循环神经网络
Recurrent Neural Network

任课教师：孔雨秋

yqkong@dlut.edu.cn

目 录 (CONTENT)

大连理工大学
Dalian University Of Technology

参考课程

大连理工大学 Dalian University of Technology

- **Andrew Ng, Recurrent Neural Network**

  https://www.bilibili.com/video/av88010803?p=1

- 李宏毅，深度学习

  https://www.bilibili.com/video/BV1JE411g7XF

大连理工大学
Dalian University Of Technology

PART ONE 01

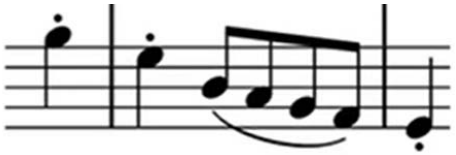
基本循环神经网络

Basic RNN

循环神经网络(Recurrent Neural Network, RNN)

大连理工大学 Dalian University of Technology

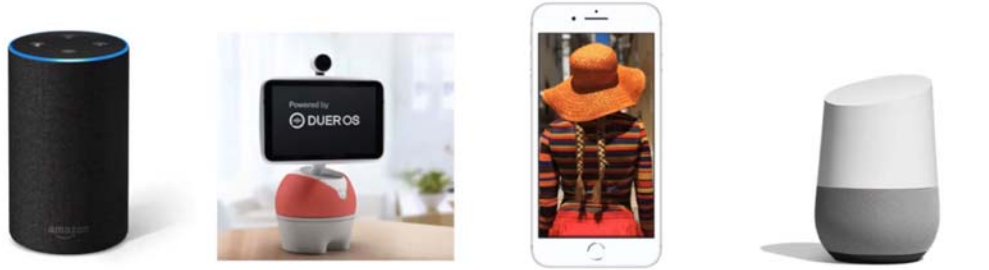
■ 处理序列数据

	x	y
语音识别		“The quick brown fox jumped over the lazy dog”
音乐生成	$\varnothing$	
情感分类	“There is nothing to like in this movie.”	★☆☆☆☆
机器翻译	“Voulez-vous chanter avec moi?”	“Do you want to sing with me?”
视频行为识别		Running
命名实体识别	“Yesterday, Harry Potter met Hermione Granger.”	“Yesterday, Harry Potter met Hermione Granger.”

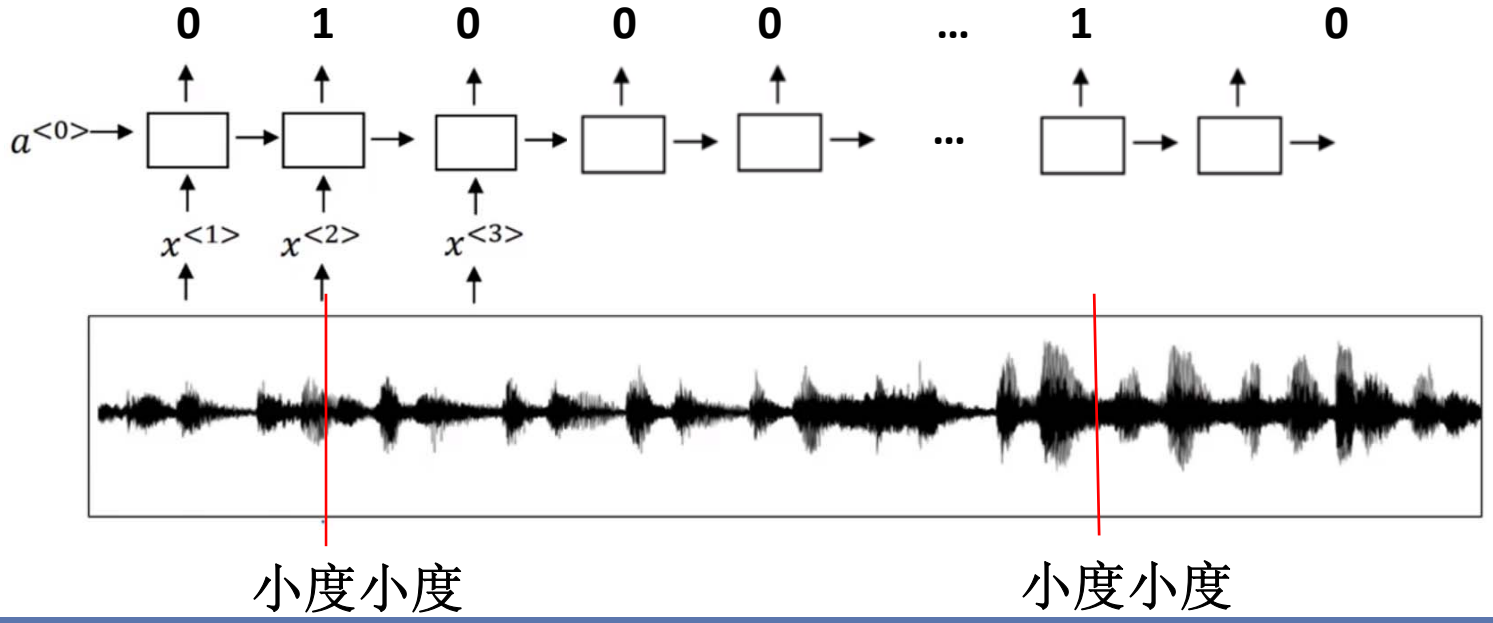
循环神经网络

大连理工大学 Dalian University of Technology

■ 触发字检测



Amazon Echo (Alexa)　Baidu DuerOS (xiaodunihao)　Apple Siri (Hey Siri)　Google Home (Okay Google)



基本循环神经网络

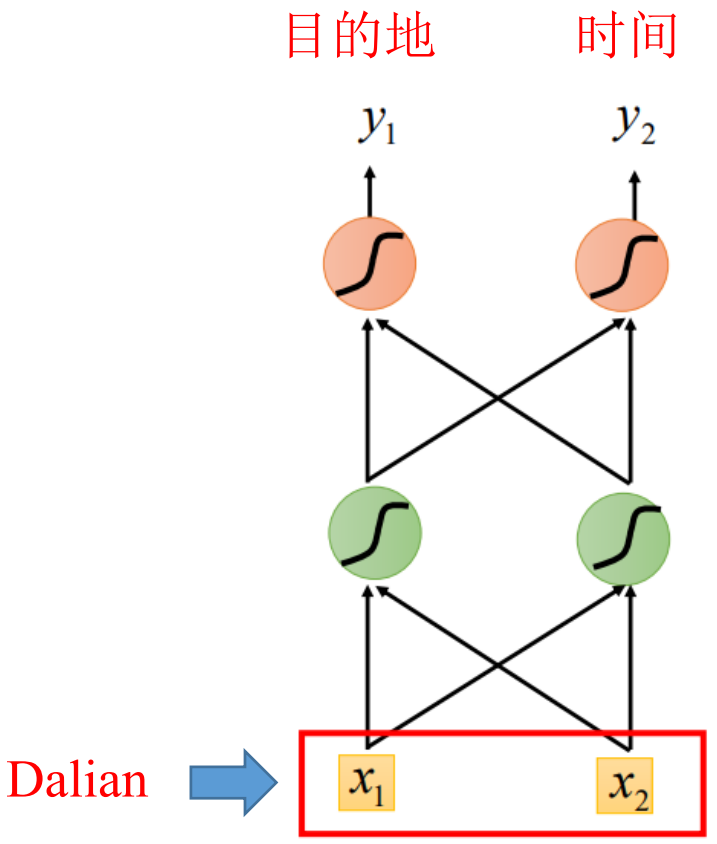
■ 前馈神经网络 V.S. 循环神经网络

Arrive Dalian on 1st December.

任务：预测目的地是哪？时间是？

输入：每次输入一个单词（特征表示）

输出：属于目的地或时间的概率



基本循环神经网络

■ 前馈神经网络 V.S. 循环神经网络

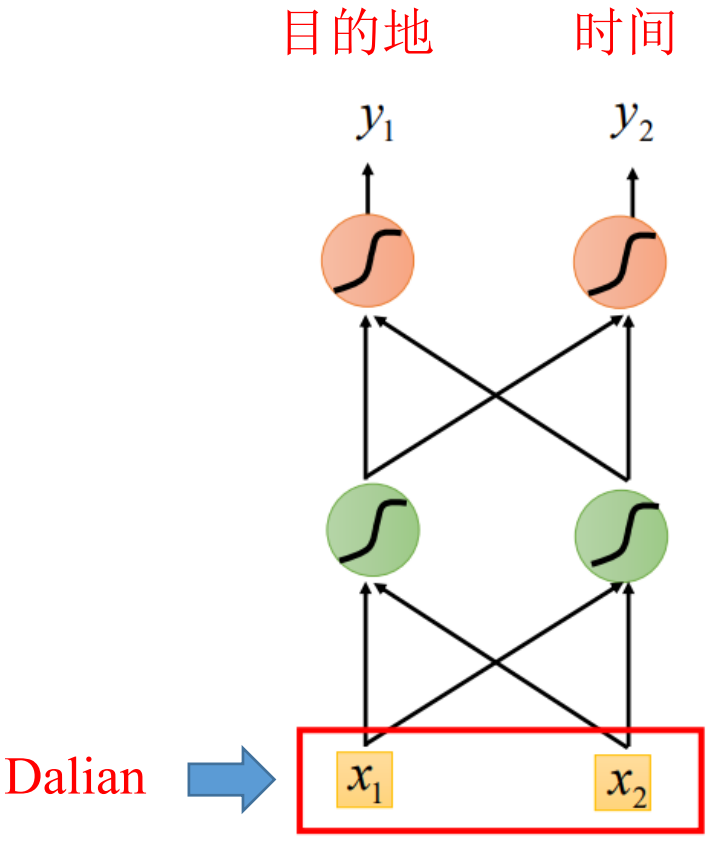
Arrive Dalian on 1st December.

