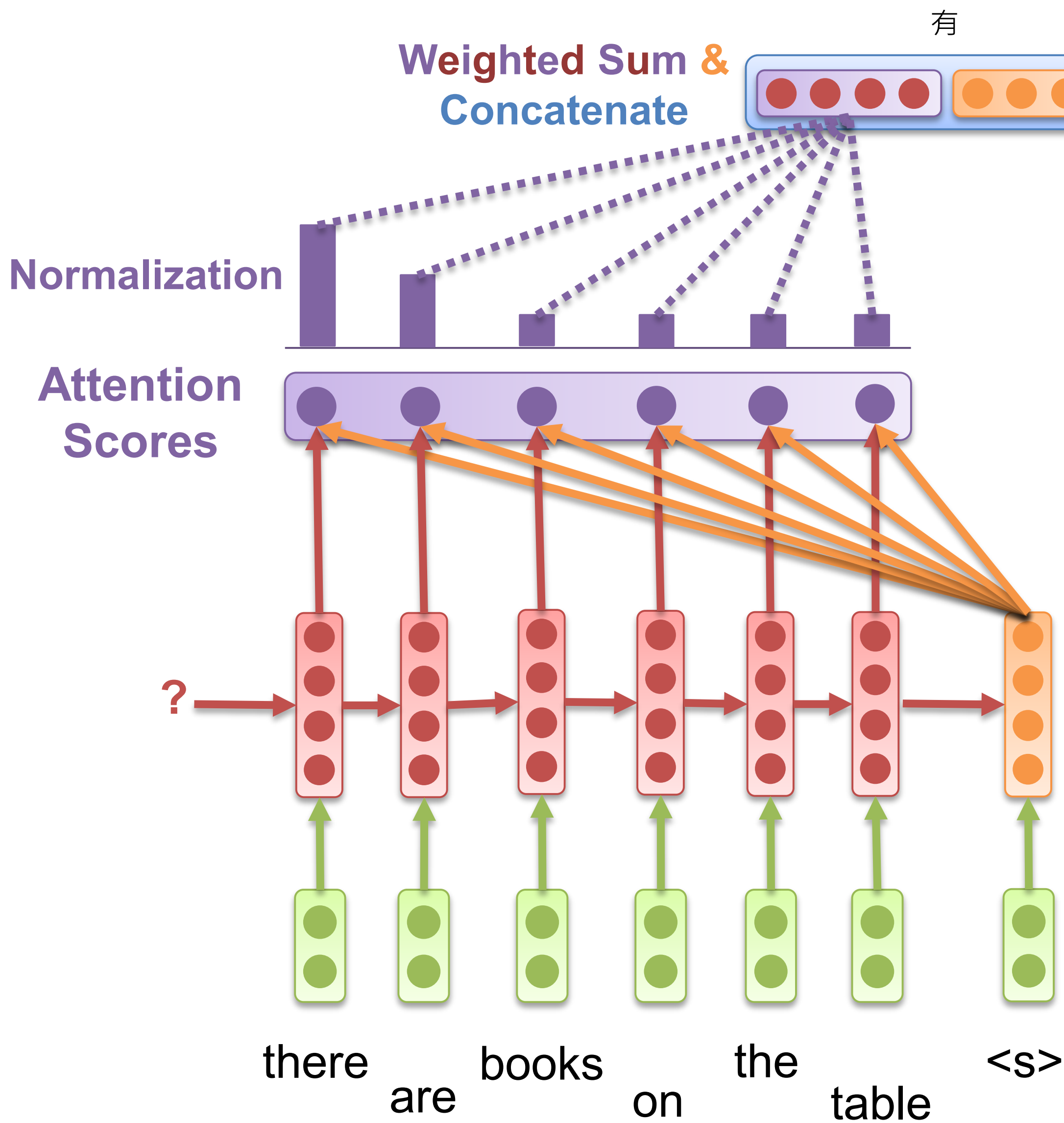


The Bottleneck Problem

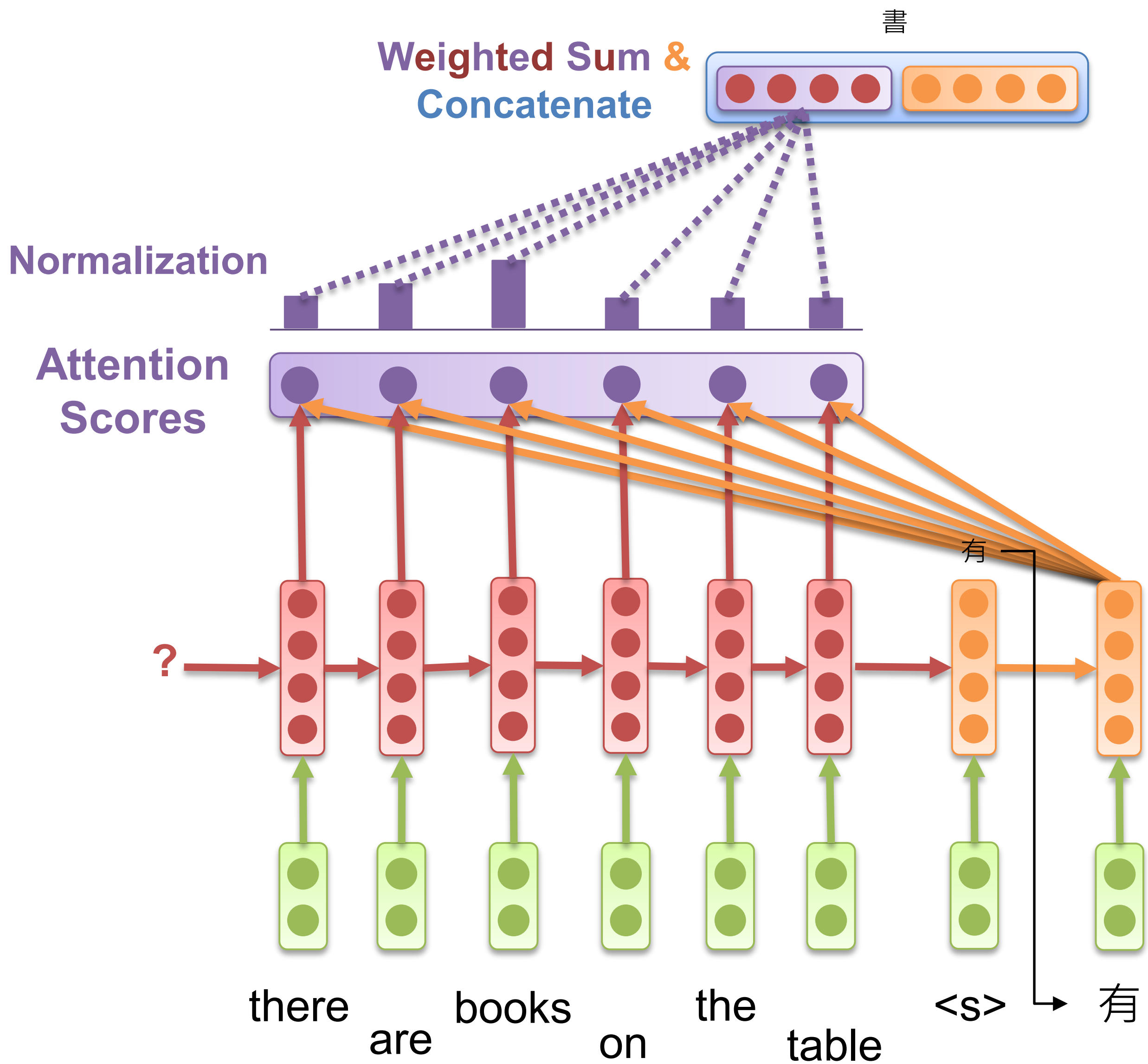
- The bottleneck feature needs to capture all information about the source sentence
  - Information bottleneck!



