中国科学院大学
University of Chinese Academy of Sciences

自然语言处理

第4讲 Transformers

王 石 资康莉 刘 瑜

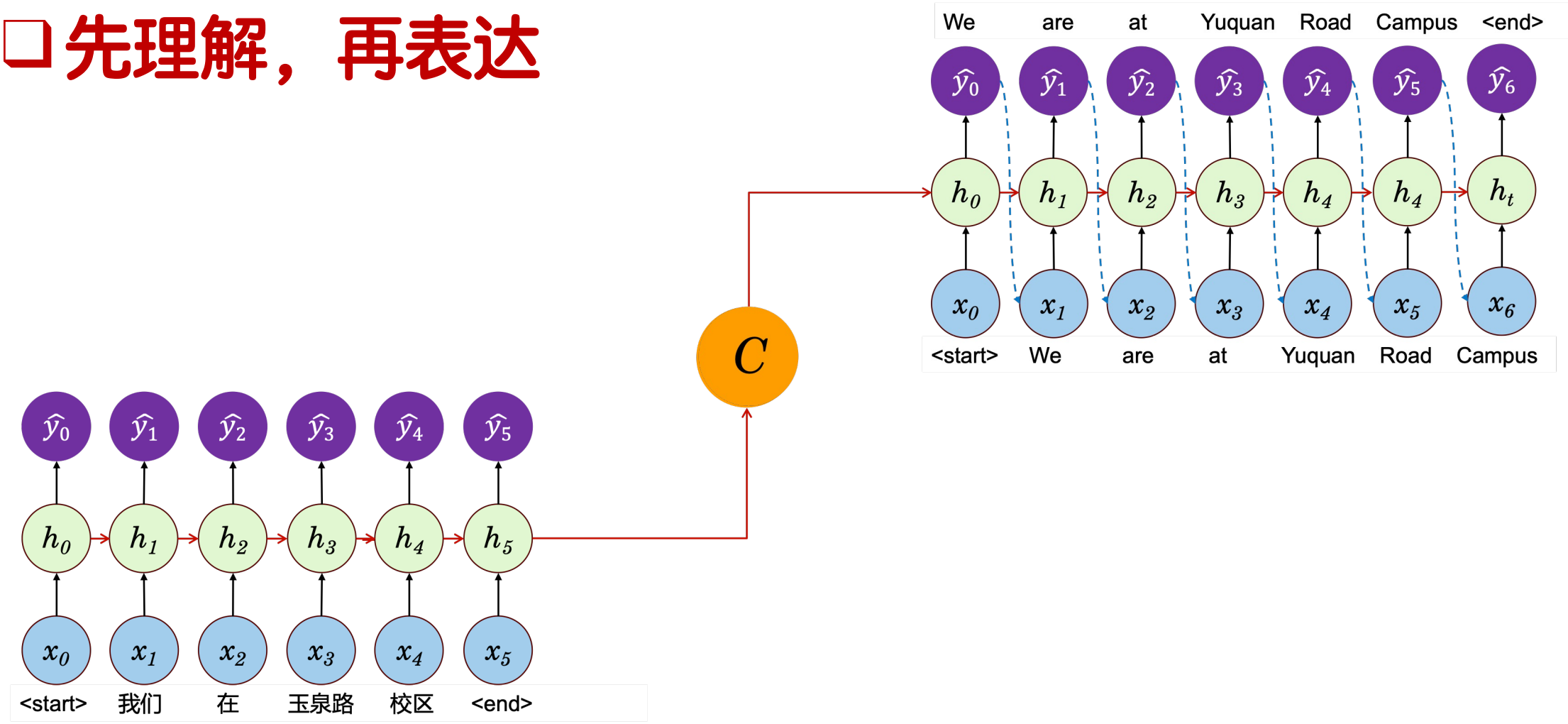
2026年春季课程

https://ictkc.github.io/teaching/

第四讲 Transformers

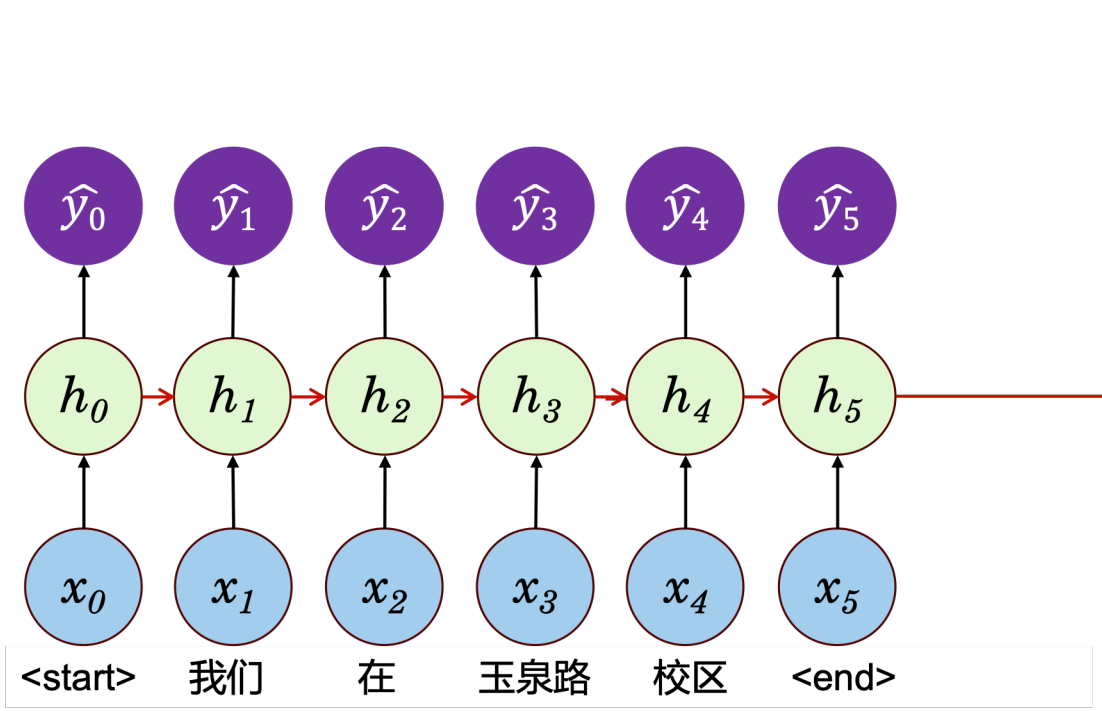
Encoder-Decoder原理

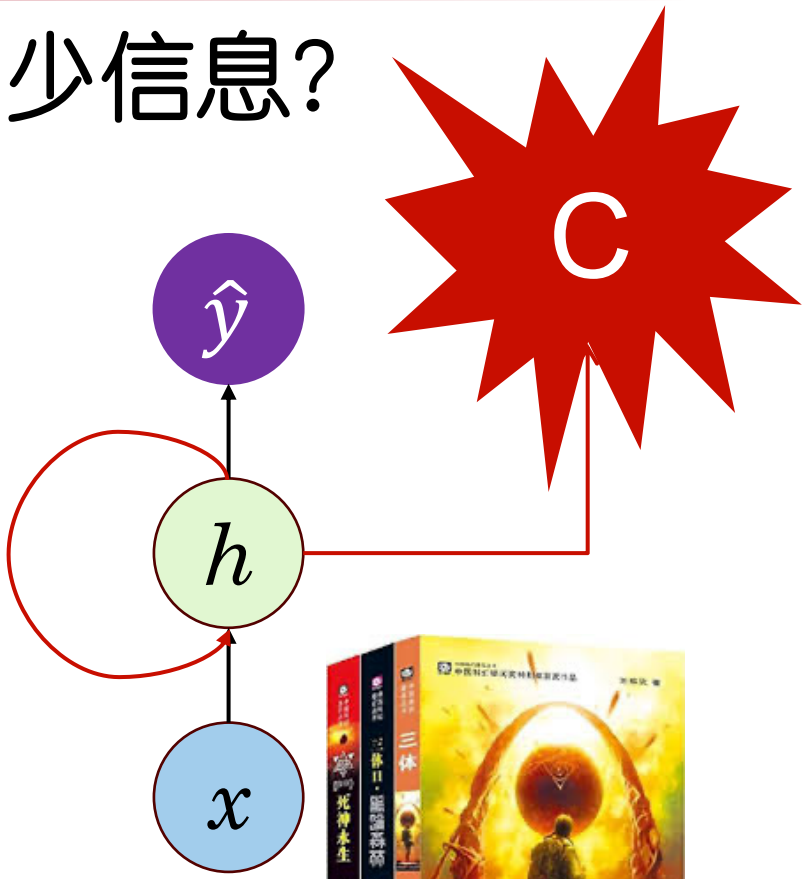
❑先理解，再表达



Encoder-Decoder缺点

- 区区一个固定长度的C，能容下多少信息？





在读一本侦探小说，在最后一页侦探说：“我将揭示罪犯的身份，他的名字是_____”

目 录

输入序列不同部分对于输出序列重要性不同

the late Emperor was taken from us before he could finish his life's work, the restoration of the Han. Today, the empire is still divided in three, and our very survival is threatened. Yet still, the officials at court and the soldiers