任务：预测目的地是哪？时间是？

Arrive Dalian on 1st December. 目的地

Leave Dalian on 1st December. 出发地

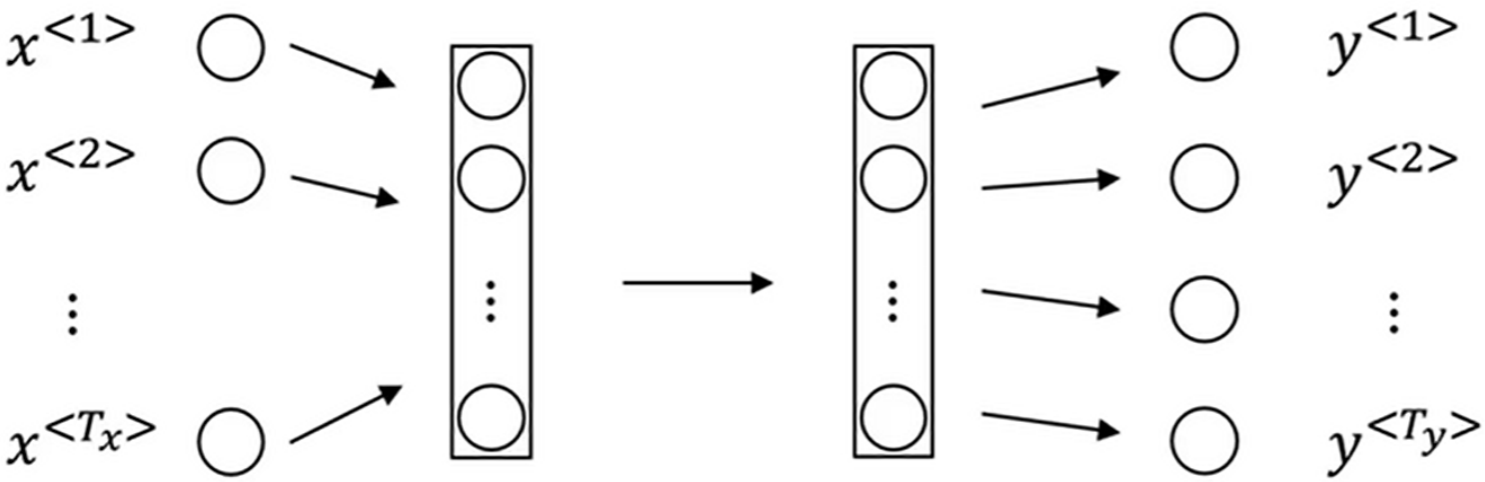
前馈神经网络不考虑上下文信息，
输入是固定的，输出就是固定的。



基本循环神经网络

大连理工大学 Dalian University of Technology

■ 前馈神经网络 V.S. 循环神经网络



- 不同的训练样本，输入、输出序列的长度不同
- 在文本序列不同位置学习到的特征不共享

基本循环神经网络

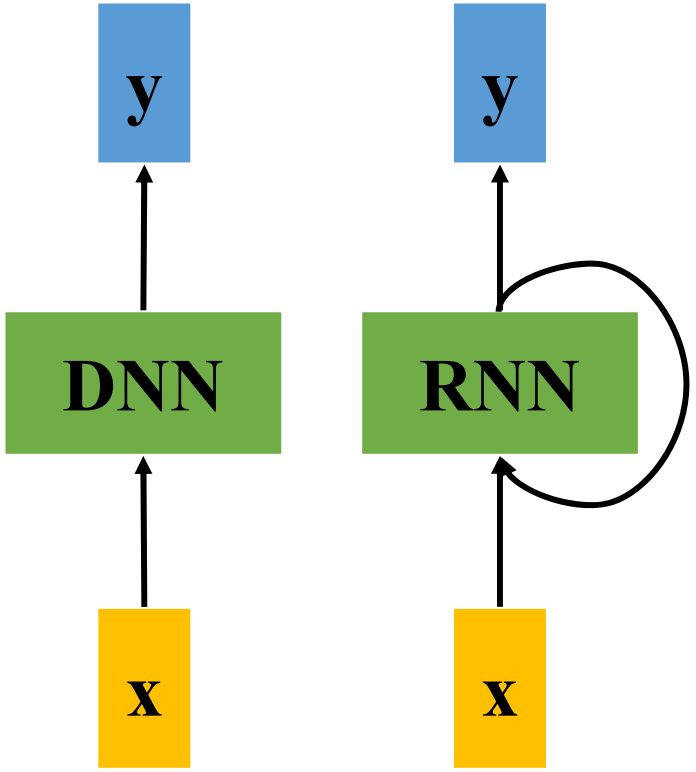
大连理工大学 Dalian University of Technology

■ 前馈神经网络 V.S. 循环神经网络

Arrive Dalian on 1st December. 目的地

Leave Dalian on 1st December. 出发地

- 神经网络需要记忆功能，捕获历史时刻的信息
- 将t时刻隐藏层的输出重新输入到隐藏层中，作为t+1时刻的输入

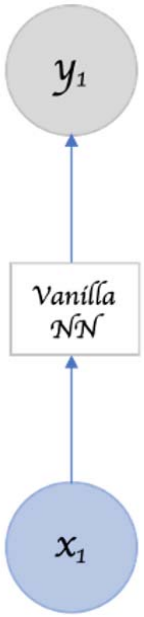


基本循环神经网络

大连理工大学 Dalian University of Technology

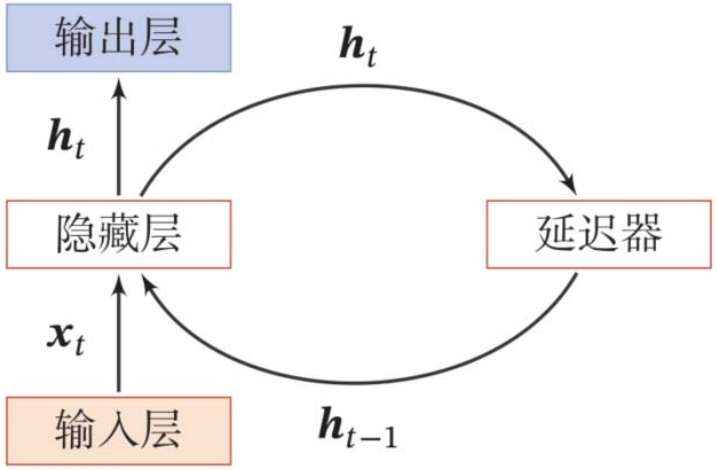
经验是智慧之父，
记忆是智慧之母。

$\mathbf{y}_t = f(\mathbf{x}_t)$



普通神经网络

$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t)$



循环神经网络

命名实体识别

大连理工大学 Dalian University of Technology

■ 符号标记

x Harry Potter and Hermione Granger invented a new spell.

$T_x = 9$ $\mathbf{x}^{<1>}$ $\mathbf{x}^{<2>}$ $\mathbf{x}^{<3>}$ ... $\mathbf{x}^{<t>}$ ... $\mathbf{x}^{<9>}$

y 1 1 0 1 1 0 0 0 0

$T_y = 9$ $\mathbf{y}^{<1>}$ $\mathbf{y}^{<2>}$ $\mathbf{y}^{<3>}$ ... $\mathbf{y}^{<t>}$ ... $\mathbf{y}^{<9>}$

第i个样本 $\mathbf{X}^{(i)} = (\mathbf{x}^{<1>}, \mathbf{x}^{<2>}, \cdots, \mathbf{x}^{<T_x^{(i)}>})$ $\mathrm{x}^{(i)<t>}$

$\mathbf{Y}^{(i)} = (\mathbf{y}^{<1>}, \mathbf{y}^{<2>}, \cdots, \mathbf{y}^{<T_y^{(i)}>})$ $\mathrm{y}^{(i)<t>}$

命名实体识别

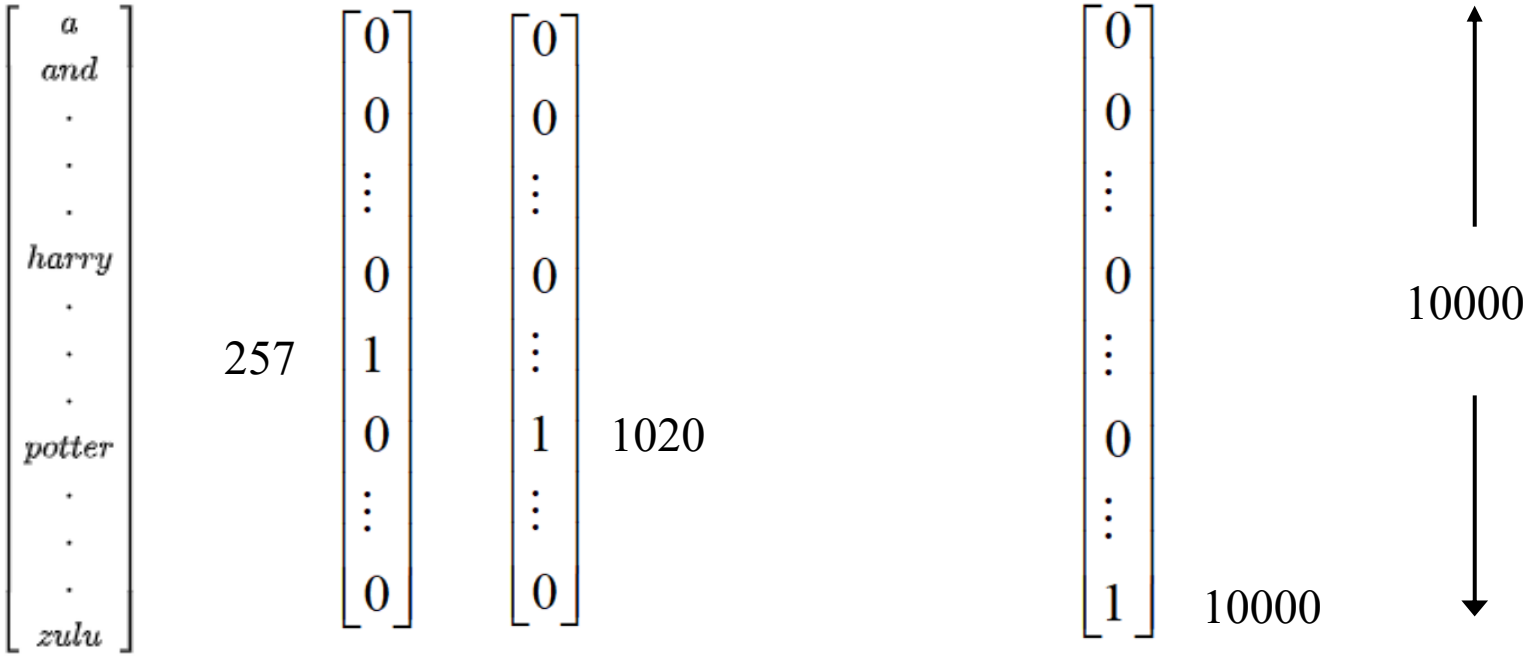
大连理工大学 Dalian University of Technology

■ 符号标记

x Harry Potter and Hermione Granger invented a new spell.