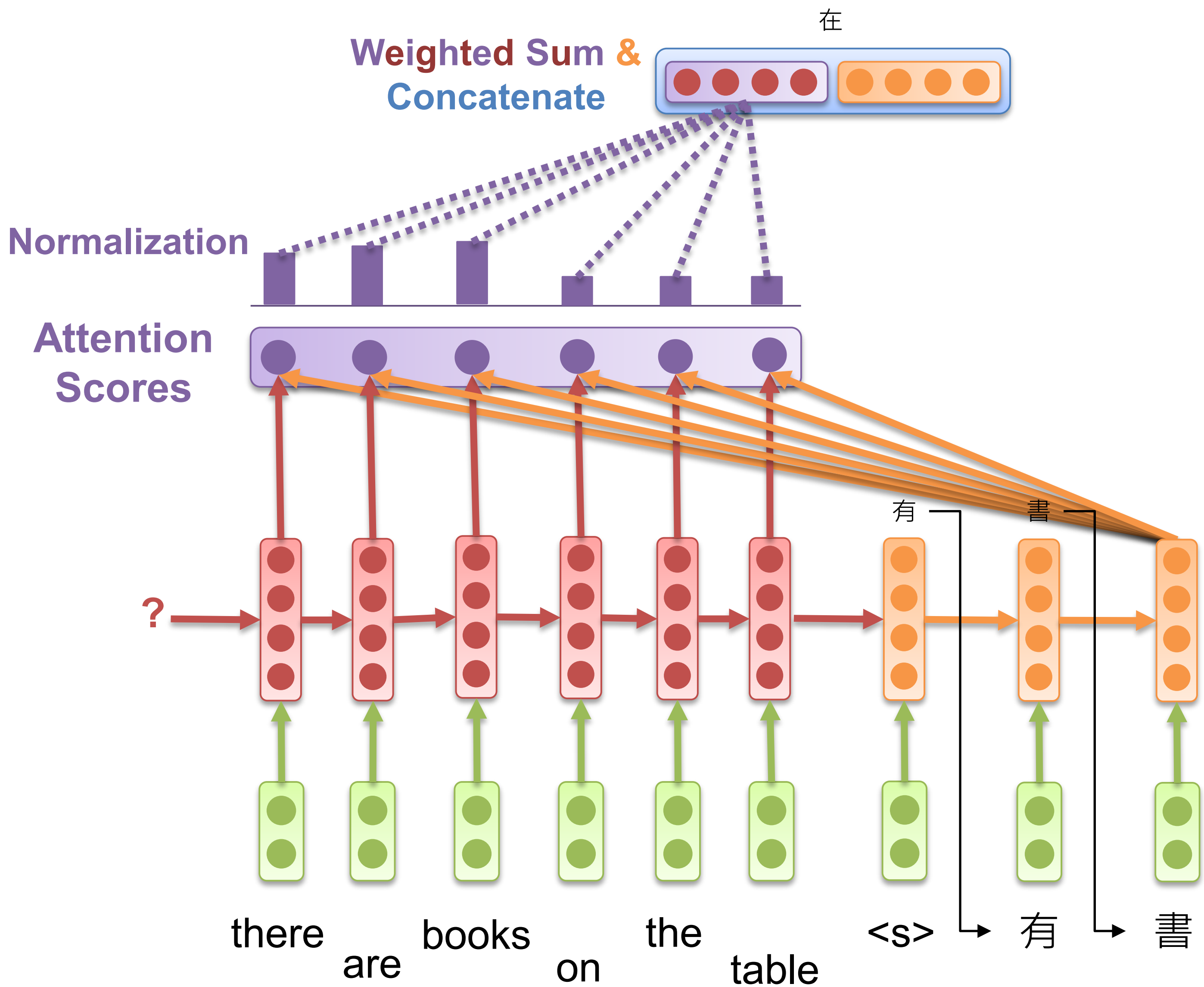
Attention Mechanism.



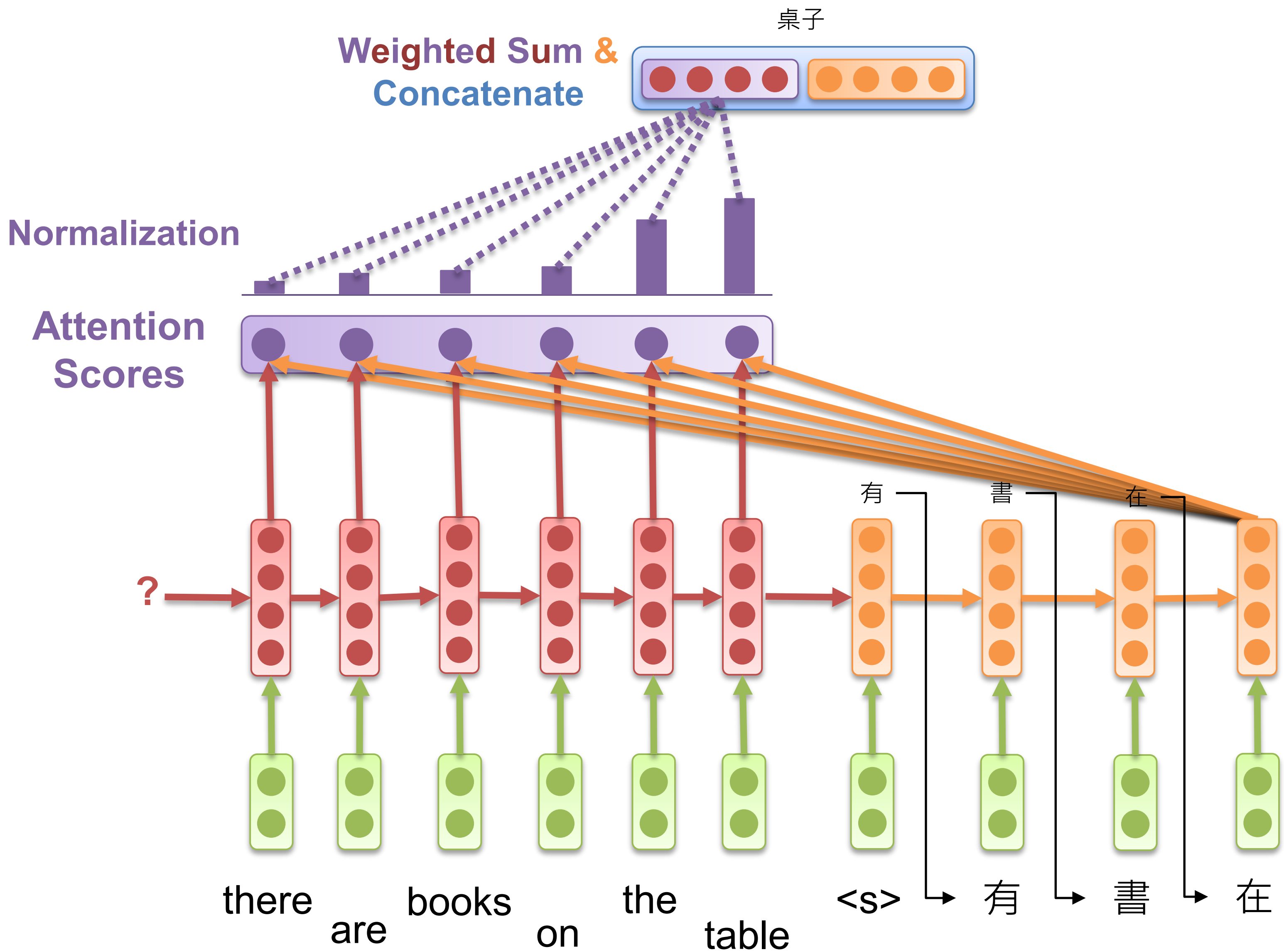
Attention Mechanism..