throughout the realm remain loyal to you, your majesty. Because they remember the late emperor, all of them, and they wish to repay his kindness in service to you. This is the moment to extend your divine influence, to honor the memory of the late Emperor and strengthen the morale of your officers. It is not the time to listen to bad advice or close your ears to the suggestions of loyal men. The emperors of the Western Han chose their courtiers wisely, and their dynasty flourished.

先帝创业未半而中道崩殂，今天下三分，益州疲弊，此诚危急存亡之秋也。然侍卫之臣不懈于内，忠志之士忘身于外者，盖追先帝之殊遇，欲报之于陛下也。诚宜开张圣听，以光先帝遗德，恢弘志士之气，不宜妄自菲薄，引喻失义，以塞忠谏之路也。

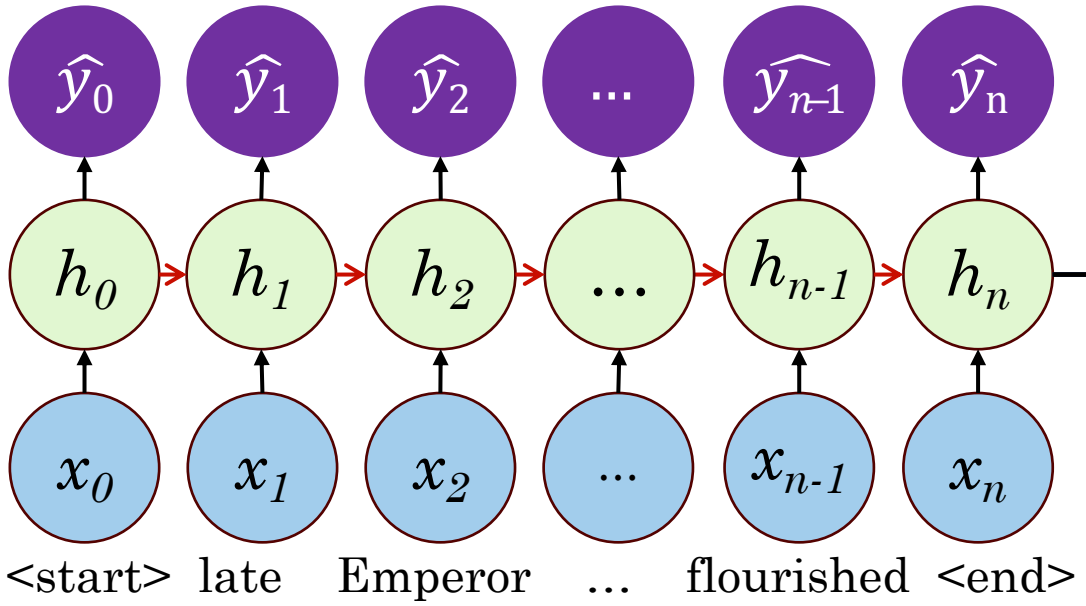
注意：不能走极端，是“开张圣听下有主有次”

