3.5 机器翻译与Transformer模型

林洲汉
上海交大NLP课题组

目录

目录

机器翻译的产生与发展

- 概念：机器翻译 (machine translation, MT) 是 用计算机把一种语言(源语言, source language) 翻译成另一种语言(目标语言, target language) 的一门技术。

信达雅：机器翻译的困难

- 信、达、雅，由中国近代启蒙思想家、翻译家严复提出的翻译理论，又称“三难原则”。出自严复译著《天演论》中的“译例言”，其讲到：“译事三难：信、达、雅。求其信已大难矣，顾信矣不达，虽译犹不译也，则达尚焉。”
  - “信”（faithfulness）指意义不悖原文，即是译文要准确，不偏离，不遗漏，也不要随意增减意思；
  - “达”（expressiveness）指不拘泥于原文形式，译文通顺明白；
  - “雅”（elegance）则指译文时选用的词语要得体，追求文章本身的古雅，简明优雅。

信达雅：机器翻译的困难

信达雅：机器翻译的困难

信达雅：机器翻译的困难

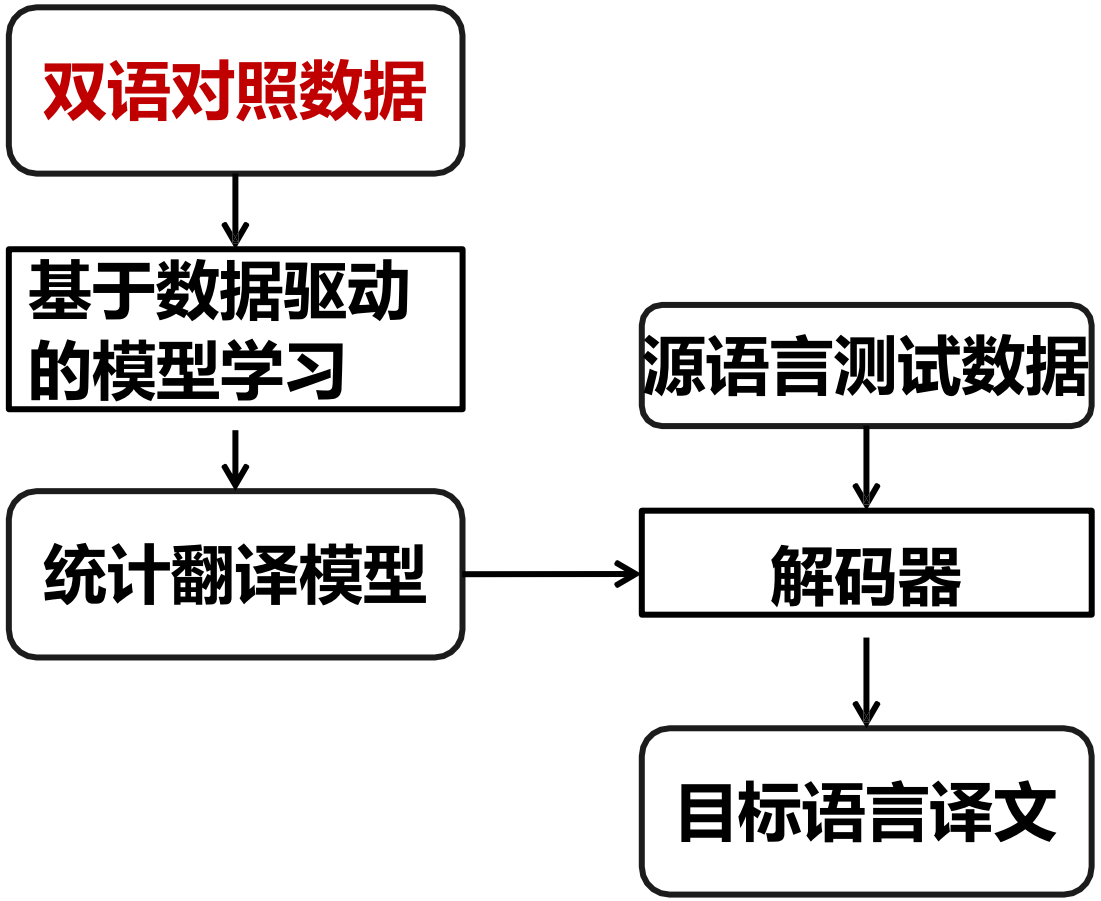
信达雅：机器翻译的困难

- 黛玉自在枕上感念宝钗……又听见窗外竹梢焦叶 之上，雨声淅沥，消寒透幕，不觉又滴下泪来。(《红楼梦》第45回)
- As she lay there alone, Dai-Yu’ s thoughts turned to Bao-chai …Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she bad begun to cry. 文学翻译家David Hawkes
- 摘自冯志伟著《机器翻译研究》，2004

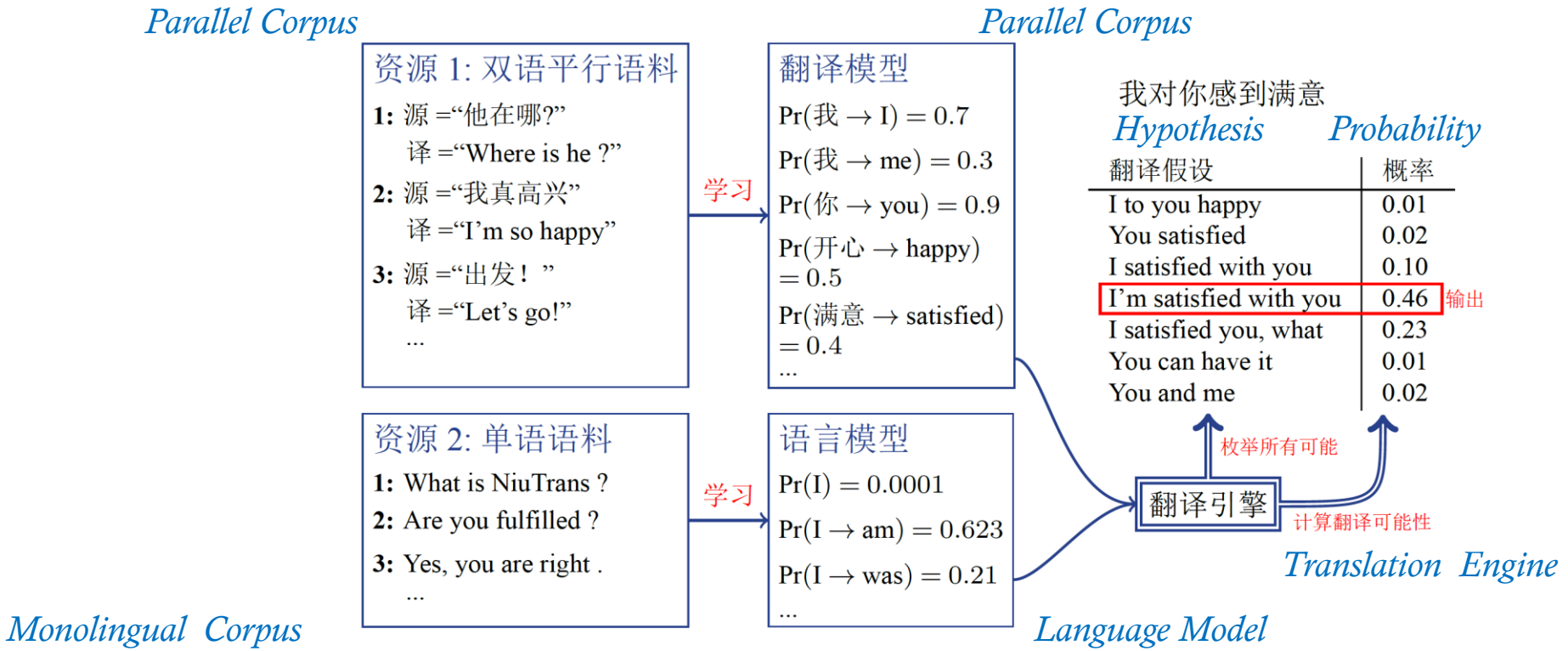
目录

统计机器翻译

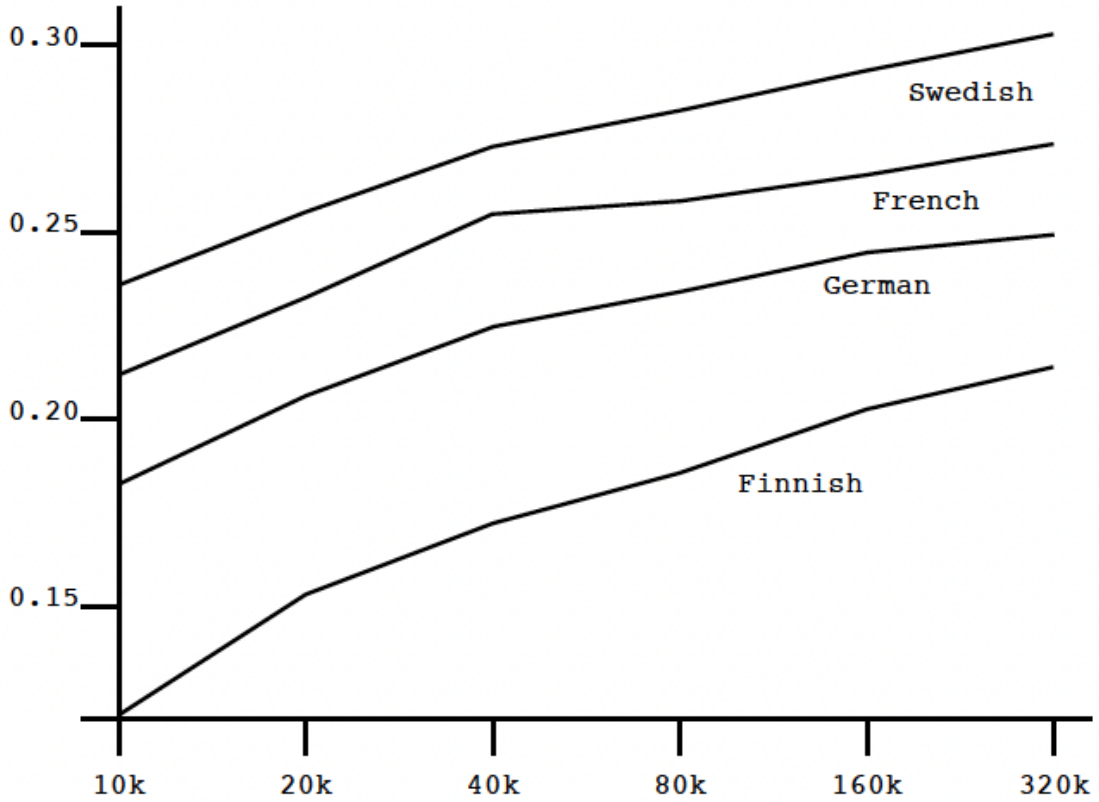
- **总体思想：**



- **Statistical Machine Translation (SMT)**
  - Parallel corpus: sentence-level alignment.
  - Monolingual corpus: *n*-grams probability.
  - To learn the translation rules statistically.



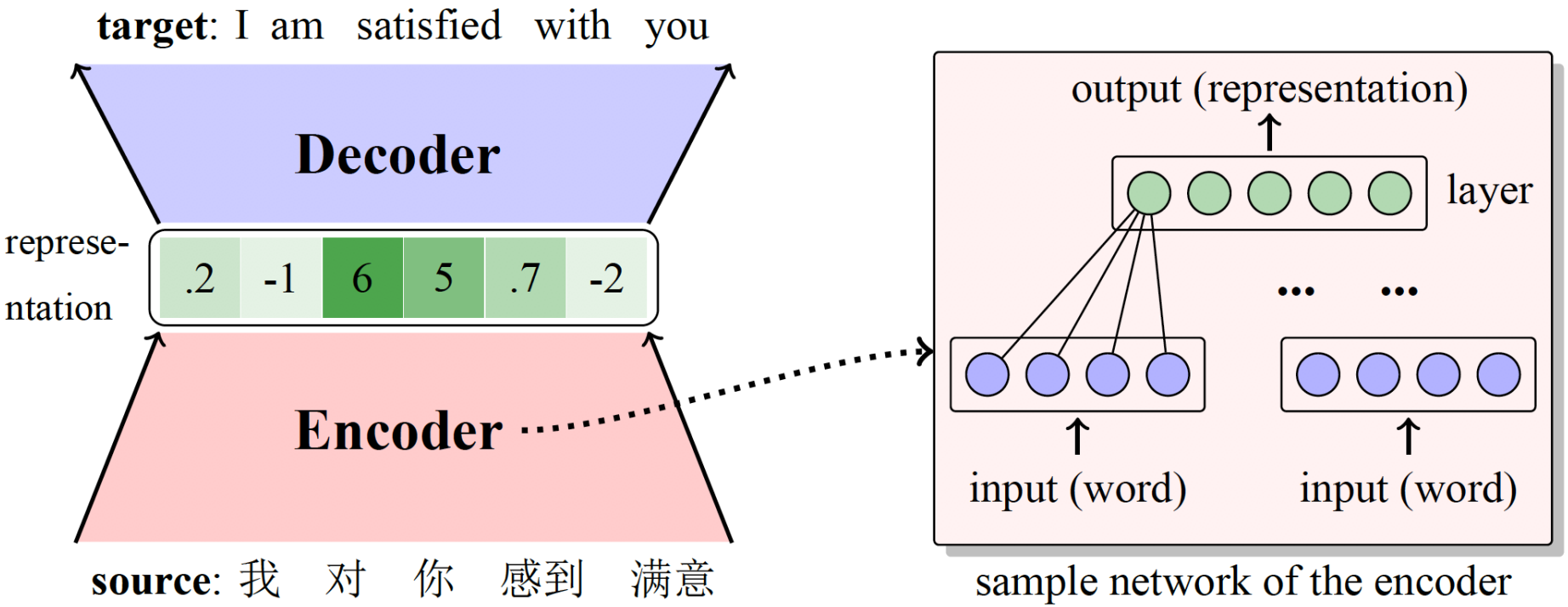
More data, better translations!



[from Koehn, 2003: Europarl]

目录

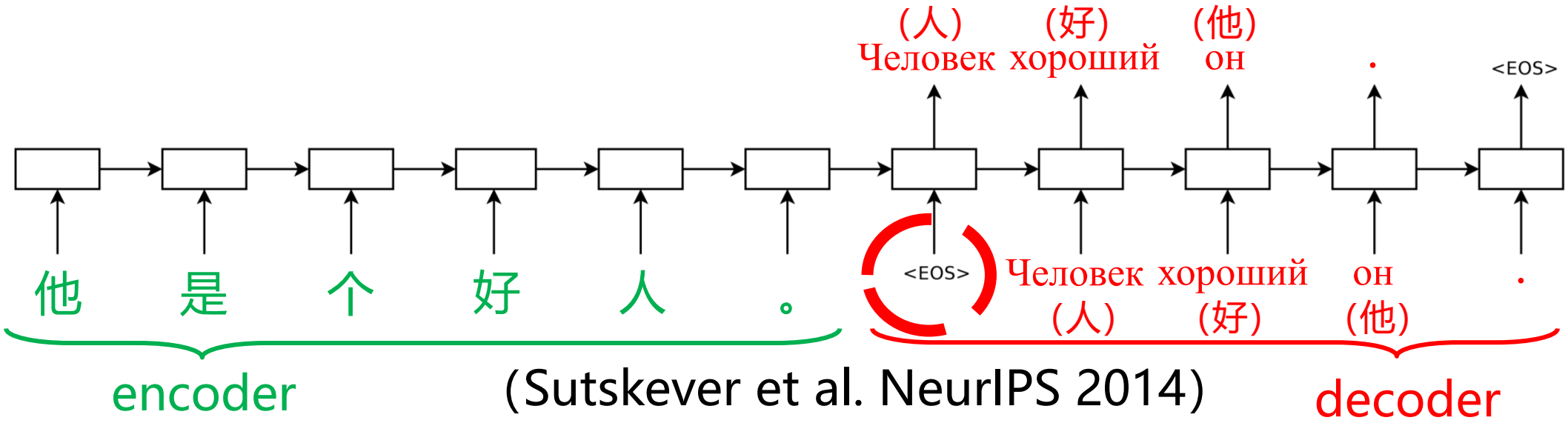
- **Neural Machine Translation (NMT):**
  - Parallel corpus as sequence-to-sequence input.
  - Rules are not necessary any more.