$\mathbf{x}^{<1>}$ $\mathbf{x}^{<2>}$ $\mathbf{x}^{<3>}$ ... $\mathbf{x}^{<t>}$ ... $\mathbf{x}^{<9>}$

Vocabulary

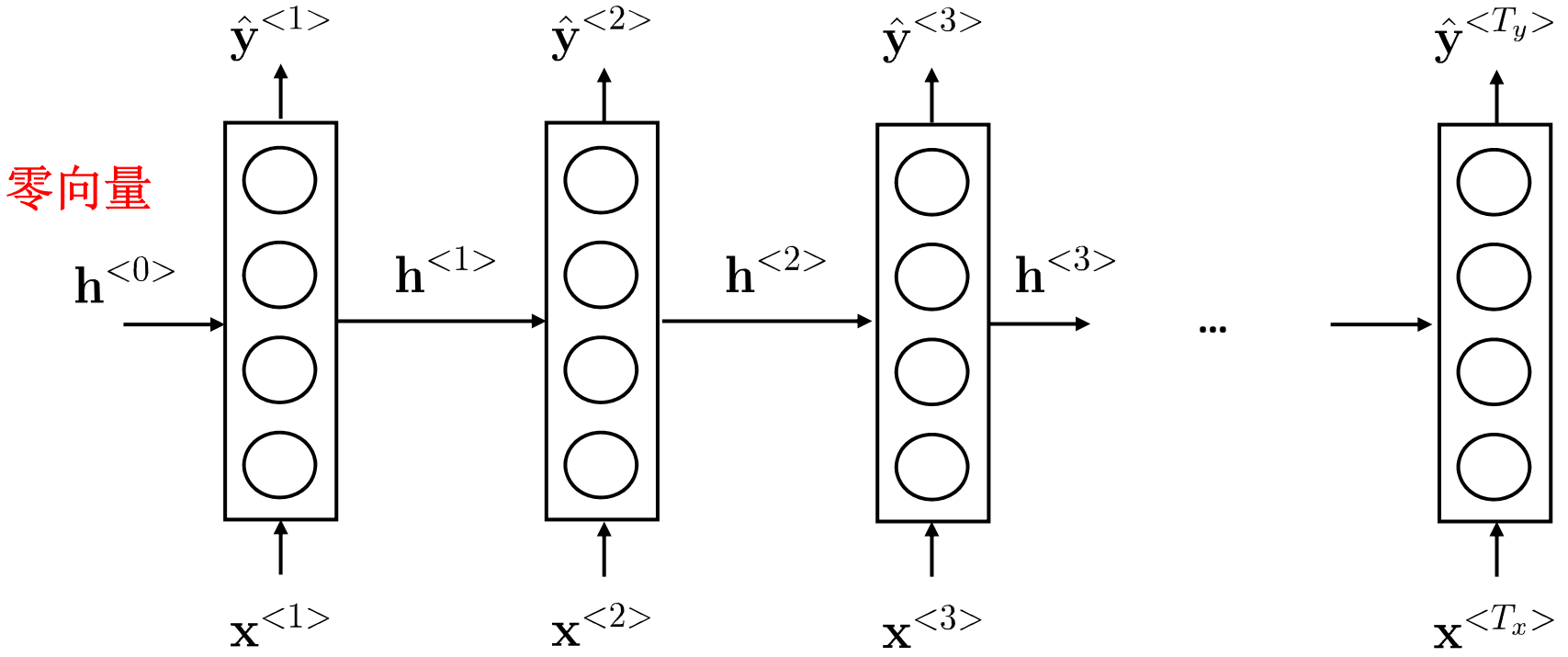


不在**Vocabulary**里的单词可用**<UNK>**标记

命名实体识别

大连理工大学 Dalian University of Technology

■ 循环神经网络

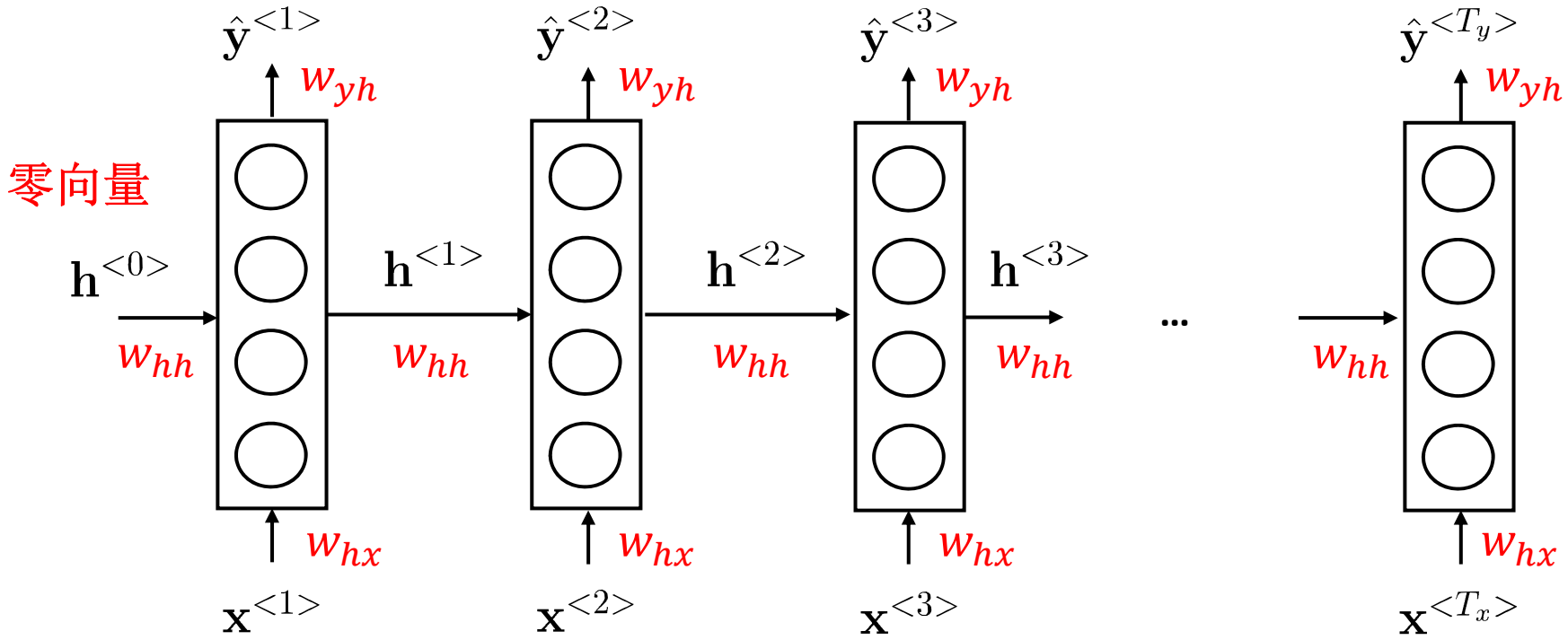


- 从左到右扫描序列数据
- 每个时间步共享参数

命名实体识别

大连理工大学 Dalian University of Technology

■ 循环神经网络

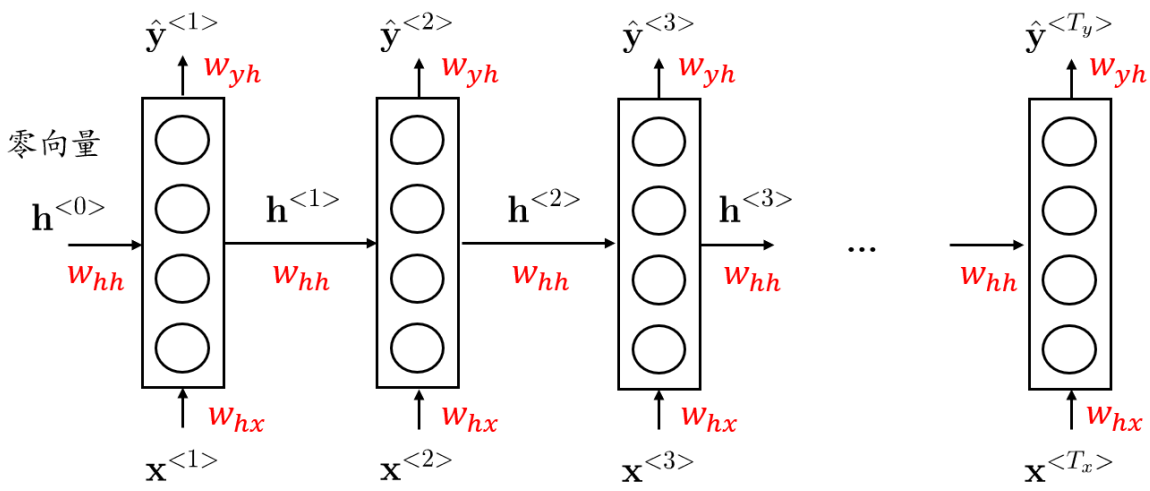


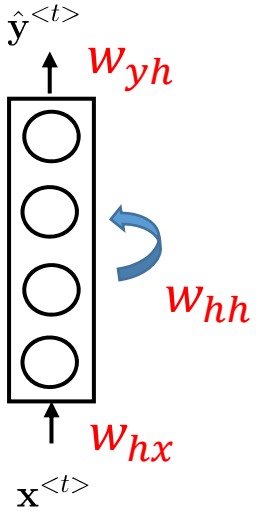
- 从左到右扫描序列数据
- 每个时间步共享参数

命名实体识别

大连理工大学 Dalian University of Technology

■ 循环神经网络





$$\mathbf{h}^{<0>} = \mathbf{0}$$

$$\begin{cases} \mathbf{h}^{<1>} = g(\mathbf{W}_{hh}\mathbf{h}^{<0>} + \mathbf{W}_{hx}\mathbf{x}^{<1>} + \mathbf{b}_h) \\ \hat{y}^{<1>} = g_2(\mathbf{W}_{yh}\mathbf{h}^{<1>} + \mathbf{b}_y) \end{cases}$$

$$\begin{cases} \mathbf{h}^{<t>} = g(\mathbf{W}_{hh}\mathbf{h}^{<t-1>} + \mathbf{W}_{hx}\mathbf{x}^{<t>} + \mathbf{b}_h) \\ \hat{y}^{<t>} = g_2(\mathbf{W}_{yh}\mathbf{h}^{<t>} + \mathbf{b}_y) \end{cases}$$

激活函数：tanh

激活函数：sigmoid（二分类）
softmax（多分类）

命名实体识别

大连理工大学 Dalian University of Technology

■ 循环神经网络

（100,100） 100

$$\begin{cases} \mathbf{h}^{<t>} = g(\mathbf{W}_{hh}\mathbf{h}^{<t-1>} + \mathbf{W}_{hx}\mathbf{x}^{<t>} + \mathbf{b}_h) \\ \hat{y}^{<t>} = g_2(\mathbf{W}_{yh}\mathbf{h}^{<t>} + \mathbf{b}_y) \end{cases} \qquad \mathbf{h}^{<t>} = g(\mathbf{W}_h[\mathbf{h}^{<t-1>}, \mathbf{x}^{<t>}] + \mathbf{b}_h)$$

