

# 彩云 小译

机器翻译与编码器 - 解码器架构

张江

# AI改变生活



# 机器翻译的演进

1949

---



Warren Weaver

# 机器翻译的演进

The spirit is  
willing but the  
flesh is weak



The wine is good  
but the meat is  
spoiled

1949

1960



Warren Weaver

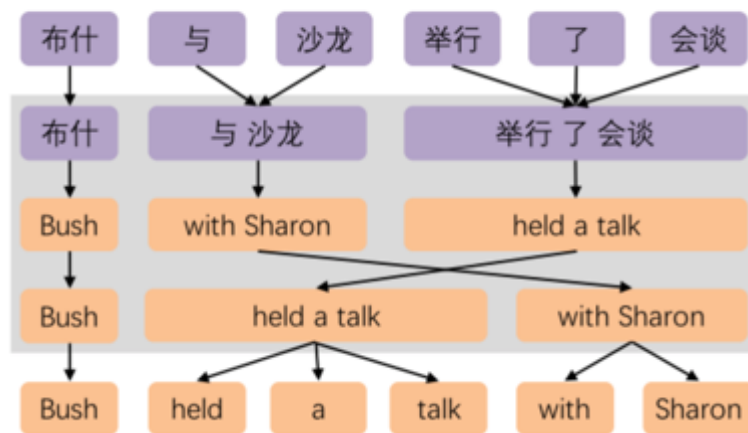
基于规  
则的机  
器翻译

美国花2千万美元  
为机器翻译挖掘了  
一个坟墓

# 机器翻译的演进

美国花2千万美元  
为机器翻译挖掘了  
一个坟墓

- 短语翻译模型：以短语为基本翻译单元



1949

1960

1990



Warren Weaver

基于规  
则的机  
器翻译

统计机器翻译  
Statistical  
Machine Learning

# 机器翻译的演进

美国花2千万美元  
为机器翻译挖掘了一个坟墓



神经机器翻译

1949

1960

1990

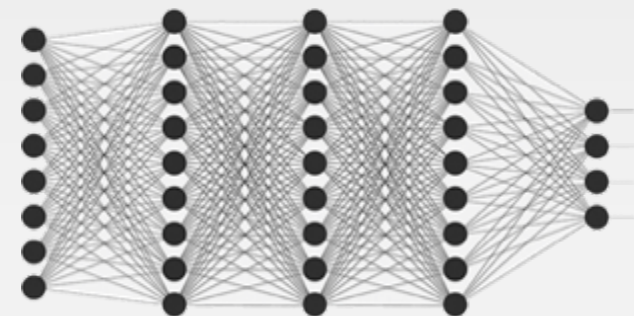
2010



Warren Weaver

基于规则的机器翻译

统计机器翻译  
Statistical  
Machine Learning



# 神经网络机器翻译

---

## Sequence to Sequence Learning with Neural Networks

---

Ilya Sutskever  
Google  
ilyasu@google.com

Oriol Vinyals  
Google  
vinyals@google.com

Quoc V. Le  
Google  
qvl@google.com

### Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

Published as a conference paper at ICLR 2015

## NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau  
Jacobs University Bremen, Germany

KyungHyun Cho   Yoshua Bengio\*  
Université de Montréal

### 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 a fixed-length vector is a bottleneck in improving the performance of this basic encoder-decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

### 1 INTRODUCTION

473v7 [cs.CL] 19 May 2016

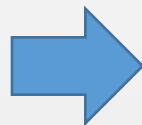
# 人类翻译的过程

The human translation process may be described as:

1. Decoding the meaning of the source text; and
2. Re-encoding this meaning in the target language.



# 人类翻译的过程

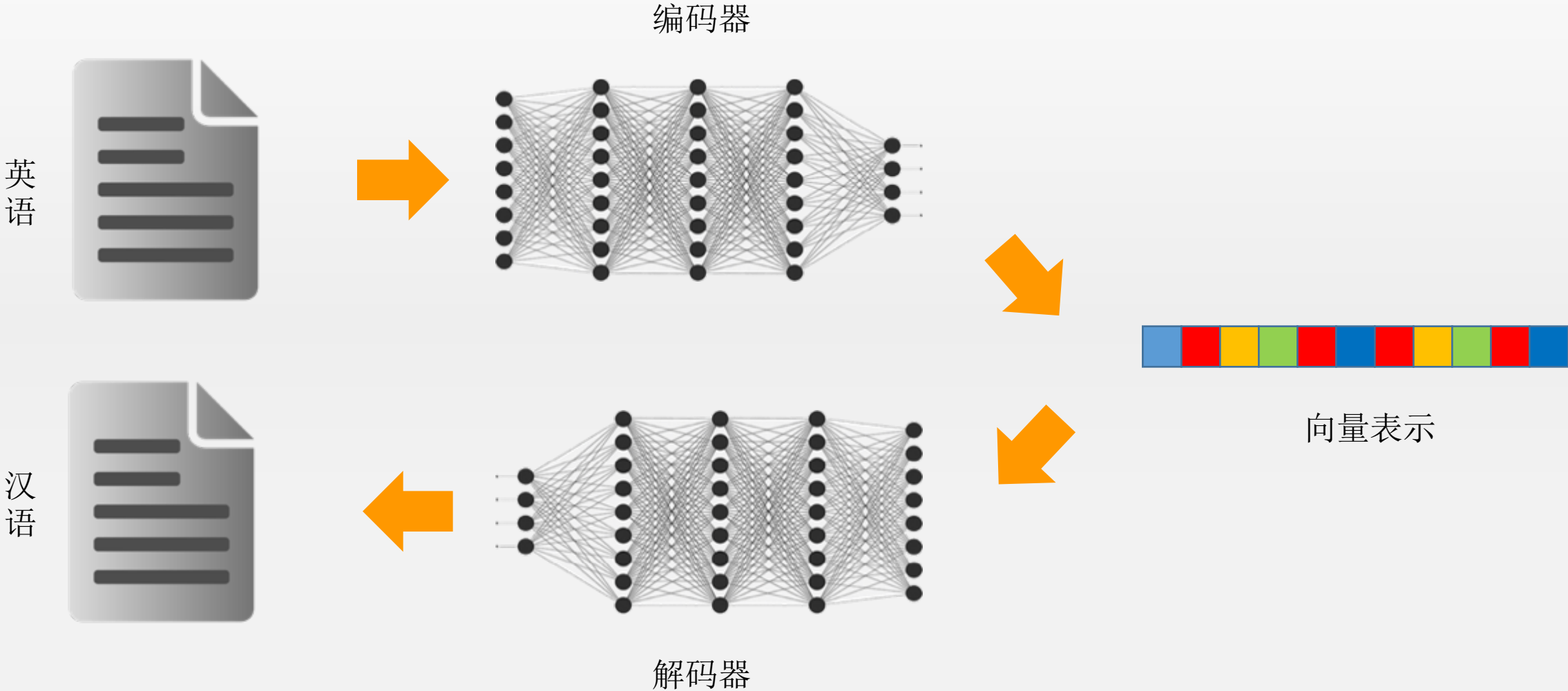


编码成内部状态

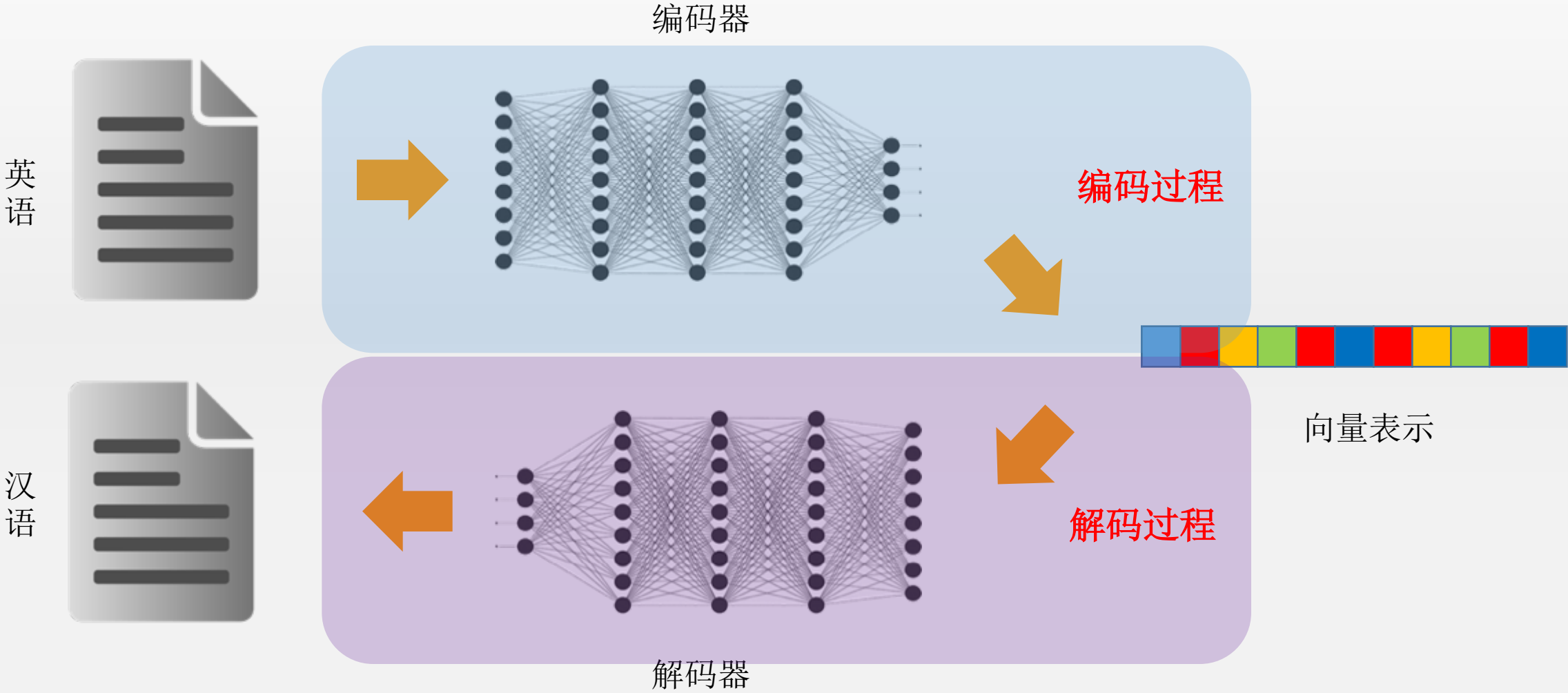


解码成目标语言

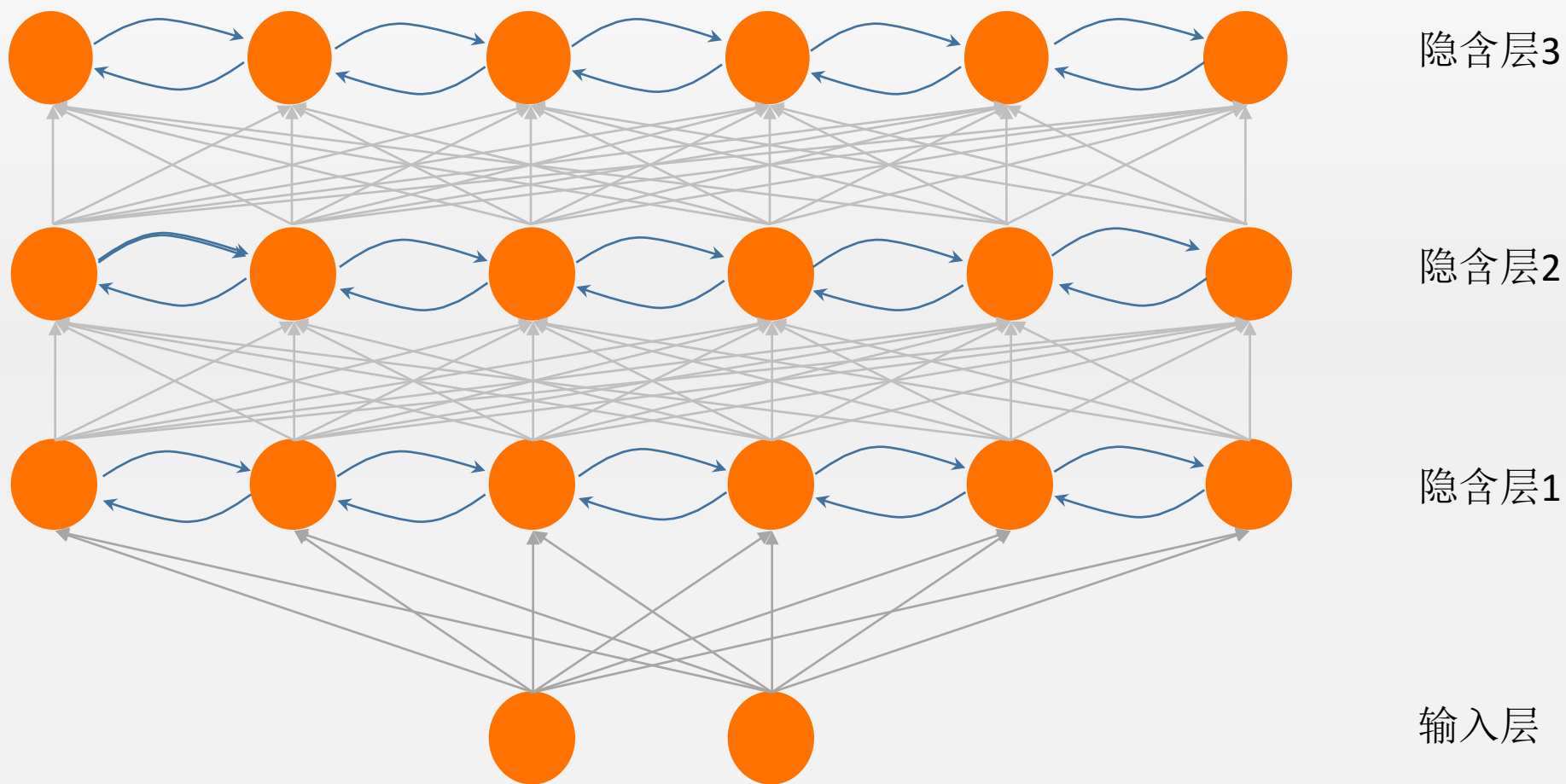
# 编码器解码器架构



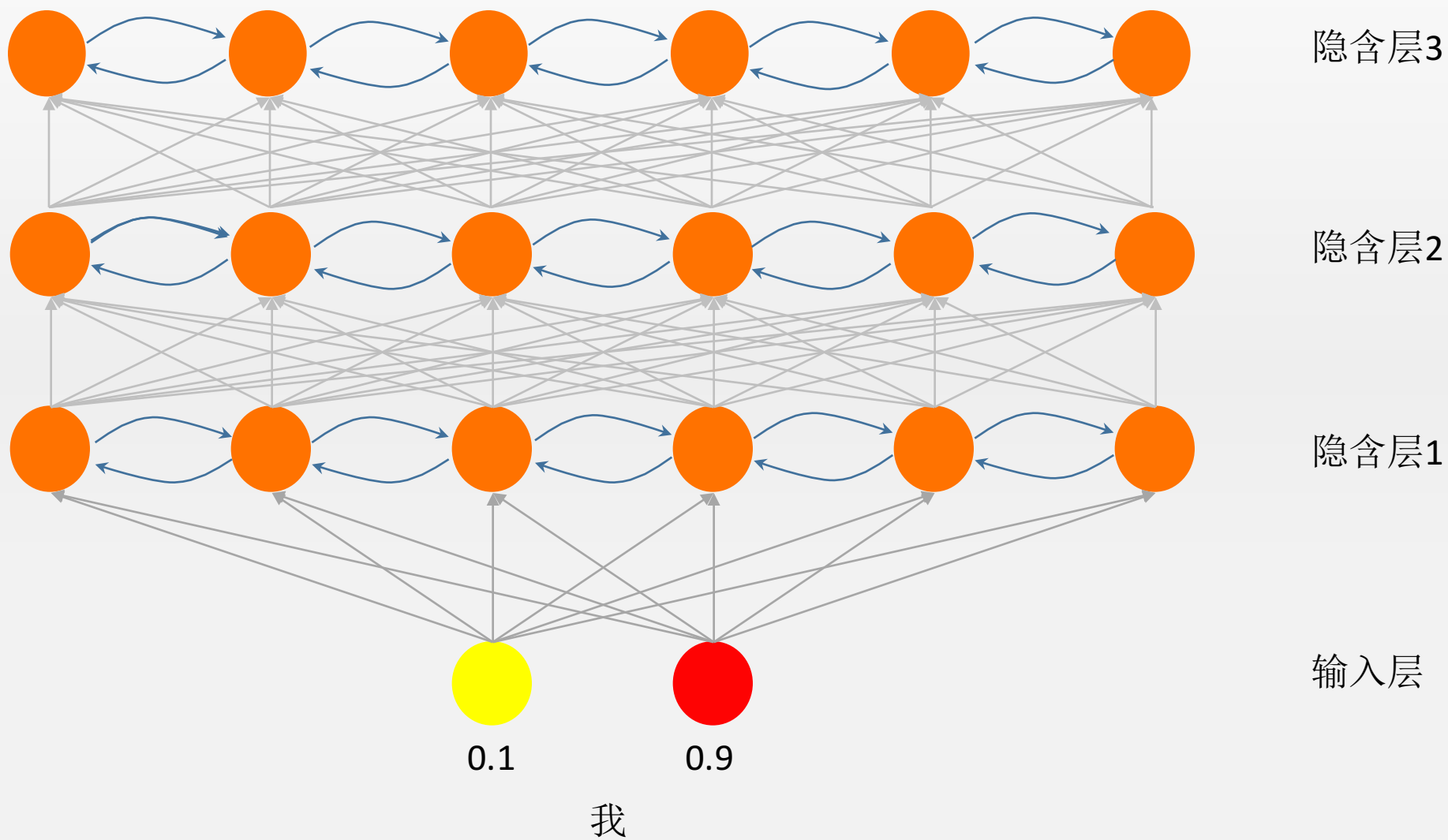
# 编码器解码器架构



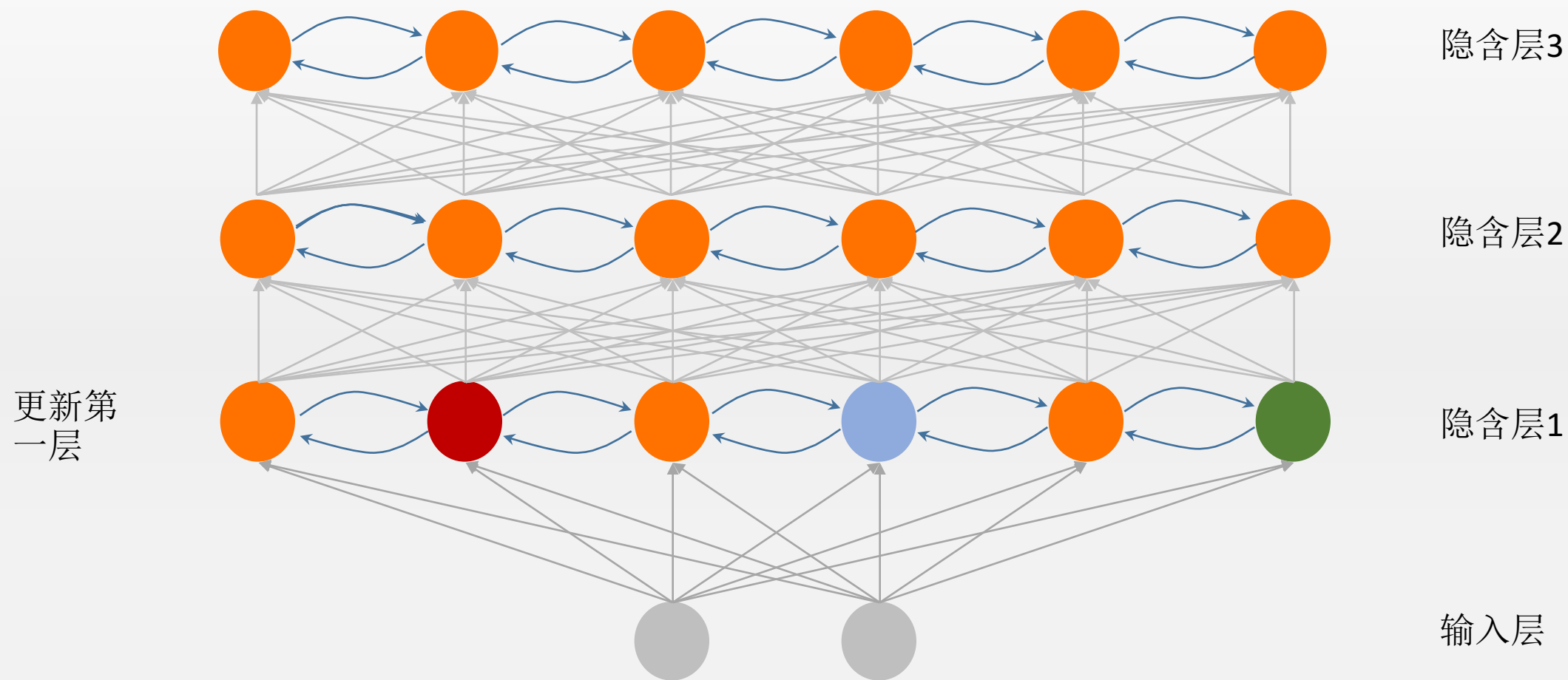
## 编码器：多层RNN



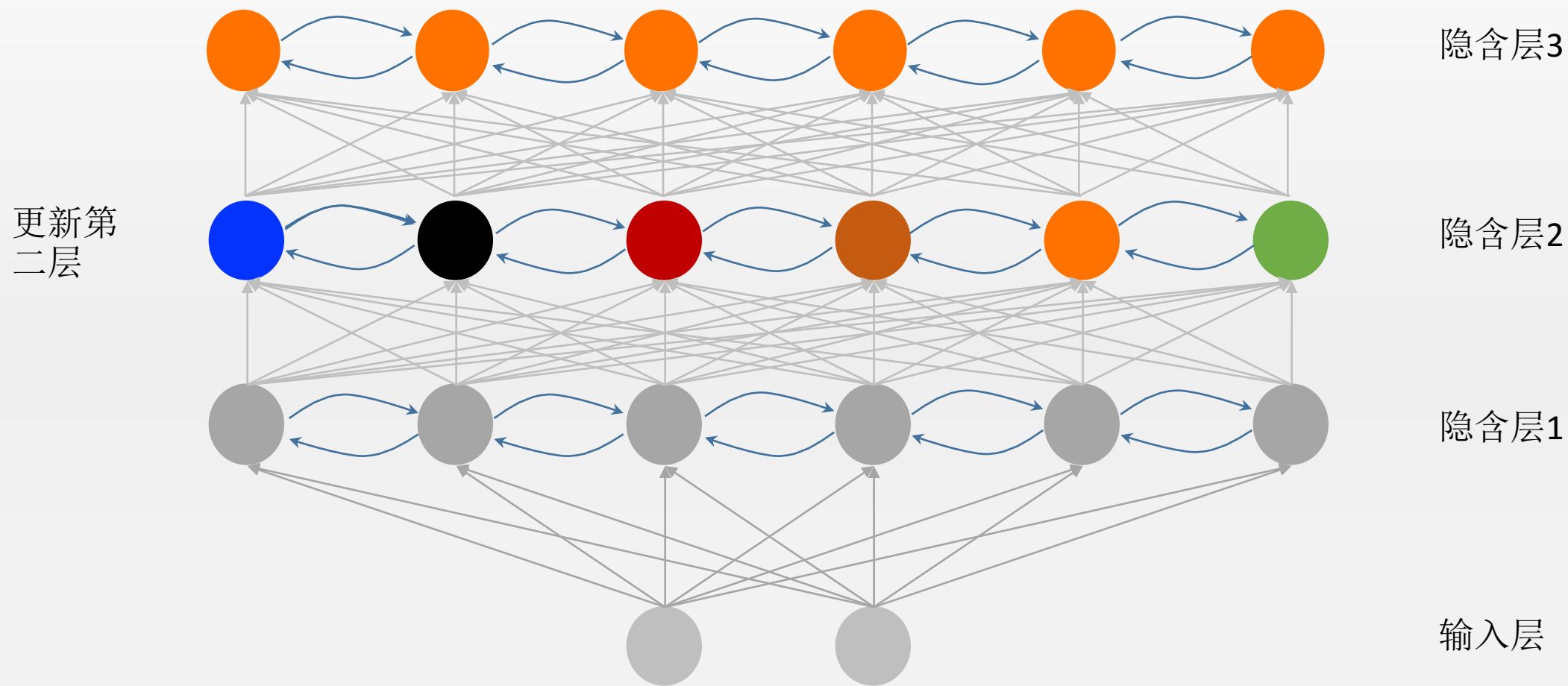