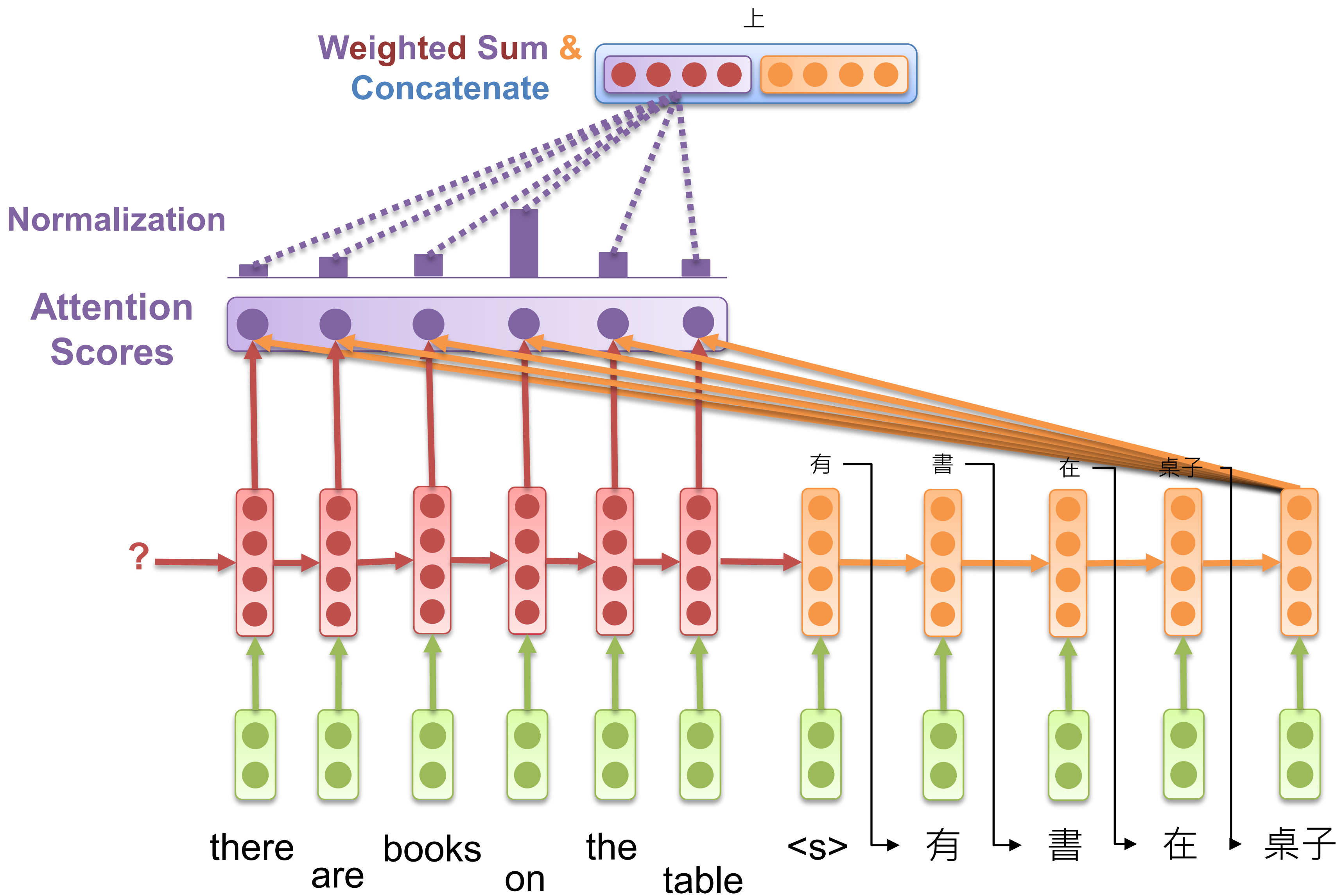
Attention Mechanism...