目录

使用LSTM/GRU进行机器翻译

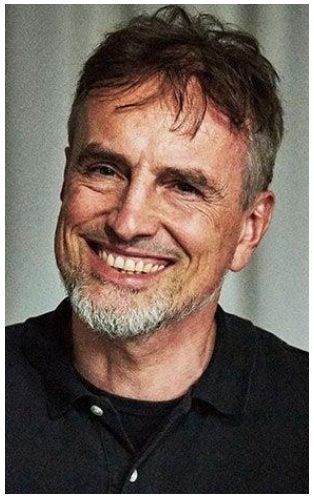
- 在神经机器翻译（NMT）发展的早期（2014年），人们直接使用LSTM来做NMT。
  - 事实上，以机器翻译为代表的一大类任务都可以表示成这种“先读取一段文本，再根据这段文本生成新的文本”的任务。这一类模型被称为seq2seq模型。
  - 通常包含一个encoder来处理源文本，一个decoder来生成目标文本。因而也被称为encoder-decoder架构。
- 当然，前面在序列标注中我们讲的双向模型、深层模型等变种，都可以在这里适用。（可以怎么融合到这个场景中去？）



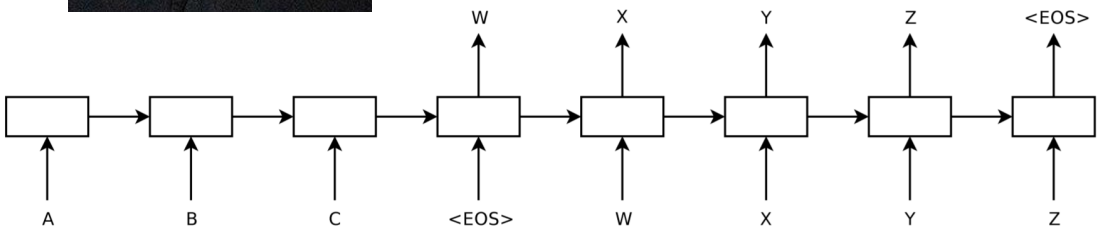
(Sutskever et al. NeurIPS 2014)

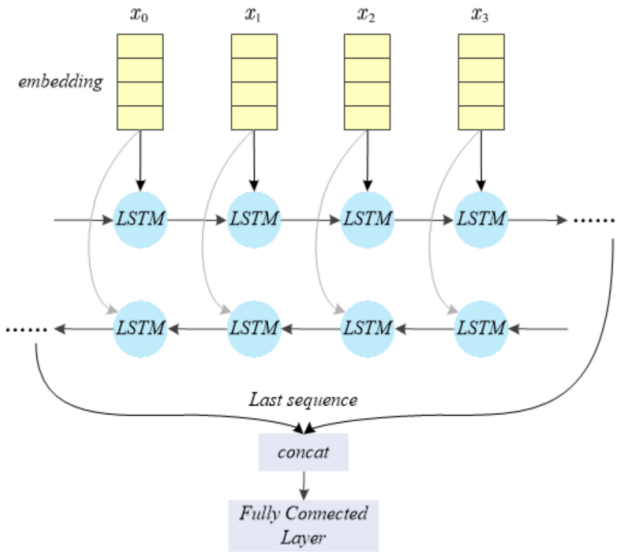
目录

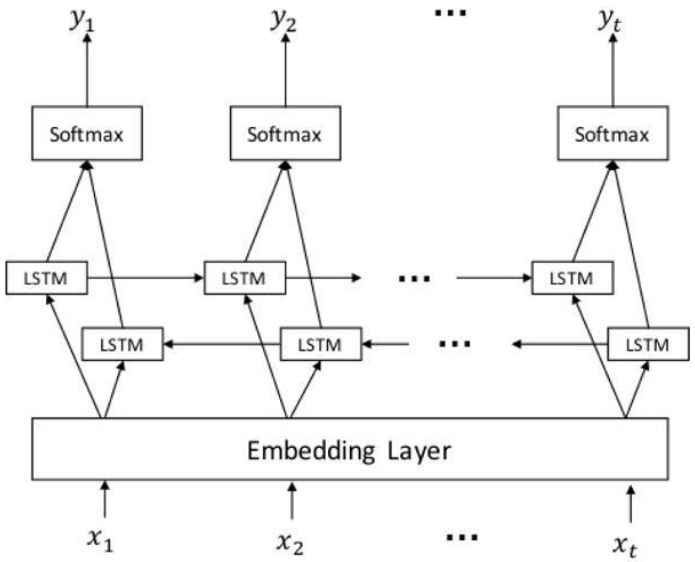
2014年是一个LSTM制霸NLP的年代



Jürgen Schmidhuber in 2014:
“Google itself might end up as one huge LSTM.”







机器翻译（Sutskever et al. 2014）　　文本分类/情感分析　　序列标注　　……

- **看起来RNN对于序列信息处理已经很合理了，就像卷积神经网络对图片一样**
- **还有什么可以进步的呢?**

LSTM/GRU等循环神经网络（RNN）中存在的问题

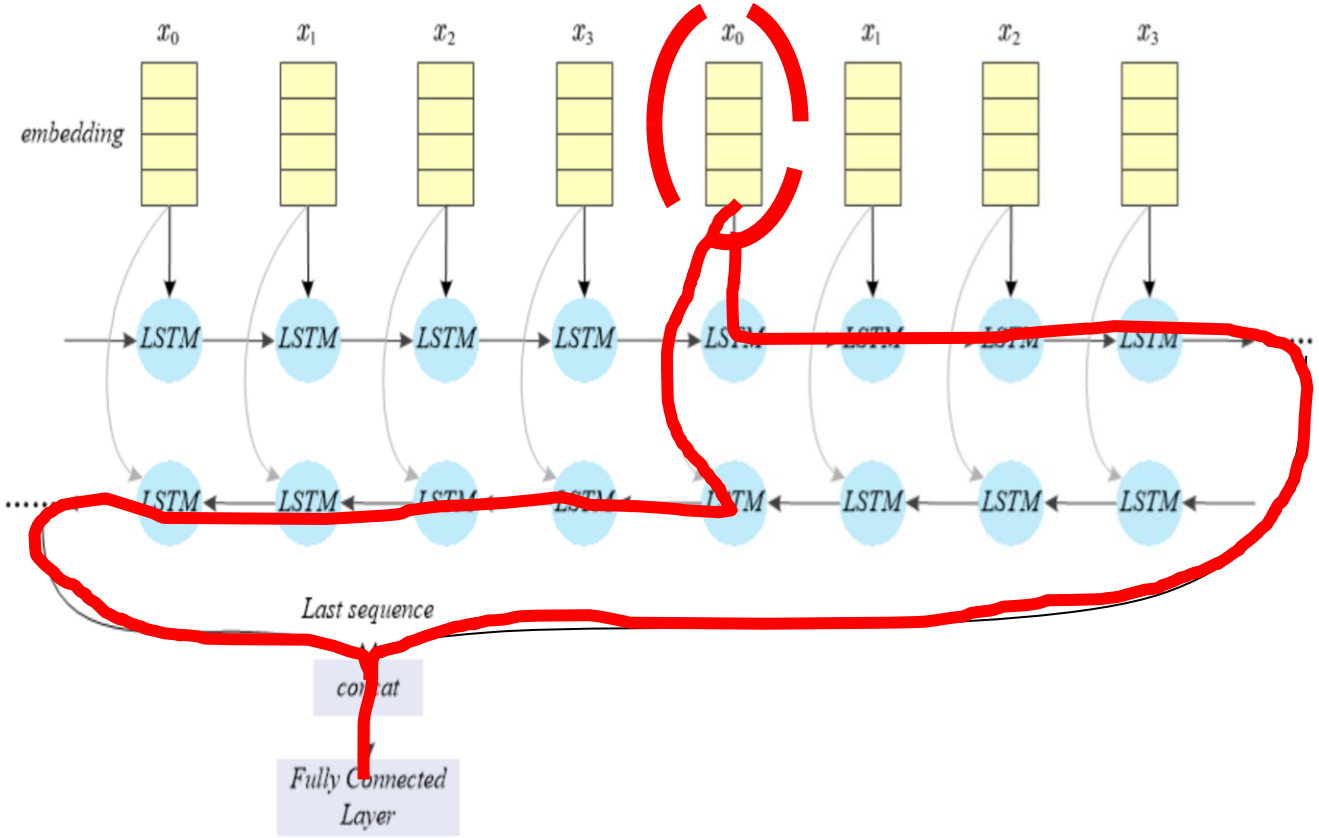
文本分类与情感分析：

假如序列中间的词是对分类结果产生影响的关键词

- 无论从正向还是反向，loss中的梯度传导到该词都需要漫长的路程。
- 由于RNN的梯度爆炸和梯度消失（gradient vanishing/exploding）问题，这样的长程关系（long-term dependency）很难被有效优化。

有工作对LSTM在所有step上的h统一使用一次Max Pooling来避免该问题

- 神奇的是，效果还真不错。。。
- 还有没有更好的解决办法?



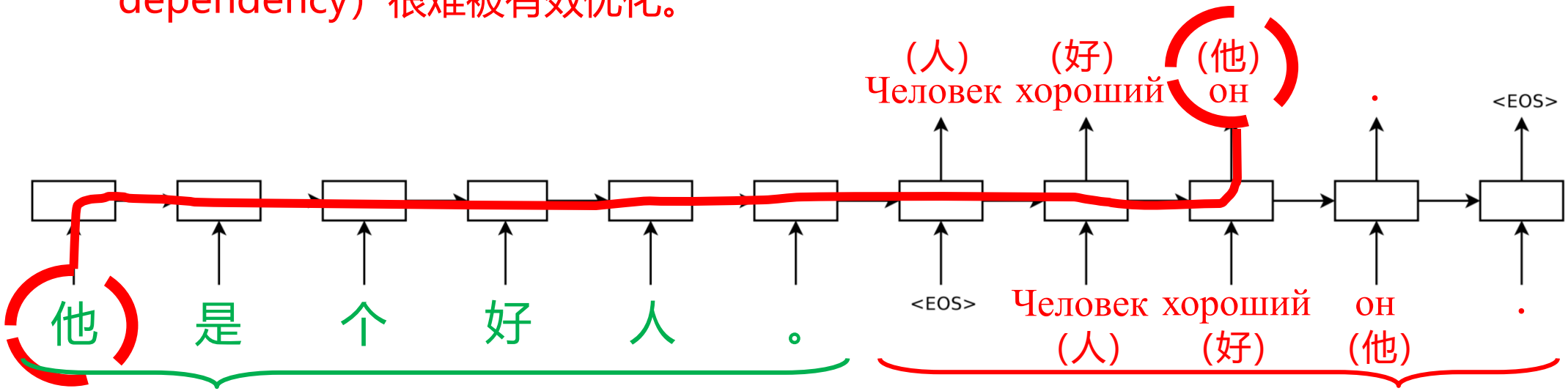
文本分类/情感分析

LSTM/GRU等循环神经网络（RNN）中存在的问题

考虑机器翻译中的一个极端例子：

“他”的语义从第一步就被存入RNN中，此后一直保存在LSTM的h和c中，直到第八步才被读取。

- 需要经过8个循环步（recurrent step）！
- 中间不断地有新的词被存入和读取，这些新词的语义与“他”关系不大，但整个过程中“他”的语义需要被很好地保存在h和c中，占用h和c的capacity。
- 同样的，来自“он”的梯度，也需要8步才能到达“他”。由于RNN的梯度爆炸和梯度消失（gradient vanishing/exploding）问题，这样的长程关系（long-term dependency）很难被有效优化。



机器翻译（Sutskever et al. NeurIPS 2014）

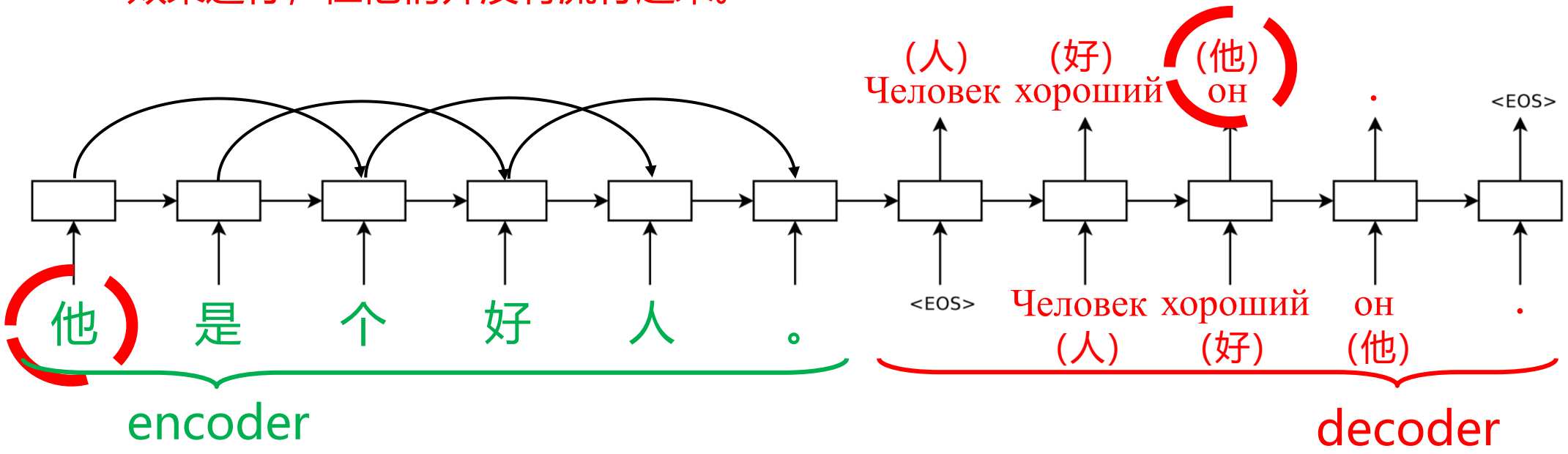
LSTM/GRU等循环神经网络（RNN）中存在的问题

考虑机器翻译中的一个极端例子：

“他”的语义从第一步就被存入RNN中，此后一直保存在LSTM的h和c中，直到第八步才被读取。

很多工作试图在RNN中添加各种各样的skip connection，来减少gradient传播所需经过的步数。

- 比如clockwork RNN (Koutník 2014)，
  Hierarchical Multiscale RNN (Junyoung 2014)
- 效果还行，但他们并没有流行起来。



注意力机制（Attention Mechanism）

最终赢得这方面尝试的成功方案，是注意力机制（Attention Mechanism）

事实上，人在翻译长句子的某一个词的时候，更关心这个词对应的原文中相关的几个词，而并不关心别的上下文。

也就是说，只要能正确找到当前词对应的原文（learning to align），翻译就成功了一大半。

体会一下：

The common belief of some linguists that each language is a perfect vehicle for the thoughts of the nation speaking it is in some ways the exact counterpart of the conviction of the Manchester school of economics that supply and demand will regulate everything for the best.