运行 $t=0$



运行 $t=0$

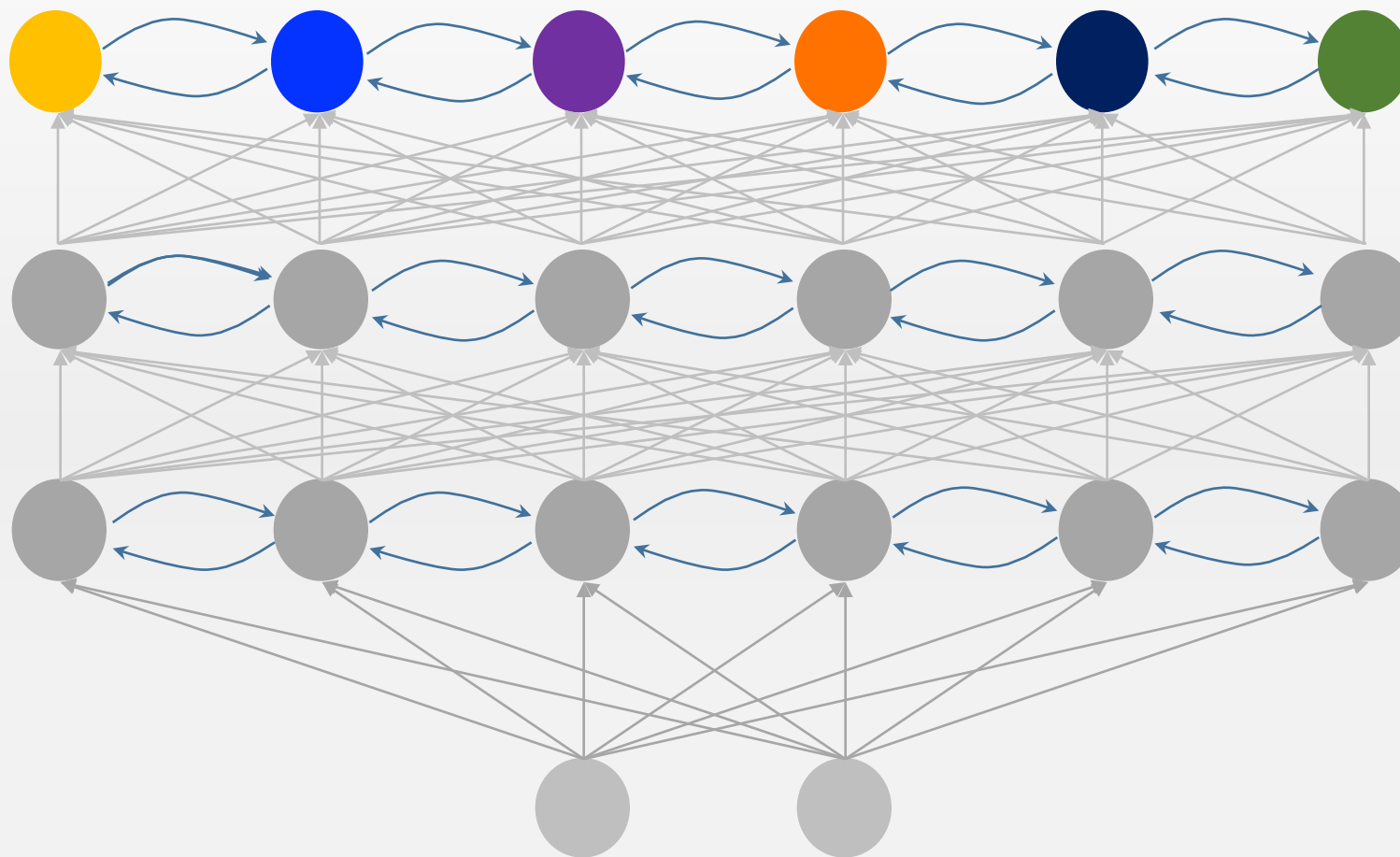


运行 $t=0$



运行 $t=0$

更新第  
三层



隐含层3

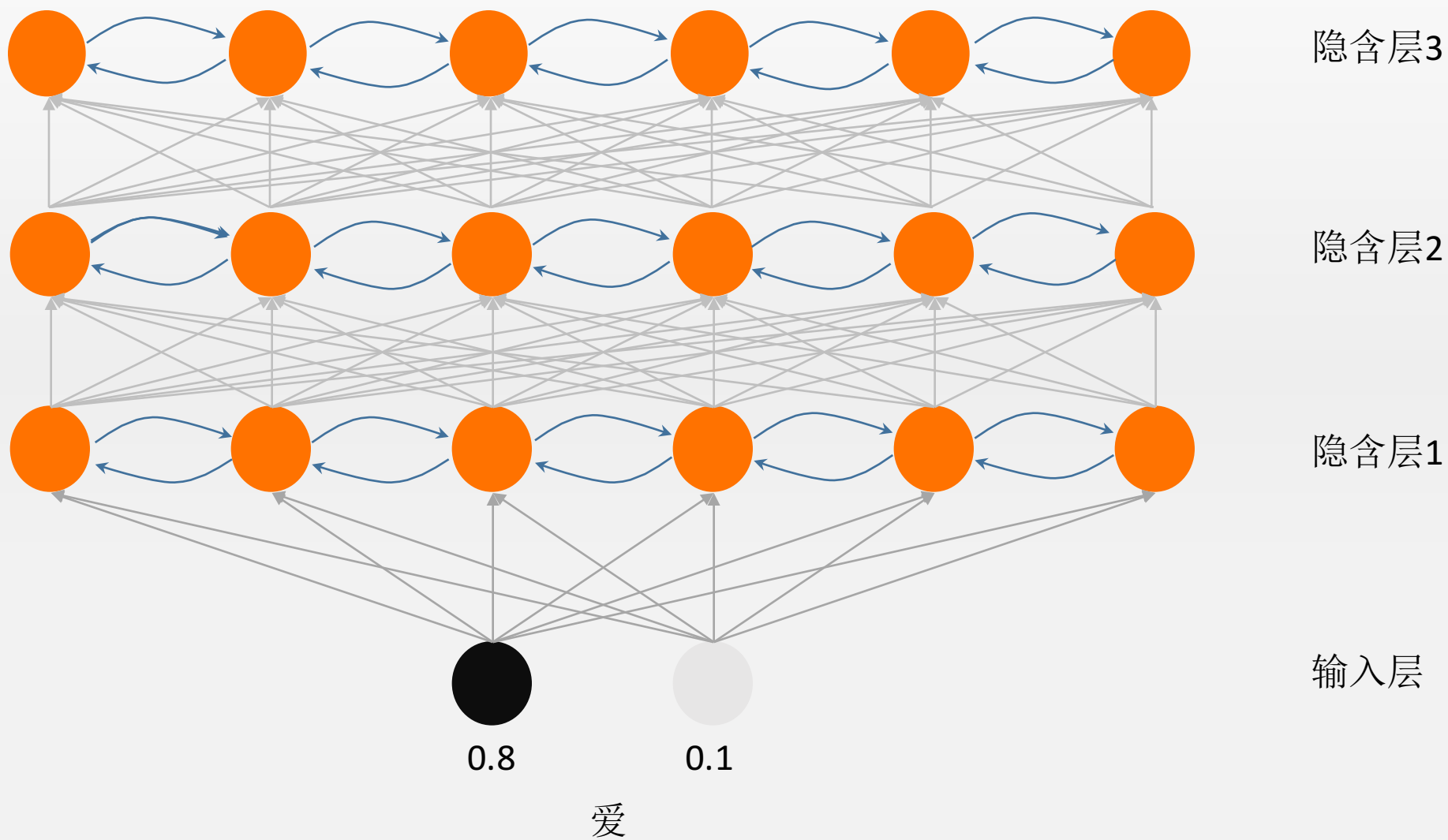
隐含层2

隐含层1

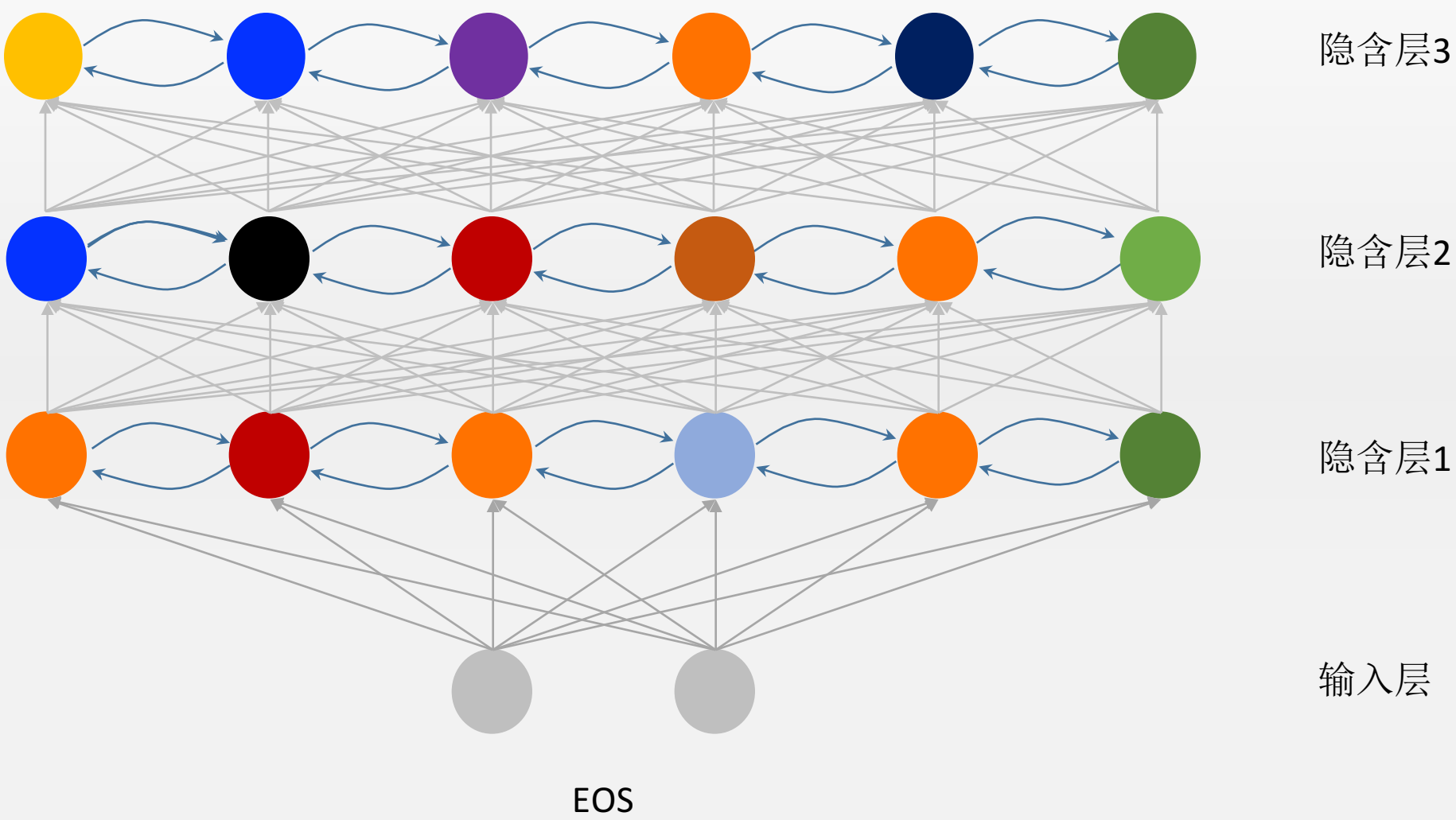
输入层



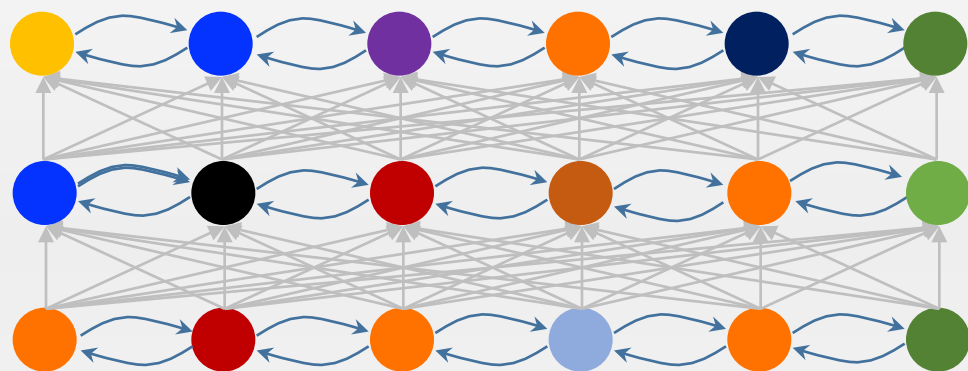
运行 $t=1$



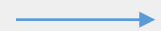
t=T后的状态编码



## 编码器的输出



$t=T$ 时刻所有层

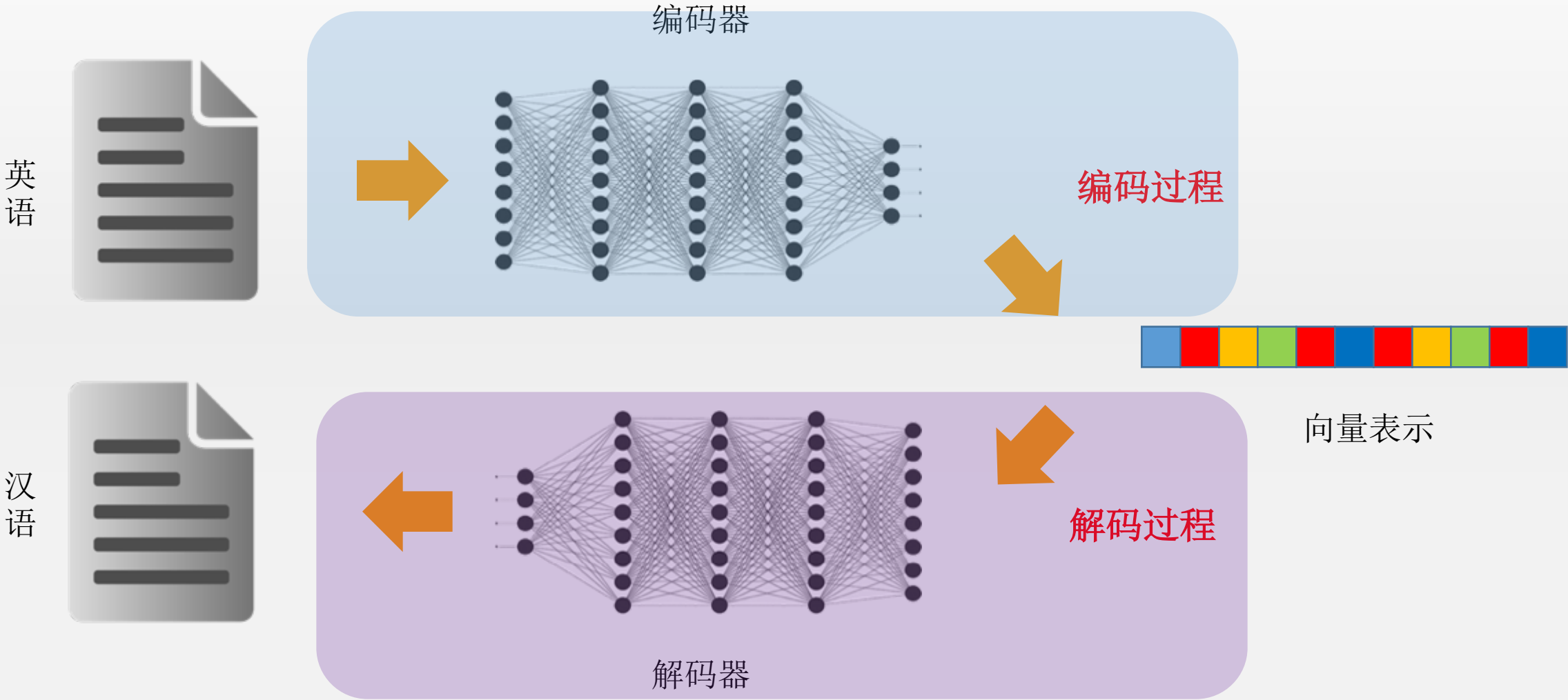


相当于

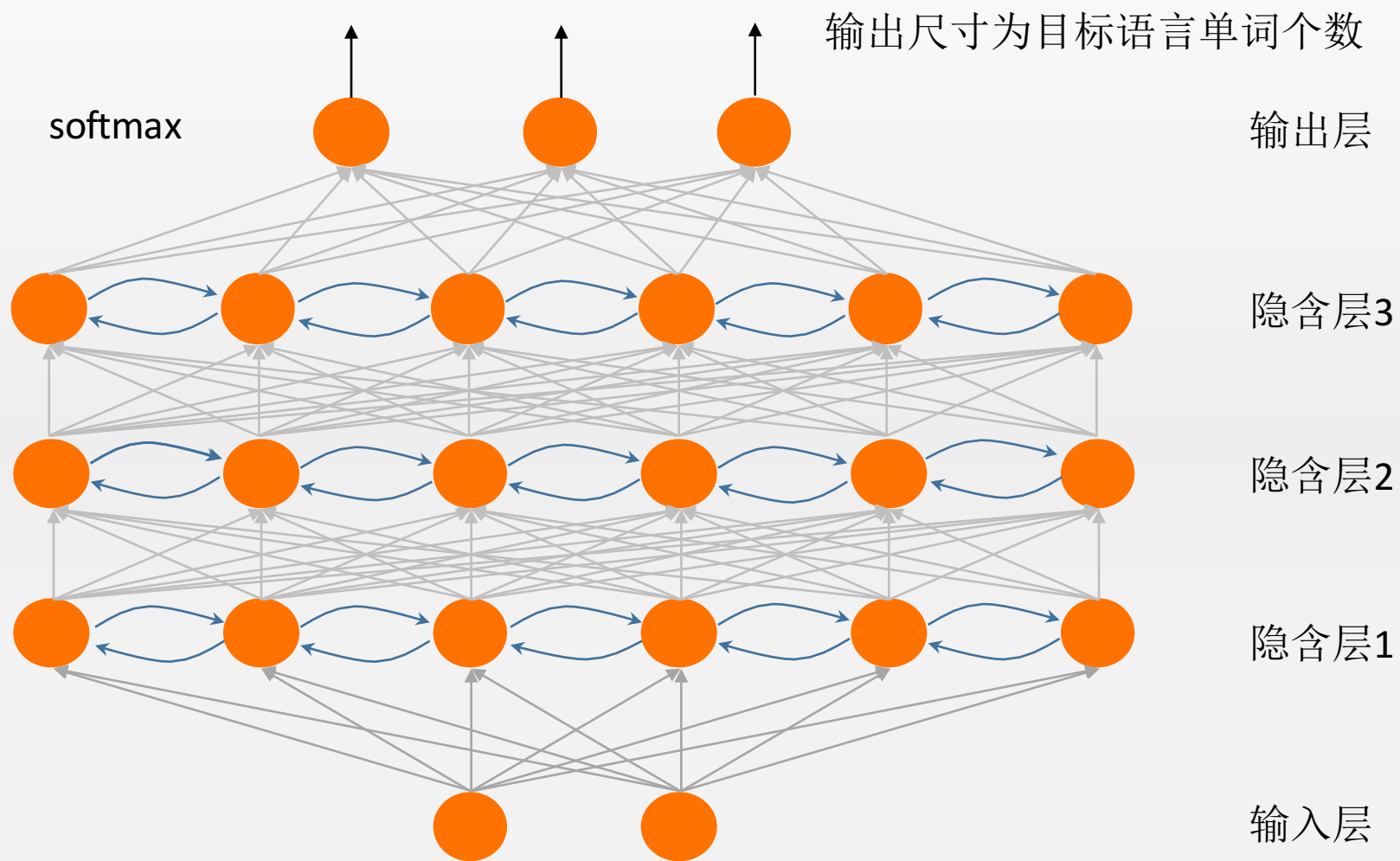


向量表示