Attention Mechanism....



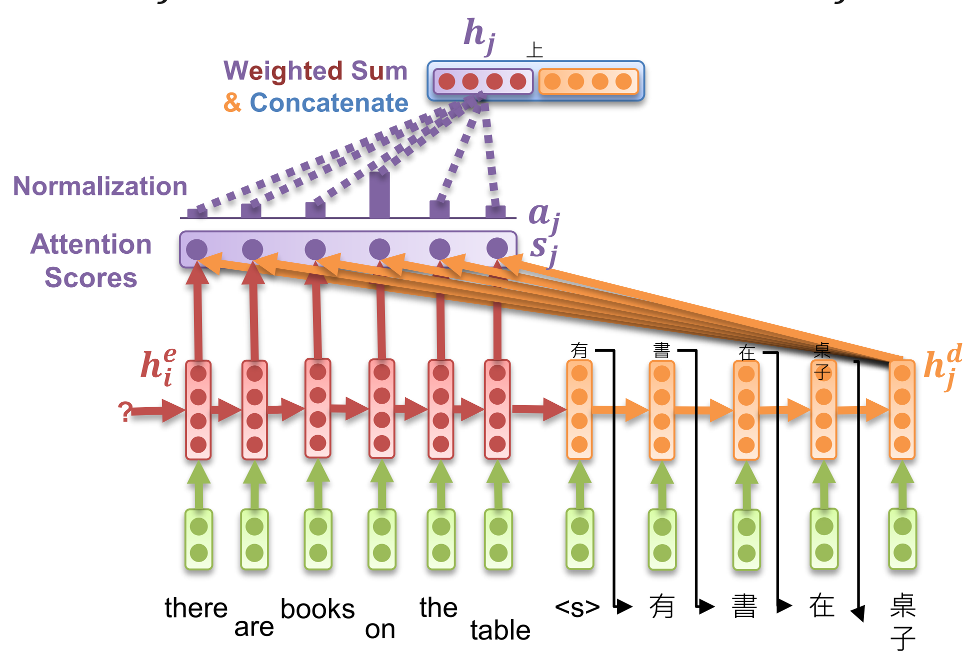
Attention Mechanism.....



Descriptions

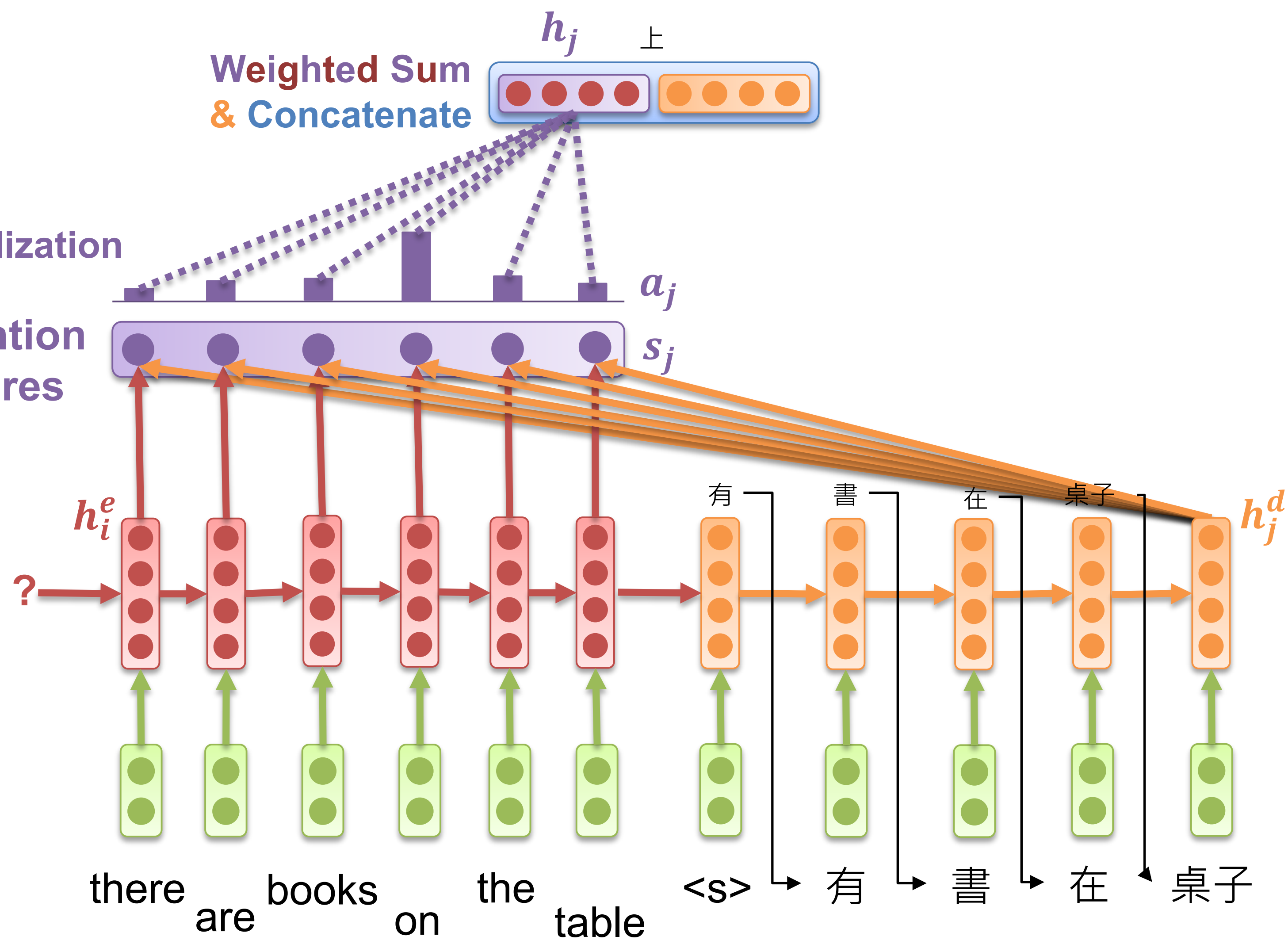
- The attention mechanism
  - The encoder states $h_1^e, h_2^e, \cdots, h_i^e, \cdots, h_I^e \in \mathbb{R}^{d_1}$
  - The decoder states $h_1^d, h_2^d, \cdots, h_j^d, \cdots, h_J^d \in \mathbb{R}^{d_2}$
  - The attention score vector at time j is $s_j \in \mathbb{R}^I$
  - Softmax is taken on s_j to get the attention distribution $a_j \in \mathbb{R}^I$
  - A new vector representation h_j is derived by referring to a_j and the encoder states