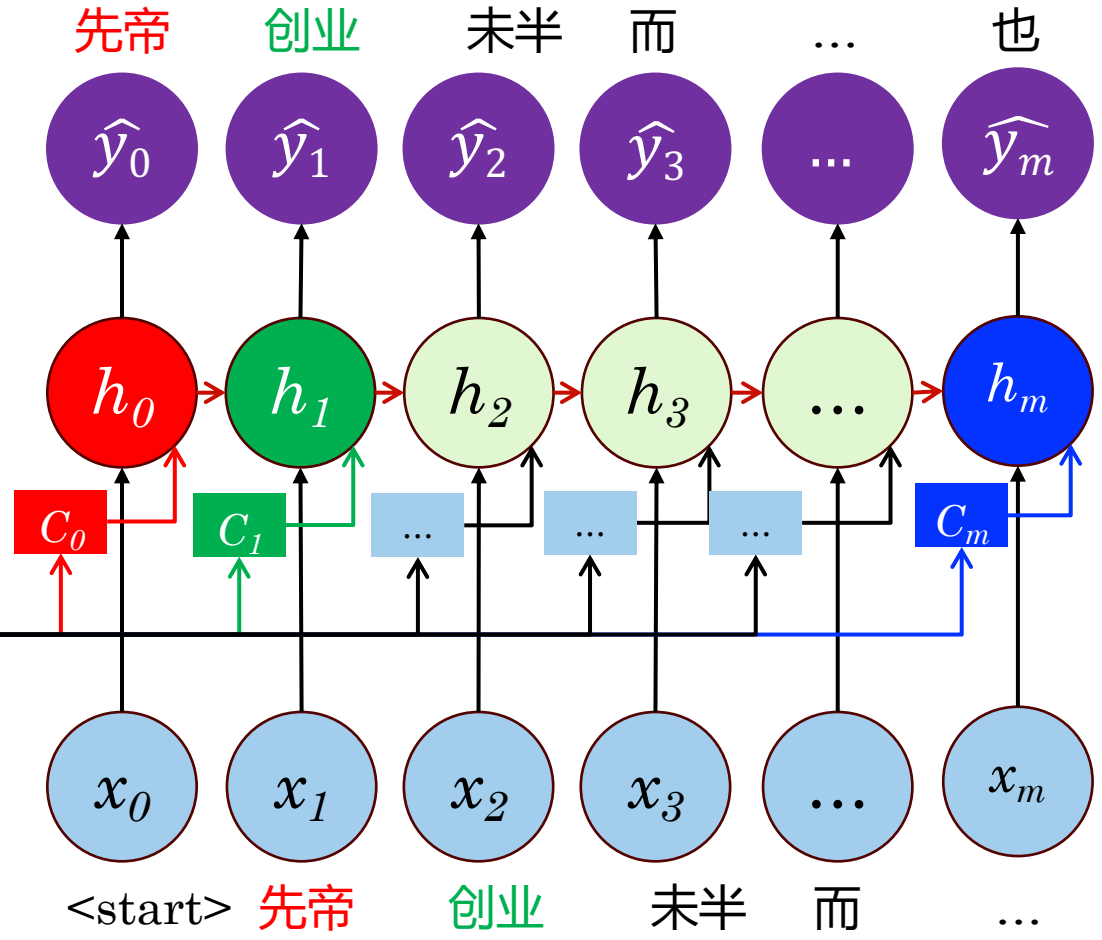
the late Emperor was taken from us before he could finish his life's work, the restoration of the Han. Today, the empire is still divided in three, and our very survival is threatened. Yet still, the officials at court and the soldiers throughout the realm remain loyal to you, your majesty. Because they remember the late emperor, all of them, and they wish to repay his kindness in service to you. This is the moment to extend your divine influence, to honor the memory of the late Emperor and strengthen the morale of your officers. It is not the time to listen to bad advice or close your ears to the suggestions of loyal men. The emperors of the Western Han chose their courtiers wisely, and their dynasty flourished.

先帝创业未半而中道崩殂，今天下三分，益州疲弊，此诚危急存亡之秋也。然侍卫之臣不懈于内，忠志之士忘身于外者，盖追先帝之殊遇，欲报之于陛下也。诚宜开张圣听，以光先帝遗德，恢弘志士之气，不宜妄自菲薄，引喻失义，以塞忠谏之路也。