（100,10000） 10000

$$\begin{cases} \mathbf{h}^{<t>} = g(\mathbf{W}_h[\mathbf{h}^{<t-1>}, \mathbf{x}^{<t>}] + \mathbf{b}_h) \\ \hat{y}^{<t>} = g(\mathbf{W}_y\mathbf{h}^{<t>} + \mathbf{b}_y) \end{cases}$$

$$\mathbf{W}_h = [\mathbf{W}_{hh}; \mathbf{W}_{hx}]$$

（100,10100）（100,100） （100,10000）

$$[\mathbf{h}^{<t-1>}, \mathbf{x}^{<t>}] = \begin{bmatrix} \mathbf{h}^{<t-1>} \\ \mathbf{x}^{<t>} \end{bmatrix} \begin{matrix} 100 \\ 10000 \end{matrix}$$

$$[\mathbf{W}_{hh}; \mathbf{W}_{hx}] \begin{bmatrix} \mathbf{h}^{<t-1>} \\ \mathbf{x}^{<t>} \end{bmatrix} = \mathbf{W}_{hh}\mathbf{h}^{<t-1>} + \mathbf{W}_{hx}\mathbf{x}^{<t>}$$

命名实体识别

大连理工大学 Dalian University of Technology

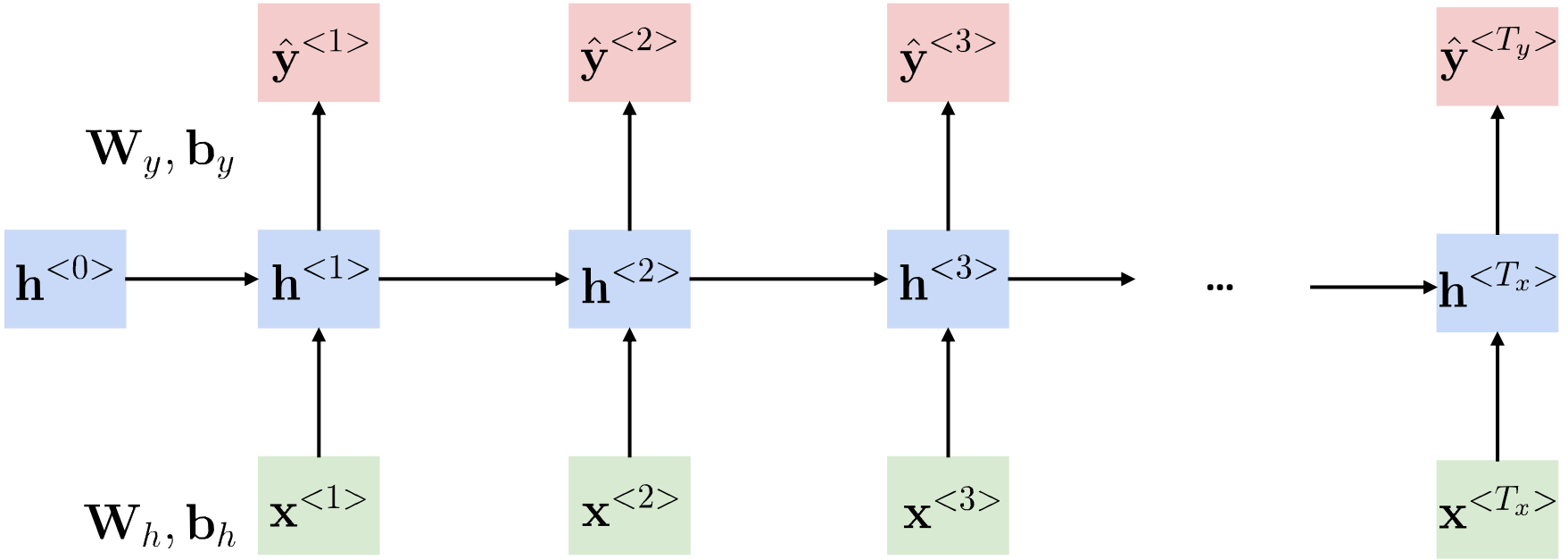
■ 循环神经网络

```
In [2]: X, W_xh = torch.randn(3, 1), torch.randn(1, 4)
        H, W_hh = torch.randn(3, 4), torch.randn(4, 4)
        torch.matmul(X, W_xh) + torch.matmul(H, W_hh)

Out[2]: tensor([[ 5.2633, -3.2288,  0.6037, -1.3321],
                [ 9.4012, -6.7830,  1.0630, -0.1809],
                [ 7.0355, -2.2361,  0.7469, -3.4667]])

In [3]: torch.matmul(torch.cat((X, H), dim=1), torch.cat((W_xh, W_hh), dim=0))

Out[3]: tensor([[ 5.2633, -3.2288,  0.6037, -1.3321],
                [ 9.4012, -6.7830,  1.0630, -0.1809],
                [ 7.0355, -2.2361,  0.7469, -3.4667]])
```

命名实体识别

大连理工大学 Dalian University of Technology

■ 循环神经网络



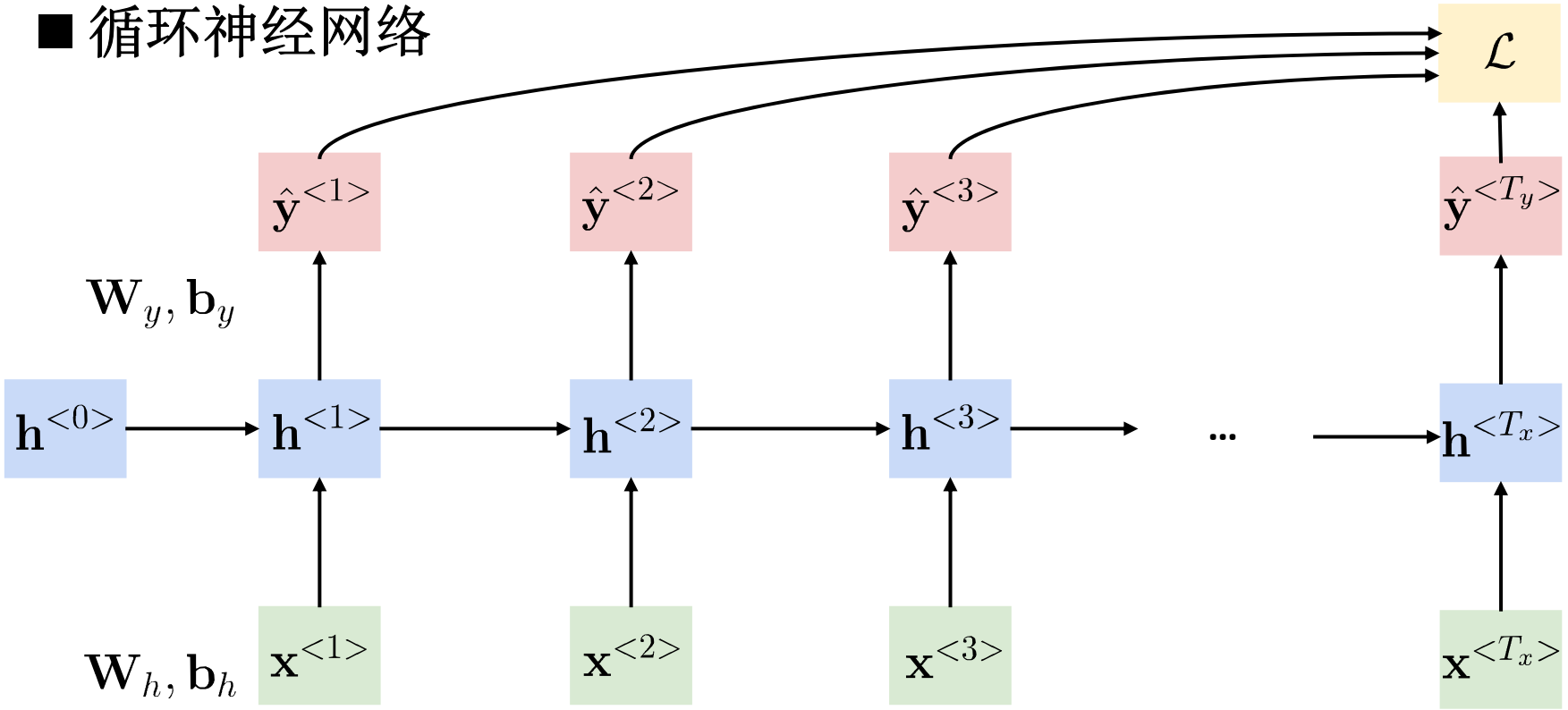
$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} \log \hat{y}^{<t>} - (1 - y^{<t>}) \log(1 - \hat{y}^{<t>})$$

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

命名实体识别

大连理工大学 Dalian University of Technology

■ 循环神经网络



$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} \log \hat{y}^{<t>} - (1 - y^{<t>}) \log(1 - \hat{y}^{<t>})$$

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

定理 6.1 – 循环神经网络的通用近似定理 [Haykin, 2009]： 如果一个完全连接的循环神经网络有足够数量的 sigmoid 型隐藏神经元，它可以以任意的准确率去近似任何一个非线性动力系统

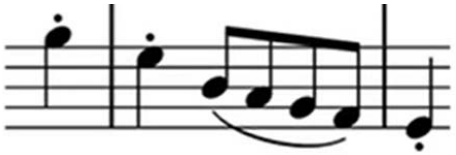
$$\boldsymbol{s}_t = g(\boldsymbol{s}_{t-1}, \boldsymbol{x}_t), \tag{6.10}$$

$$\boldsymbol{y}_t = o(\boldsymbol{s}_t), \tag{6.11}$$

其中 $\boldsymbol{s}_t$ 为每个时刻的隐状态，$\boldsymbol{x}_t$ 是外部输入，$g(\cdot)$ 是可测的状态转换函数，$o(\cdot)$ 是连续输出函数，并且对状态空间的紧致性没有限制.

不同结构的RNN

大连理工大学 Dalian University of Technology

	x		y
语音识别		→	“The quick brown fox jumped over the lazy dog”
音乐生成	$\varnothing$	→	
情感分类	“There is nothing to like in this movie.”	→	★☆☆☆☆
机器翻译	“Voulez-vous chanter avec moi?”	→	“Do you want to sing with me?”
视频行为识别		→	Running
命名实体识别	“Yesterday, Harry Potter met Hermione Granger.”	→	“Yesterday, Harry Potter met Hermione Granger.”