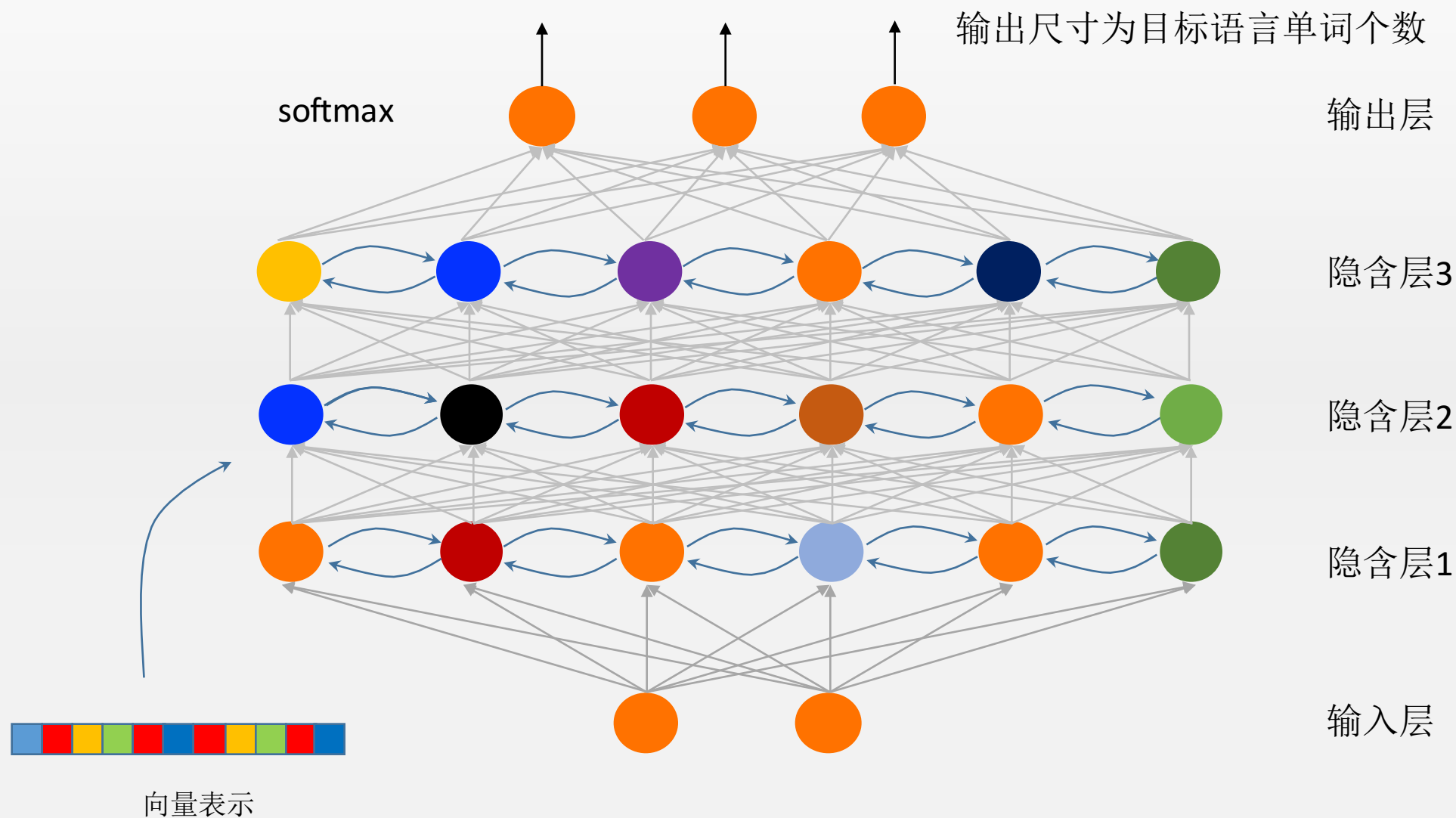
# 编码器解码器架构



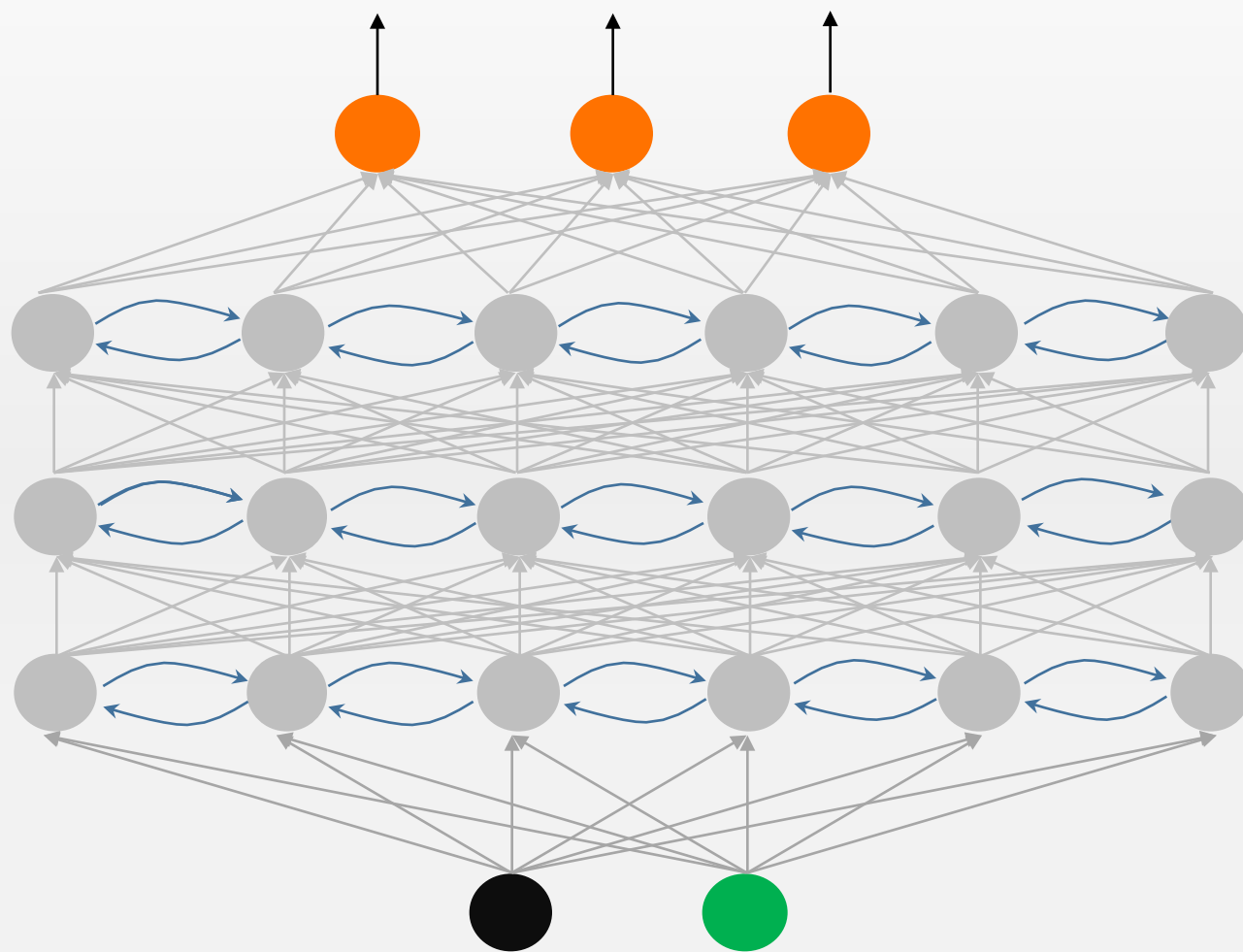
## 解码器：相同层次的RNN



# 解码器的初始状态：拷贝编码器的状态



$t=0$



SOS

输出层

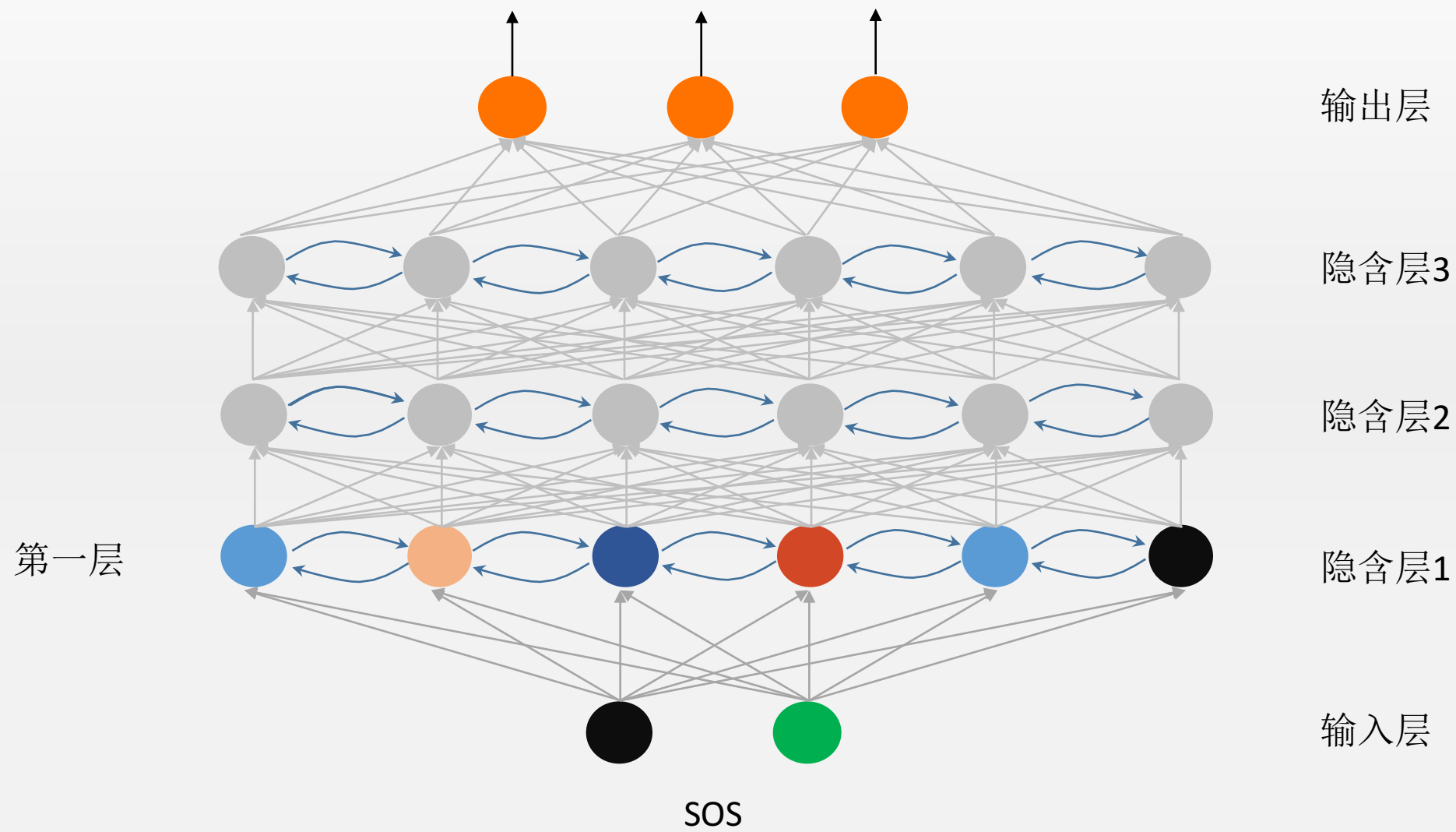
隐含层3

隐含层2

隐含层1

输入层

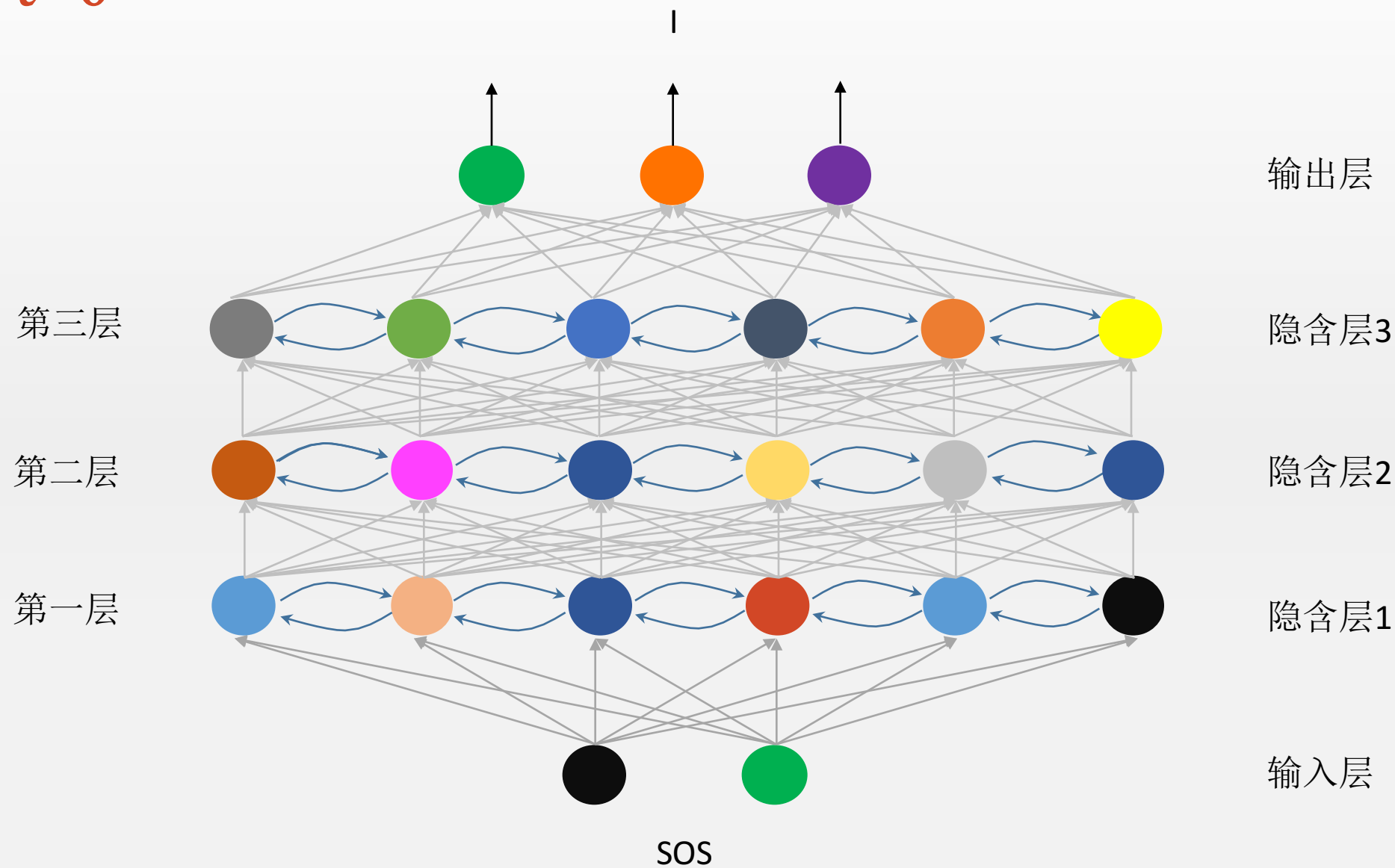
t=0





标准答案: I love this game<EOS>

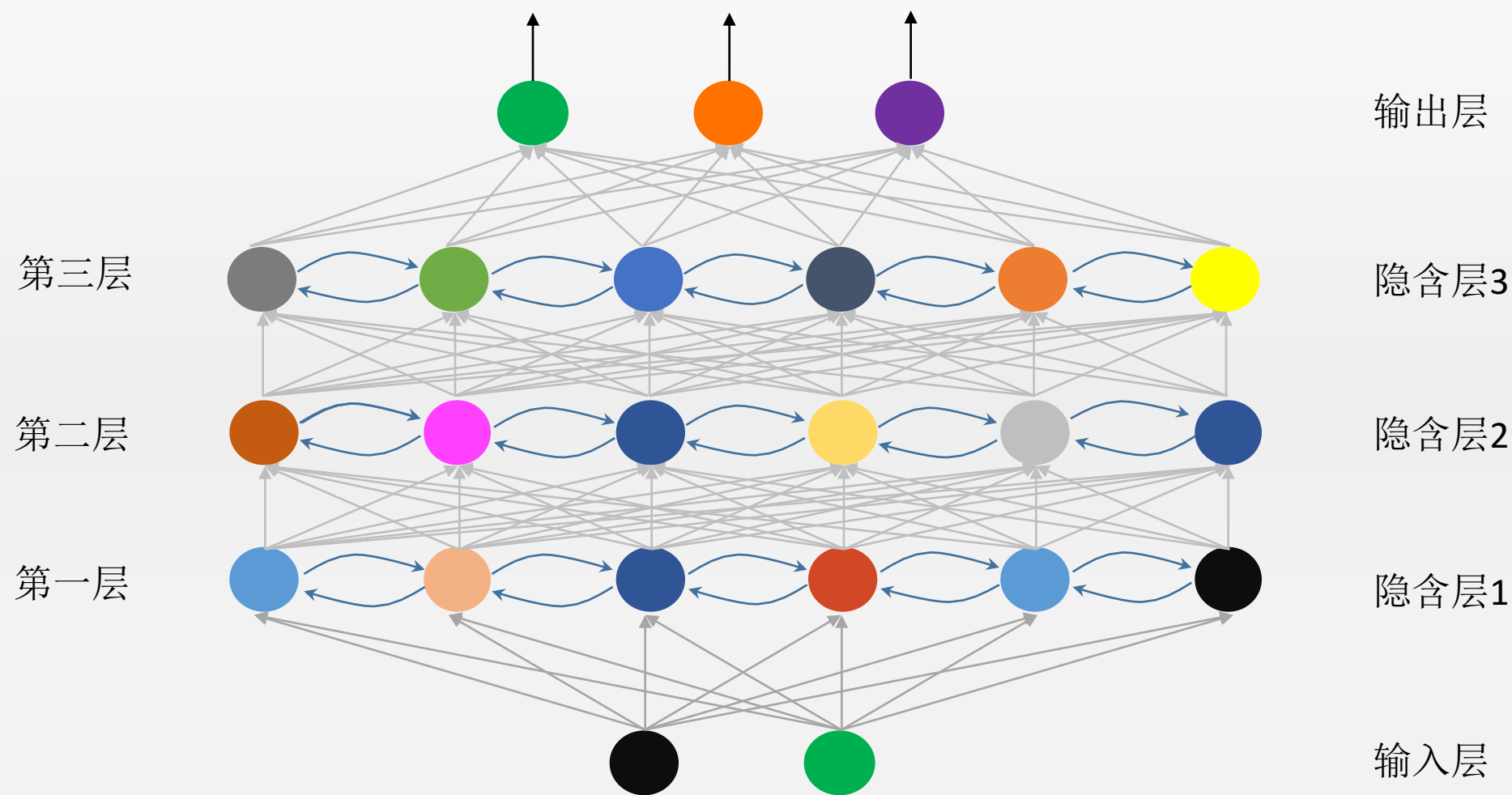
t=0



标准答案: I love this game<EOS>

t=1

like

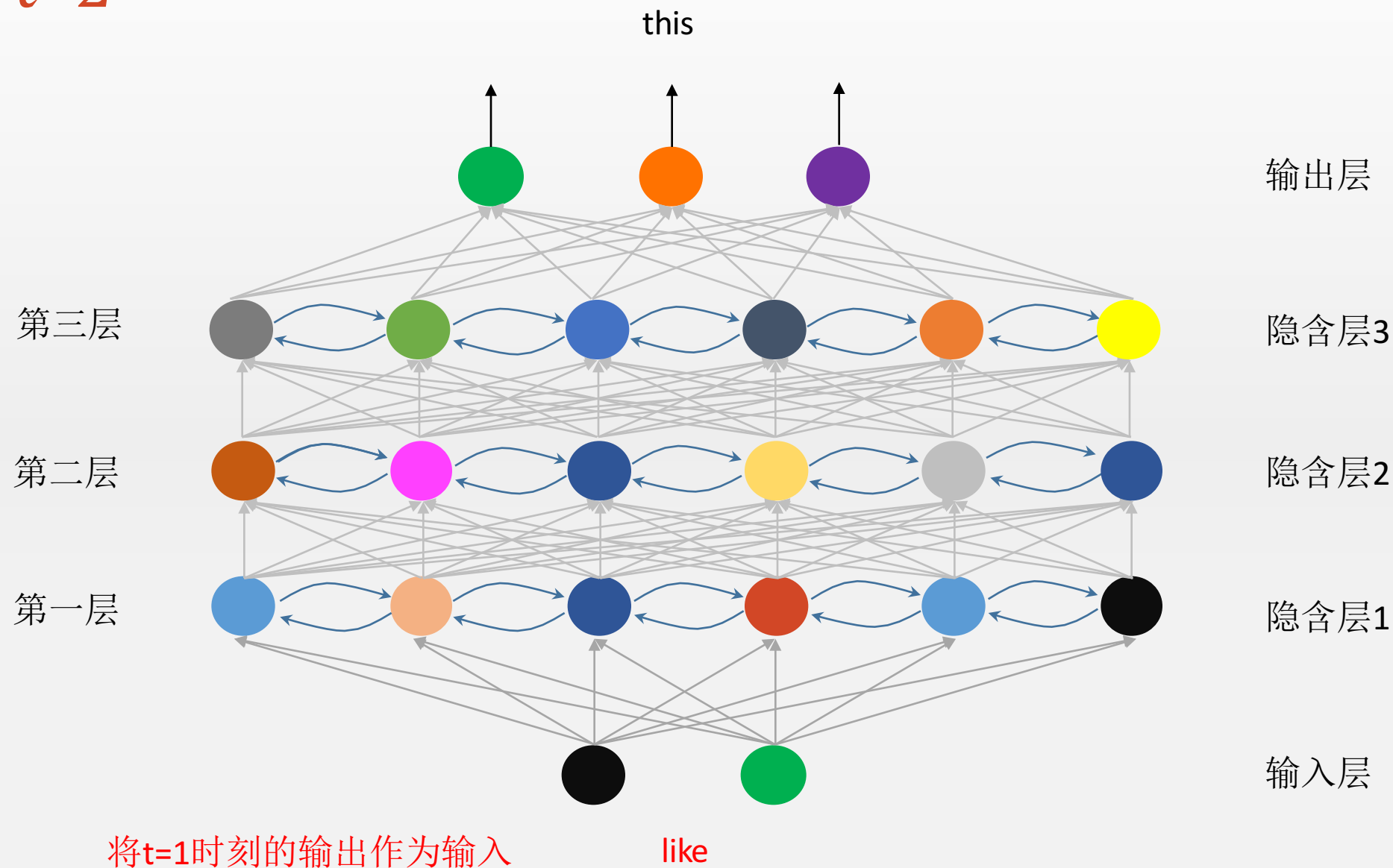