注意力机制：解码不同词时，用不同C

- 解码器在i时刻不输入固定C，而是不同的C_i
- C_i自动选取与$\hat{y}_i$最相关的上下文

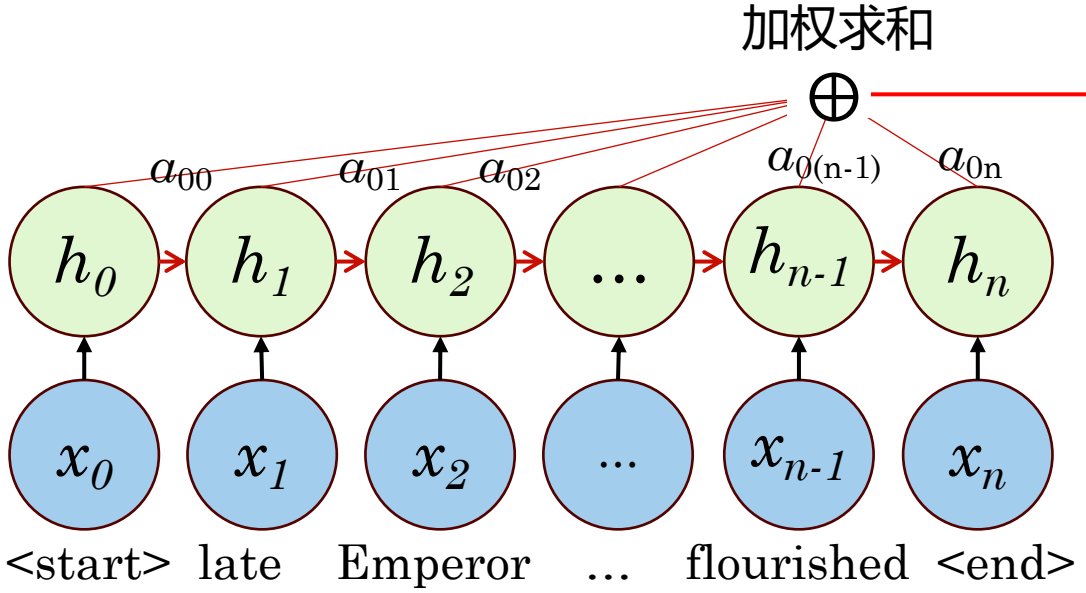




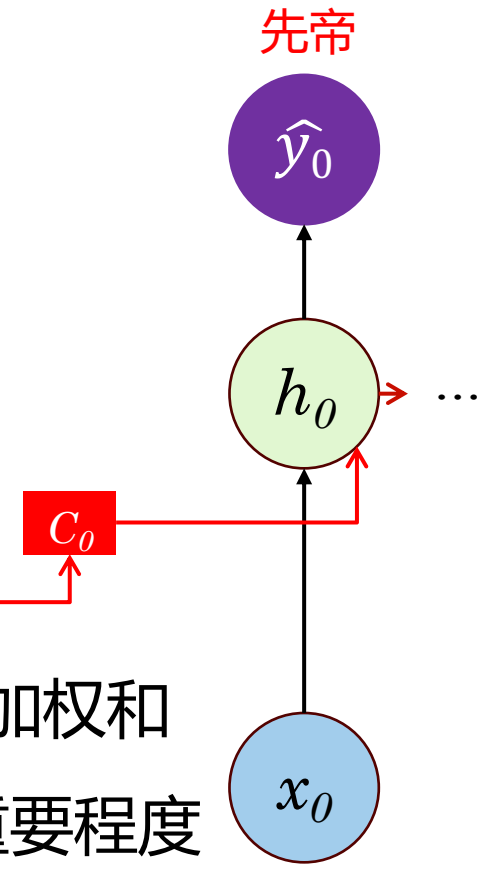
C_i如何计算

C_i特征

1. 来源于所有的隐变量
2. 不同隐变量有不同权重

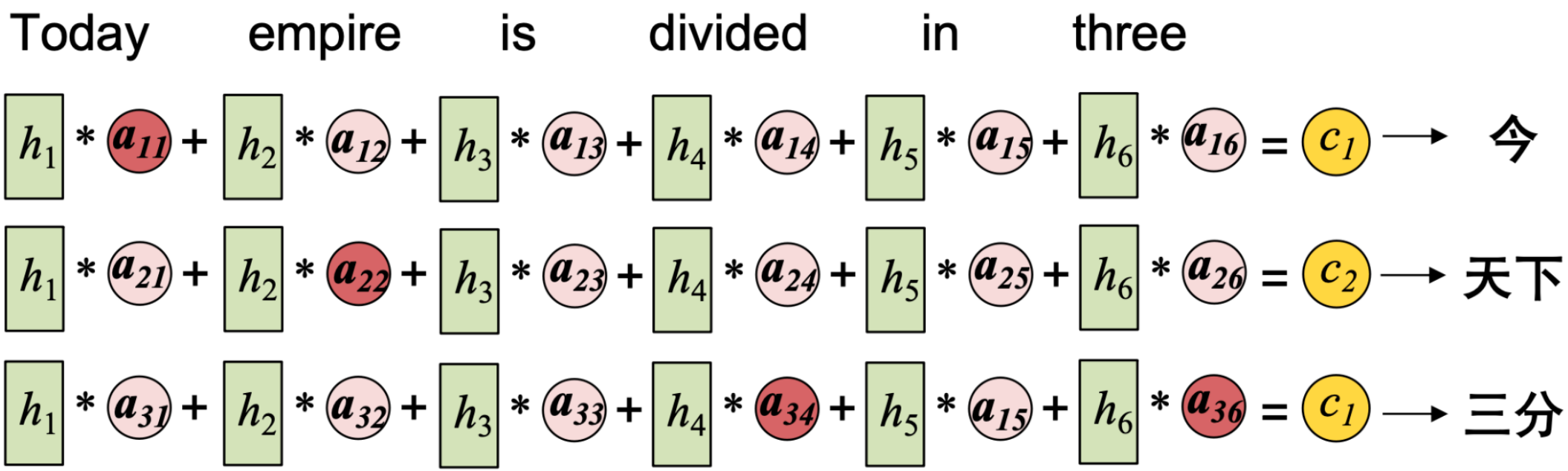


$$C_i = \sum_{j=0}^{n} a_{ij} h_j$$



- C_i是编码器中隐状态的加权和
- a_{ij}是输入x_j对输出$\hat{y}_i$的重要程度

C_i和α_{ij}的关系示例



α_{ij}如何计算?

α_{ij}如何计算

α_{ij}特点分析

1. 具有归一化特征（所有输入对$\hat{y}_i$的权重之和为1，$\sum_{j=0}^{n} a_{ij} = 1$）
2. 与编码器隐藏层h_j有关 + 与解码器隐藏层h'_{i-1}有关
3. 属于模型参数，在训练中产生