怎么在神经网络中体现？

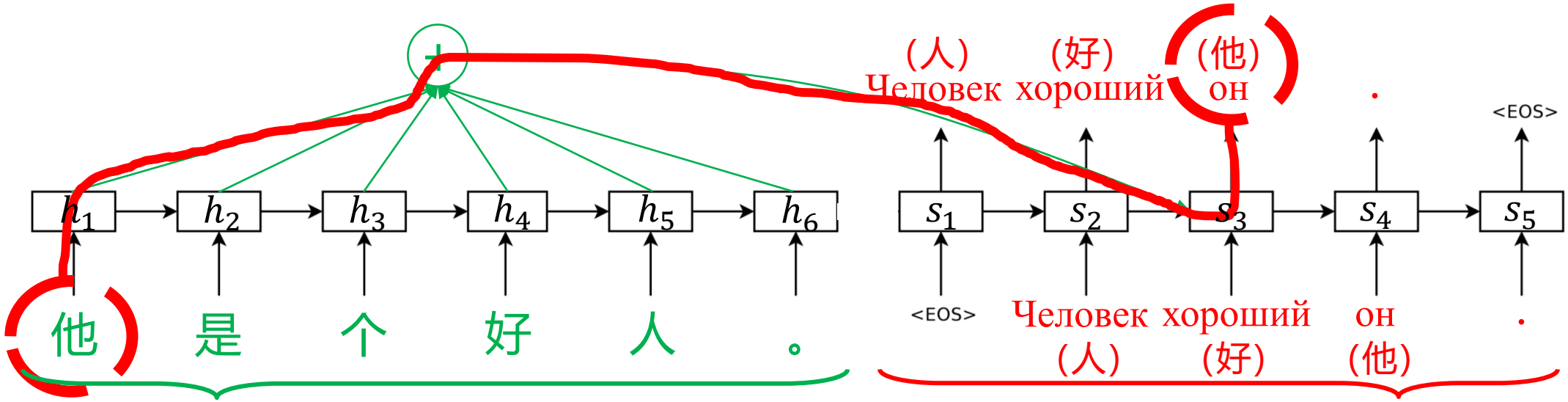
注意力机制（Attention Mechanism）

考虑机器翻译中的一个极端例子：

~~“他”的语义从第一步就被存入RNN中，此后一直保存在LSTM的h和c中，直到第八步才被读取。~~

Decoder在预测“он”时，不光使用前一步的RNN state（s_{i-1}），
而且使用encoder的所有RNN state（即$h_1 \sim h_N$）的线性组合。

每个h_i处所分到的线性组合权重由s_{i-1}与h_i共同决定。



机器翻译（Dzmitry et al. ICLR 2014）

注意力机制（Attention Mechanism）

普通RNN

$$s_i = f(s_{i-1}, y_{i-1})$$

f是任意非线性函数，
代表RNN中一步的运算。

带有attention的RNN

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{N} \alpha_{ij} h_j$$

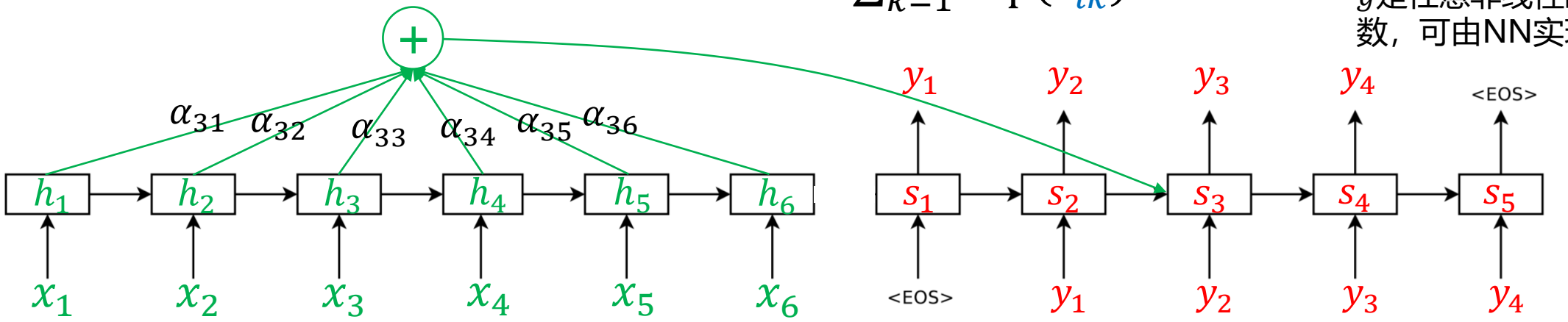
$$\alpha_{ij} \overset{\text{softmax}}{=} \frac{\exp(e_{ij})}{\sum_{k=1}^{N} \exp(e_{ik})}$$

$$e_{ij} = g(s_{i-1}, h_i)$$

g是任意非线性函数，可由NN实现。

Dzmitry Bahdanau

Yoshua Bengio



encoder

机器翻译（Dzmitry et al. ICLR 2014）

decoder

注意力机制（Attention Mechanism）

更多细节：

- Dzmitry的原文中encoder使用的是BiGRU。
- c_i在decoder的GRU内部，像s_{i-1}一样，乘上一个矩阵后与其他两个元素相加。
- G的实现是一个两层的全连接NN， tanh作为隐层激活函数。（还可以用什么？）

带有attention的RNN

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{N} \alpha_{ij} h_j$$

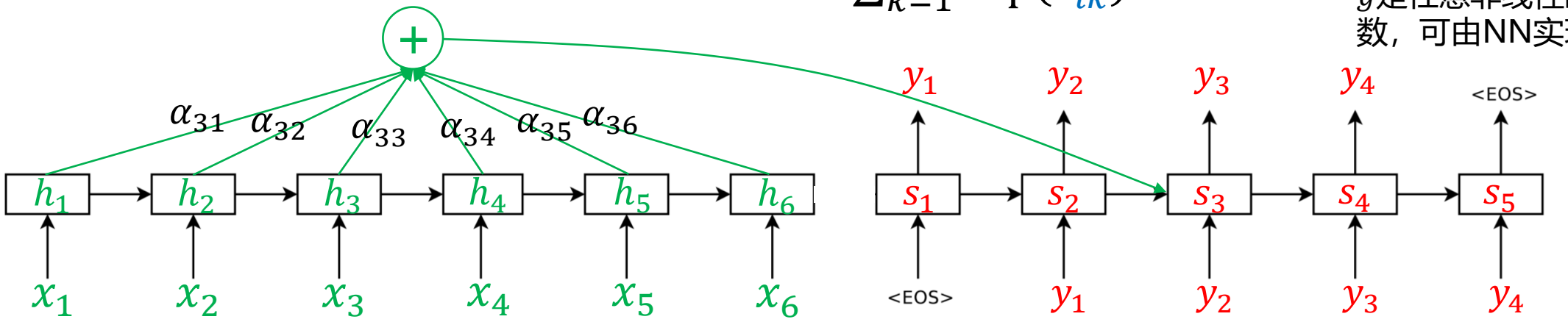
$$\alpha_{ij} \overset{\text{softmax}}{=} \frac{\exp(e_{ij})}{\sum_{k=1}^{N} \exp(e_{ik})}$$

$$e_{ij} = g(s_{i-1}, h_i)$$

g是任意非线性函数，可由NN实现。

Dzmitry Bahdanau

Yoshua Bengio



encoder　　机器翻译（Dzmitry et al. ICLR 2014）　　decoder

目录

自注意力机制（Self-attention）

- **Attention用于机器翻译取得了巨大的成功。**
- **在没有decoder的时候，比如做文本分类、情感分析的时候，如何使用attention?**

自注意力机制（Self-attention）

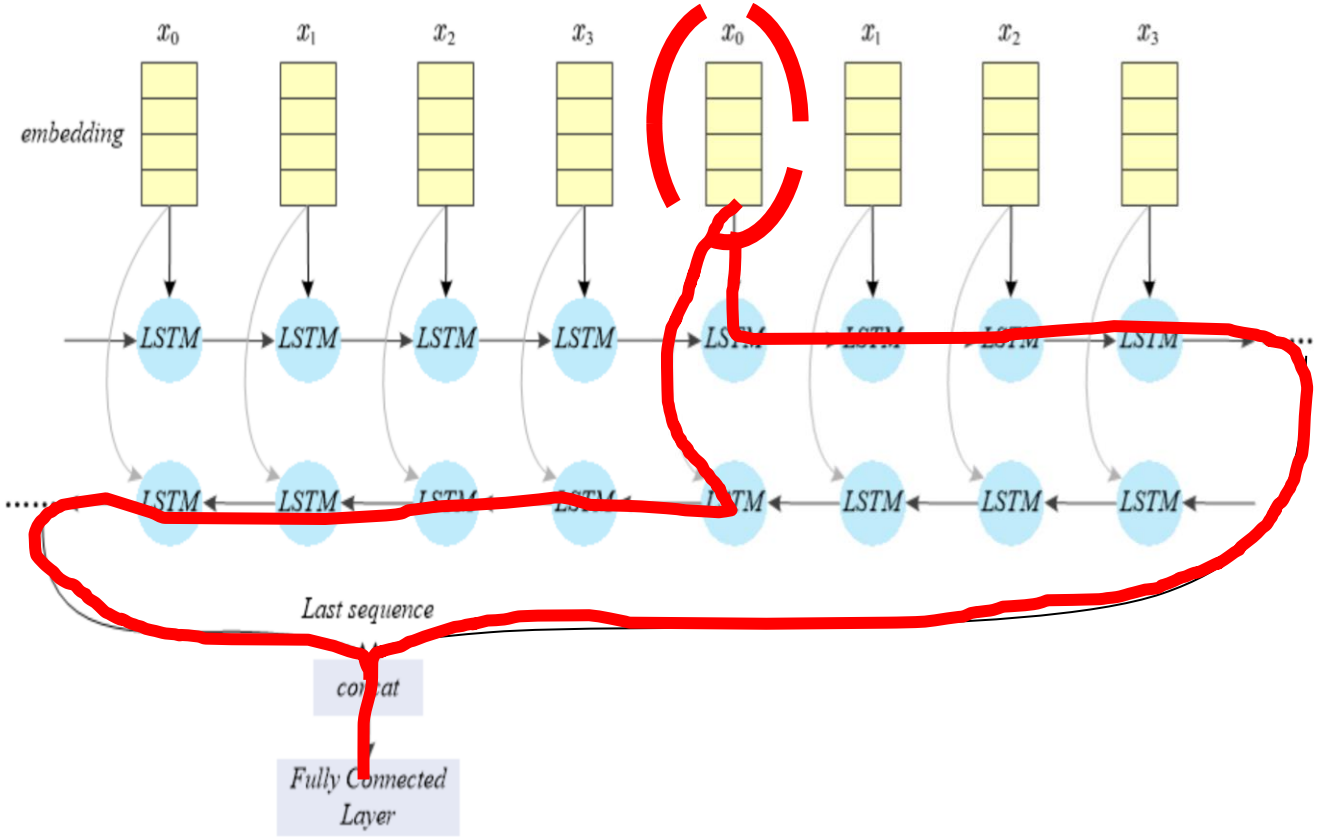
文本分类与情感分析：

假如序列中间的词是对分类结果产生影响的关键词

- 无论从正向还是反向，loss中的梯度传导到该词都需要漫长的路程。
- 由于RNN的梯度爆炸和梯度消失（gradient vanishing/exploding）问题，这样的长程关系（long-term dependency）很难被有效优化。

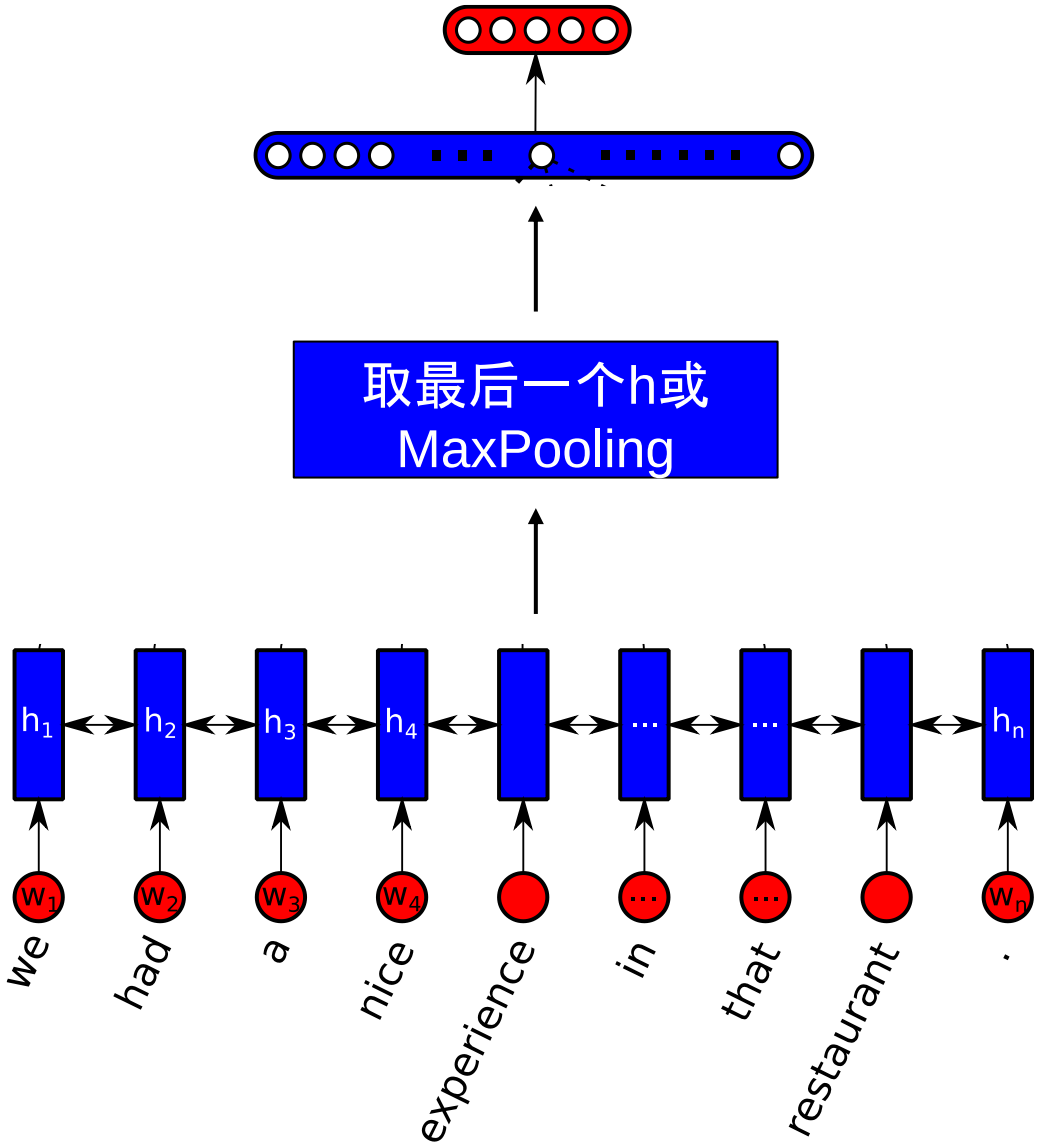
有工作对LSTM在所有step上的h统一使用一次Max Pooling来避免该问题

- 神奇的是，效果还真不错。。。
- 还有没有更好的解决办法?