将t=0时刻的输出作为输入

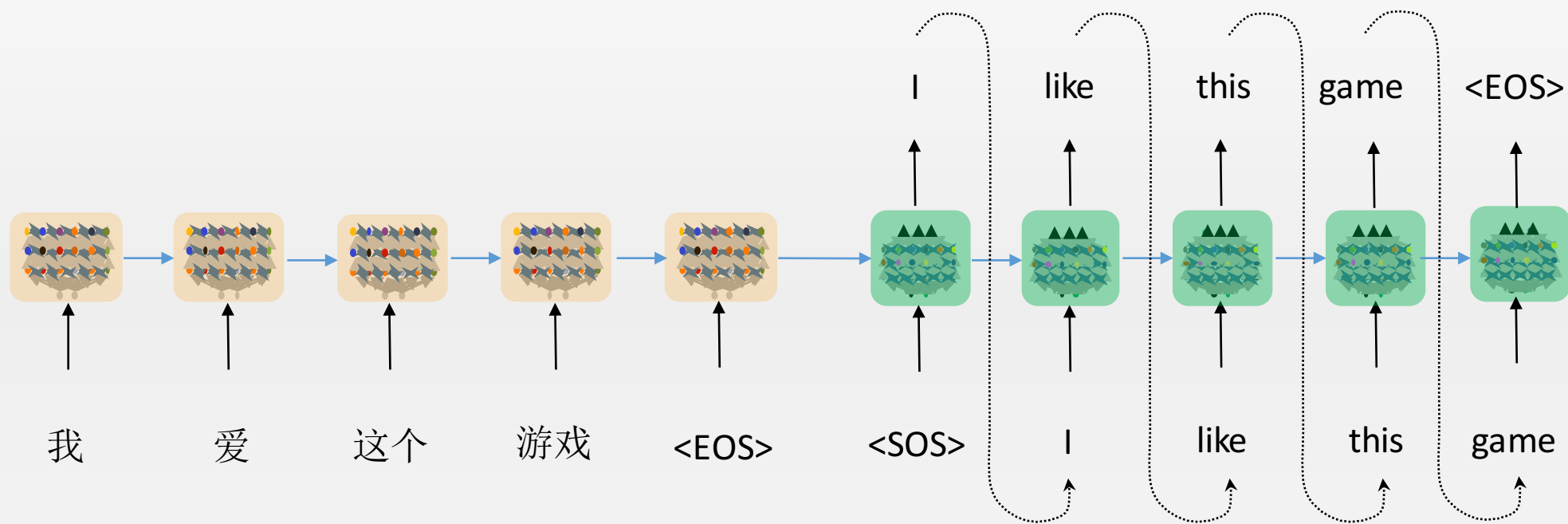
|

标准答案: I love this game<EOS>

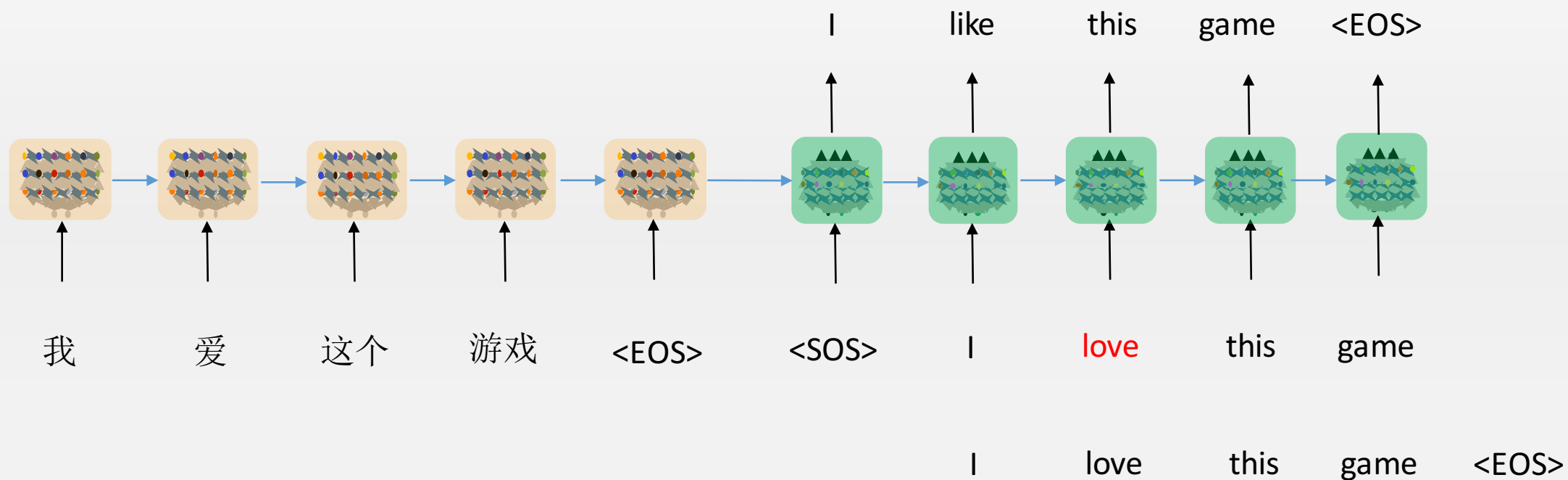
t=2



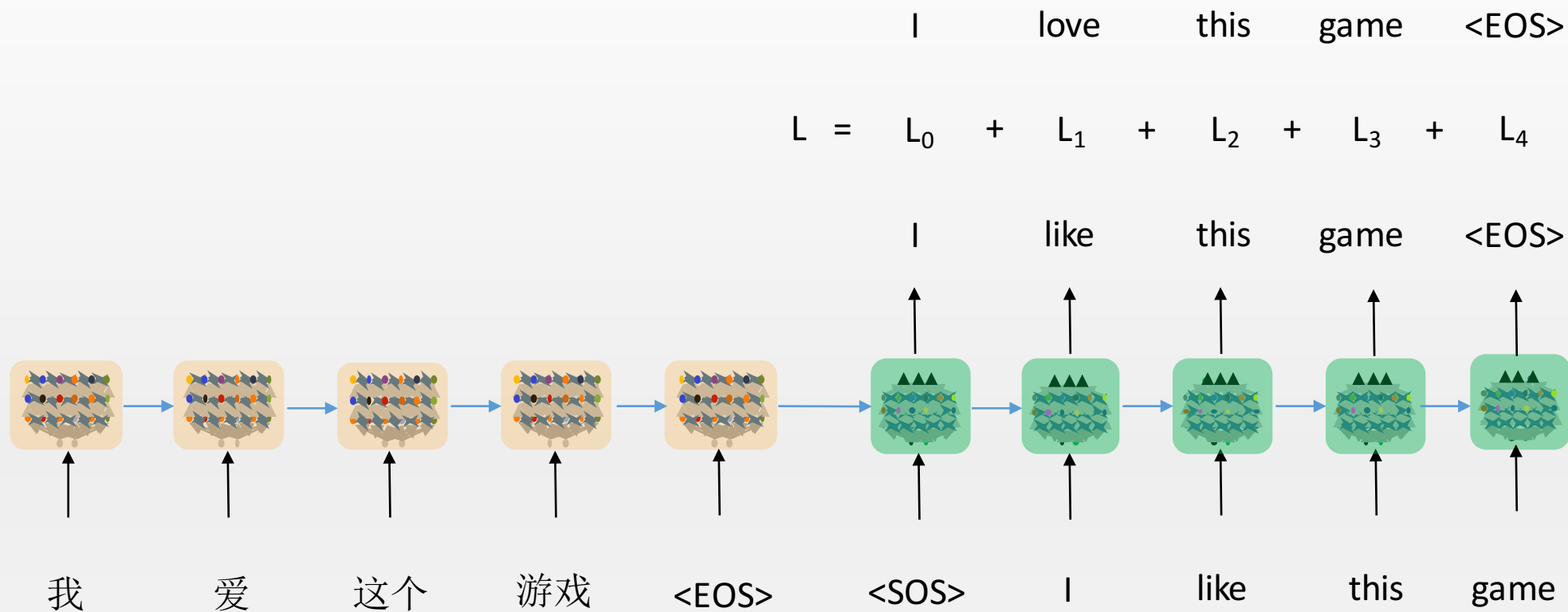
## 整个流程



## 整个流程



# 损失函数

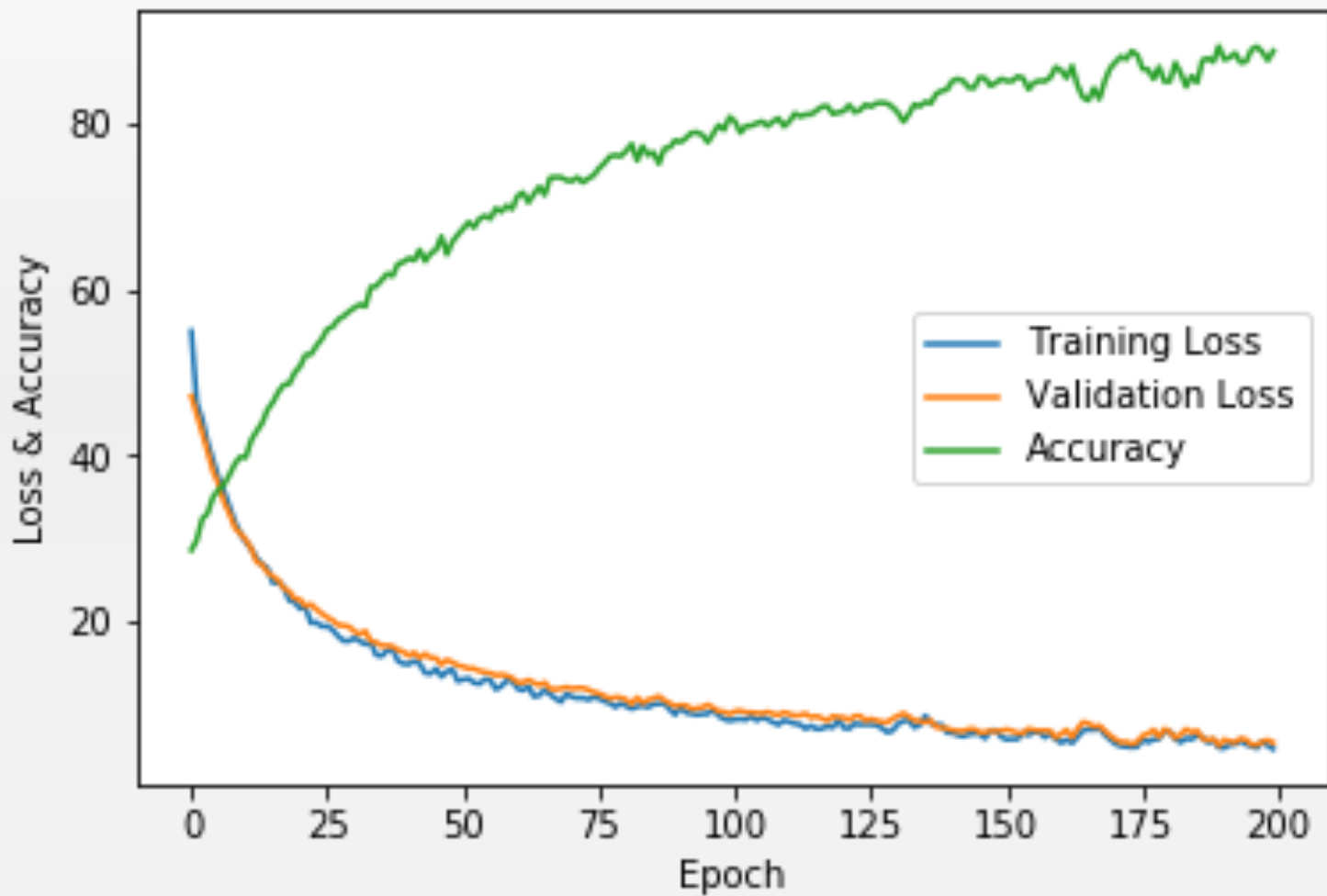


## 编码器 - 解码器架构小结

- 编码器和解码器都是多层RNN网络
- 输入完整源句子后，编码器的所有隐含层单元的状态就是对源句子的编码
- 将编码器的隐含层单元付给解码器的隐含层单元作为初始状态完成编码的传递
- 通过评估解码器每一步的输出损失求和得到整体架构的损失函数
- 利用backward和梯度下降算法可以自动同时训练编码器 - 解码器