不同结构的RNN

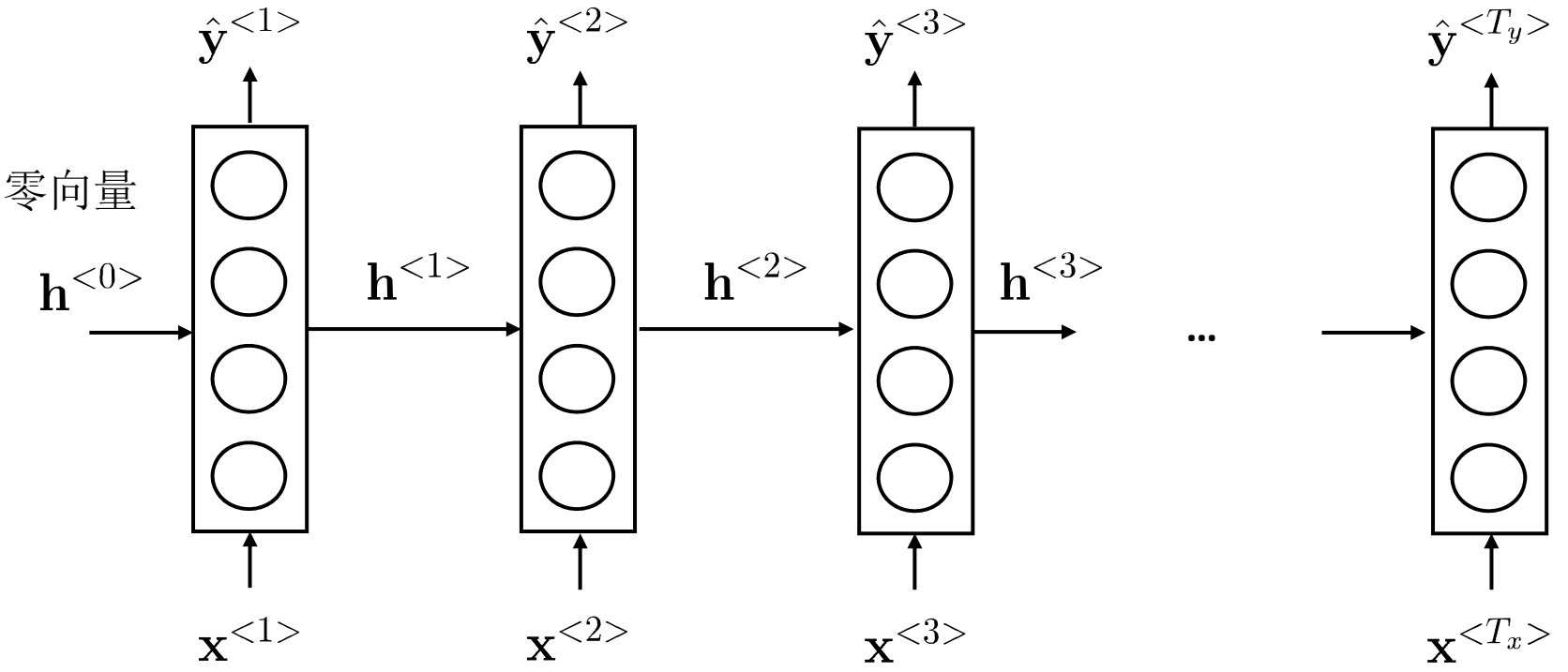
■ **Many-to-many （命名实体识别）**

$$T_x = T_y$$

$$\hat{y}_t = g(\boldsymbol{h}_t), \qquad \forall t \in [1, T]$$

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE, Baidu ORG, and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space. The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE. Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL, with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE.

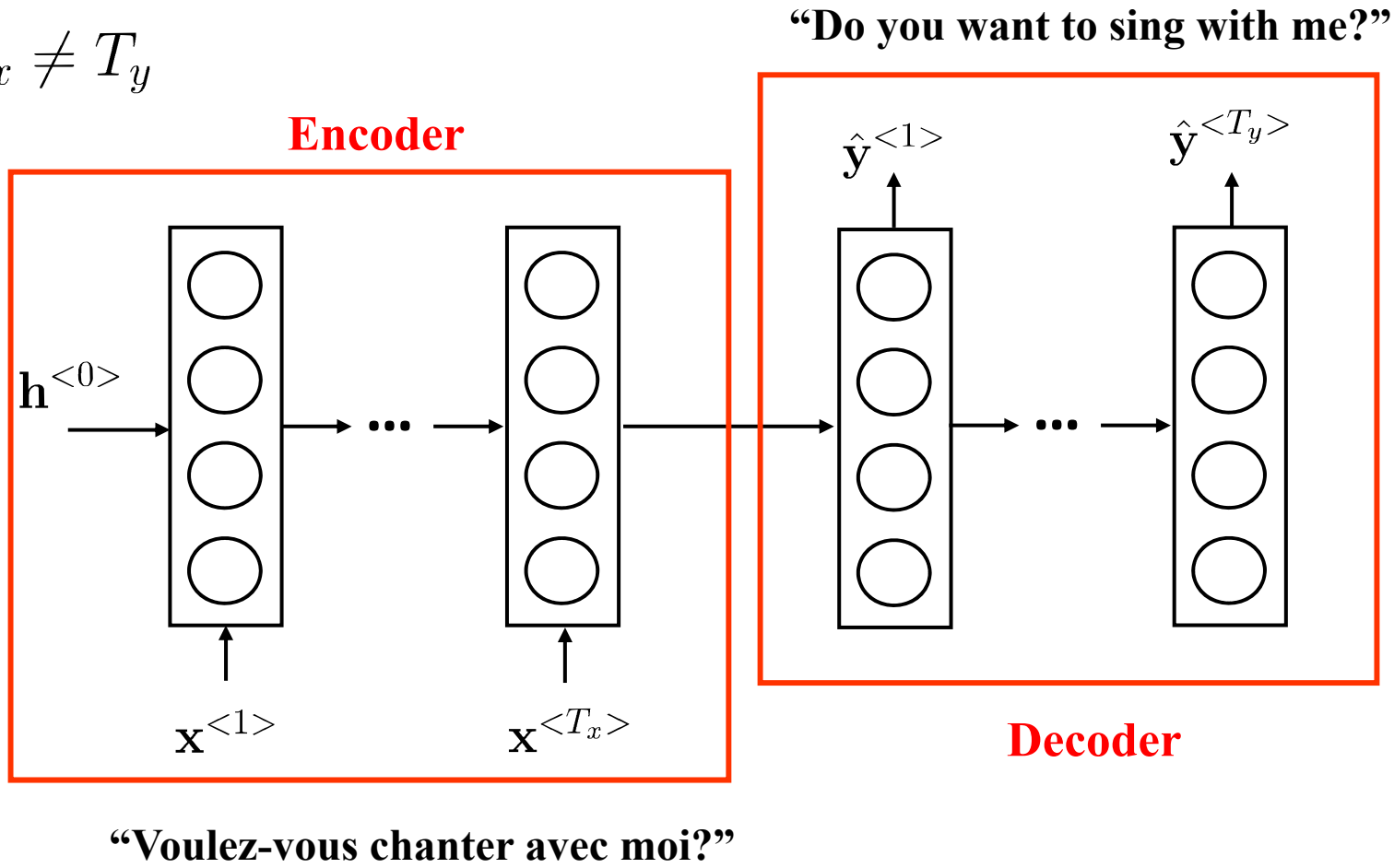
To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG, IBM ORG, and Microsoft ORG.



不同结构的RNN

■ **Many-to-many （机器翻译）**

$T_x \neq T_y$



不同结构的RNN

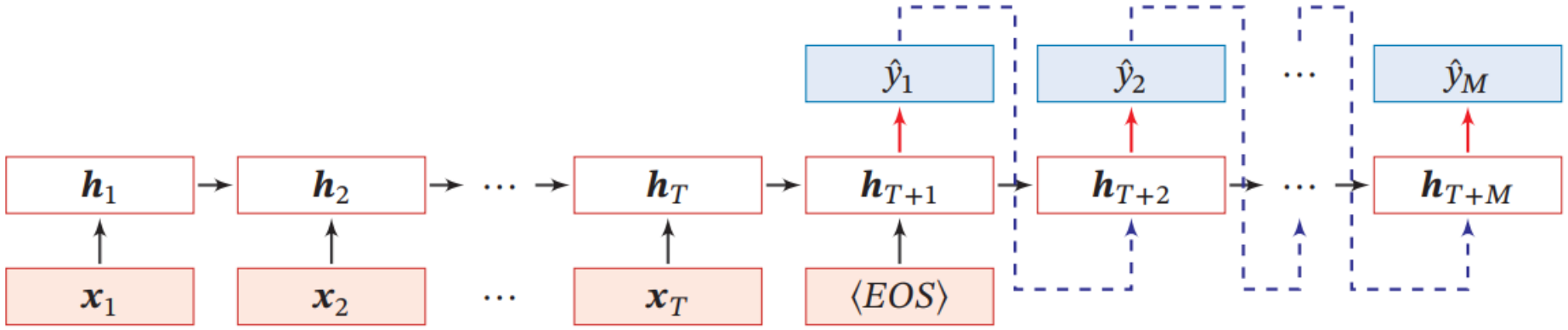
大连理工大学 Dalian University of Technology

■ **Many-to-many （机器翻译）**

$$\boldsymbol{h}_t = f_1(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t), \qquad \forall t \in [1, T]$$

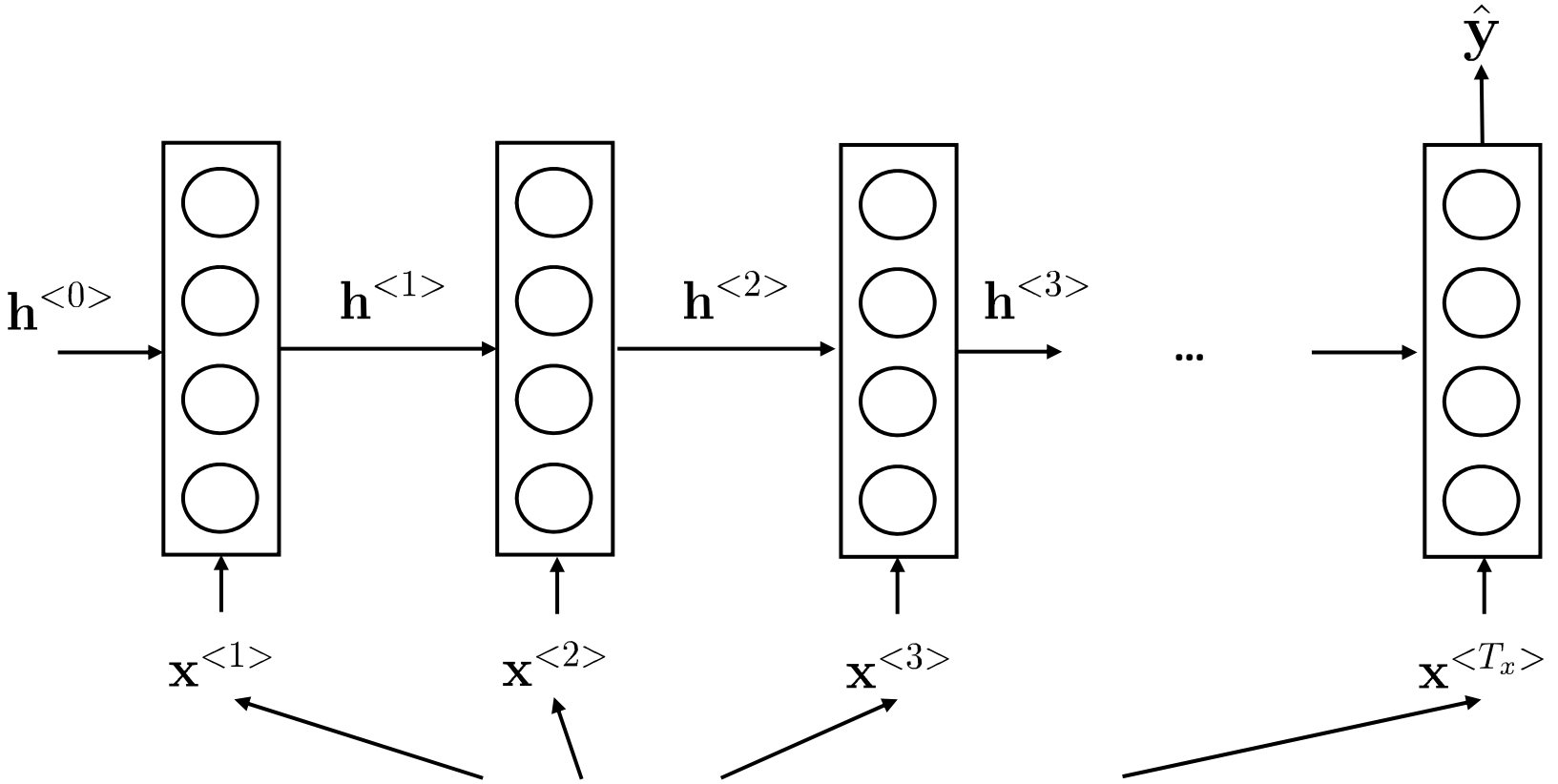
$$\boldsymbol{h}_{T+t} = f_2(\boldsymbol{h}_{T+t-1}, \hat{\boldsymbol{y}}_{t-1}), \qquad \forall t \in [1, M]$$