$$h_j = \sum_{i=1}^{I} a_j^i \, h_i^e$$



The Attention Scores.

- There are several ways for us to compute the attention scores



The Attention Scores..

- Basic dot-product Attention
  - Assume $d_1 = d_2$

$$s^i = h_i^e \cdot h_j^d$$

- Multiplicative Attention
  - $W \in \mathbb{R}^{d_1 \times d_2}$ is a learned parameter

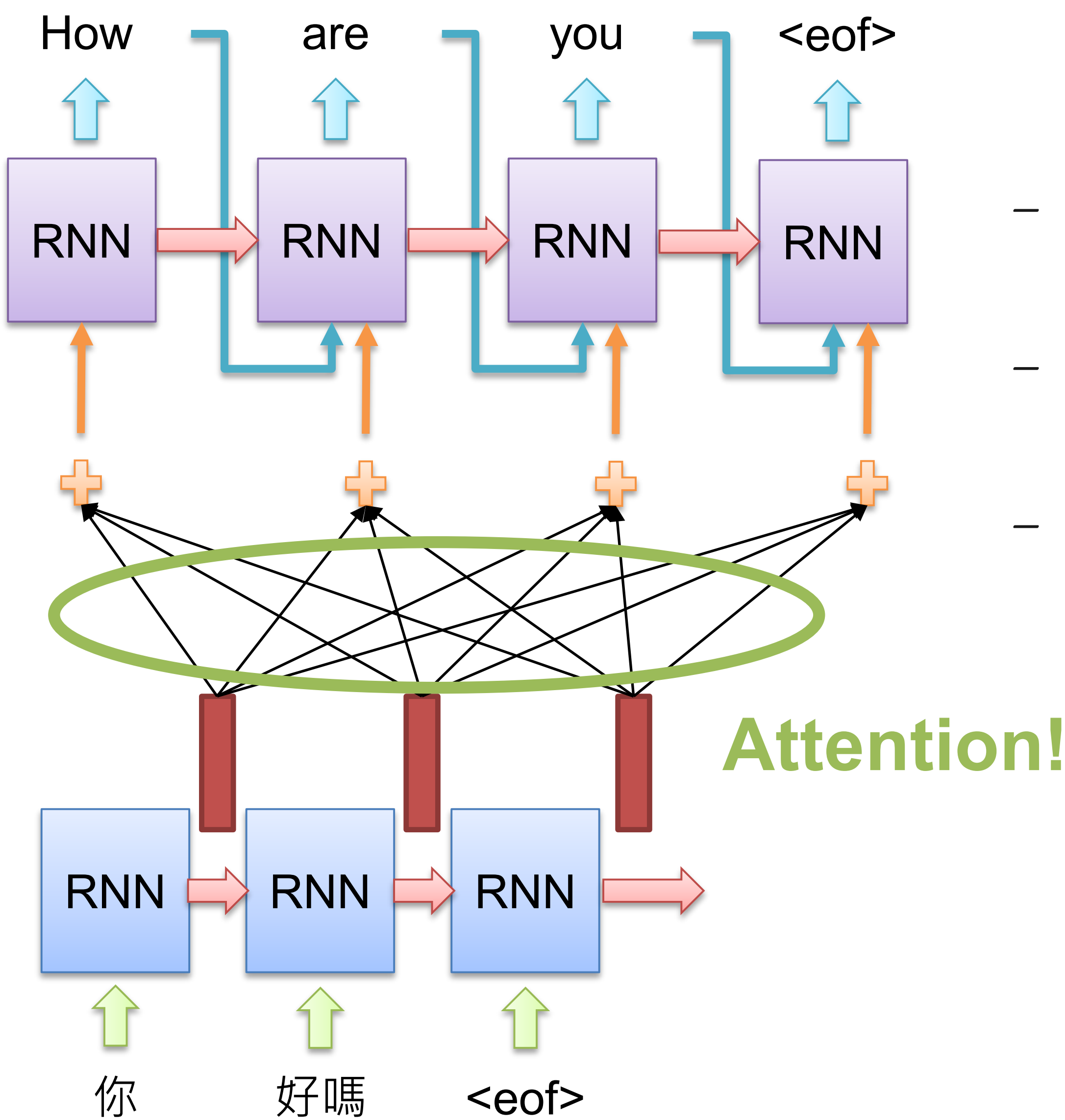
$$s^i = (h_i^e)^{\mathrm{T}} W h_j^d$$

- Additive Attention
  - $W_1 \in \mathbb{R}^{d_3 \times d_1}$, $W_2 \in \mathbb{R}^{d_3 \times d_2}$, and $W_3 \in \mathbb{R}^{d_3}$ are learned parameters