文本分类/情感分析

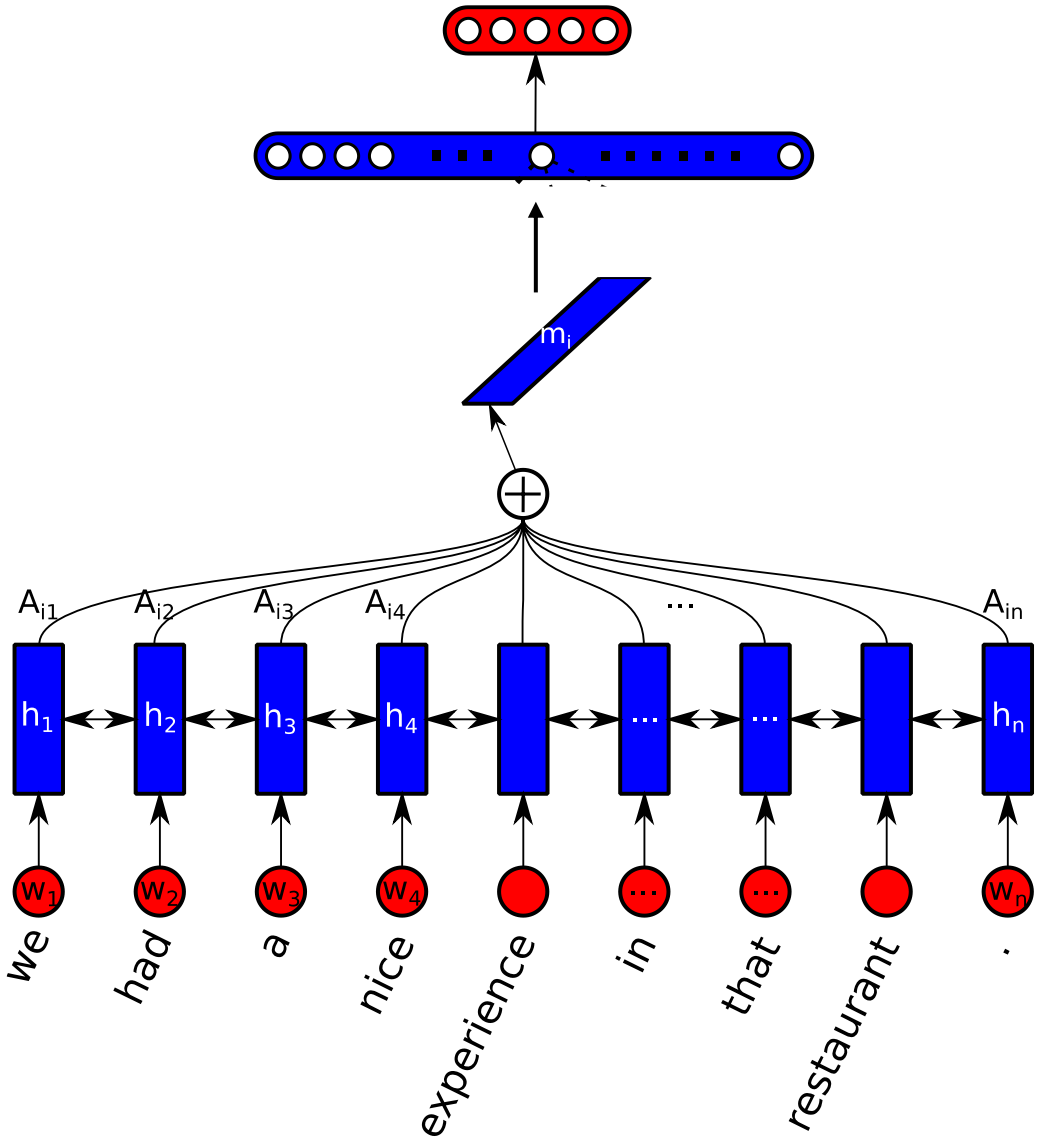
自注意力机制（Self-attention）



以情感分析为例，将一个句子分为5个不同的类别。

左边是attention出现之前常用的情感分析模型。

自注意力机制（Self-attention）



自注意力（Lin et al. ICLR 2016）

“取最后一个h或MaxPooling” → “每个h_i的加权求和”

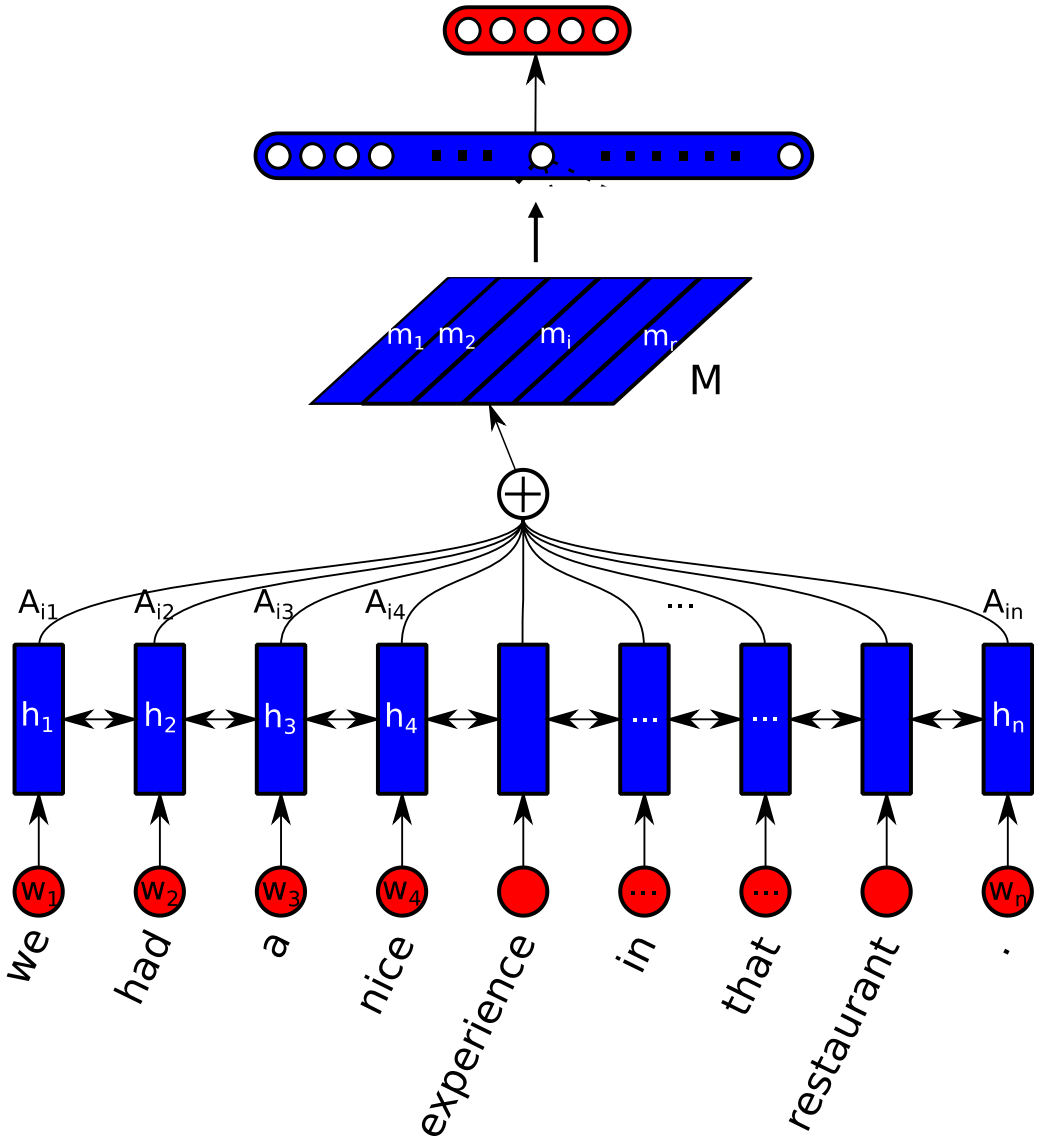
加权的权值A_j由一个两层的全连接NN输出，再通过softmax得到。

$$A_j = \frac{\exp(e_j)}{\sum_{k=1}^{N}\exp(e_k)}$$

$$e_j = g(h_1, h_2, \cdots, h_N)$$

$$= W_2\tanh(W_1[h_1, h_2, \cdots, h_N])$$

自注意力机制（Self-attention）



自注意力（Lin et al. ICLR 2016）

“取最后一个h或MaxPooling” → “每个h_i的加权求和”

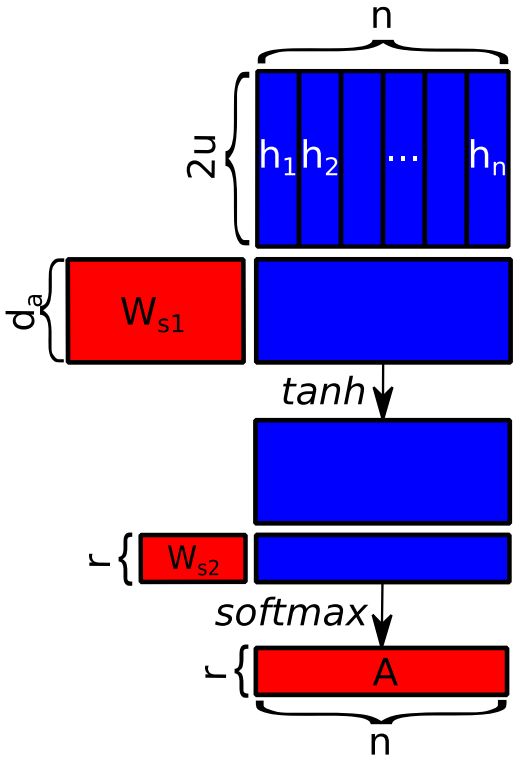
加权的权值A_j由一个两层的全连接NN输出，再通过softmax得到。

多头自注意力：不止计算一个加权平均，而是同时计算多个。

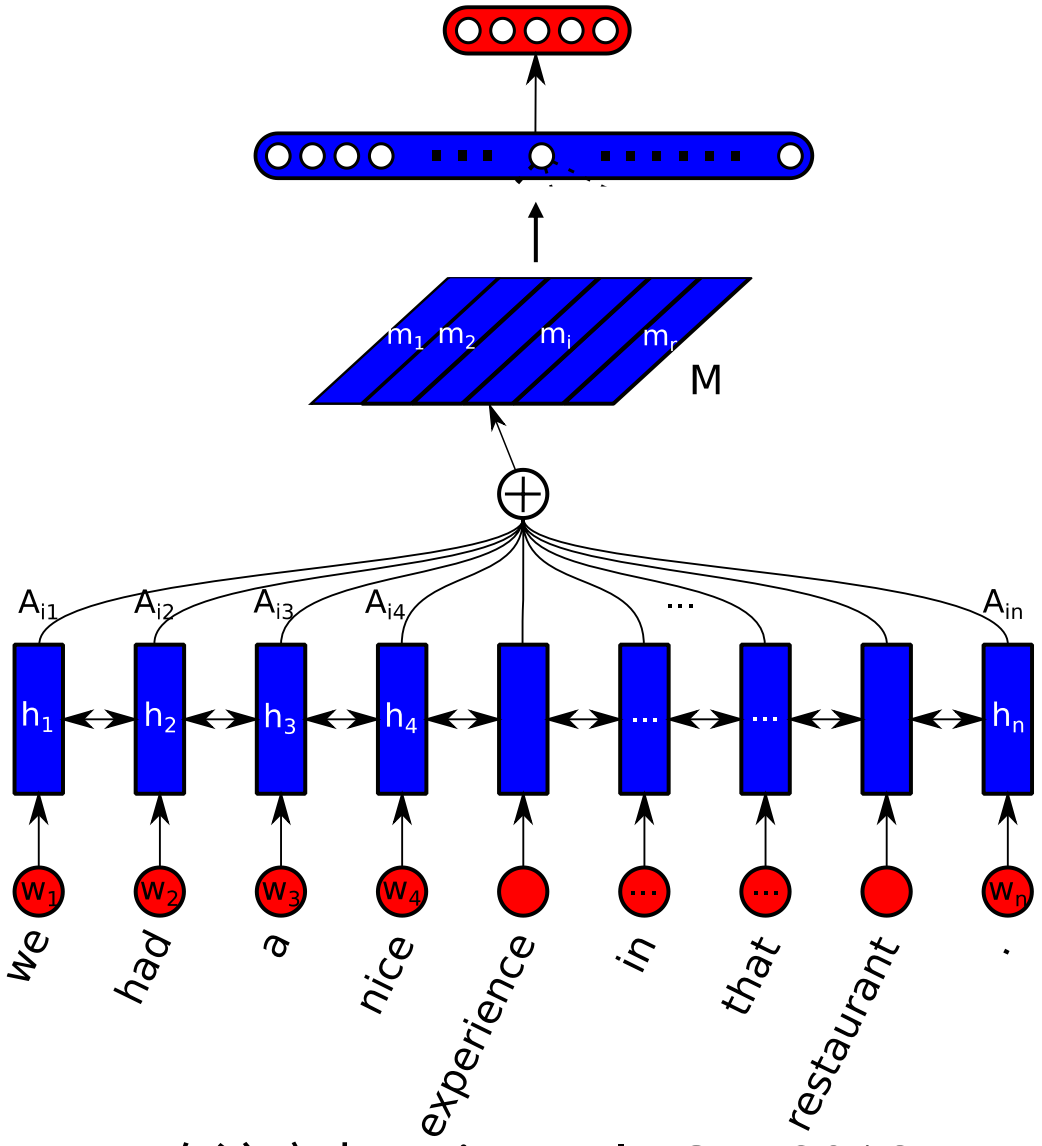
$$A_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{N}\exp(e_{ik})}$$

$$e_{ij} = g(h_1, h_2, \cdots, h_N)$$

g的具体形式 →



自注意力机制（Self-attention）



自注意力（Lin et al. ICLR 2016）

“取最后一个h或MaxPooling” → “每个h_i的加权求和”

加权的权值A_j由一个两层的全连接NN输出，再通过softmax得到。

多头自注意力：不止计算一个加权平均，而是同时计算多个。

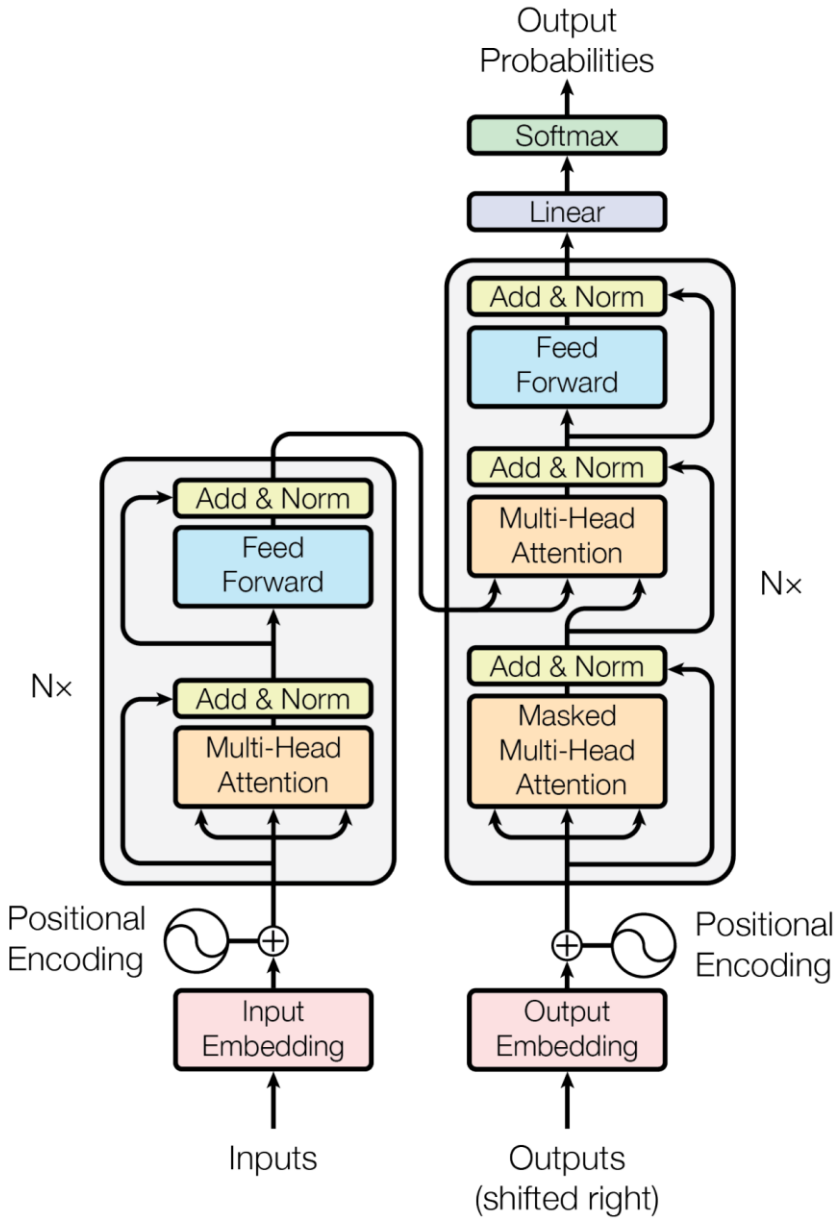
更进一步的，如果我们作以下三方面改变：

1. 干脆撤掉底下的RNN，
2. 对每一个h_i计算一组多头自注意力机制所得到的向量集，
3. 利用额外的positional embedding弥补撤掉RNN所引起的位置信息缺失