$$\hat{y}_t = g(\boldsymbol{h}_{T+t}), \qquad \forall t \in [1, M]$$



不同结构的RNN

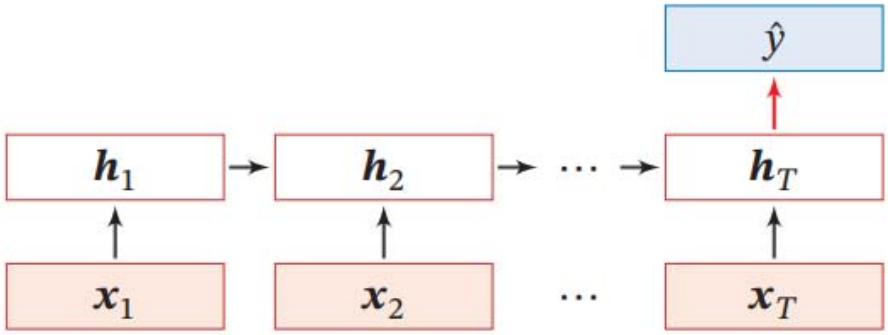
■ **Many-to-one （情感分类）**



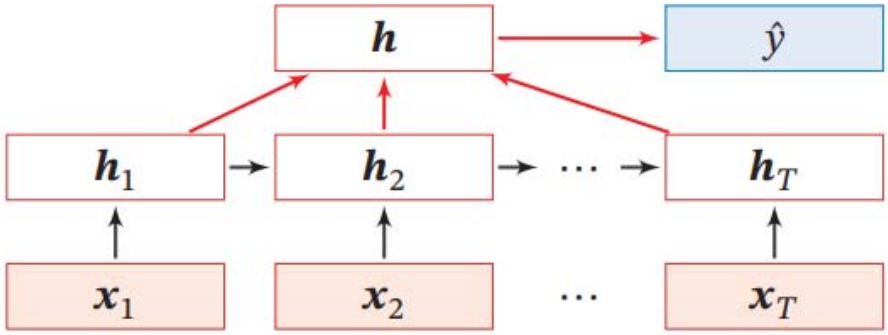
“There is nothing to like in this movie.”

不同结构的RNN

■ **Many-to-one （情感分类）**



(a) 正常模式



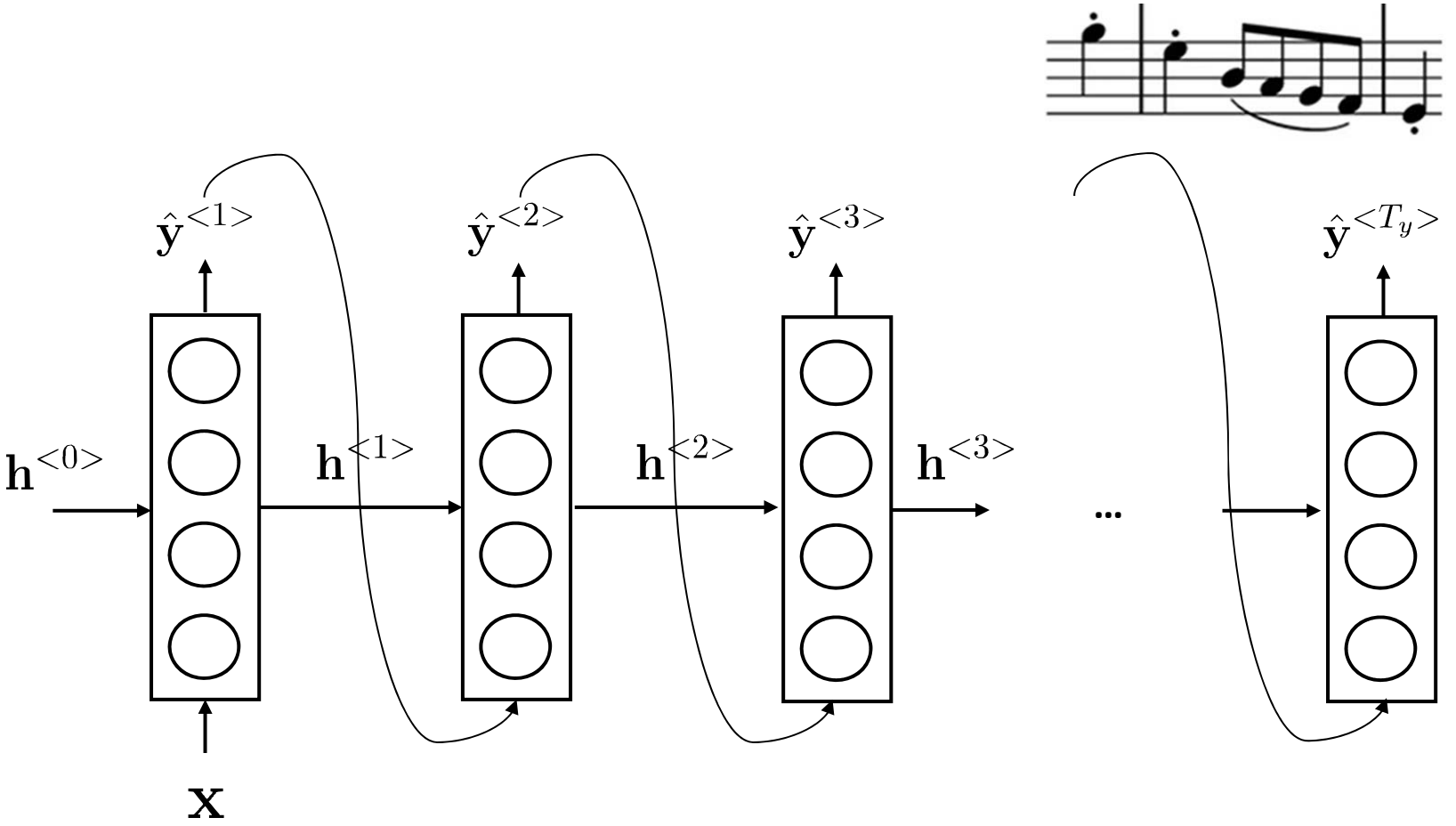
(b) 按时间进行平均采样模式

$$\hat{y} = g(\boldsymbol{h}_T)$$

$$\hat{y} = g\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{h}_t\right)$$