$$s^i = W_3^{\mathrm{T}} \tanh(W_1 h_i^e + W_2 h_j^d)$$

The encoder states $h_1^e, h_2^e, \cdots, h_i^e, \cdots, h_I^e \in \mathbb{R}^{d_1}$
The decoder states $h_1^d, h_2^d, \cdots, h_j^d, \cdots, h_J^d \in \mathbb{R}^{d_2}$

Attention-based Modeling

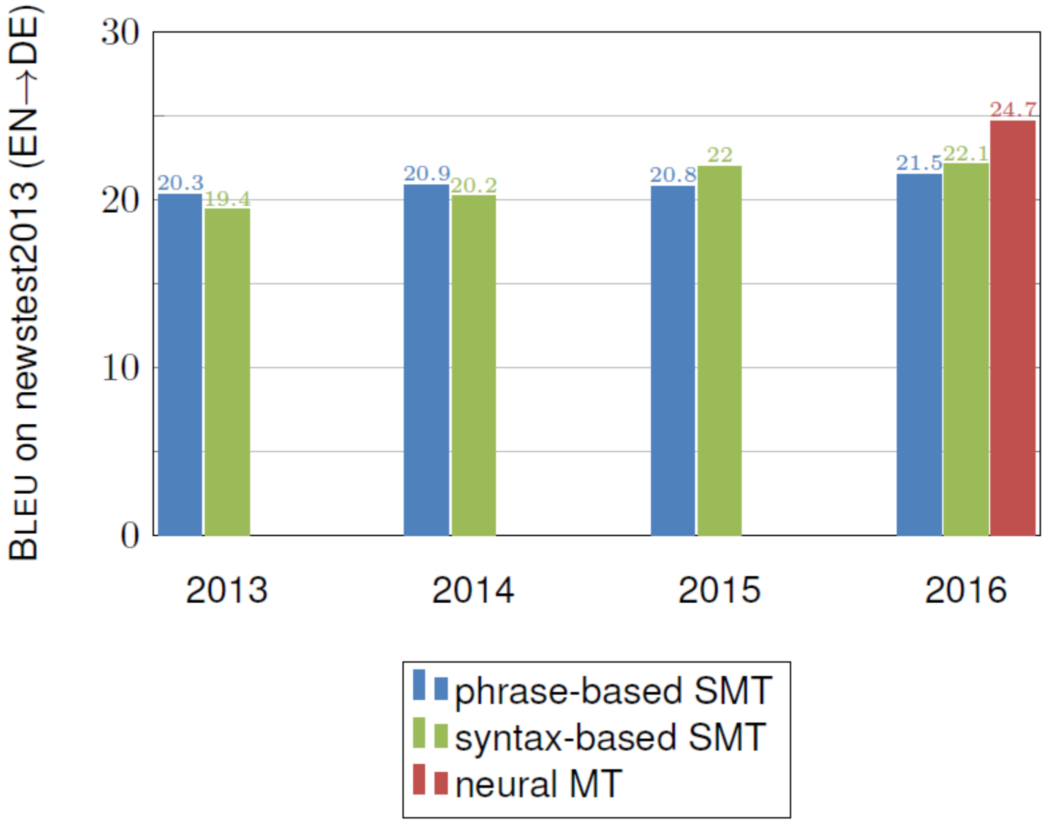


- Location-based Modeling
  - Handwriting synthesis
- Content-based Modeling
  - Machine Translation
- Hybrid Attention Modeling
  - Speech Recognition

Amazing!