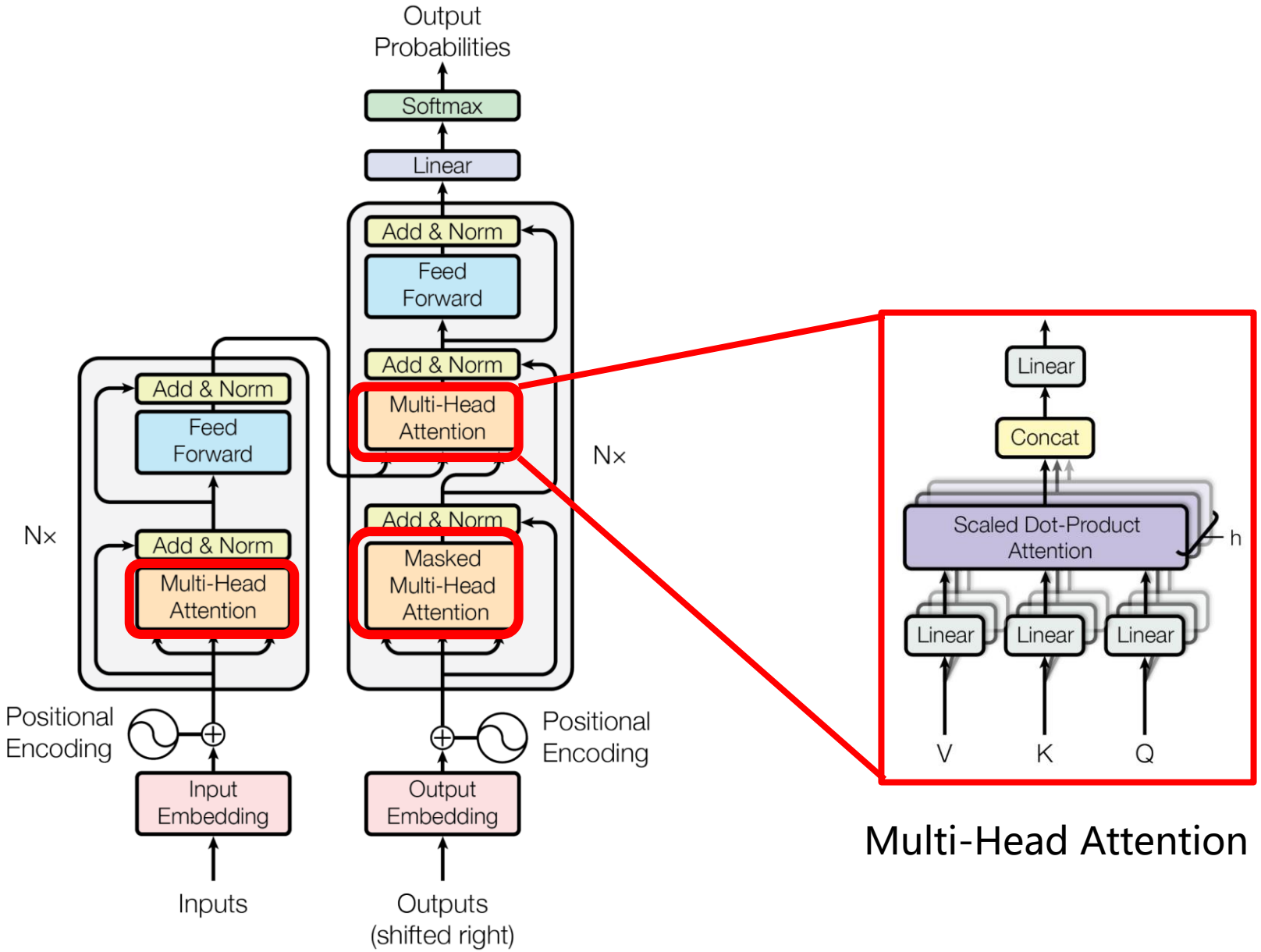
那么连RNN都可以不需要了，整个模型只使用了attention。我们就得到了Transformer。

目录

整体框架

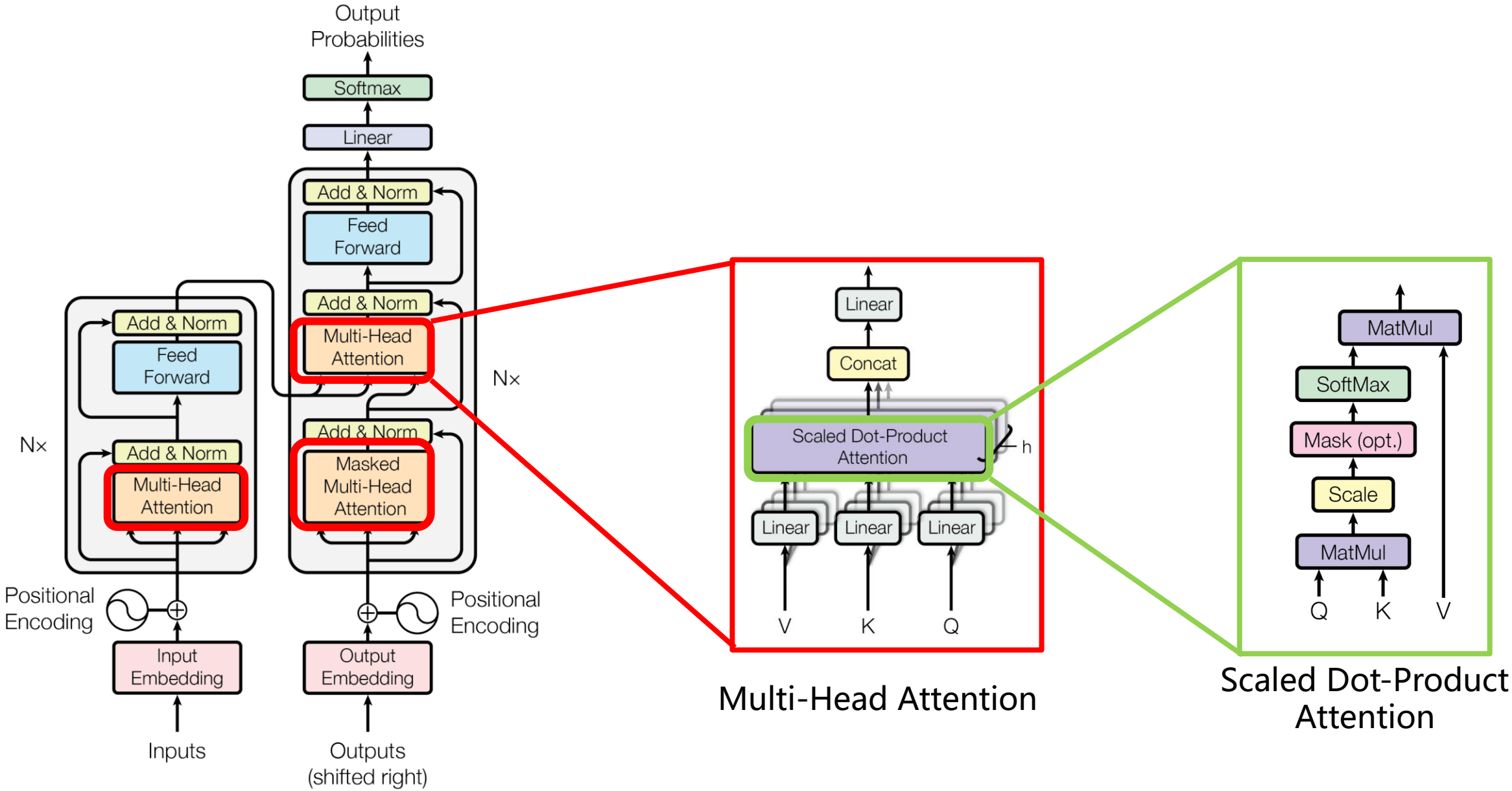

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

整体框架



Multi-Head Attention

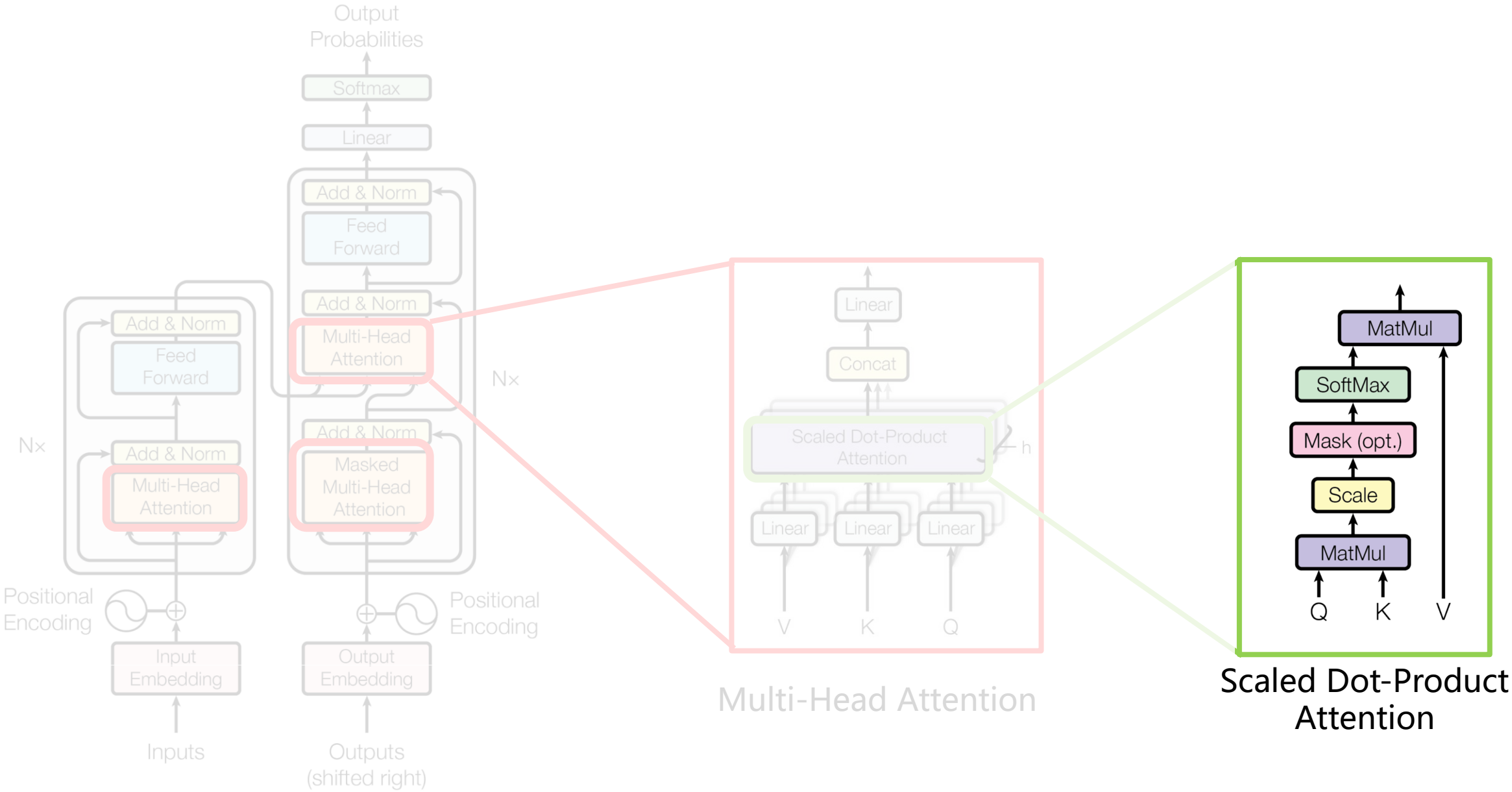
整体框架



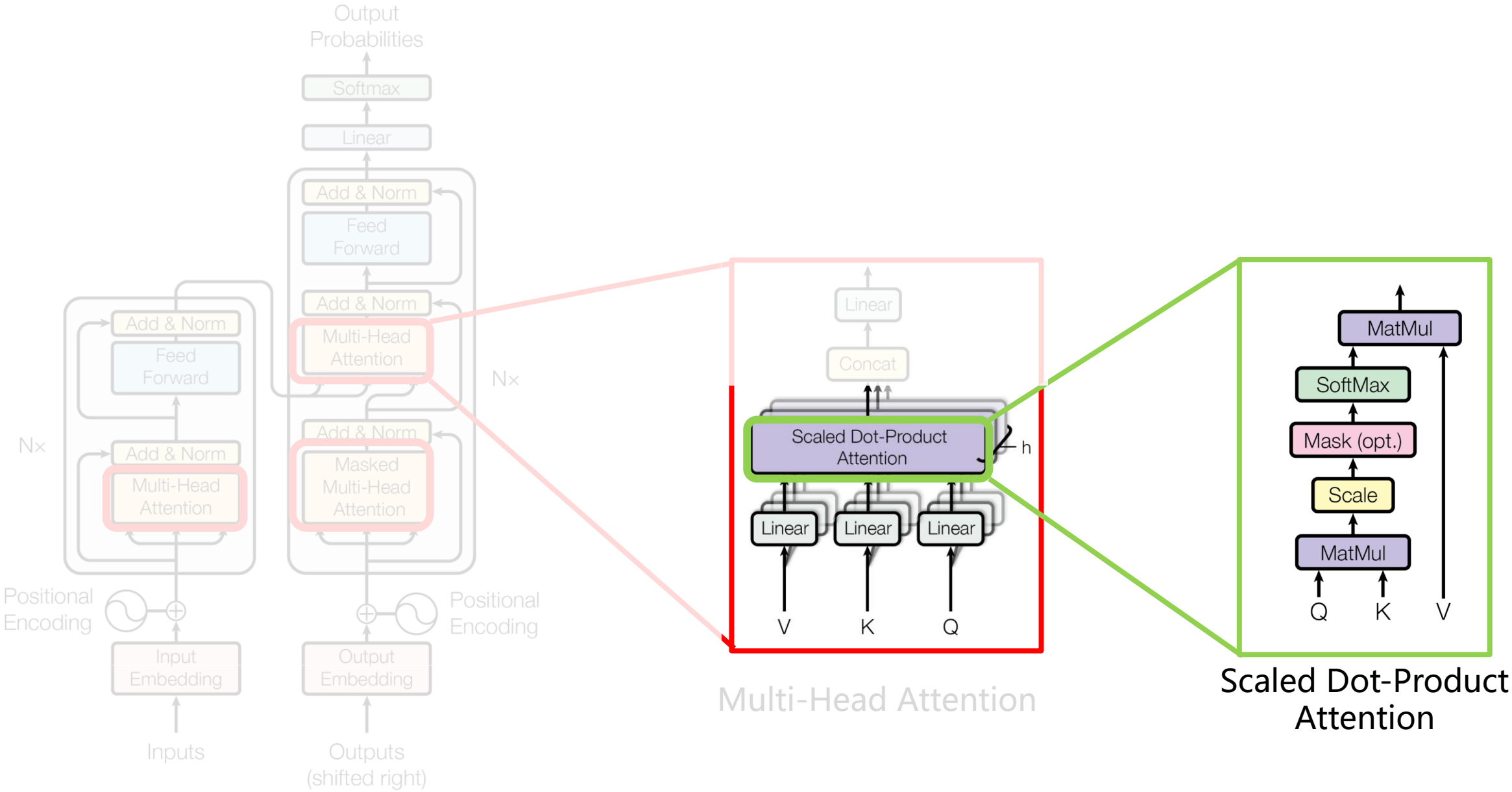
Multi-Head Attention

Scaled Dot-Product Attention

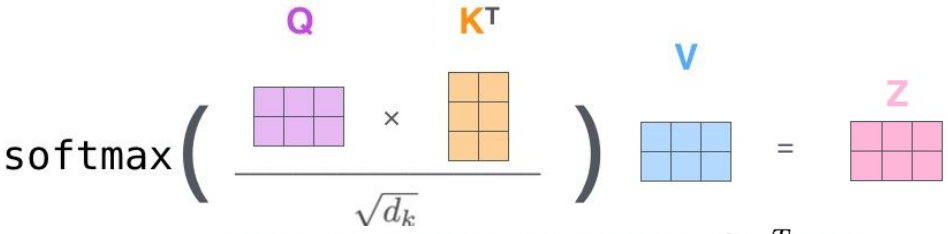
整体框架


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
Multi-Head Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Scaled Dot-Product Attention