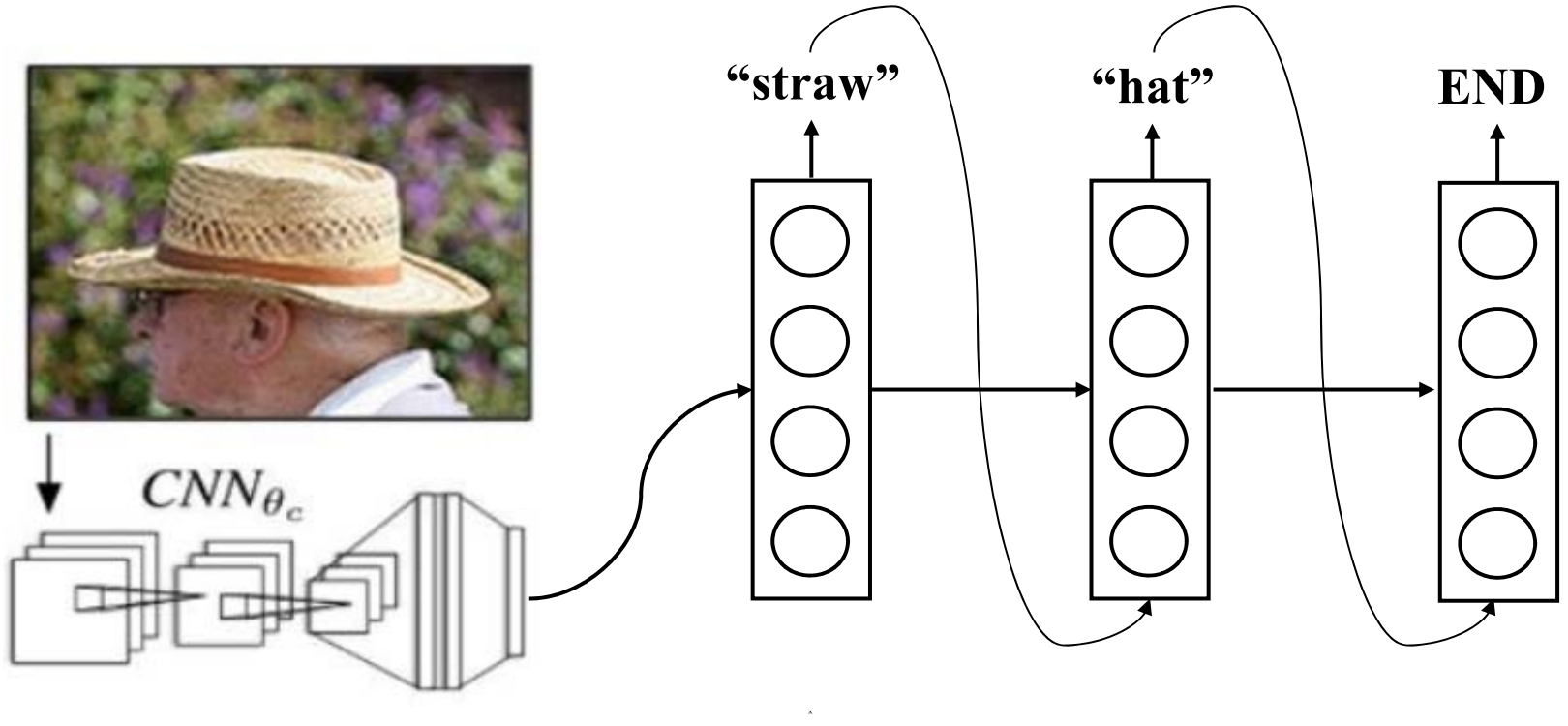
不同结构的RNN

■ **One-to-many （音乐生成）**



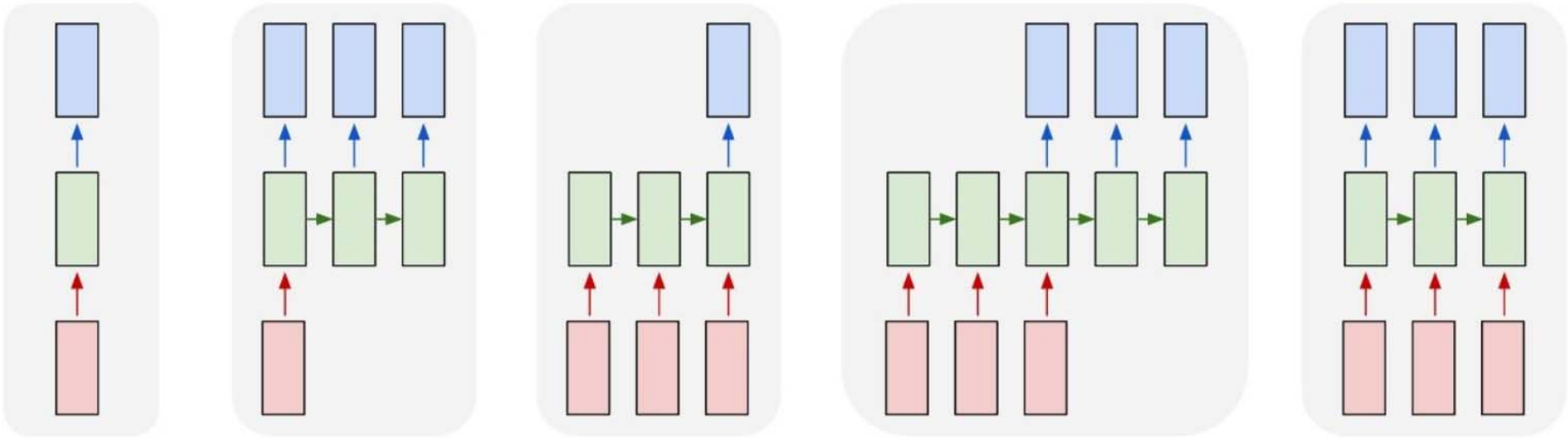
不同结构的RNN

■ **One-to-many （图像加字幕）**



不同结构的RNN

One-to-one One-to-many Many-to-one Many-to-many Many-to-many



语言模型

- 语言模型 (Language model)：是自然语言处理的重要技术，常用于语音识别和机器翻译等。
- 语言模型给一个序列（句子）分配概率，找到最可能出现的词序列。同时估计某个序列中各单词出现的概率。

- 语音识别系统：**The apple and pair salad.**

  The apple and pear salad.

  P(The apple and pair salad)=3.2e-13

  P(The apple and pear salad)=5.7e-10

  P(Sentence)=? ➡ $P(\hat{y}^{<1>}, \hat{y}^{<2>}, \cdots, \hat{y}^{<T_y>})$

语言模型

- 构建语言模型
  - 训练集：大型文本语料库
  - 特征表示：**one-hot向量**

Cats average 15 hours of sleep a day.

$y^{<1>}$ $y^{<2>}$ …

- 用**RNN**构建词序列的概率模型

Vocabulary

<EOS

句尾标记

<UNK>

标点

<EOS>

语言模型

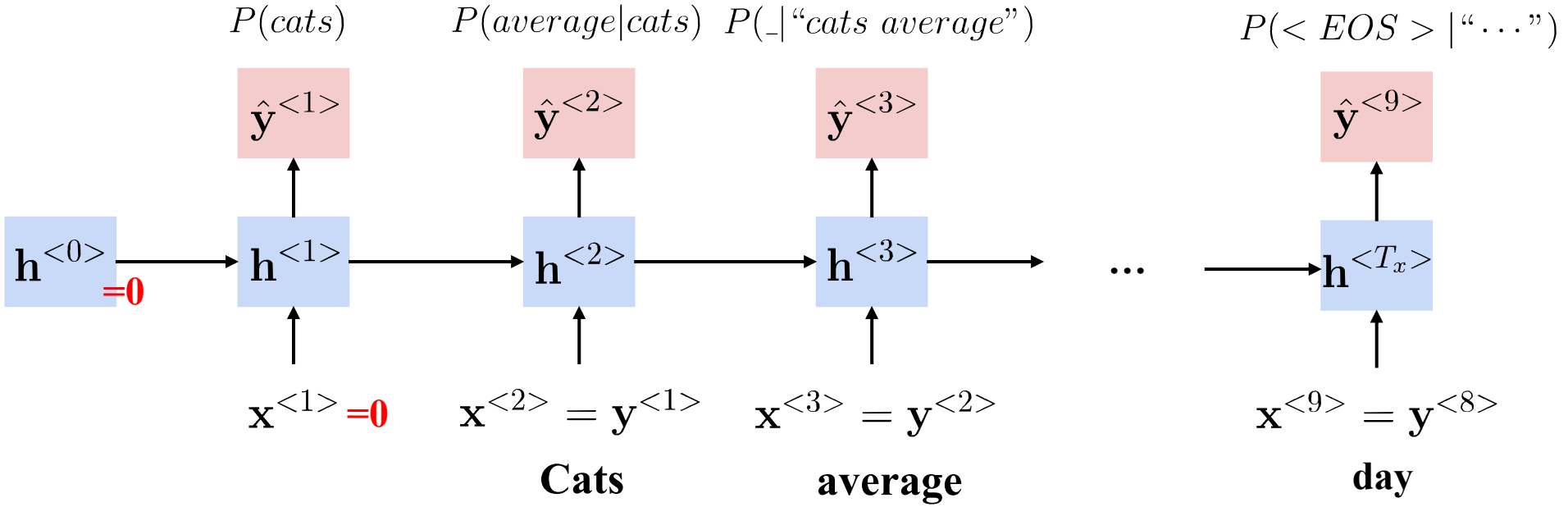
大连理工大学 Dalian University of Technology

Cats average 15 hours of sleep a day.

$$P(\mathbf{y}^{<1>}, \mathbf{y}^{<2>}, \mathbf{y}^{<3>}) = P(\mathbf{y}^{<1>}) P(\mathbf{y}^{<2>} | \mathbf{y}^{<1>}) \cdot P(\mathbf{y}^{<3>} | \mathbf{y}^{<2>}, \mathbf{y}^{<1>})$$

$$\Rightarrow \quad P(\mathbf{y}^{<1>}, \mathbf{y}^{<2>}, \cdots, \mathbf{y}^{<T>}) = \prod_{t=1}^{T} P(\mathbf{y}^{<t>} | \mathbf{y}^{<1>}, \mathbf{y}^{<2>}, \cdots, \mathbf{y}^{<t-1>})$$

$$P(a), P(and), \cdots, P(cats), \cdots, , P(zulu), P(< UNK >), P(< EOS >)$$

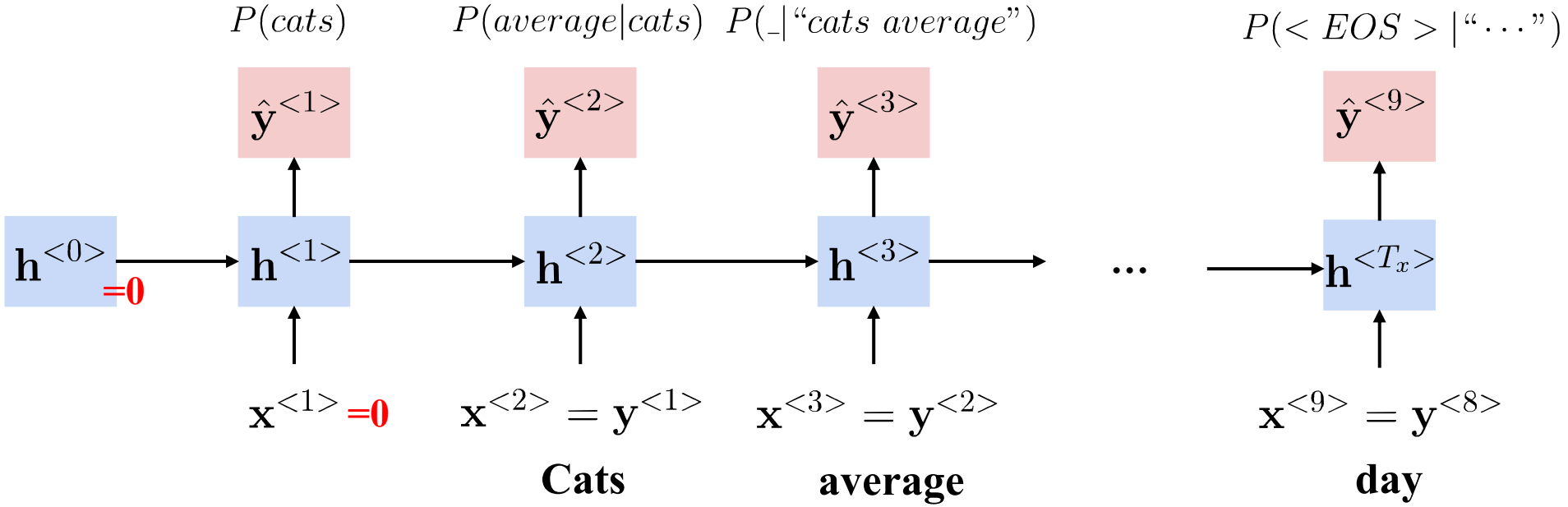


语言模型

大连理工大学 Dalian University of Technology

Cats average 15 hours of sleep a day.