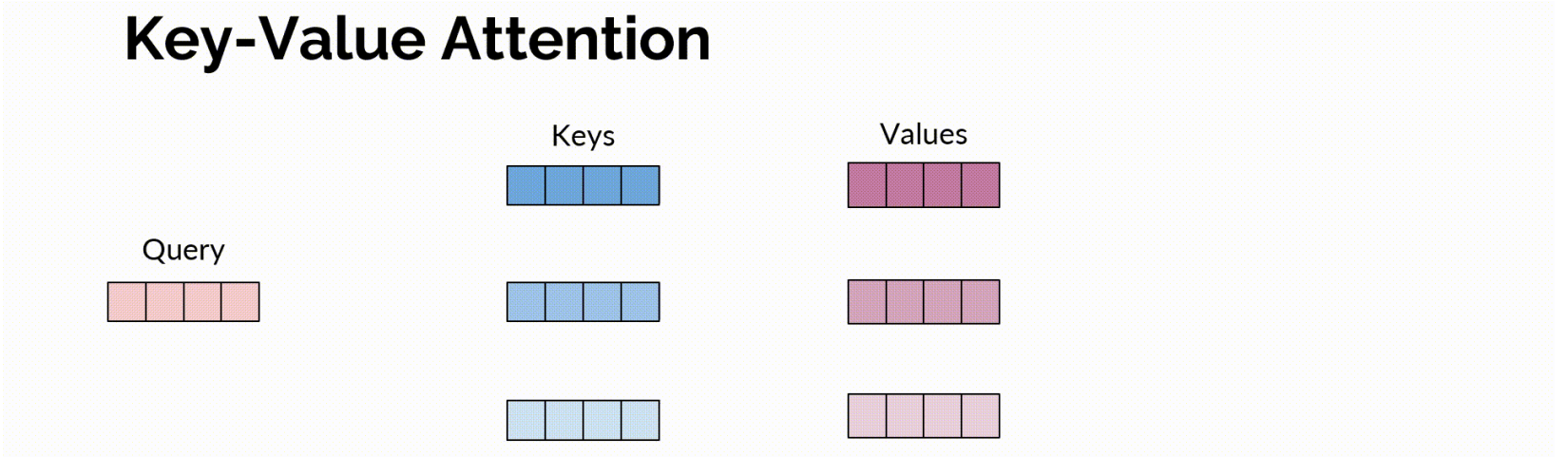
$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=0}^{n} exp(e_{ik})}$$ ← Softmax归一化

$$e_{ij} = \varphi(h'_{i-1}, h_j)$$ ← 对齐模型，可训练

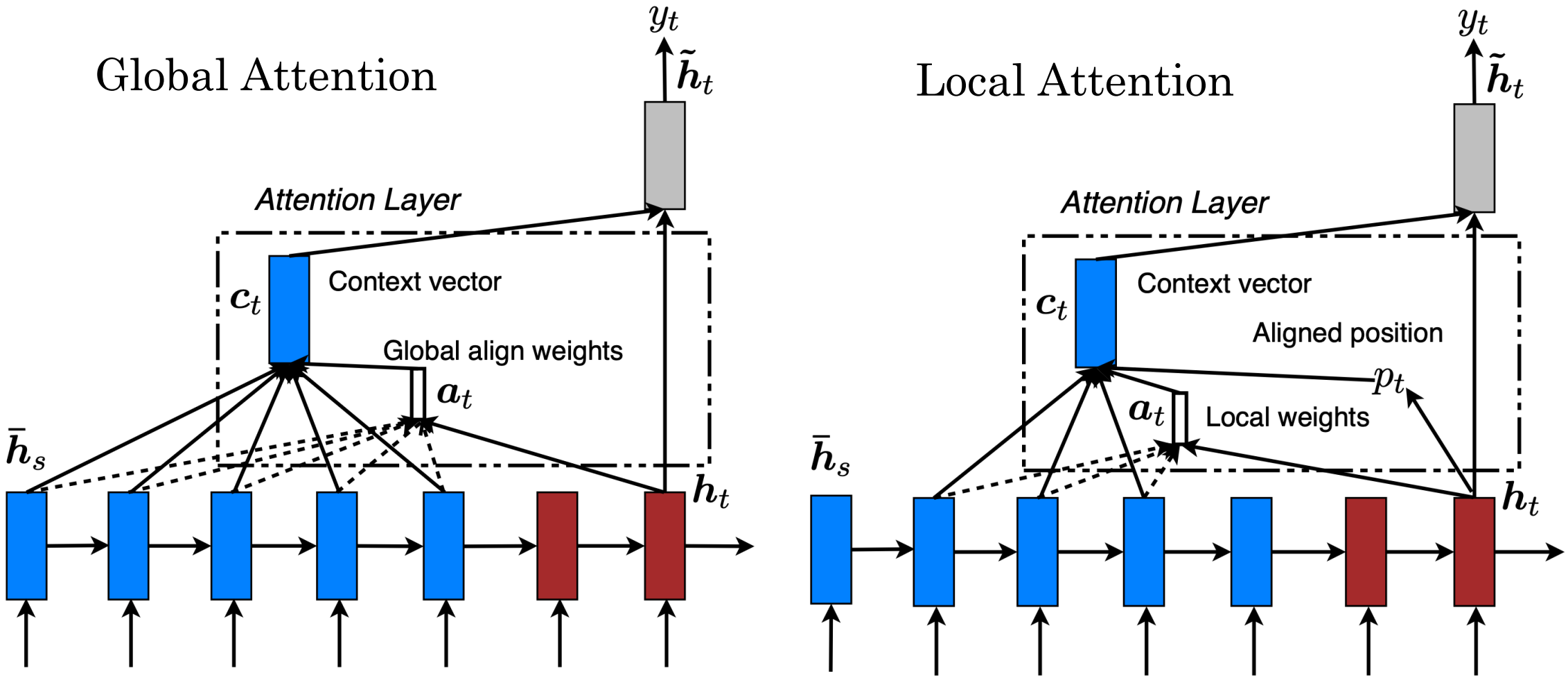
基于注意力匹配的相关要素及记法

对于一次<x_i, $\hat{y}_i$>匹配过程，其涉及到的注意力相关要素包括：

1. 输出词 $\hat{y}_i$ 称之为查询（Query），如“先帝”
2. 对于h_i，其匹配度 α_{ij}为寻找x_i的键（Key） <Q,K,V>三元组
3. x_i的语义表示为为寻找x_i的指（value）



Global Attention *v.s.* Local Attention



Luong, Minh-Thang, Hieu Pham,etc. "Effective approaches to attention-based neural machine translation." *EMNLP* 2015.

基于注意力的机器翻译应用

[PDF] Neural machine translation by jointly learning to align and translate

D Bahdanau, K Cho, Y Bengio

arXiv preprint arXiv:1409.0473, 2014 • peerj.com

Abstract

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of

展开

☆ 保存 引用 被引用次数：42458 相关文章 所有 28 个版本 »

Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder–decoder for statistical machine translation." *EMNLP*. 2014.