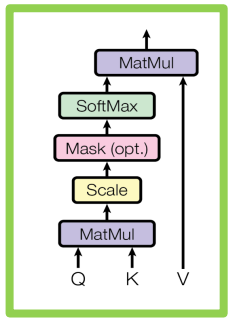
整体框架


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
Multi-Head Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Scaled Dot-Product Attention

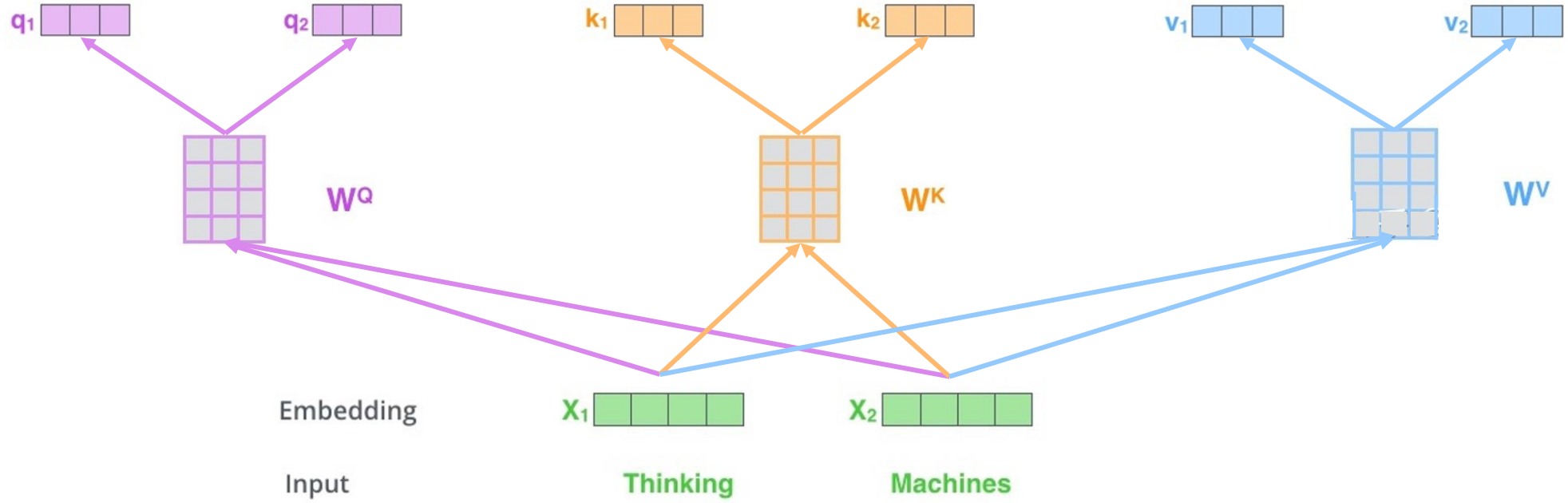
Q, K, V and Scaled Dot-Product Attention



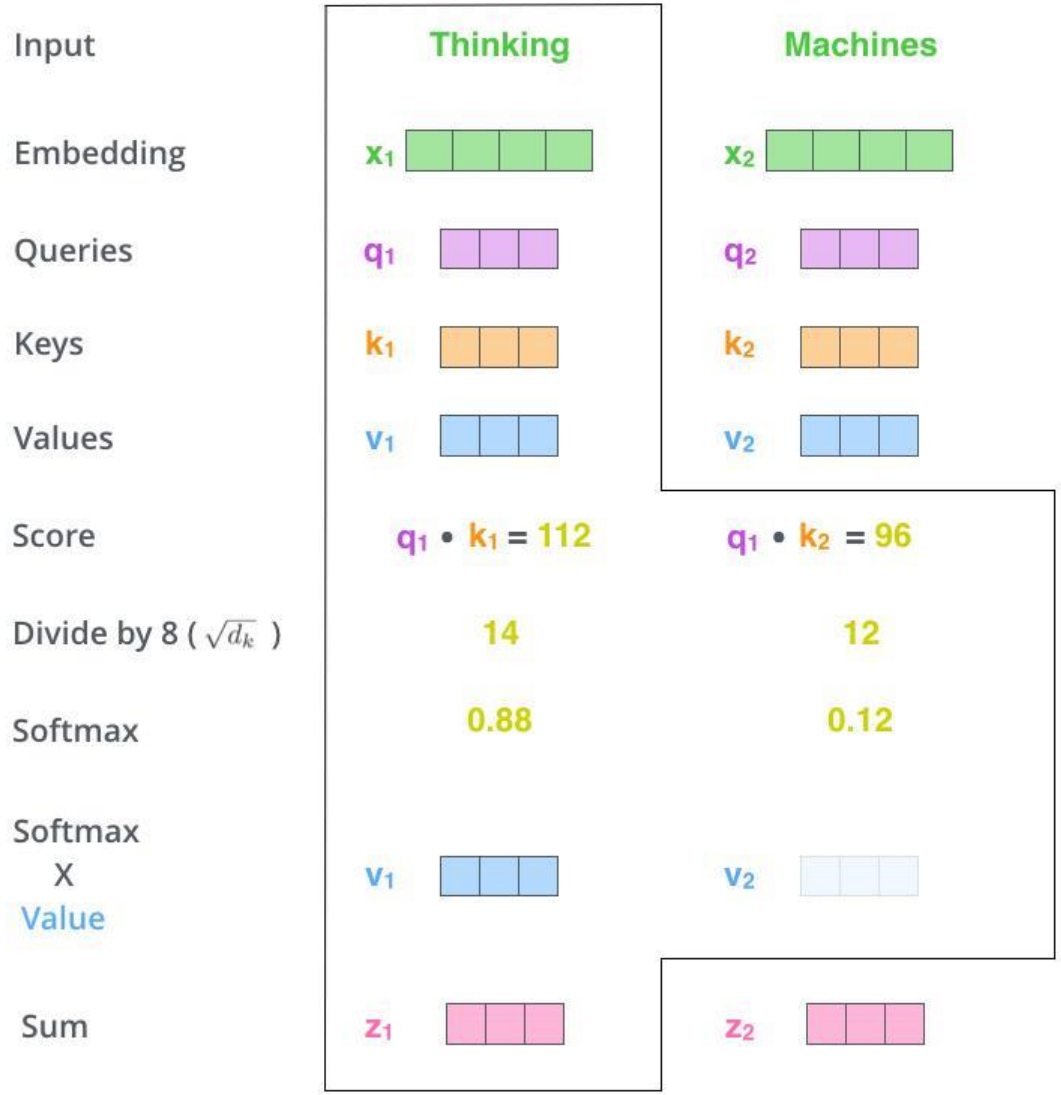
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