# 训练结果

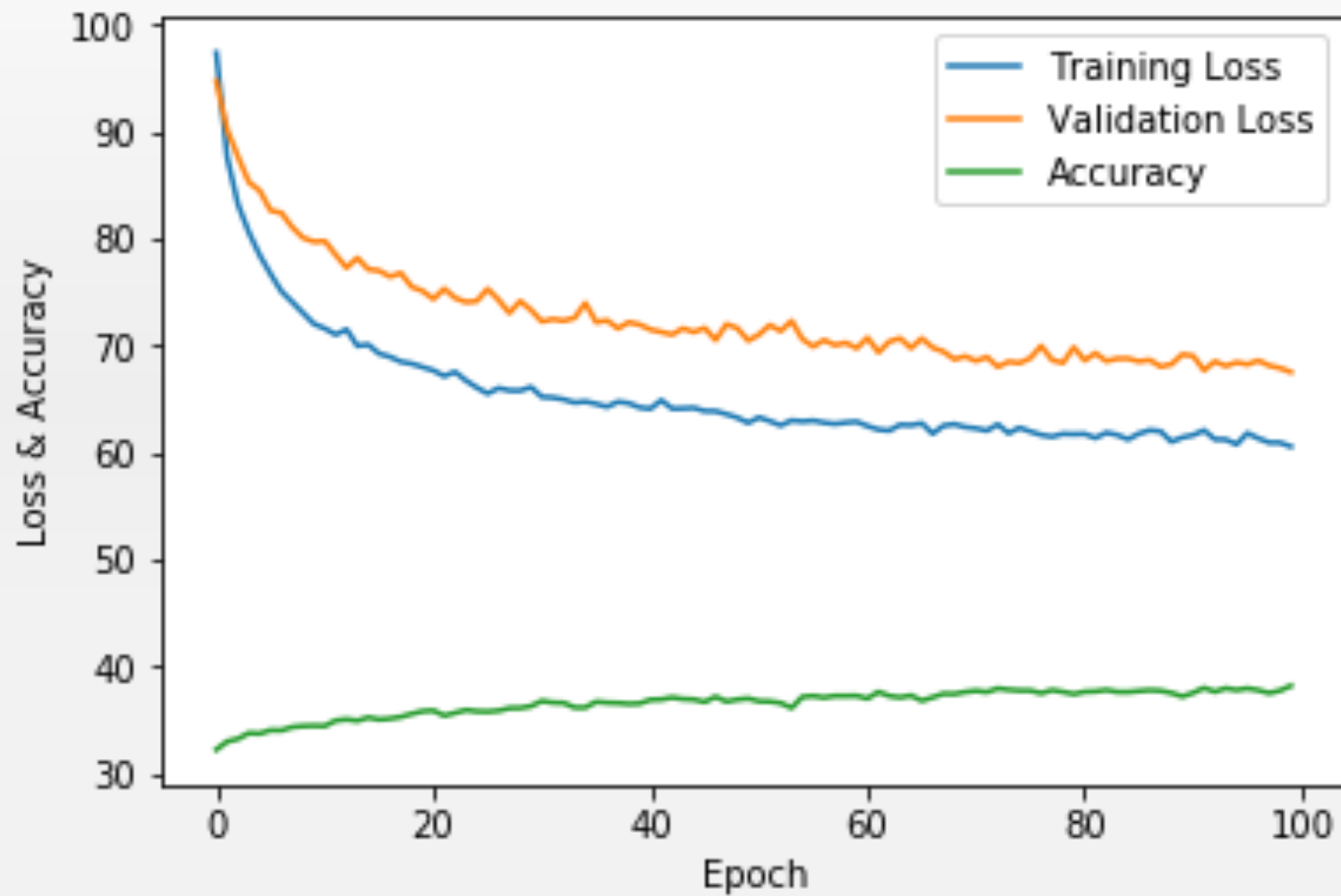
单元数16，单词截断长度：10





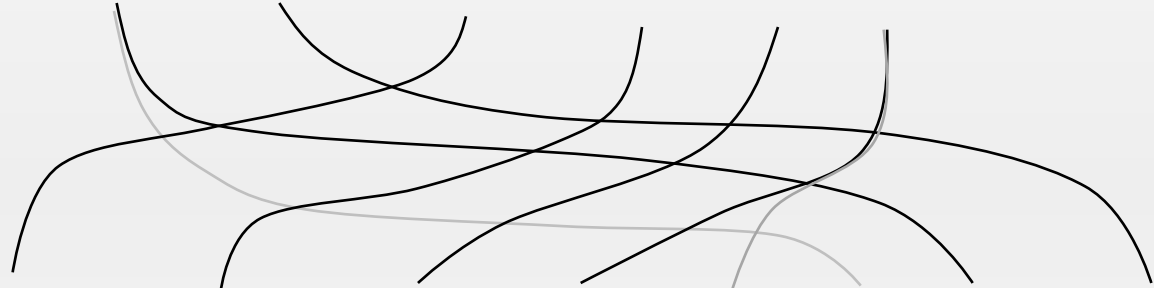
# 训练结果

单元数16，单词截断长度：20



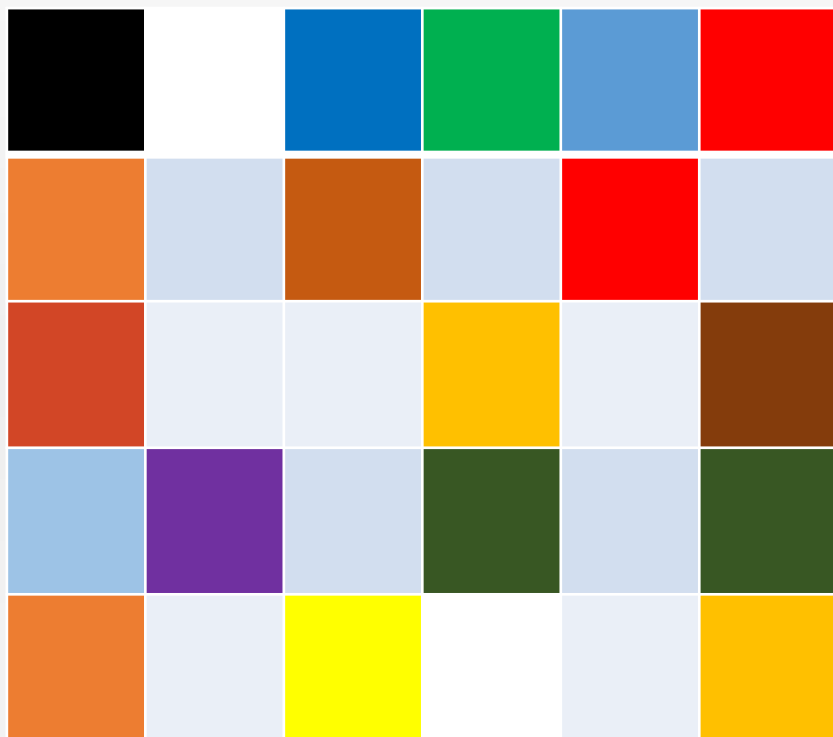
但这还远远不够

最近 几年，经济 增长 在 放缓

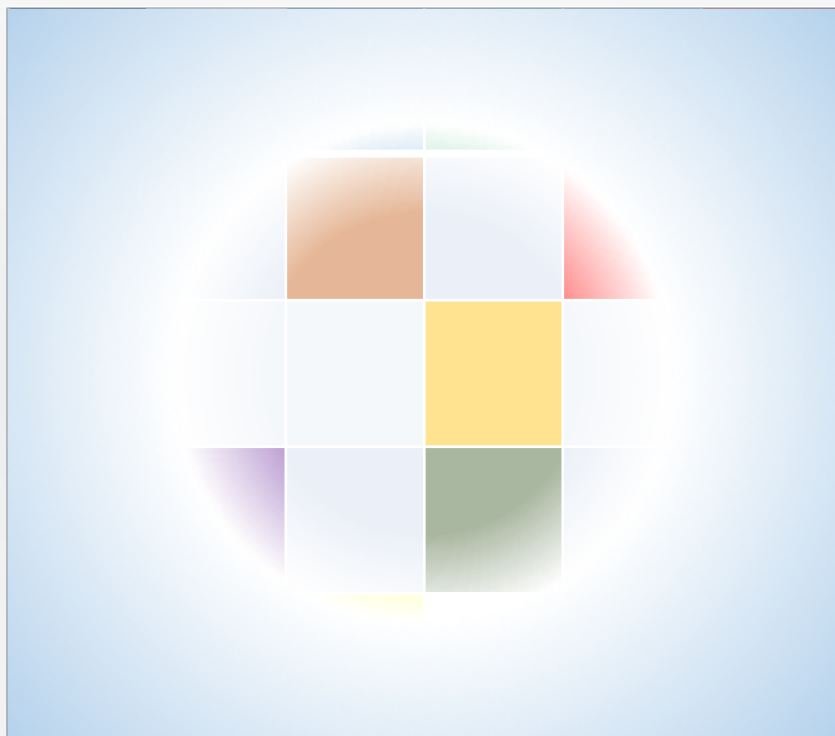


Economic growth has slowed down in recent years

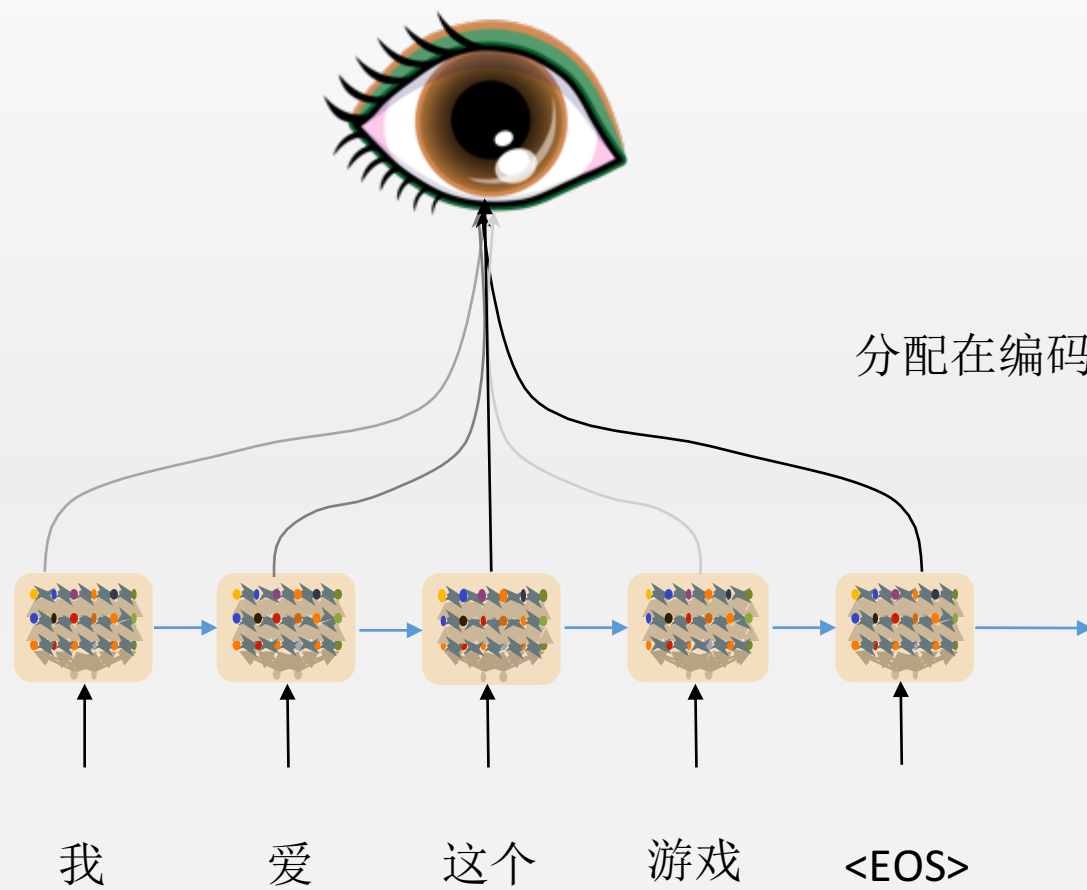
# 注意力机制



# 注意力机制

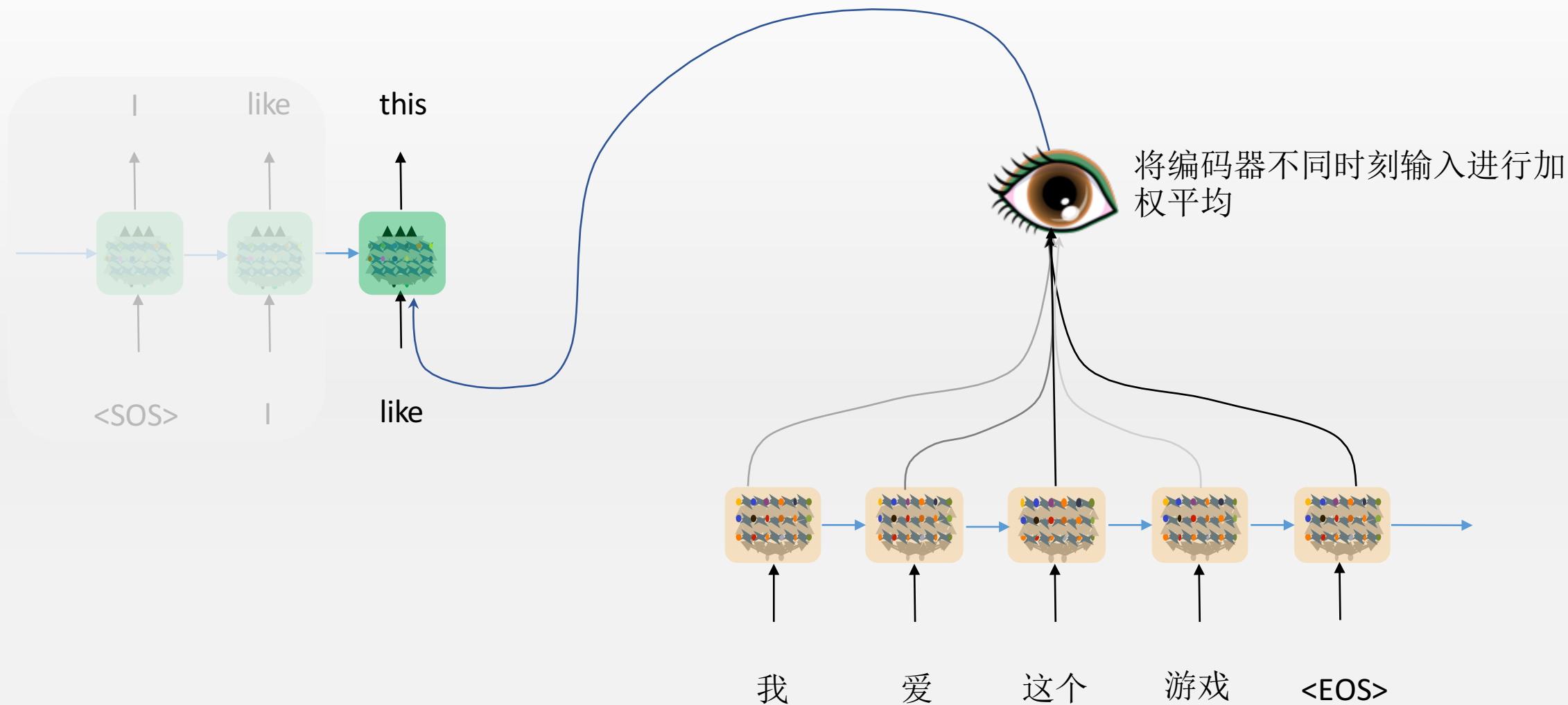


# 注意力机制



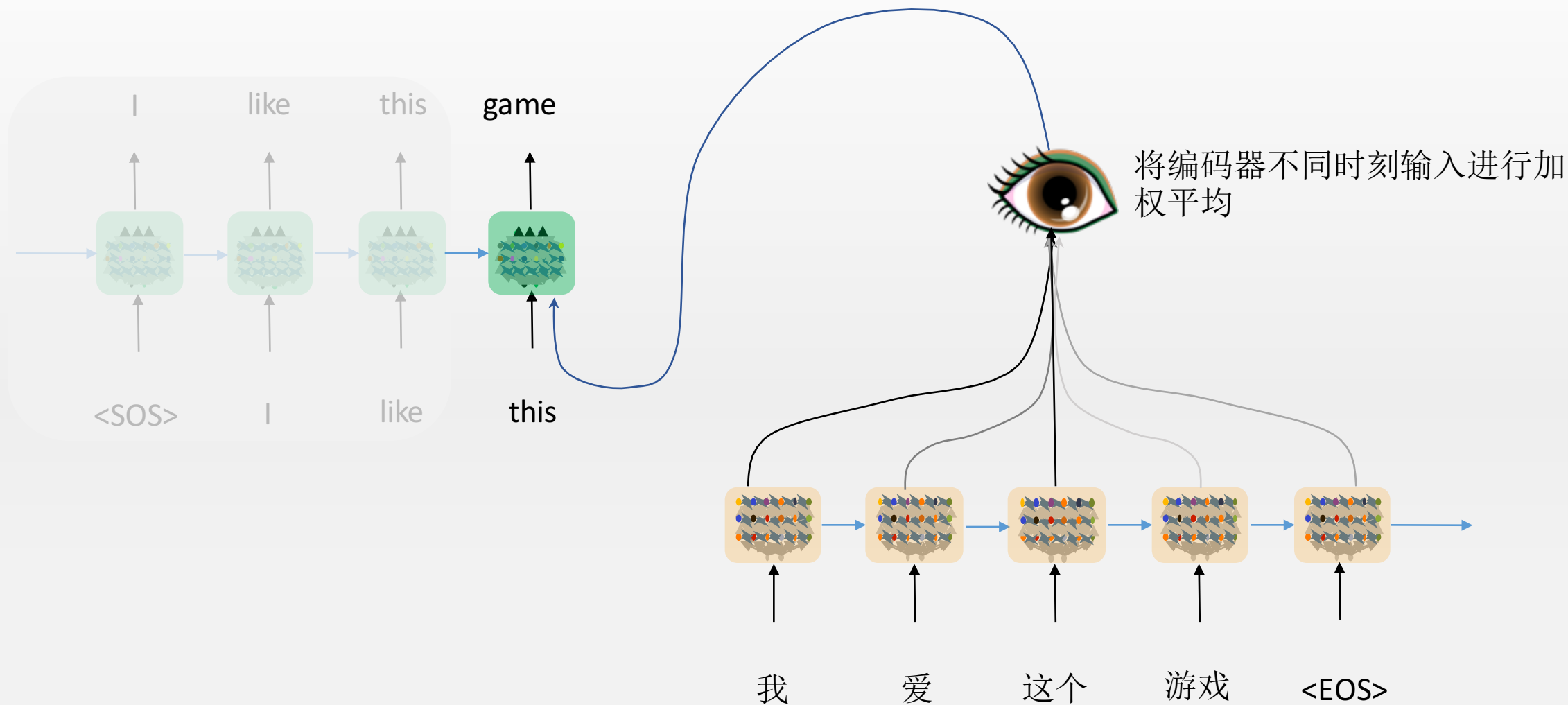
# 注意力机制

结果输入给当前时刻 $t$ 的解码器



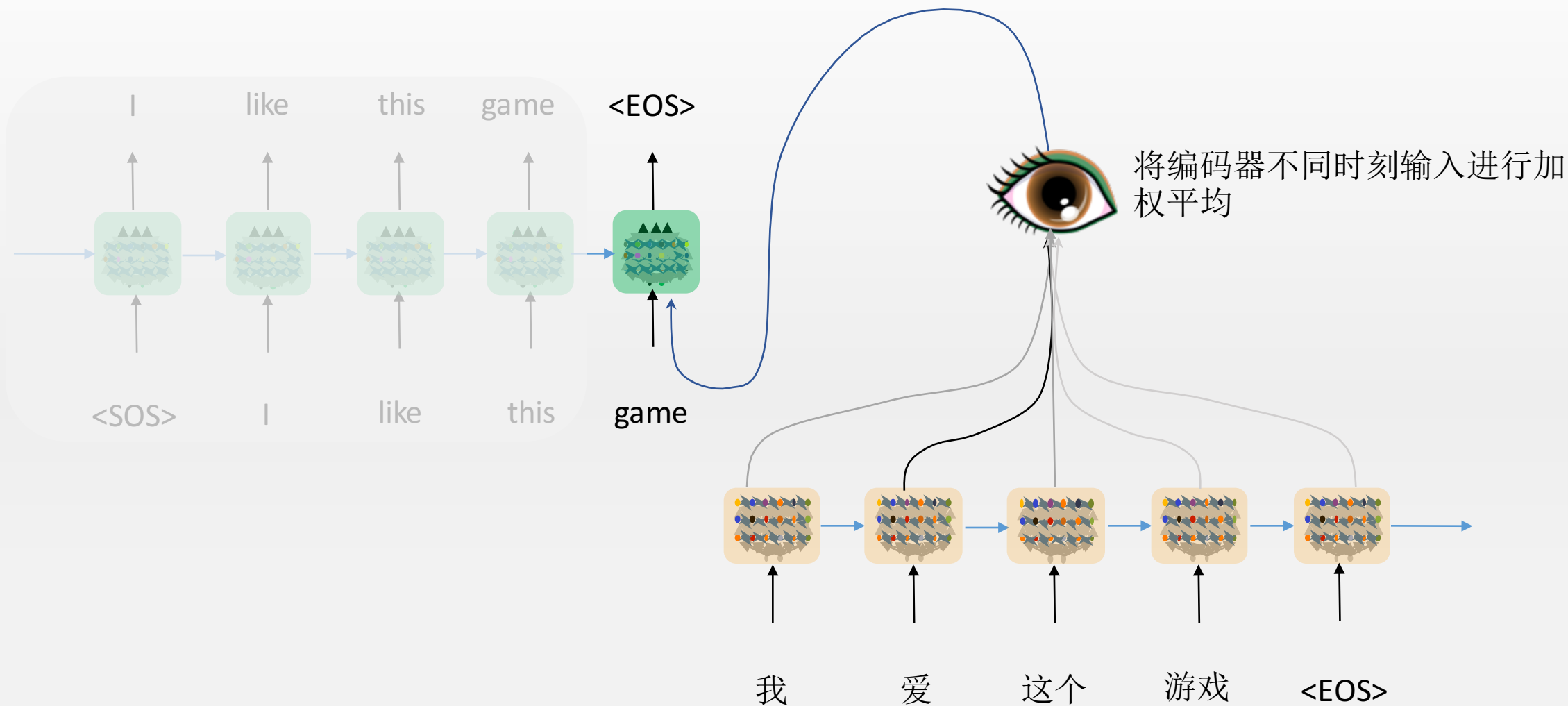
# 注意力机制

结果输入给当前时刻 $t$ 的解码器



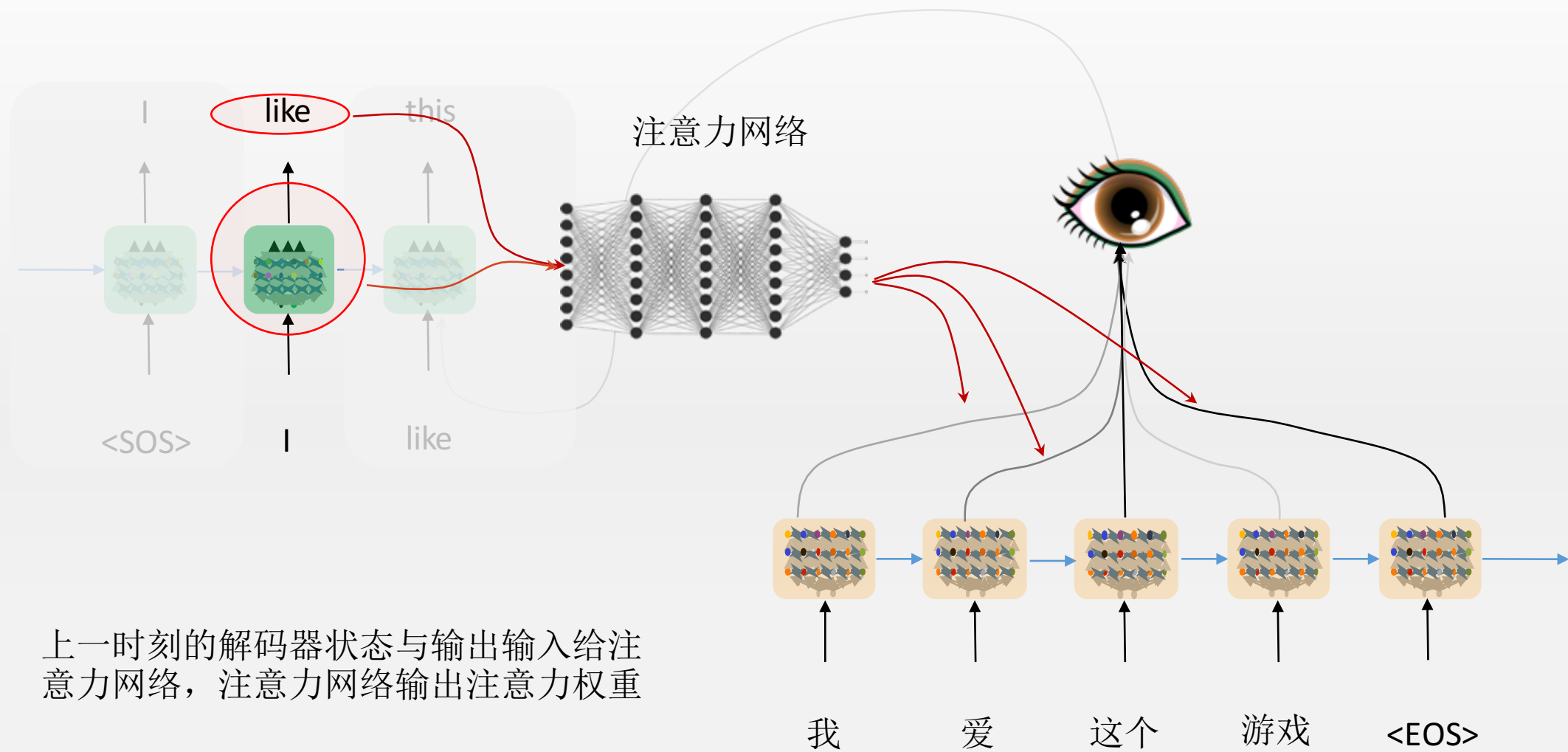
# 注意力机制

结果输入给当前时刻 $t$ 的解码器

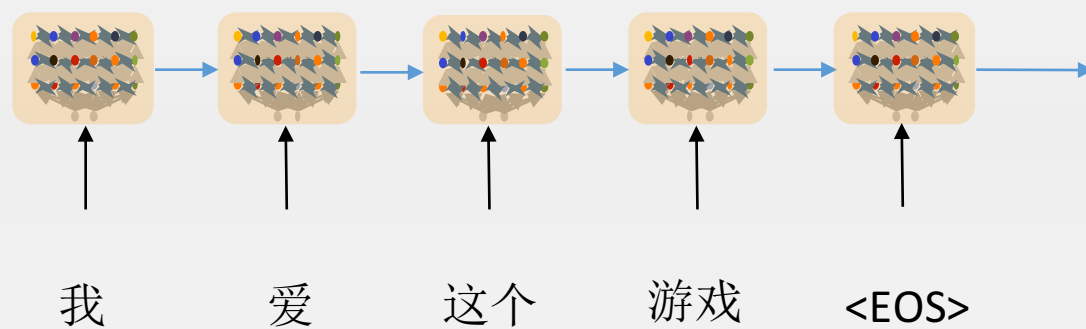




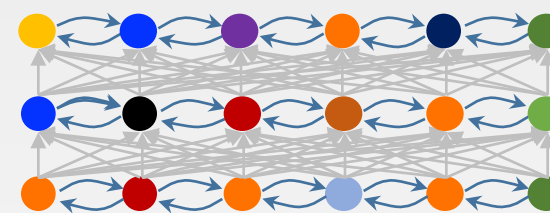
# 注意力机制



# 整个流程



向量表示1

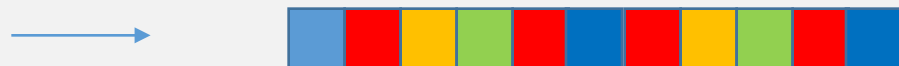
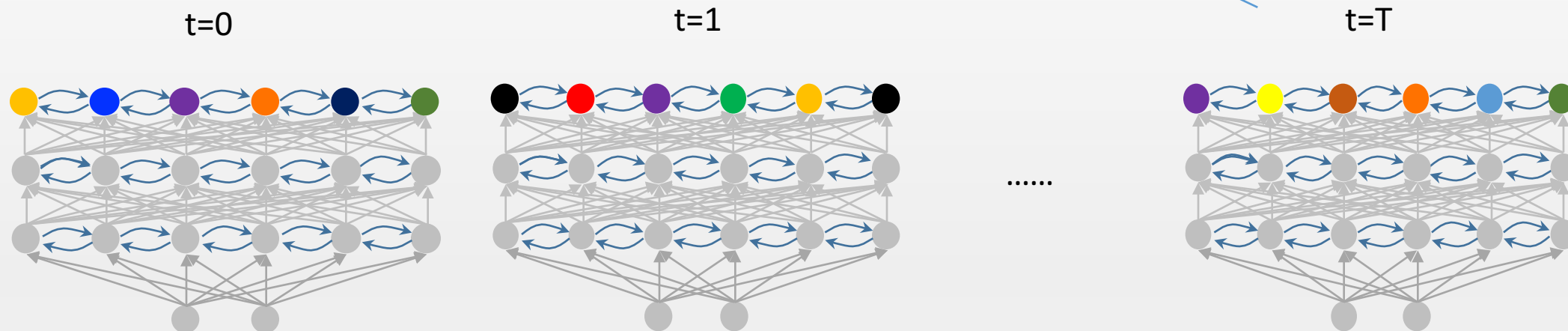


向量表示2

t=T时刻编码器所有层的状态传递给解码器

# 输出

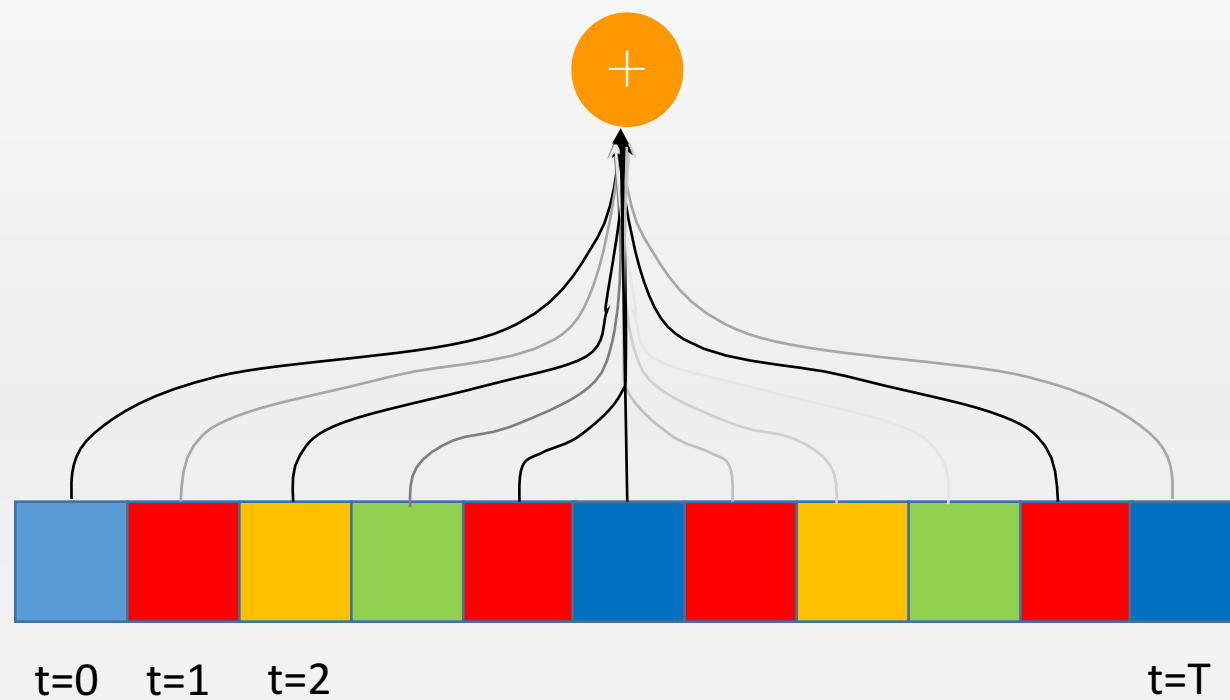
所有时刻的最后一层



向量表示

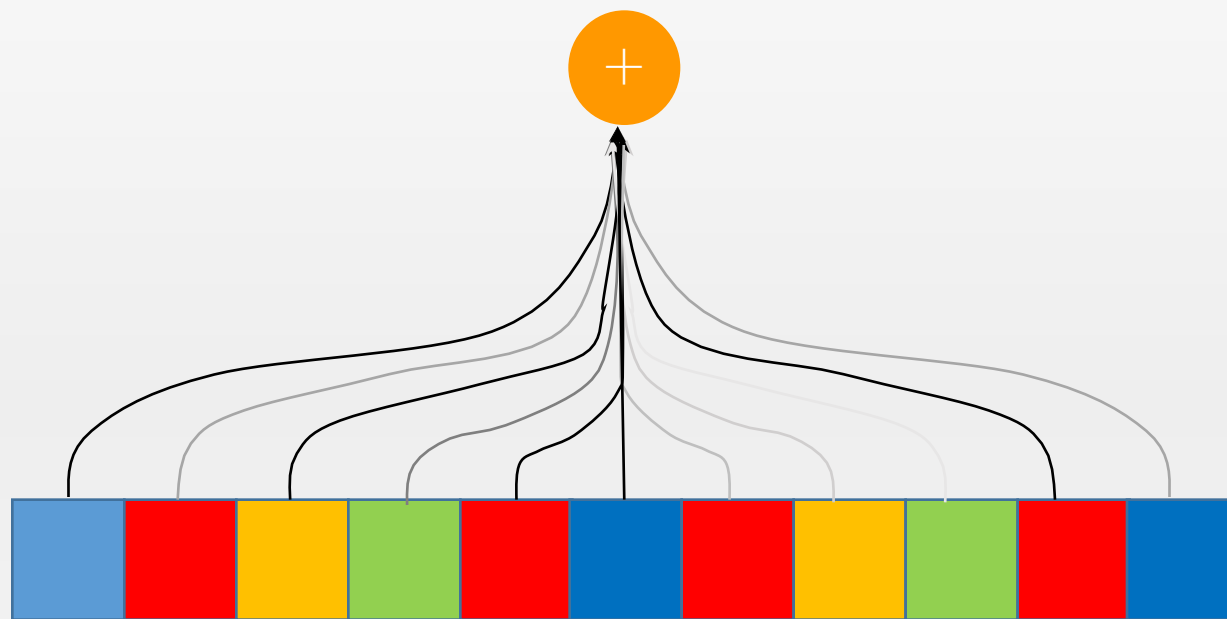
# 注意力机制

编码器的:



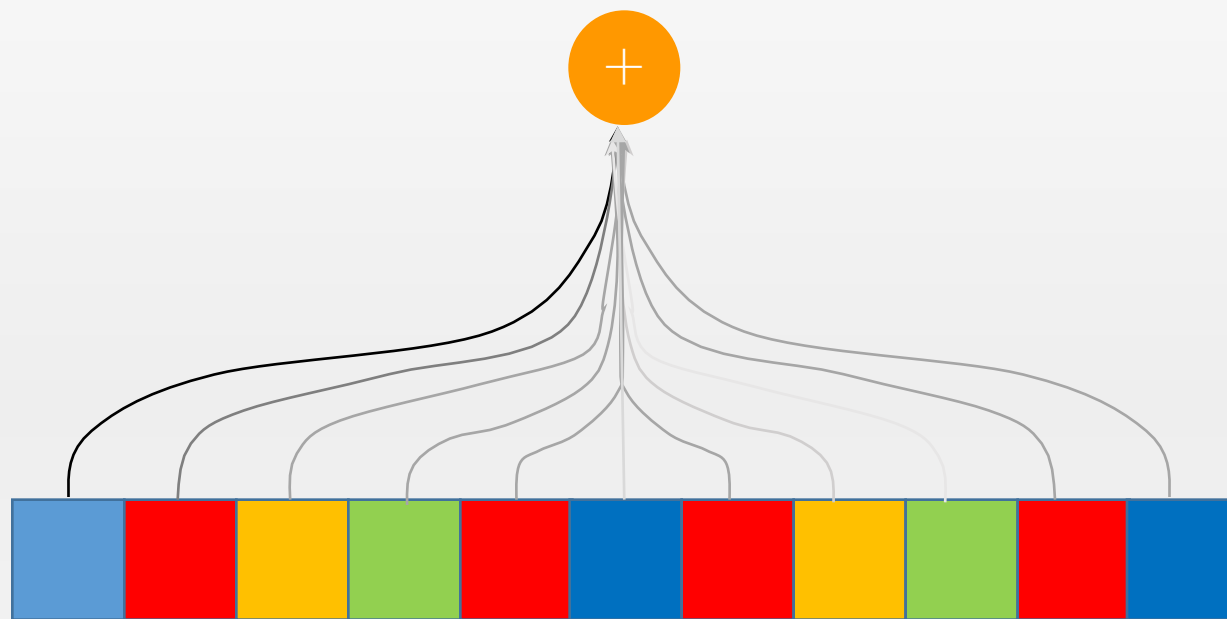
向量表示

## 注意力机制 $t=0$



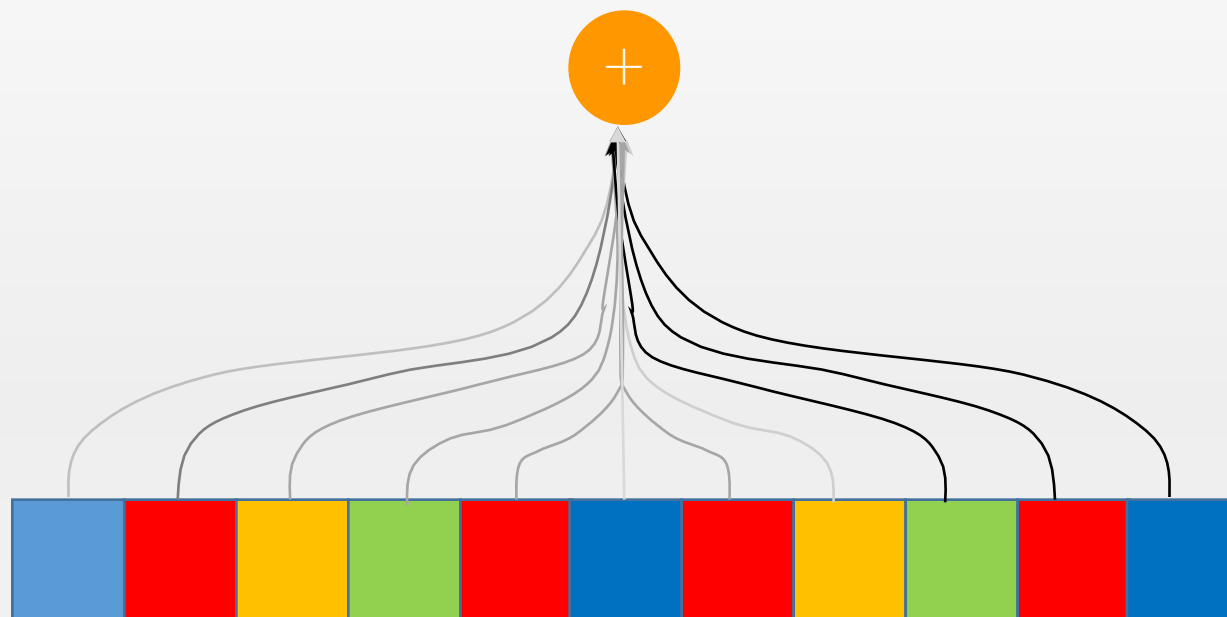
向量表示

## 注意力机制 $t=1$



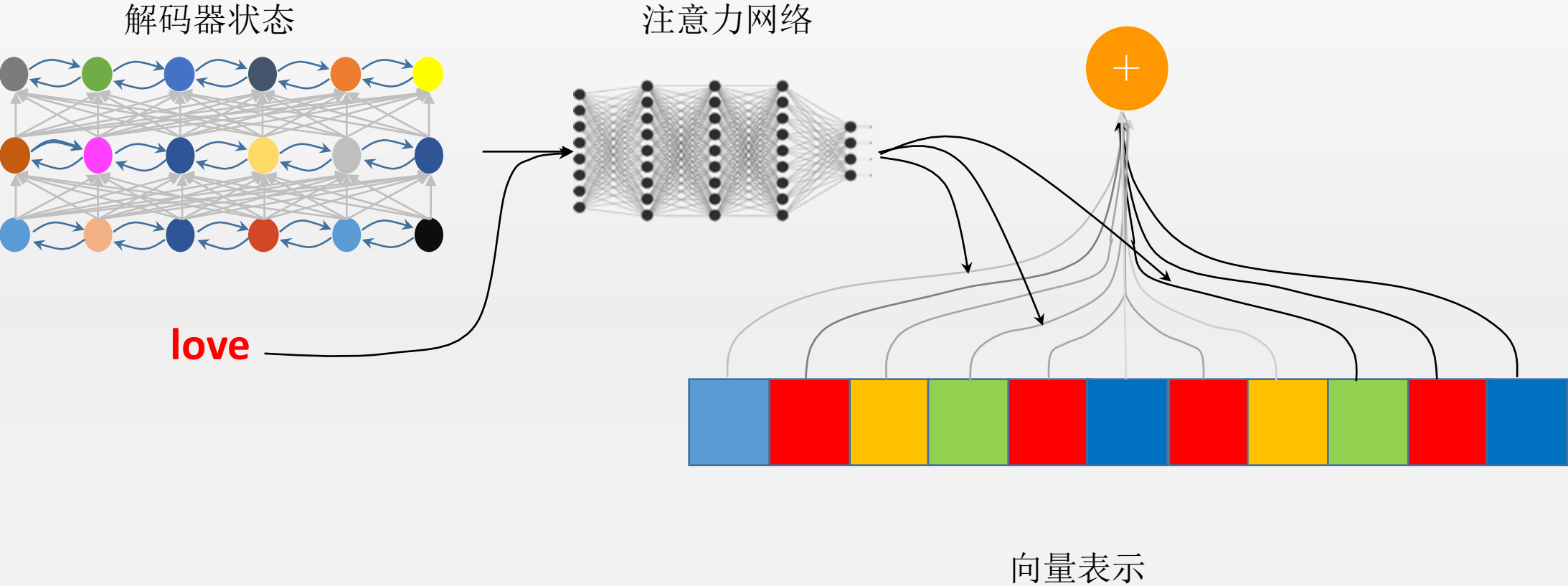
向量表示

## 注意力机制 $t=2$



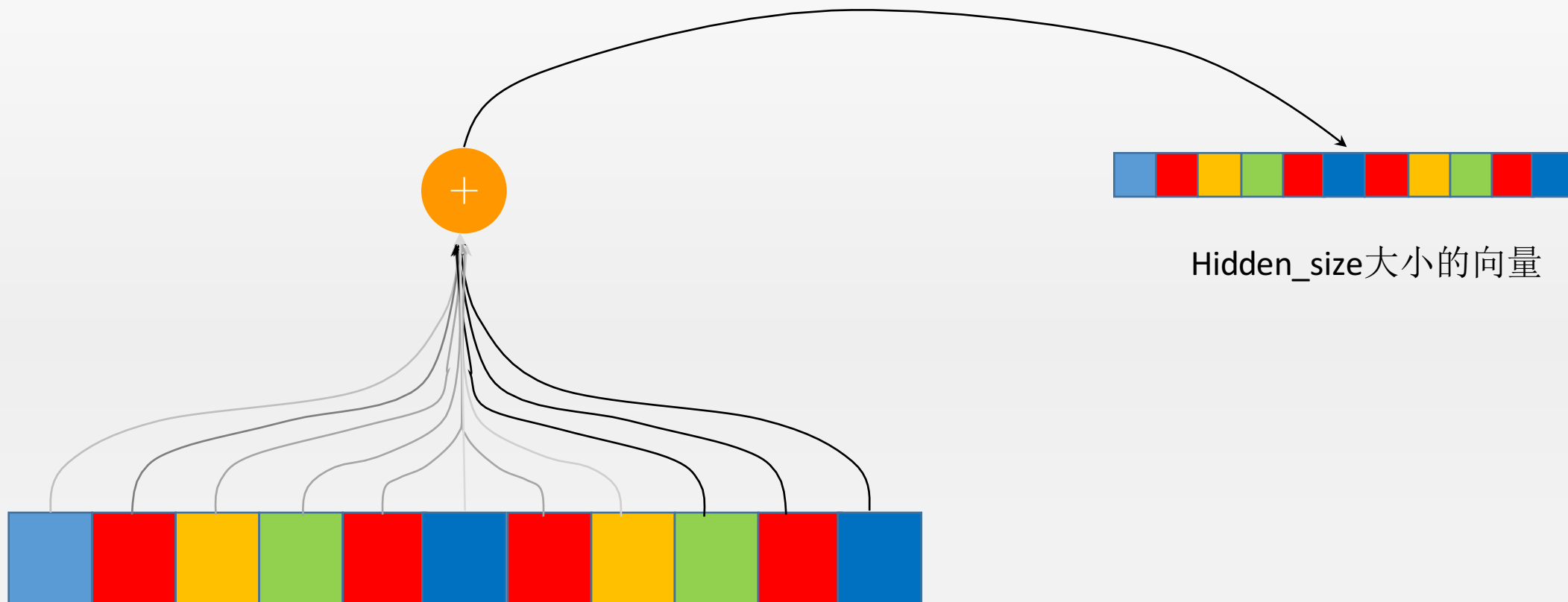
向量表示

# 注意力由解码器和输入字符决定





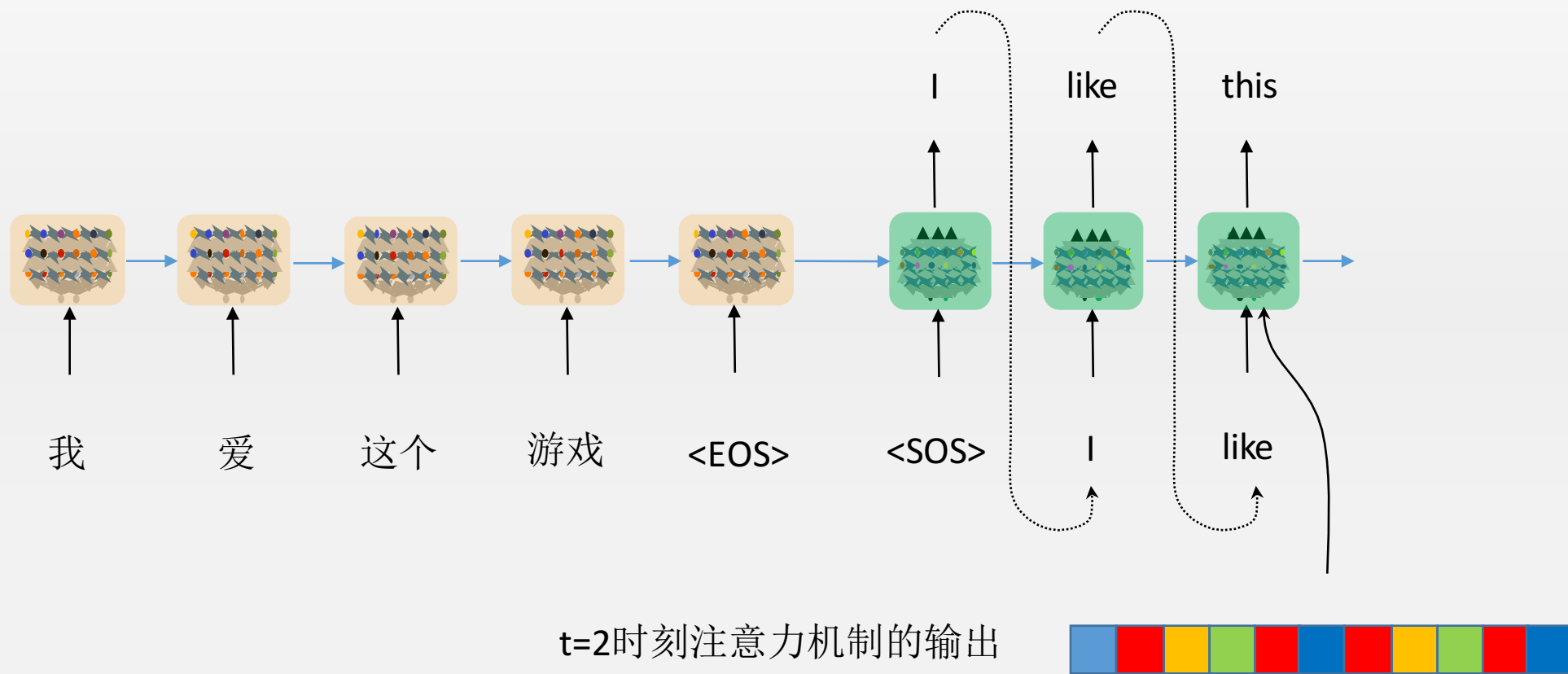
# 输出



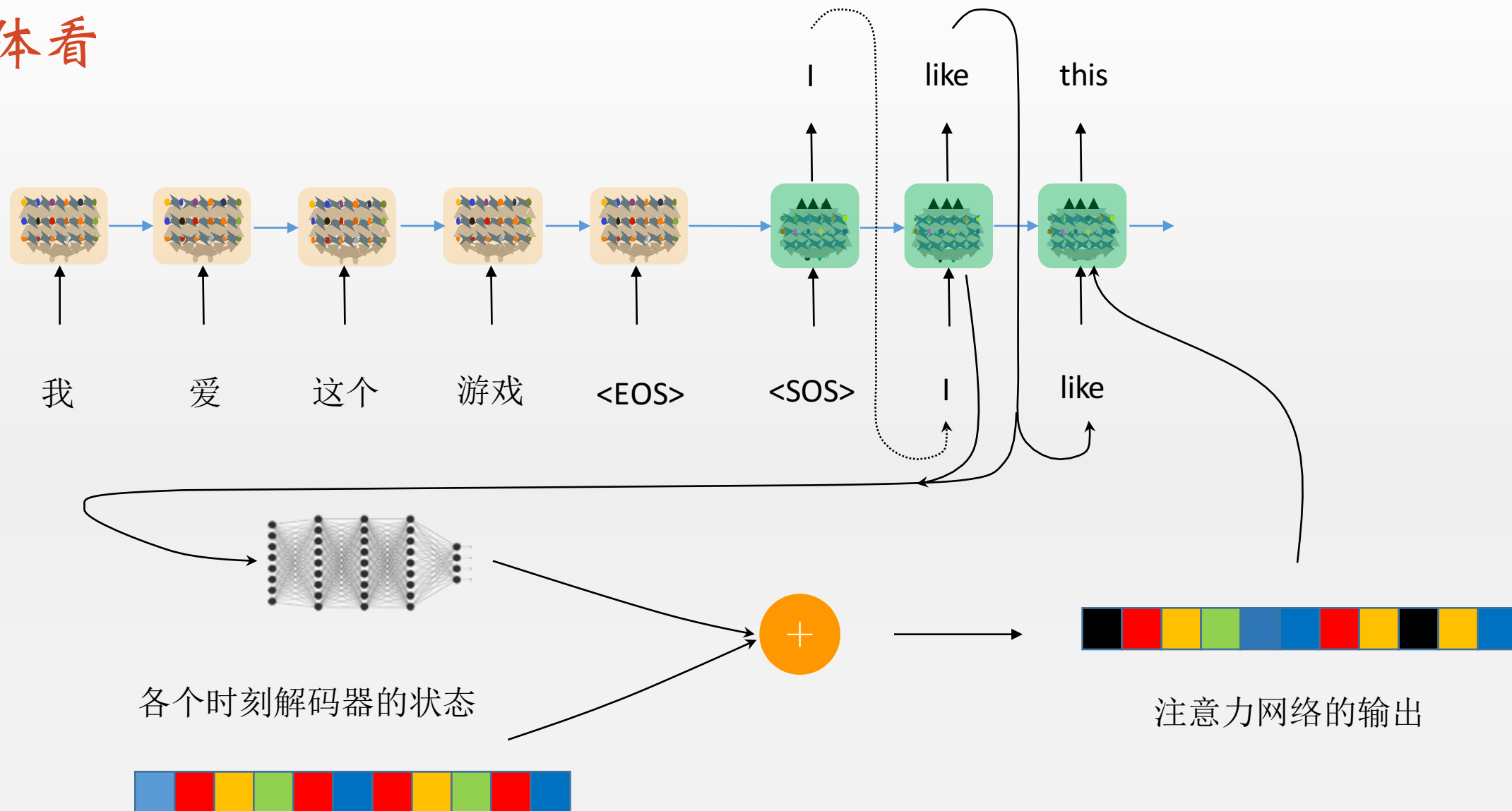
向量表示

Hidden\_size大小的向量

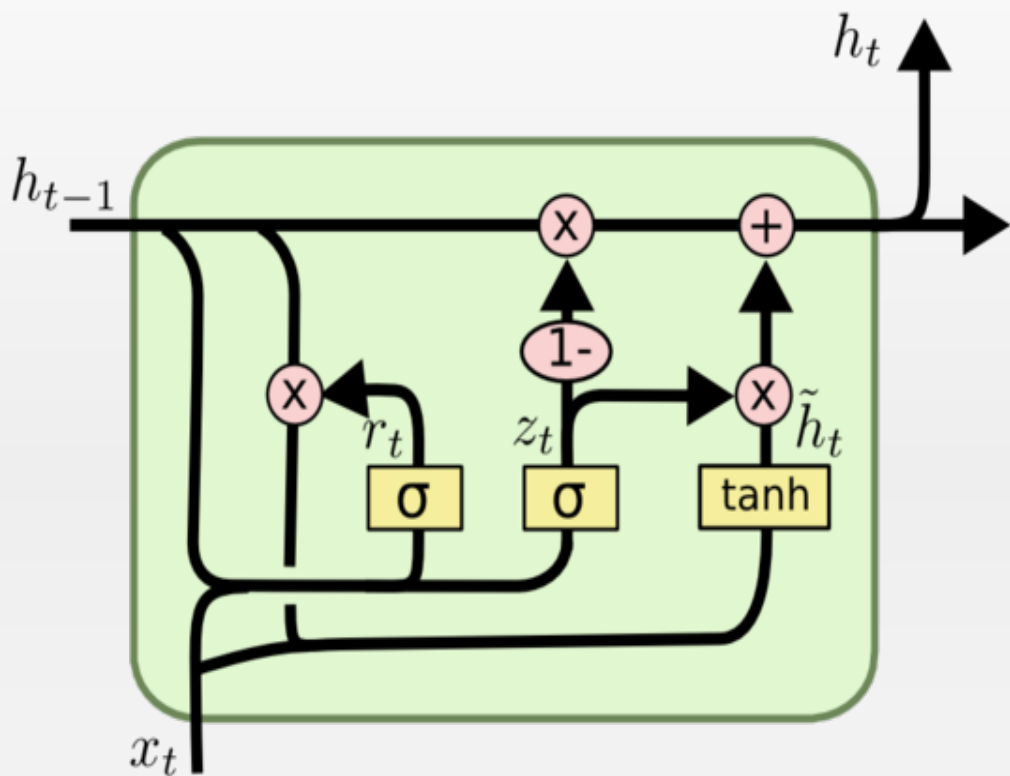
下一步



# 总体看



## GRU单元



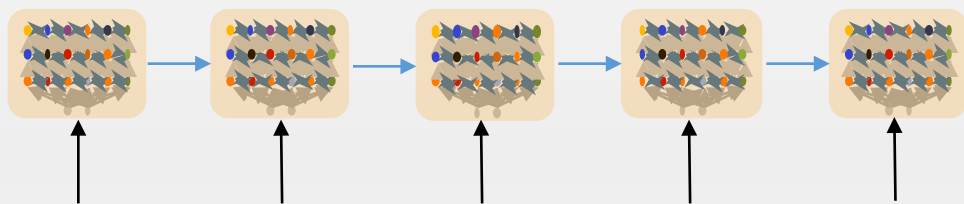
$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

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

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

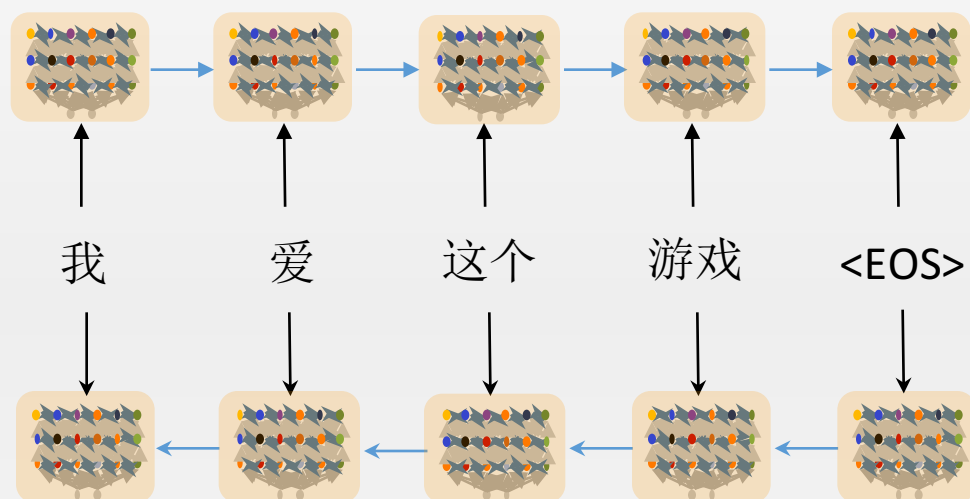
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