基于注意力的机器翻译应用

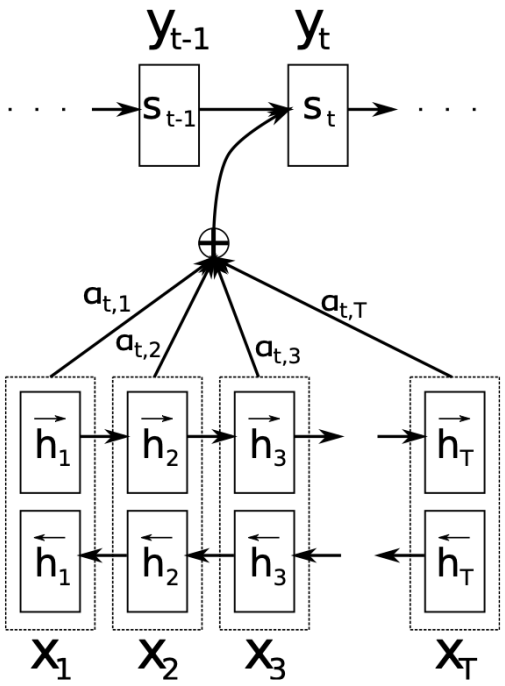
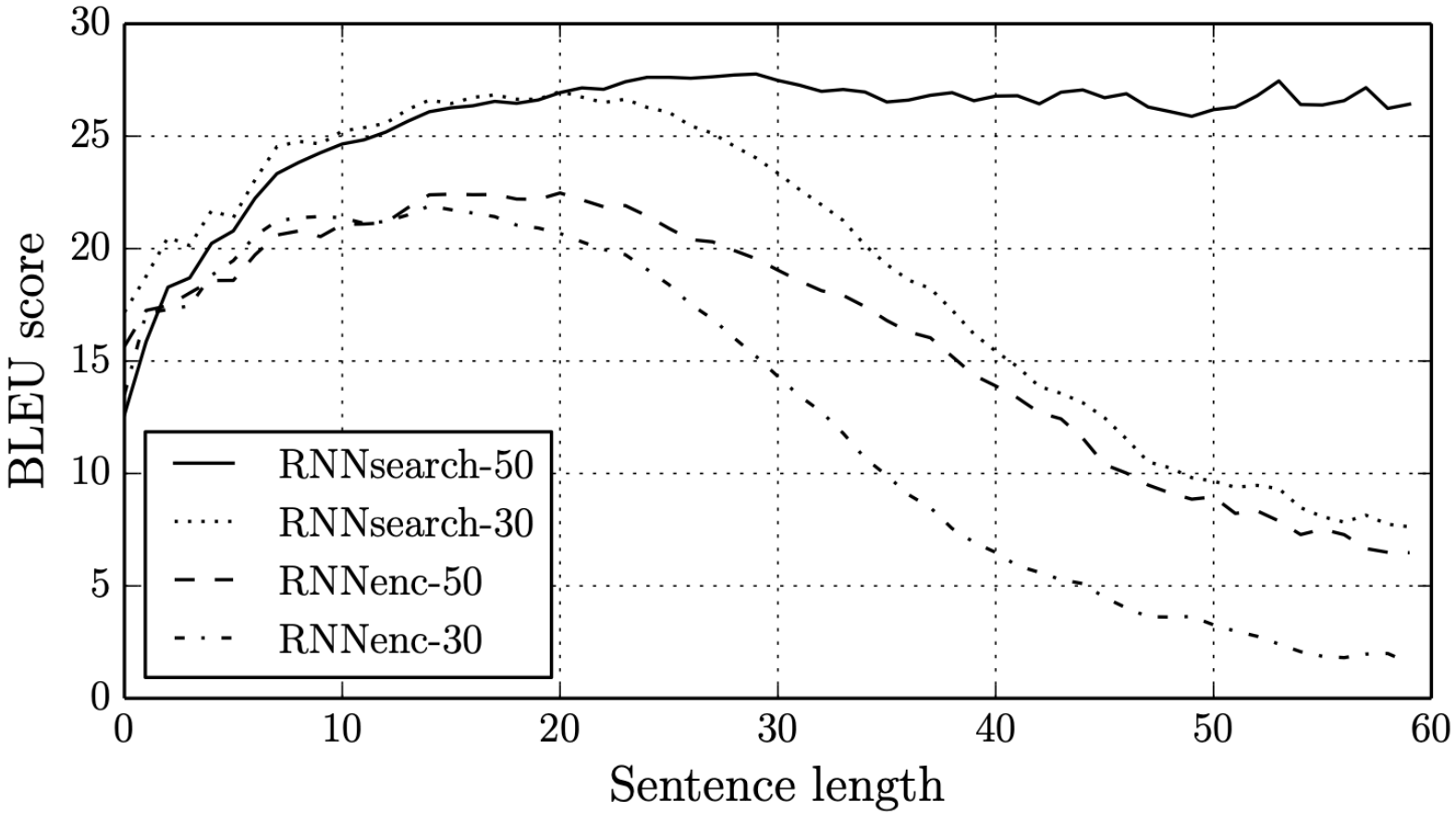
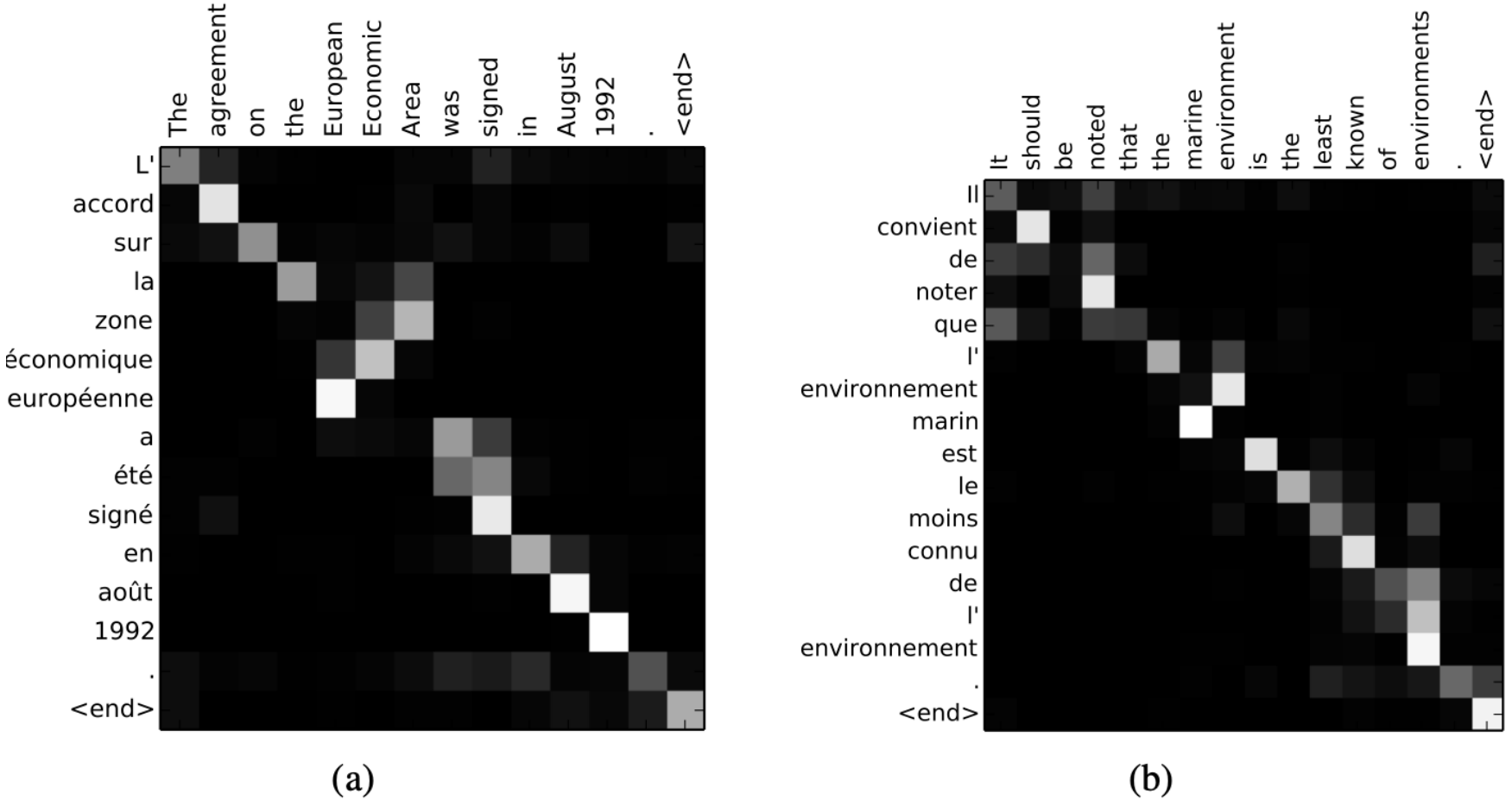