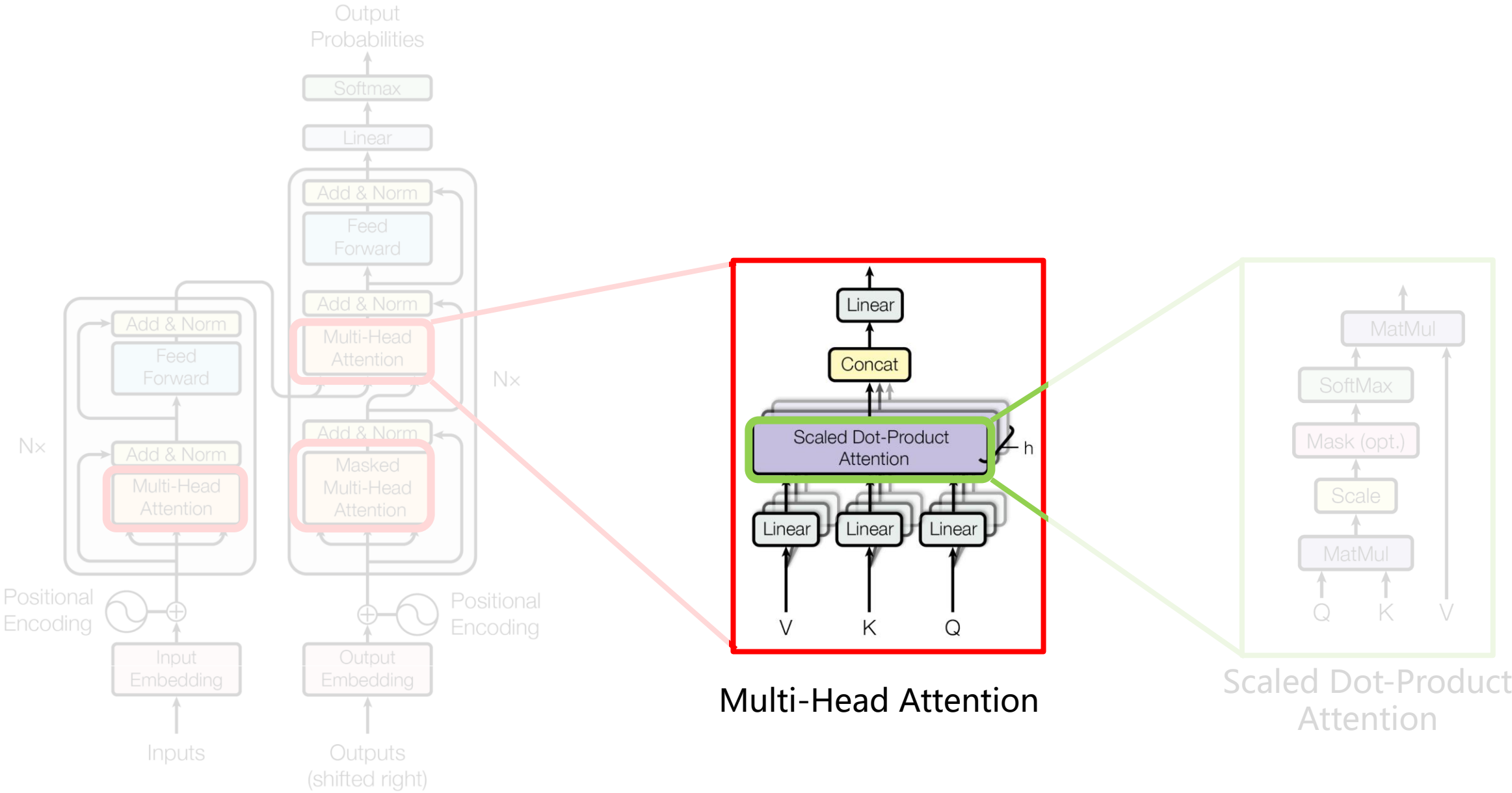
Scaled Dot-Product Attention



Q, K, V and Scaled Dot-Product Attention



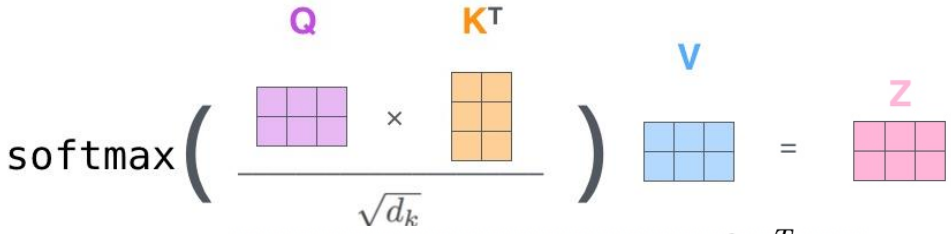
整体框架



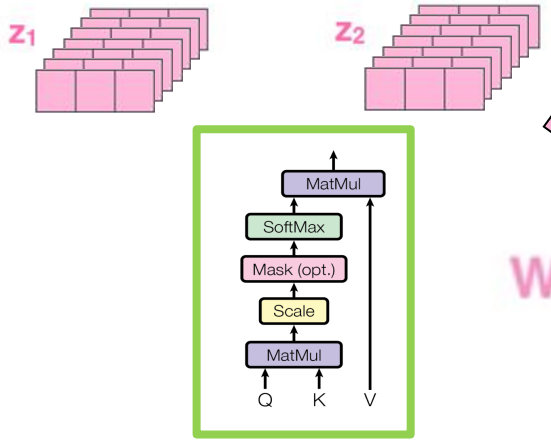
Multi-Head Attention

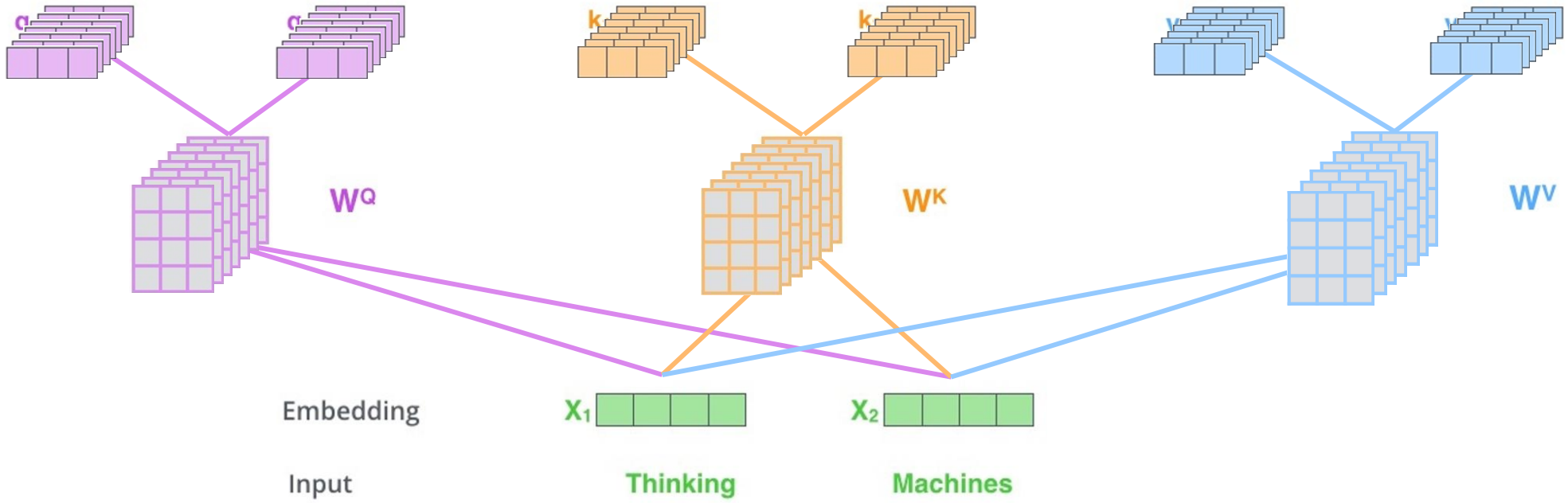
Scaled Dot-Product Attention

Q, K, V and Scaled Dot-Product Attention **with Multiple Heads**

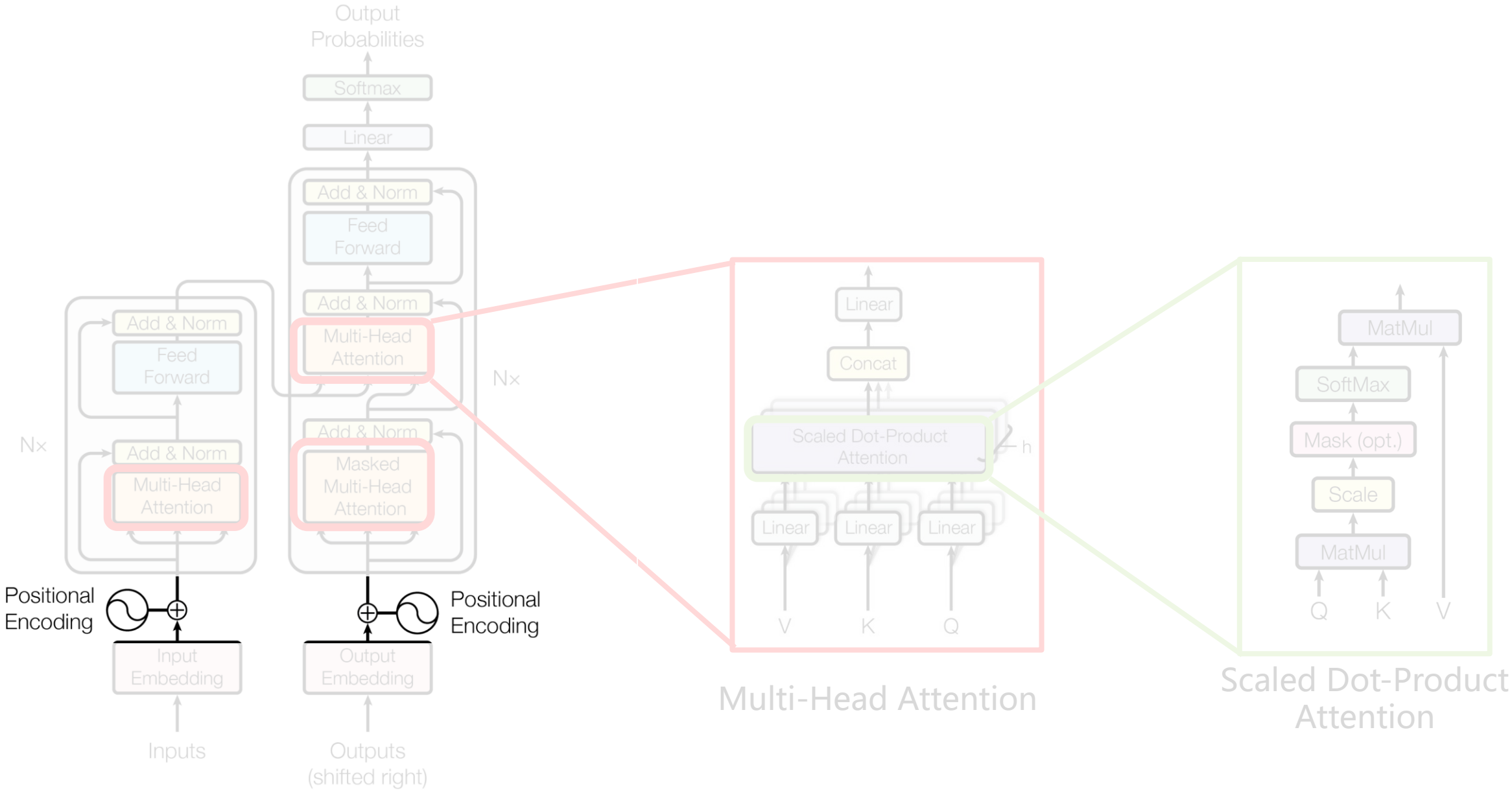


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

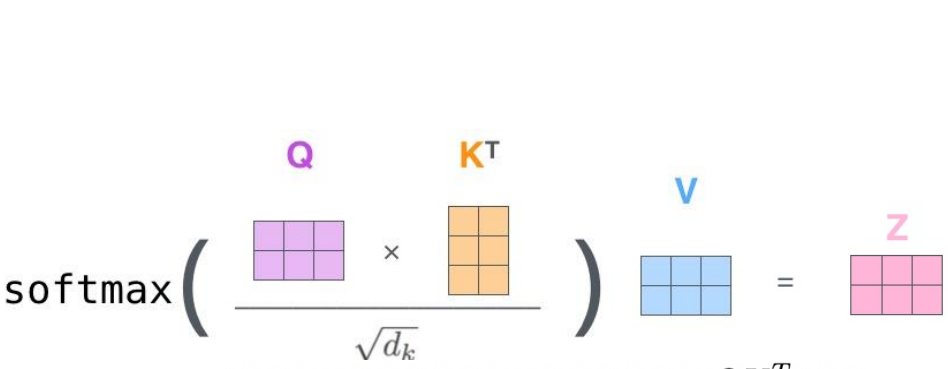




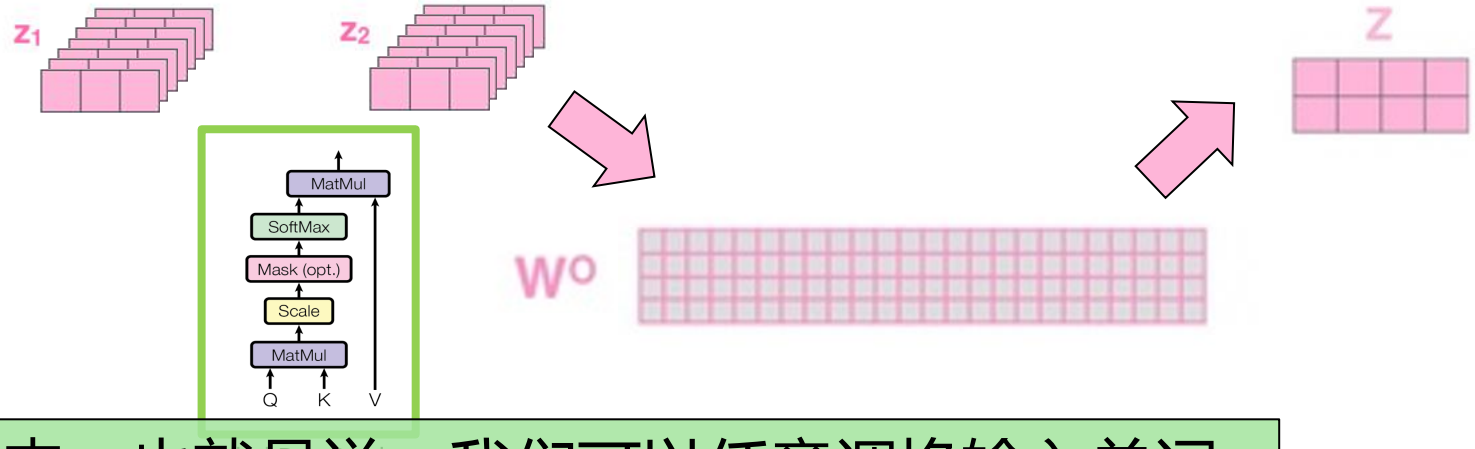
整体框架


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Multi-Head Attention
Scaled Dot-Product Attention

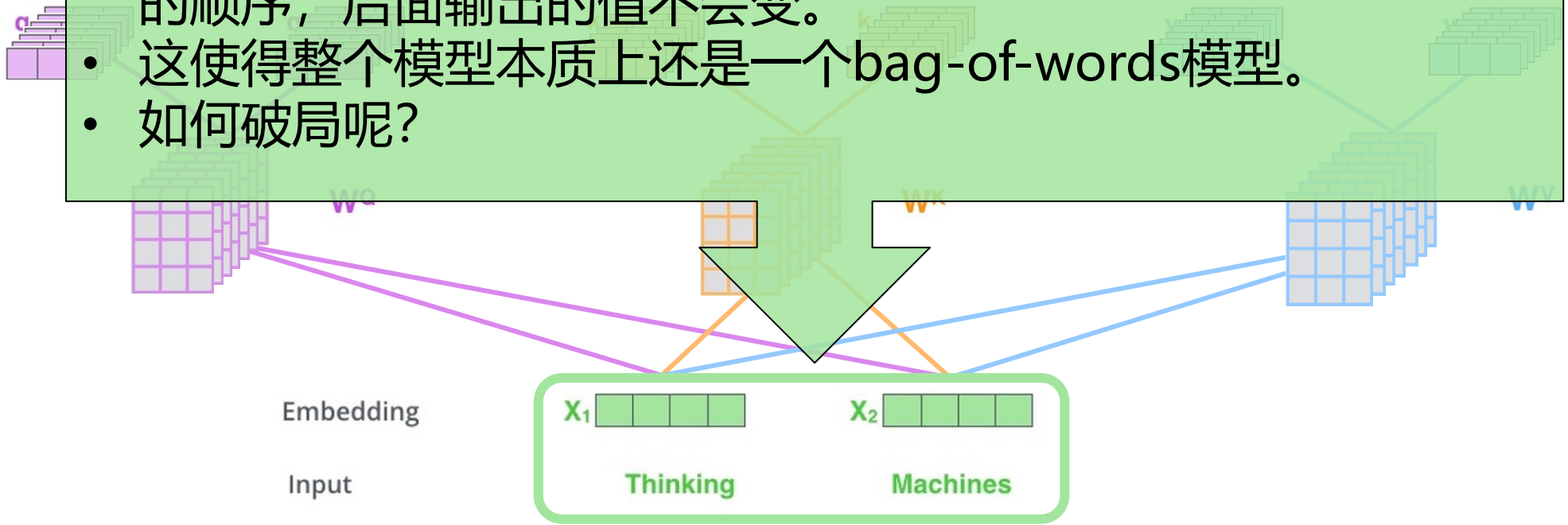
Q, K, V and Scaled Dot-Product Attention **with Multiple Heads**



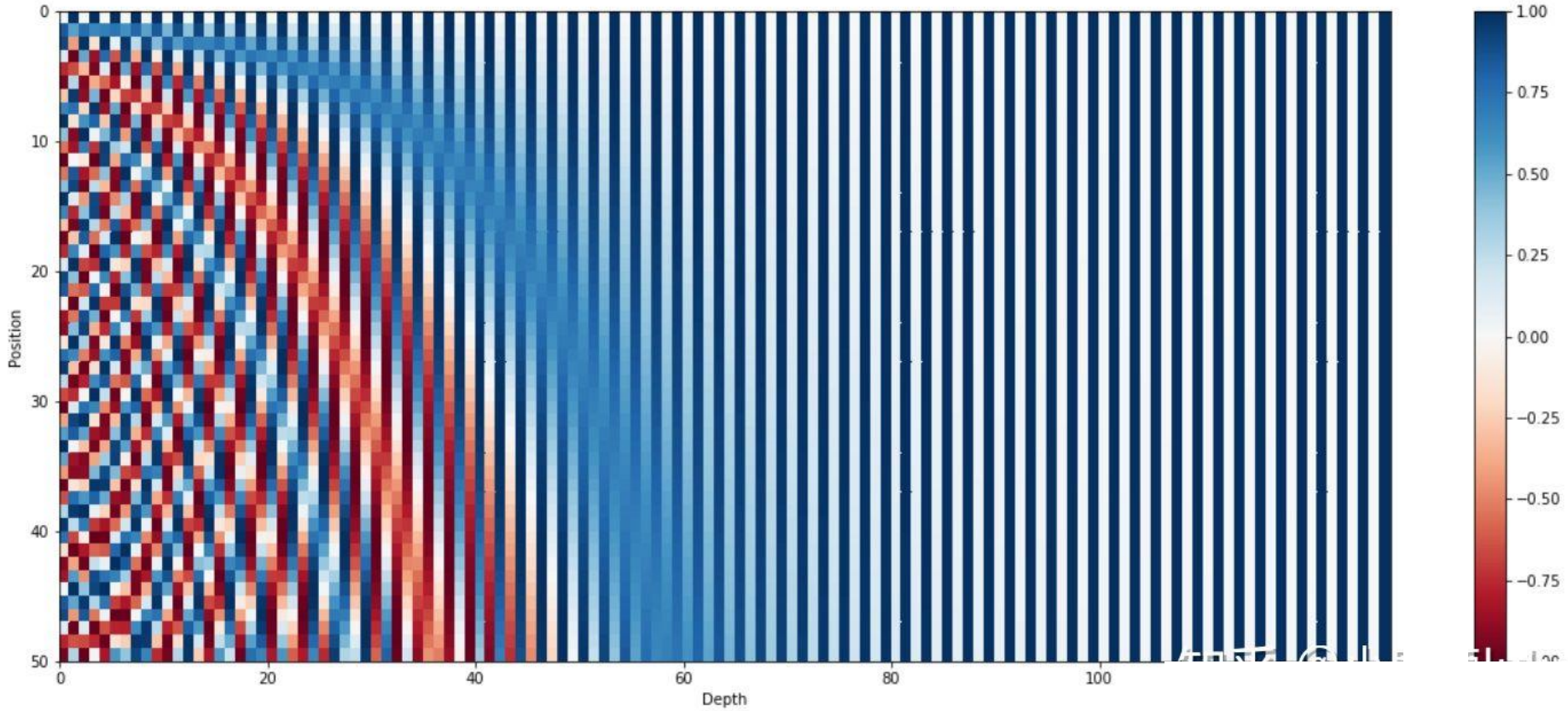
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



- 词序没有被考虑进来。也就是说，我们可以任意调换输入单词的顺序，后面输出的值不会变。
- 这使得整个模型本质上还是一个bag-of-words模型。
- 如何破局呢？



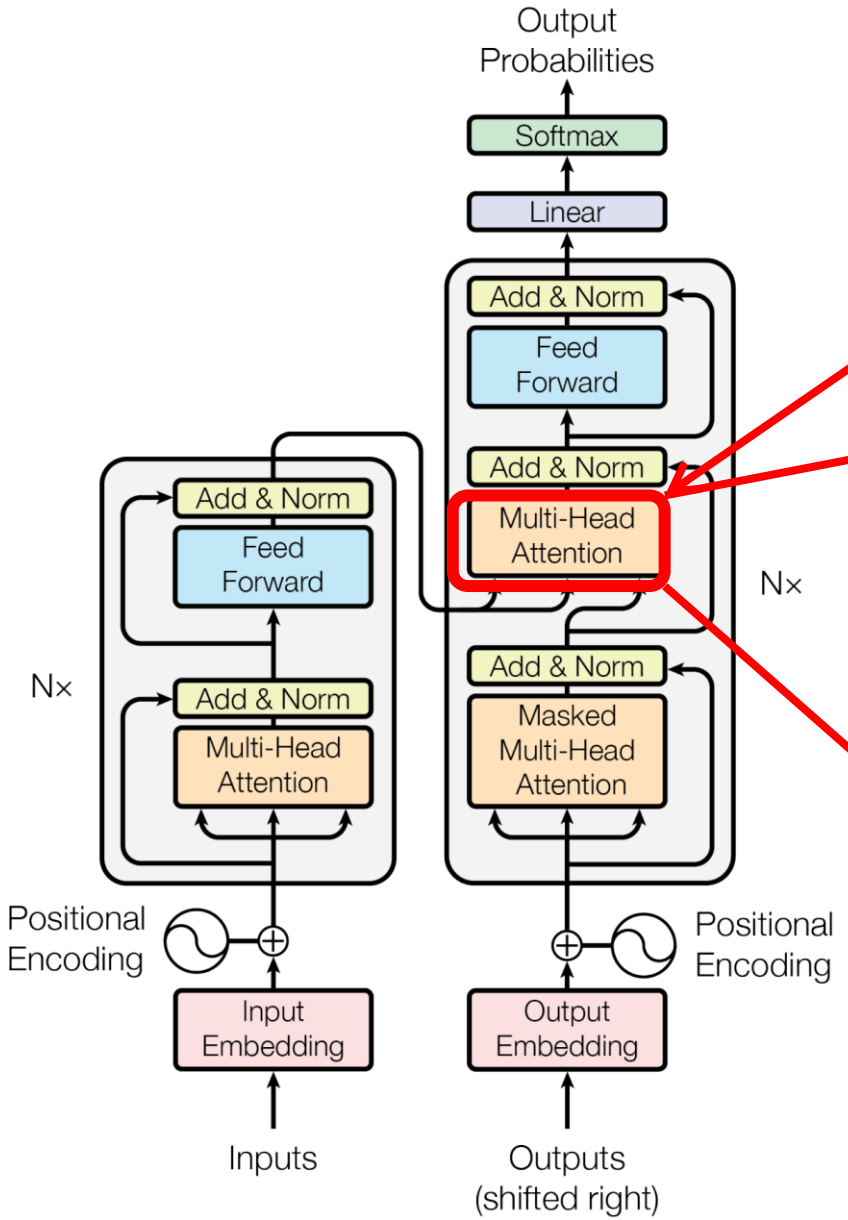
Position Encoding (PE)