# 双向GRU/LSTM

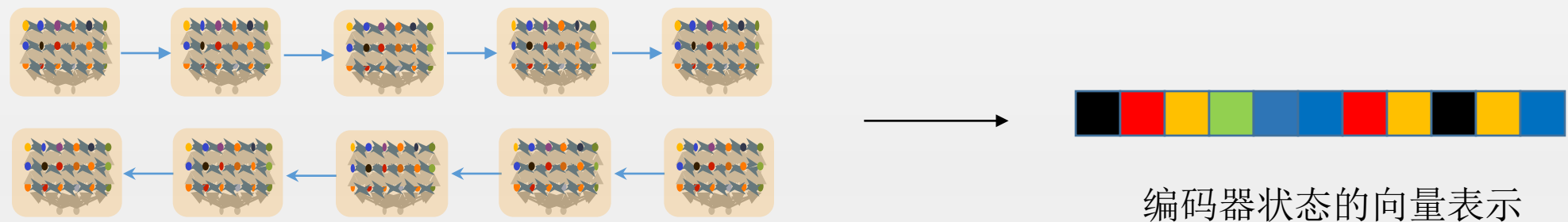


$t=s$ 时刻的状态取决于过去  
但翻译的时候，我们也要考虑未来

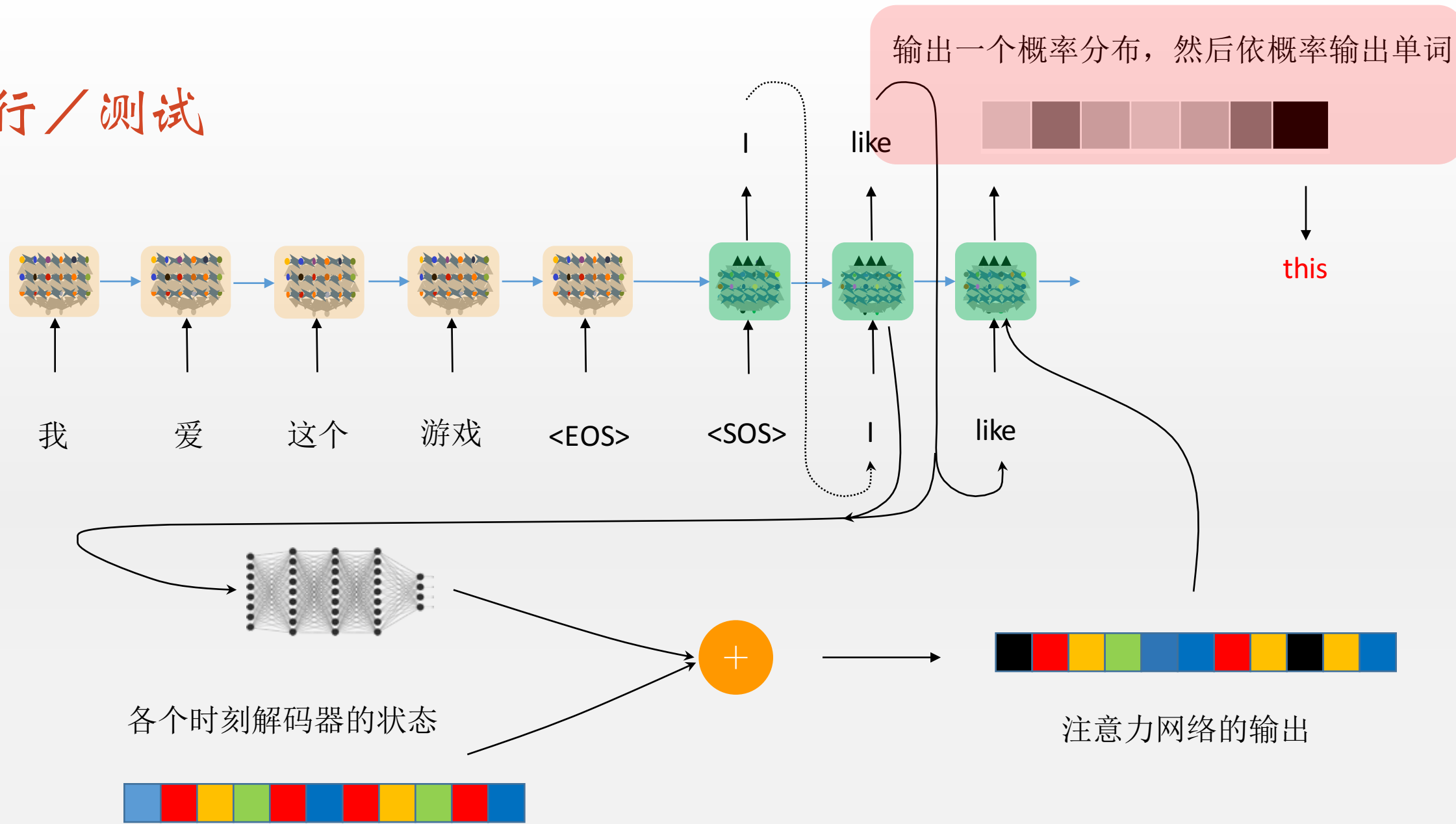
# 双向GRU/LSTM



# 双向GRU/LSTM

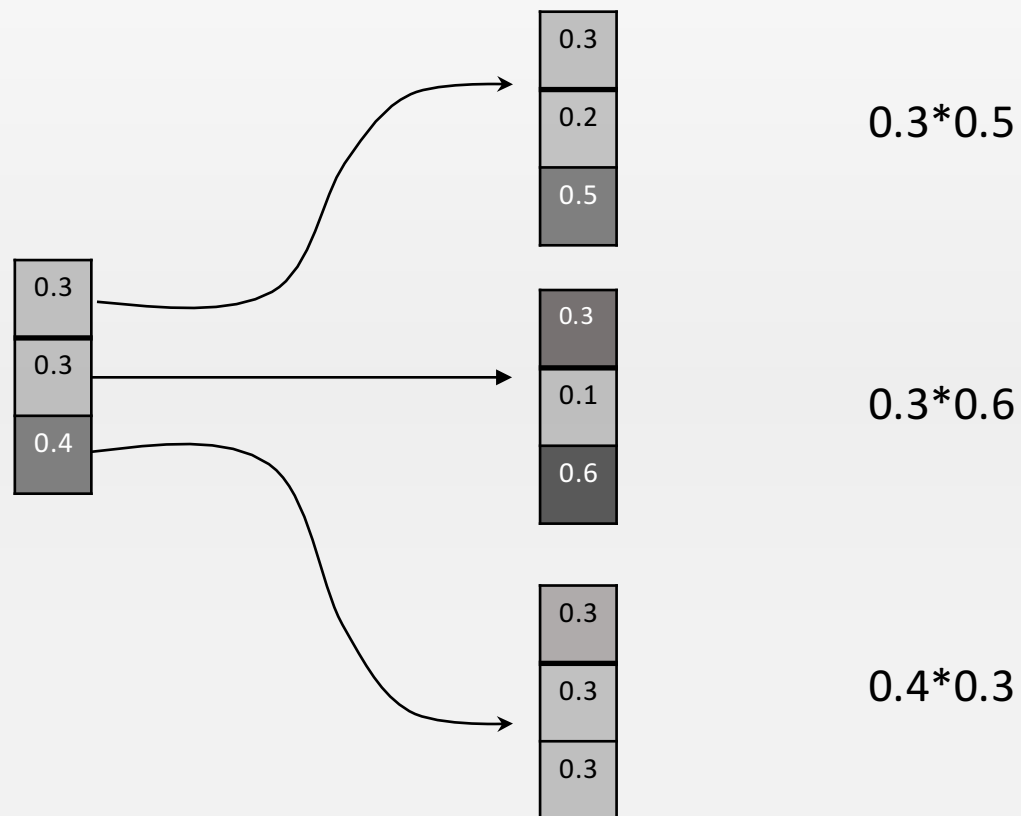


运行 / 测试

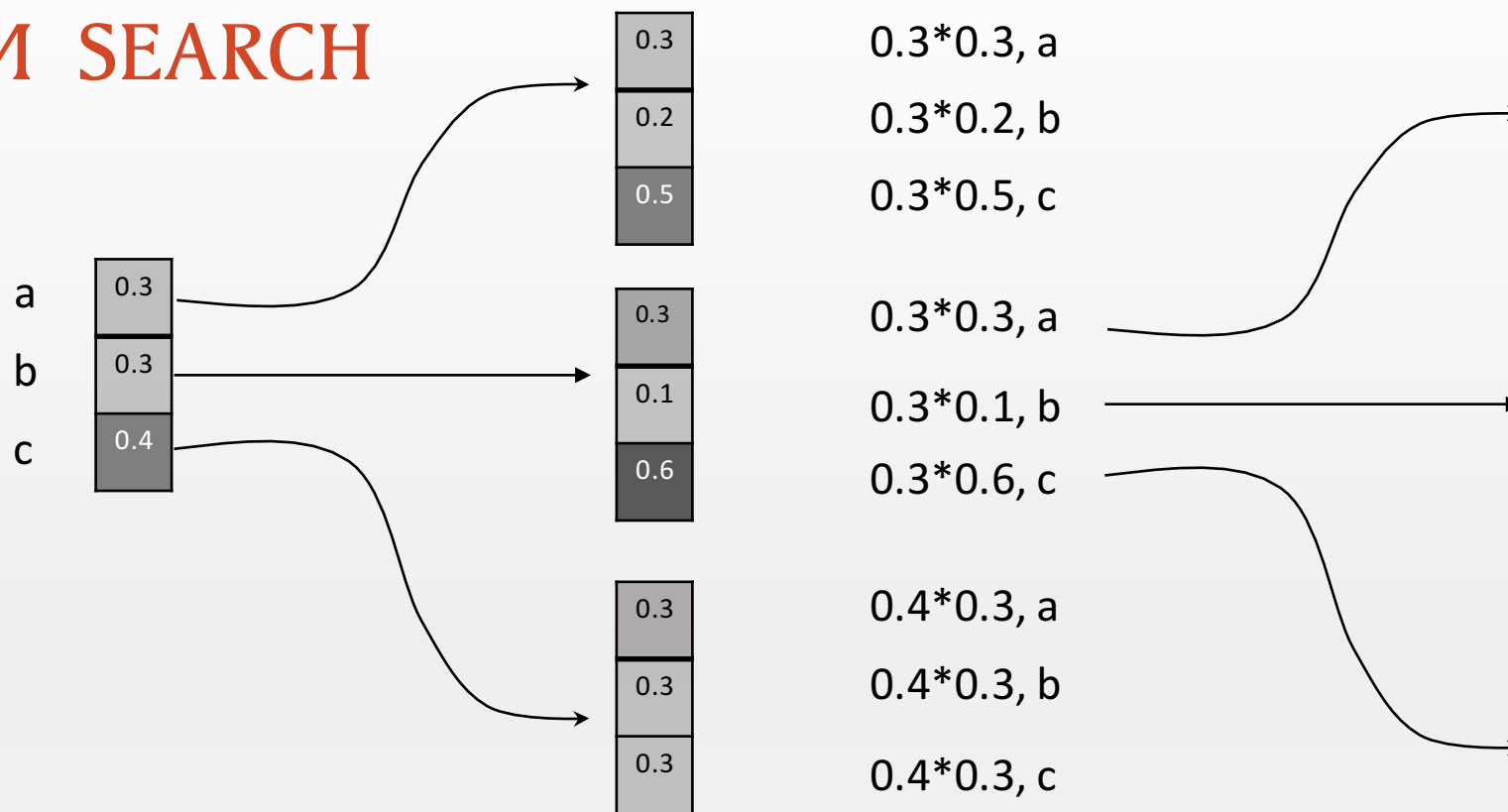




## 每步概率最大不能保证整体概率最大



# BEAM SEARCH



每一步选出最大概率的前M条路径

t=

a	c	r	p	l	y	u	a	g	z
f	d	e	e	q	w	i	s	h	x
c	f	t	g	p	q	o	d	j	c
c	e	l	s	y	a	p	f	l	v
0	1	2	3	4	5	6	7	8	9

每个格点都带着走到这一格点的路径概率和本次单词选择

# BLEU: 对翻译结果的评估

计算备选中每一个单词是否在参考中出现；然后计算所有这些词的占比

备选	the	the	the	the	the	the	the
参考1	the	cat	is	on	the	mat	
参考2	there	is	a	cat	on	the	mat

$$P = \frac{m}{w_t} = \frac{7}{7} = 1$$

$$m_{\max} = 2 \quad P = \frac{2}{7}$$

$m$ : 备选中出现于参考中的词的个数

$w_t$ : 备选词的总数

$m_{\max}$ : 备选中出现于参考中的词的个数，不能超过该词在参考中的最多个数

$w_t$ : 备选词的总数

# BLEU: 对翻译结果的评估

将n元组作为统计单位，通常，n=4

对同样的备选：the the cat，以及参考：the cat is on the mat

模型	元组的集合	评分	the
1-gram	the,the,cat	$(1+1+1)/3=1$	is
2-gram	the the, cat	$(0+1)/2=1/2$	a

# BLEU: 对翻译结果的评估

对整个句子库:

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

c: 备选集长度  
r: 参考文集长度

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

N: N-gram的最大长度, 一般取4  
 $p_n$ : n-gram的p值  
 $w_n$ :  $1/N$

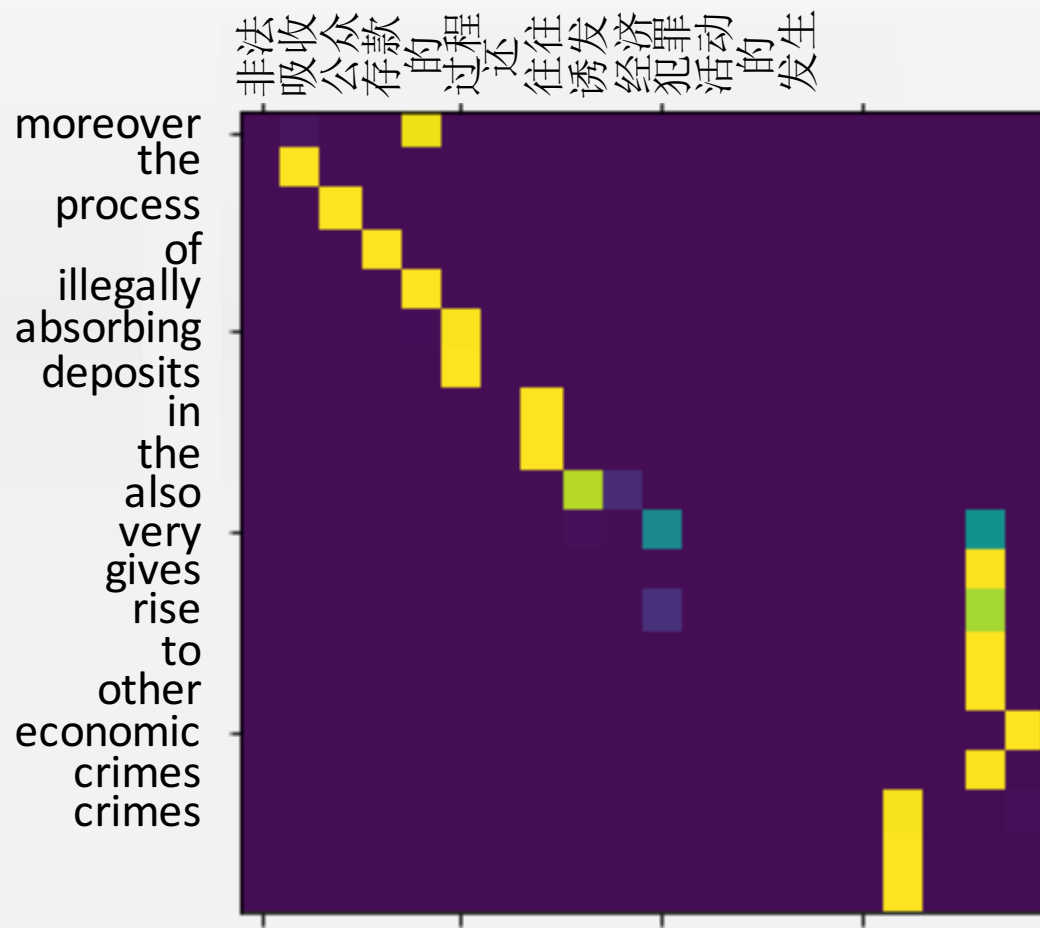
# 机器翻译的结果

结果待补充

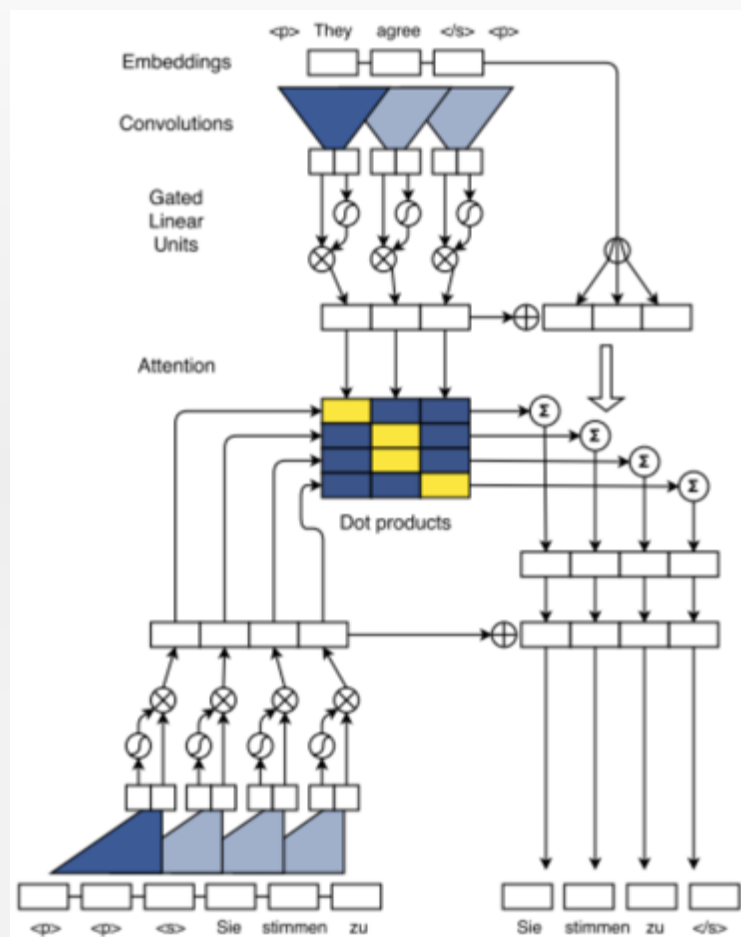
## 注意力机制是否工作？

非法吸收公众存款的过程还往往诱发其他经济犯罪活动的发生。

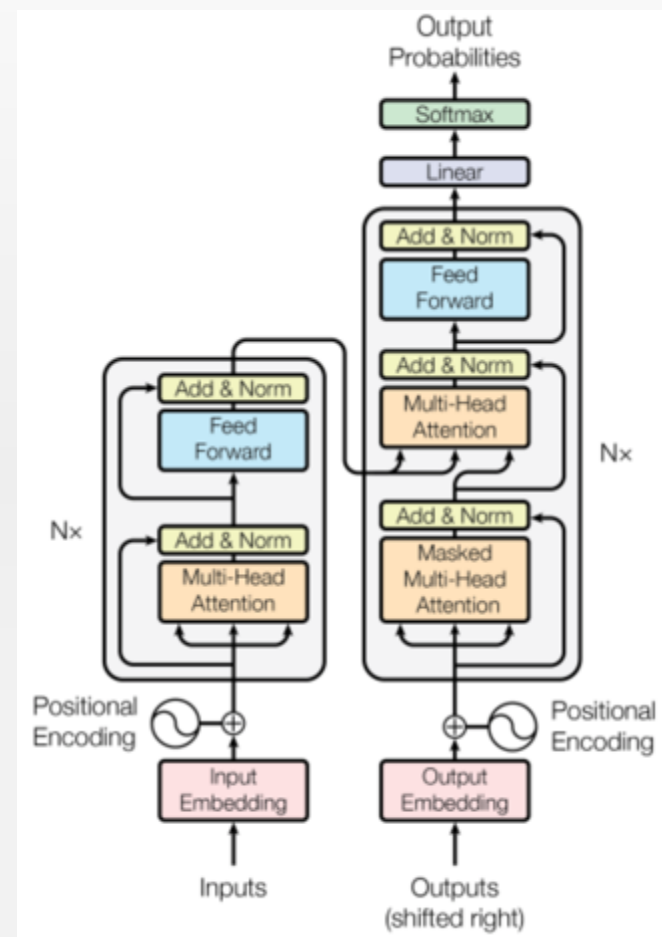
moreover the process of illegally absorbing deposits in the also very gives rise to other economic crimes crimes.



## 更新的机器翻译进展



Facebook



Google

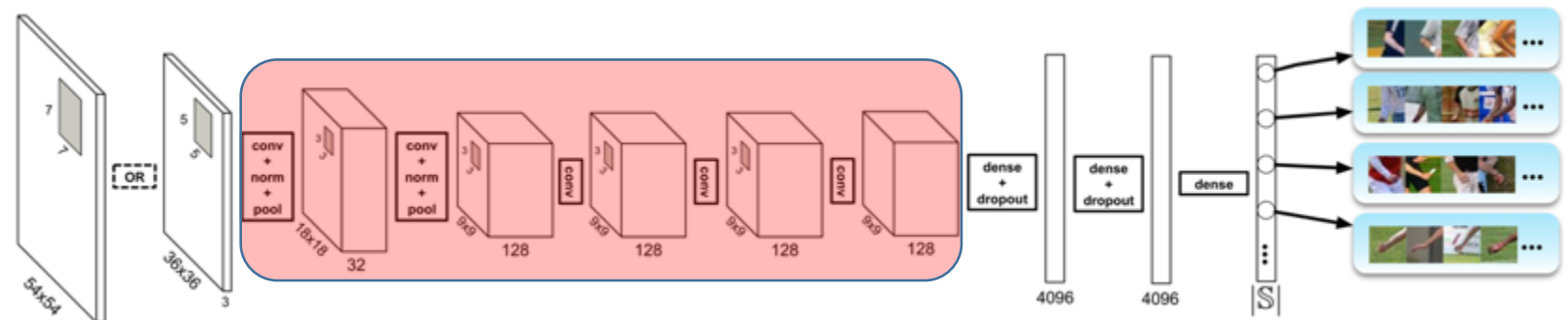


# 编码器 - 解码器架构

- 编码器和解码器分离的部件
- 每个部件都可以替换掉
- 编码器可以是卷积神经网络

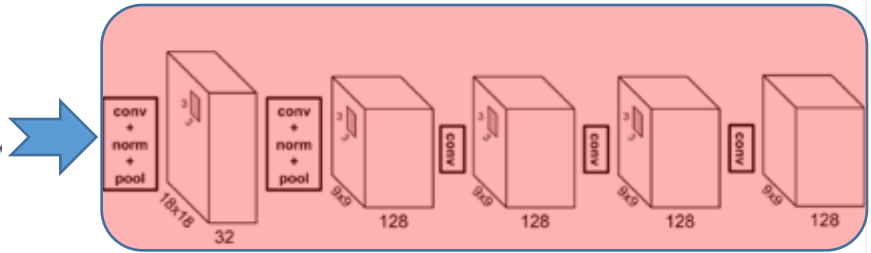
# 编码器 - 解码器架构

图像  
识别  
网络

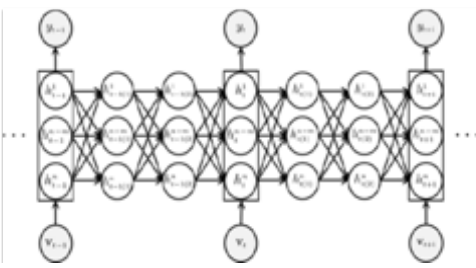


CNN网络      迁移

看图  
说话  
网络



★  
嫁接



一群人在菜市场卖菜

# 要点重述

- 机器翻译概述
- 编码器 - 解码器架构
- 注意力机制
- GRU单元
- 双向RNN

# 作业：英 - 中翻译

- 尝试更改本课程代码，训练一个英翻中的模型
- 尝试用中文翻译成英文，再从英文翻译回中文，并量化评估翻译效果（自定义指标）

# 敬请期待



• 张江