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.



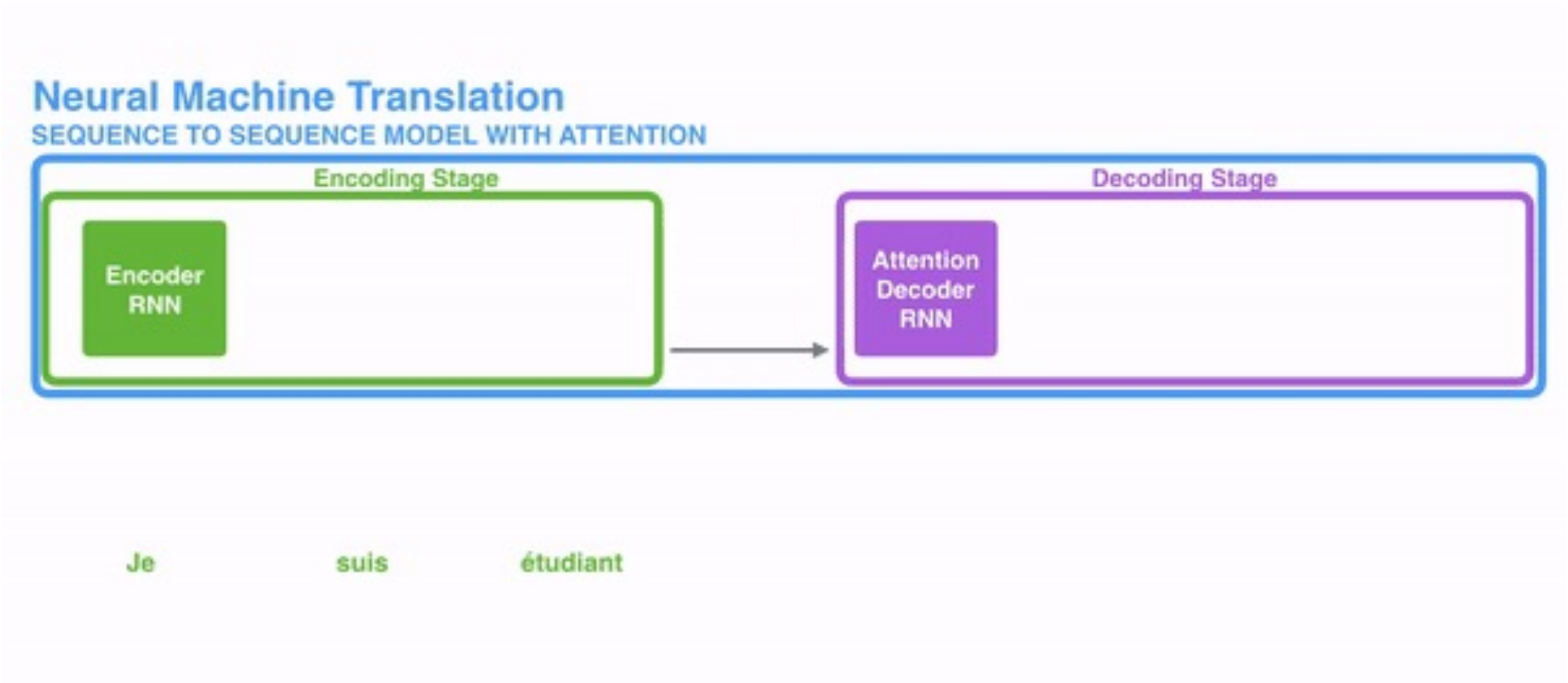
基于注意力的机器翻译应用

展示注意力机制的双语对齐（英语-法语）效果，像素灰度对齐权重

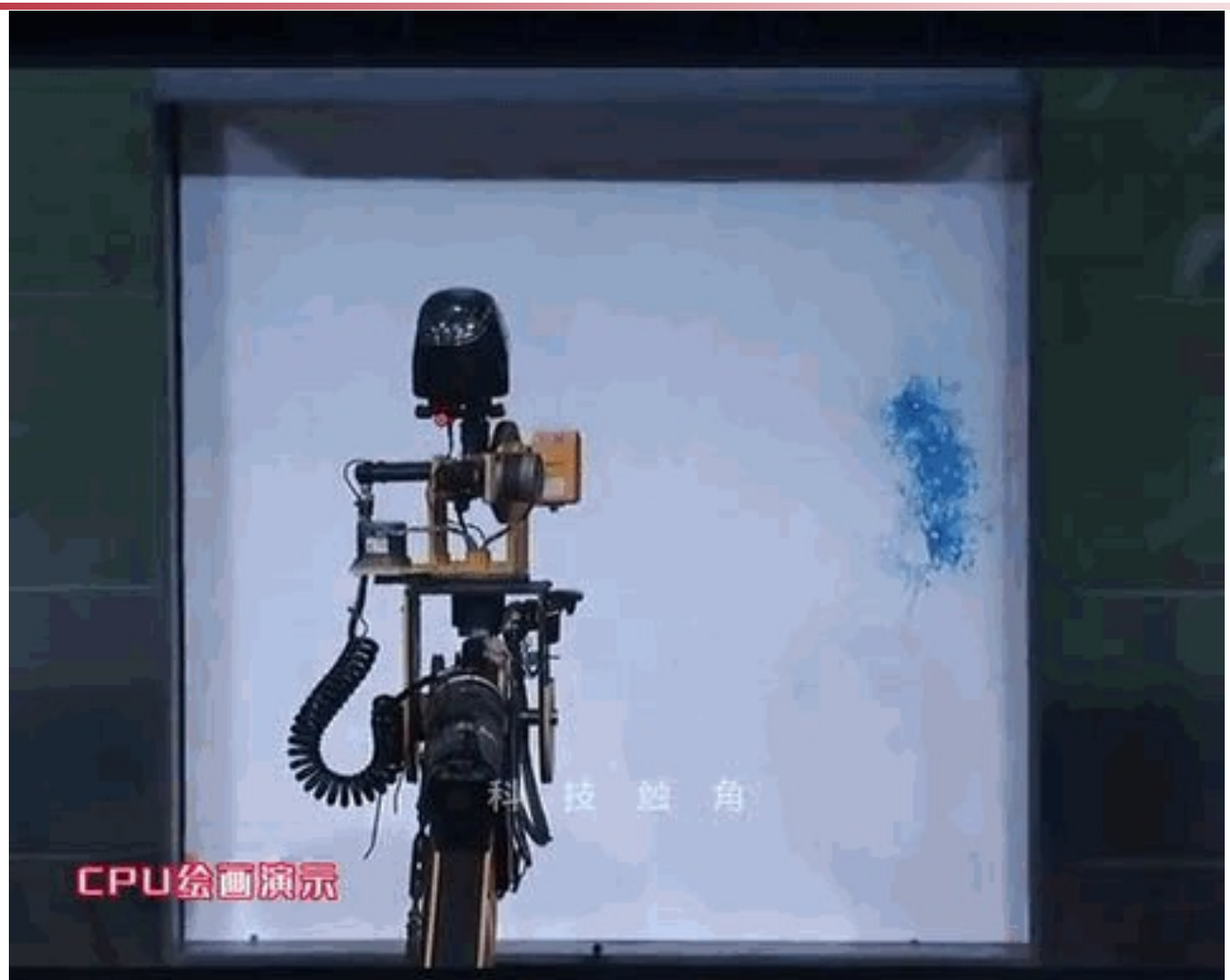


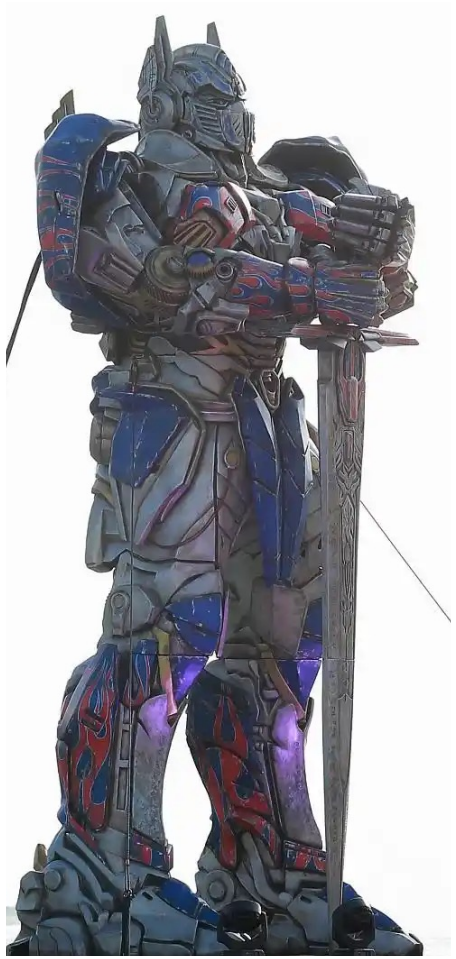
基于注意力的RNN不足之处

RNN只能串行处理，逐字理解/生成，处理效率低



能不能一次性得到所有的Ci？





1. 带注意力的Encoder-Decoder
2. Transformers
3.
4.

参考文献

- Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder–decoder for statistical machine translation." *EMNLP*. 2014.
- Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

题外：


THANKS