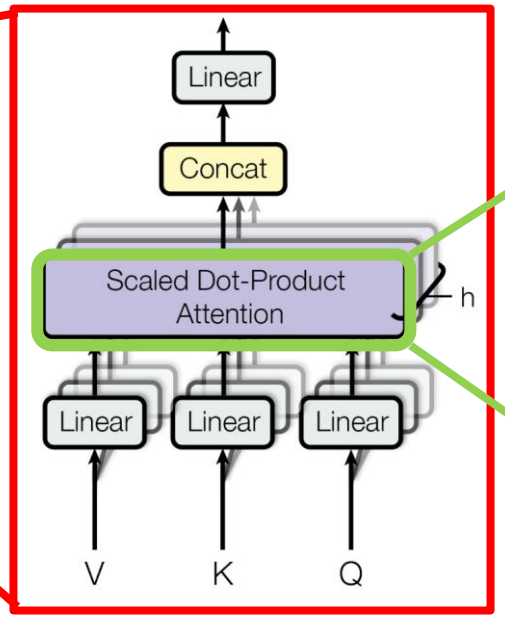
$$\mathbf{pe}_{(j,2i)} = \sin(j/10000^{2i/d_{model}}),$$

$$\mathbf{pe}_{(j,2i+1)} = \cos(j/10000^{2i/d_{model}}),$$

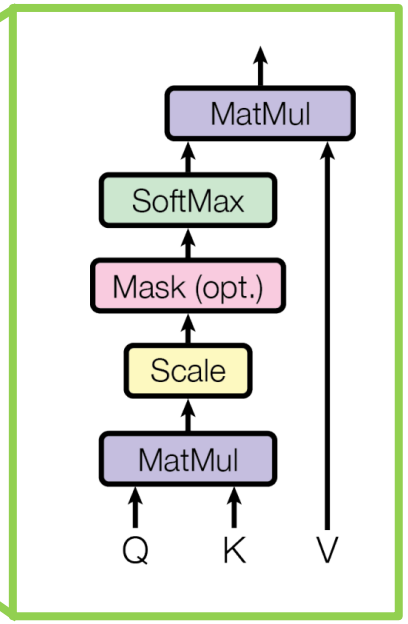
整体框架



对于这个位置的Attention，Q来自于decoder的hidden states，而K和V来自于encoder最后一层的hidden states。
计算步骤与其余两处一样。

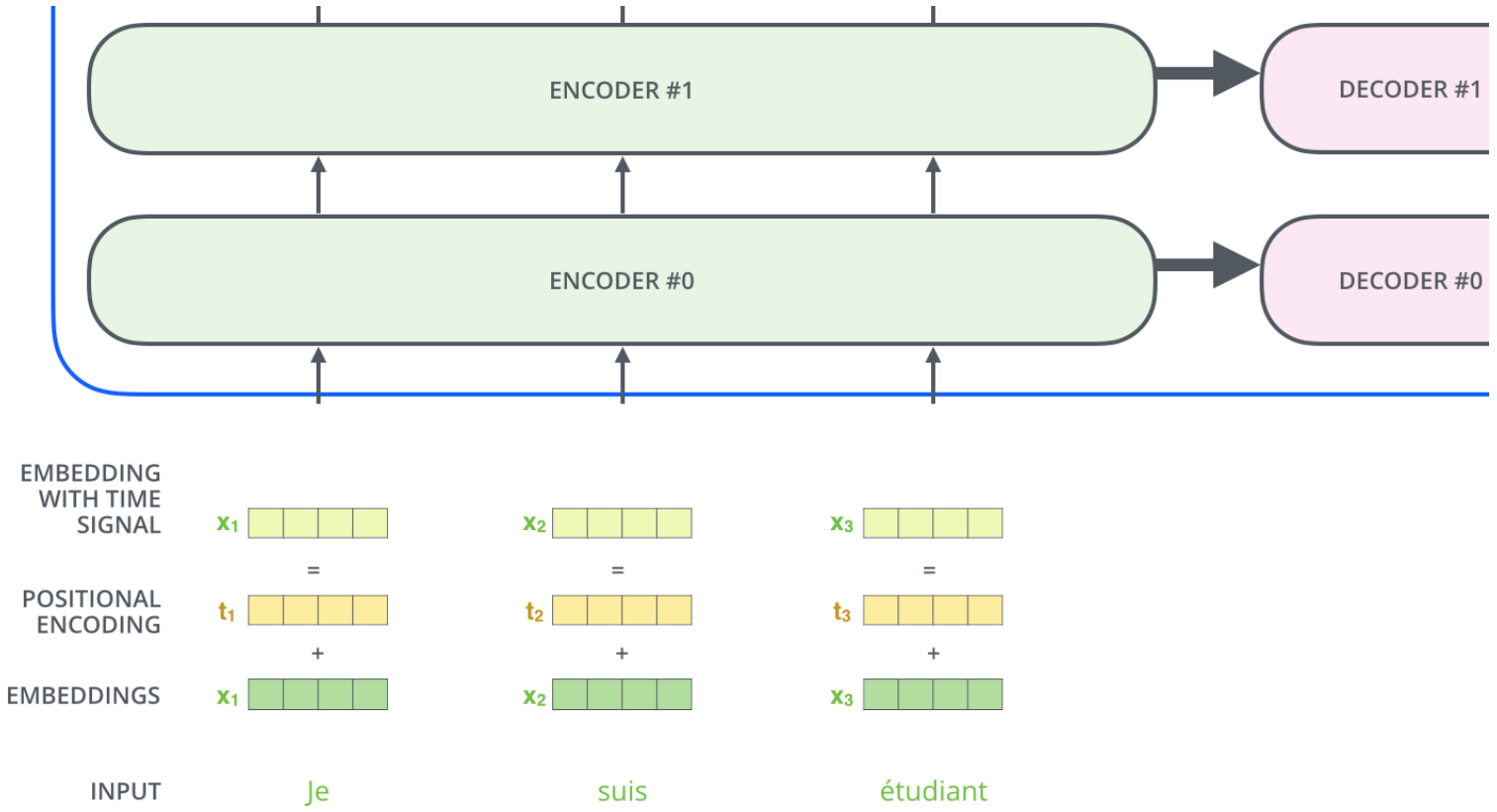


Multi-Head Attention

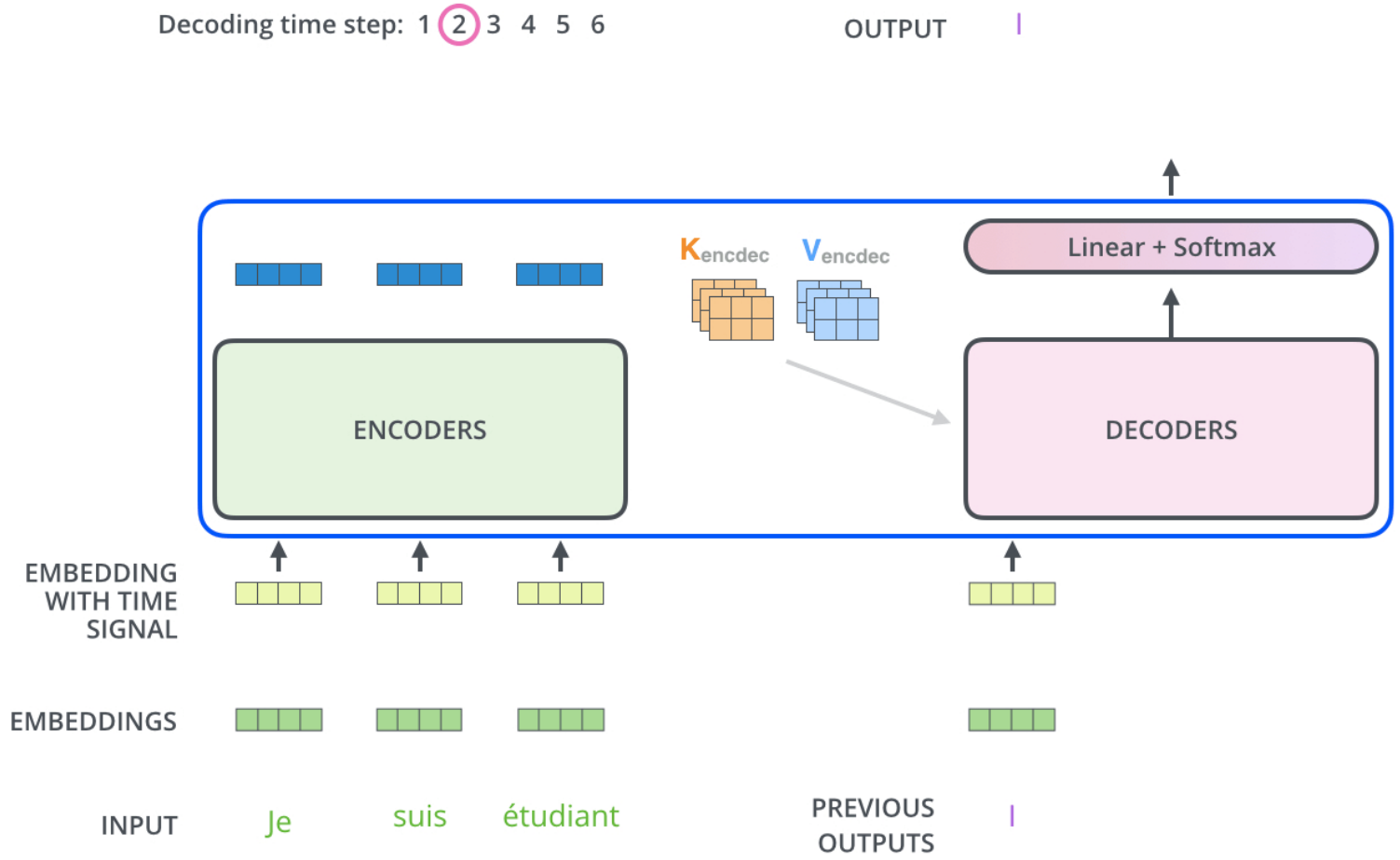


Scaled Dot-Product Attention

Encoder


ENCODER #1
DECODER #1
ENCODER #0
DECODER #0
EMBEDDING WITH TIME SIGNAL
POSITIONAL ENCODING
EMBEDDINGS
INPUT
Je
suis
étudiant

Decoder


Decoding time step: 1 2 3 4 5 6
OUTPUT
I
Linear + Softmax
K encdec
V encdec
ENCODERS
DECODERS
EMBEDDING WITH TIME SIGNAL
EMBEDDINGS
INPUT
Je
suis
étudiant
PREVIOUS OUTPUTS
I

目录

评价指标

- 靠人来看翻译结果当然是一种评价翻译质量最直接的方式。但是对于机器翻译而言，这种方法
  - – 太慢!
  - – 不能反映不同模型间精细微小的性能差别
  - – 不适合大规模地作为机器翻译模型翻译质量评价的方式。
- 如何制定自动评价指标，来定量的反映翻译质量的好坏?
  - 历史上人们使用过两种方式:
    - – Word Error Rate
    - – BLEU since 2002
  - BLEU出现之后，已经成为通用的、标准化的机器翻译模型评价指标。
- BLEU in short: Overlap with reference translations

评价指标：BLEU score

- **考虑如下两个由同一个中文句子，经中英翻译得到的英文句子：**
- Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.
- Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

- **同时我们数据集中，存在三个标准翻译：**
- Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
- Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference 3: It is the practical guide for the army always to heed the directions of the party.

评价指标：BLEU score

- **考虑如下两个由同一个中文句子，经中英翻译得到的英文句子：**
- Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.
- Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

- **同时我们数据集中，存在三个标准翻译：**
- Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
- Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference 3: It is the practical guide for the army always to heed the directions of the party.

评价指标：BLEU score

Unigram Precision:

$$\frac{\text{模型预测的句子中，与reference中重合的unigram个数}}{\text{模型预测的句子所包含的unigram总数}}$$

Modified Unigram Precision: 将分子中每个unigram计入的次数设置上限，不得超过该unigram在某一个reference中重复出现的最大次数。

Candidate: the the the the the the the.
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

Unigram Precision **1**
Modified Unigram Precision **2/7**

评价指标：BLEU score

- **考虑如下两个由同一个中文句子，经中英翻译得到的英文句子：**
- Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party. **unigram precision 17/18**
- Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct. **unigram precision 8/14**

- **同时我们数据集中，存在三个标准翻译：**
- Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
- Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference 3: It is the practical guide for the army always to heed the directions of the party.

评价指标：BLEU score

N-gram Precision:

$$\frac{\text{模型预测的句子中，与reference中重合的N-gram个数}}{\text{模型预测的句子所包含的N-gram总数}}$$

Modified N-gram Precision: 将分子中每个N-gram计入的次数设置上限，不得超过该N-gram在某一个reference中重复出现的最大次数。

- Unigram Precision 更多反映翻译的***充分度***
- N-gram Precision 更多反映翻译的***流畅度***

评价指标：BLEU score

- **考虑如下两个由同一个中文句子，经中英翻译得到的英文句子：**
- Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.
- Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

unigram precision 17/18
bigram precision 10/17

unigram precision 8/14
bigram precision 1/13

- Unigram Precision 更多反映翻译的***充分度***
- N-gram Precision 更多反映翻译的***流畅度***

- **同时**
- Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
- Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference 3: It is the practical guide for the army always to heed the directions of the party.

评价指标：BLEU score

- **但是我们怎么将不同N-gram的 precision scores组合起来呢?**
- 最直接的想法就是取平均；然而unigram的分数太高了，取平均会导致其它N-gram的重要性被淹没掉。所以这里我们使用的是几何平均。
- 同时再乘上一个额外的参数（BP）来惩罚过短的模型输出。
- 比如，BLEU-k指的就是取了1~k-gram的几何平均后得到的值。
- 如果没有指定是BLEU几，那默认指的就是BLEU-4。

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

c: 模型预测的句子的长度
r: 标准翻译中句子的长度

目录

常用实现

Moses: Open Source Toolkit

- **Open source** statistical machine translation system (developed from scratch 2006)
  - state-of-the-art *phrase-based* approach
  - novel methods: *factored translation models*, *confusion network decoding*
  - support for *very large models* through *memory-efficient* data structures
- Documentation, source code, binaries **available** at `http://www.statmt.org/moses/`
- Development also **supported by**
  - EC-funded *TC-STAR* project
  - *US* funding agencies DARPA, NSF
  - universities (Edinburgh, Maryland, MIT, ITC-irst, RWTH Aachen, ...)

常用实现

https://github.com/facebookresearch/fairseq

https://github.com/OpenNMT/OpenNMT-py

Hugging Face

https://huggingface.co/