- Neural Machine Translation went from a fringe research activity in 2014 to the leading standard method in 2016
  - 2014: First seq2seq paper published
  - 2016: Google Translate switches from SMT to NMT

- This is amazing!
  - SMT systems, built by hundreds of engineers over many years, outperformed by NMT systems trained by a handful of engineers in a few months



http://languagelog.ldc.upenn.edu/nll/?p=35120#more-35120

But.

http://languagelog.ldc.upenn.edu/nll/?p=35120#more-35120

But..

英文 中文 日文 偵測語言 | 中文(繁體) 英文 中文(簡體) | 翻譯

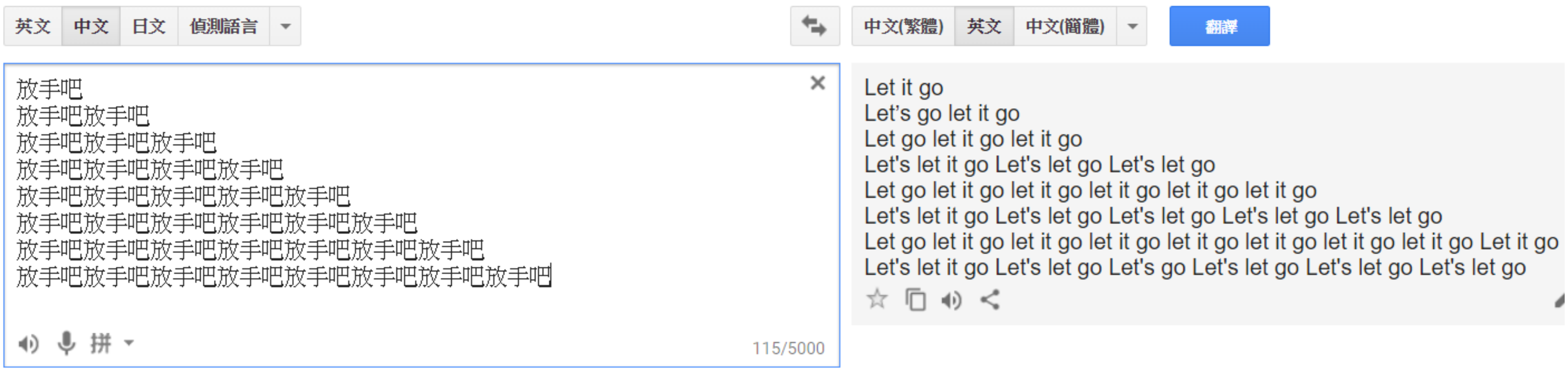
がまま
がままがまま
がままがままがまま
がままがままがままがまま
がままがままがままがままがまま
がままがままがままがままがままがまま
がままがままがままがままがままがままがまま
がままがままがままがままがままがままがままがまま
がままがままがままがままがままがままがままがままがまま
がままがままがままがままがままがままがままがままがままがまま

174/5000

放手吧
就這樣
就這樣
它將保持原樣
它將保持原樣
它將保持原樣
事實上，它保持原樣
事實上，它將保持原樣
事實上，它將保持原樣
事實上，它會保持原樣，但它會保持原樣

Fàngshǒu ba
jiù zhèyàng
jiù zhèyàng
tā jiāng bǎochí yuányàng

But...


英文 中文 日文 偵測語言
中文(繁體) 英文 中文(簡體)
翻譯
放手吧
放手吧放手吧
放手吧放手吧放手吧
放手吧放手吧放手吧放手吧
放手吧放手吧放手吧放手吧放手吧
放手吧放手吧放手吧放手吧放手吧放手吧
放手吧放手吧放手吧放手吧放手吧放手吧放手吧
放手吧放手吧放手吧放手吧放手吧放手吧放手吧放手吧
拼
115/5000
Let it go
Let's go let it go
Let go let it go let it go
Let's let it go Let's let go Let's let go
Let go let it go let it go let it go let it go let it go
Let's let it go Let's let go Let's let go Let's let go Let's let go
Let go let it go let it go let it go let it go let it go let it go let it go Let it go
Let's let it go Let's let go Let's go Let's let go Let's let go Let's let go

Questions?

kychen@mail.ntust.edu.tw