$$P(a), P(and), \cdots, P(cats), \cdots, , P(zulu), P(< UNK >), P(< EOS >)$$

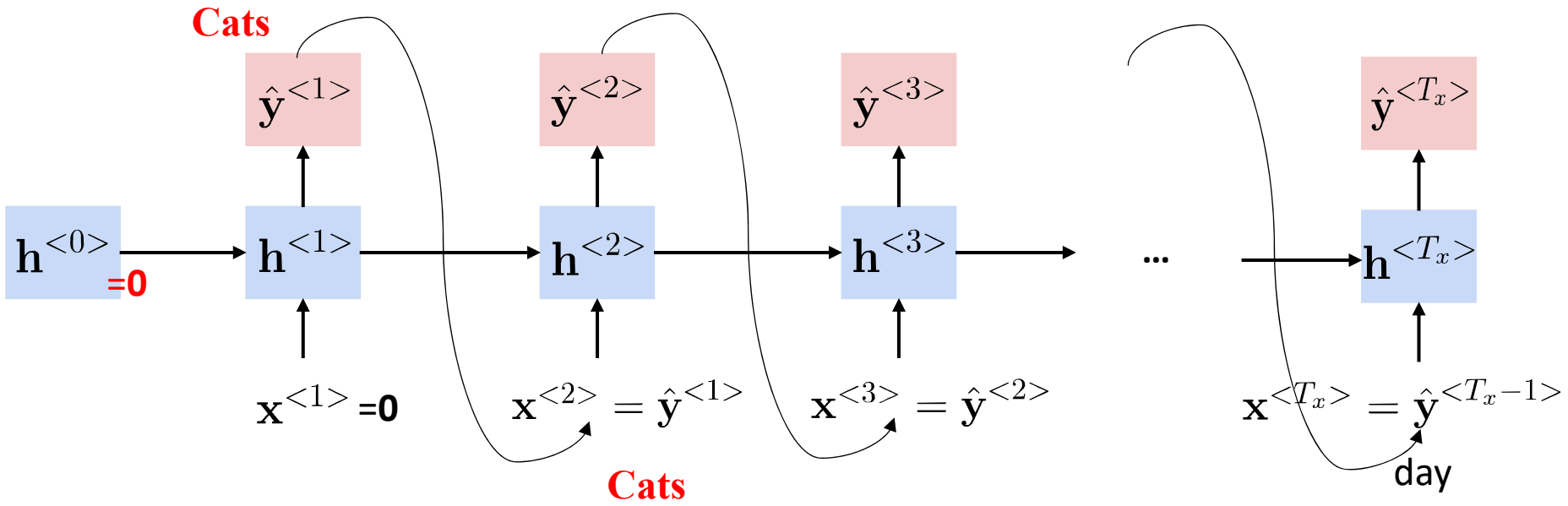


$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -\sum_i y_i^{<t>} \log \hat{y}_i^{<t>}$$

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

语言模型

- 用训练好的RNN生成句子
  - 在每个时间步下，根据输出的概率分布进行采样。
  - 将采样得到的单词作为下一个时间步的输入



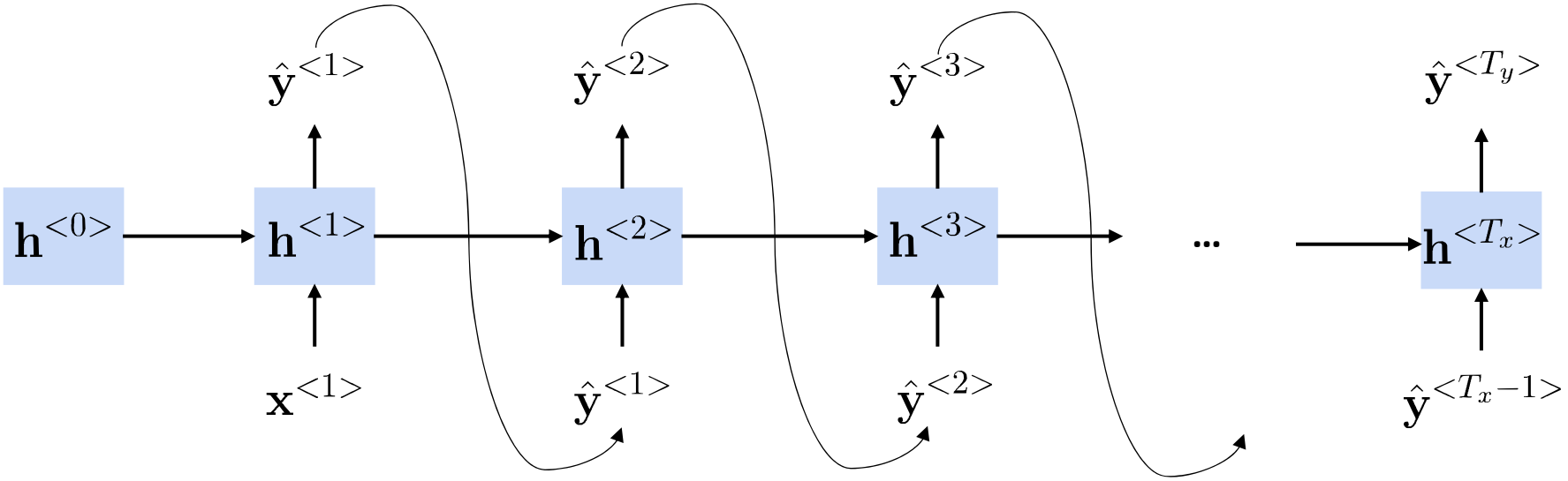
语言模型

大连理工大学 Dalian University of Technology

- 字符级别语言模型

Vocabulary=[a, and, …, zulu, <UNK>]

Vocabulary=[a, b, c, …, z, , ., ?, :, 0, 1, 2, …]



语言模型

- 文本生成

News

President enrique peña nieto, announced sench's sulk former coming football langston paring.

"I was not at all surprised," said hich langston.

"Concussion epidemic", to be examined.

The gray football the told some and this has on the uefa icon, should money as.

Shakespeare

The mortal moon hath her eclipse in love.

And subject of this thou art another this fold.

When besser be my love to me see sabl's.

For whose are ruse of mine eyes heaves.

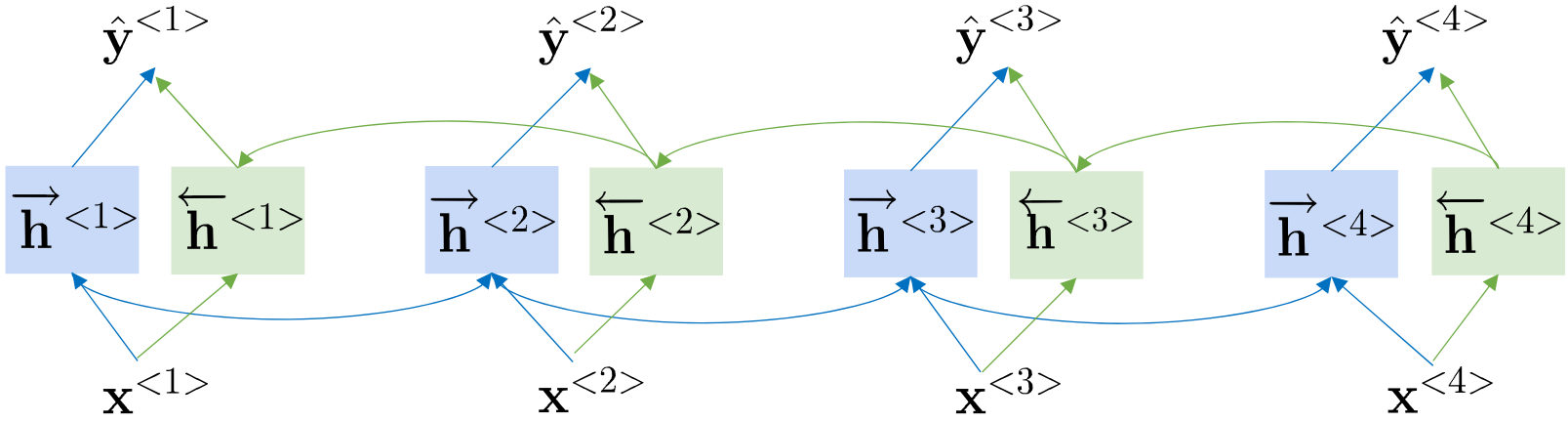
更多结构的RNN

大连理工大学 Dalian University of Technology

- 双向循环神经网络 **(Bi-directional RNN)**
  - 获取历史数据的信息 + 获取未来数据的信息

He said, “Teddy bears are on sale!”

He said, “Teddy Rooseverlt was a great President!”

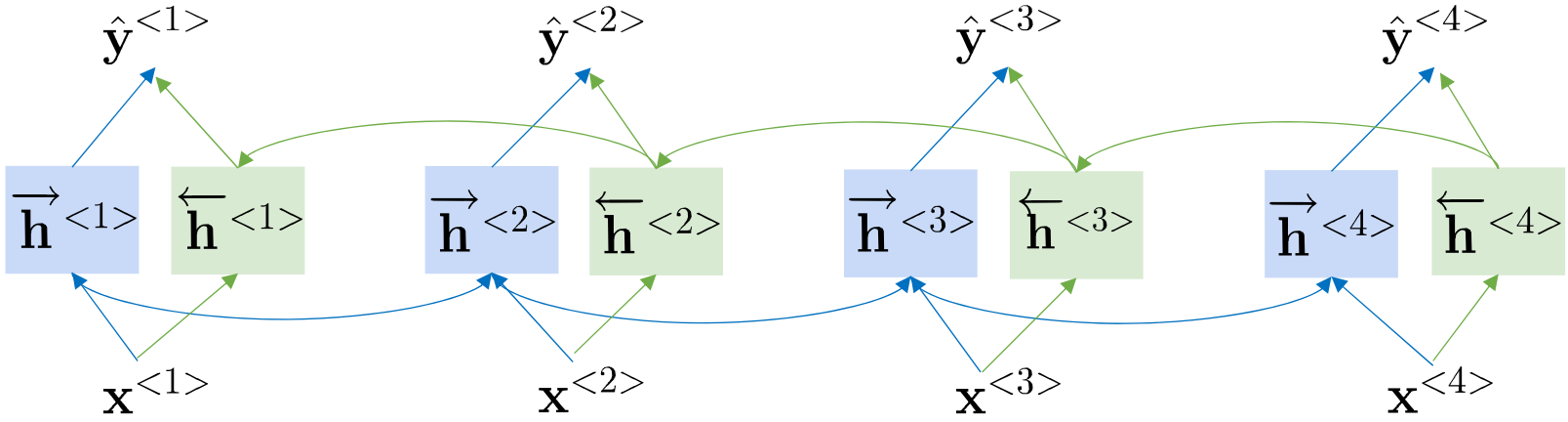


$$\hat{y}^{<t>} = g(\mathbf{W}_y[\overrightarrow{\mathbf{h}}^{<t>}, \overleftarrow{\mathbf{h}}^{<t>}] + \mathbf{b}_y)$$

更多结构的RNN

- 双向循环神经网络 **(Bi-directional RNN)**
  - 获取历史数据的信息 + 获取未来数据的信息

- 缺陷：需要获取完整的序列（句子）才能进行预测，不能进行实时处理

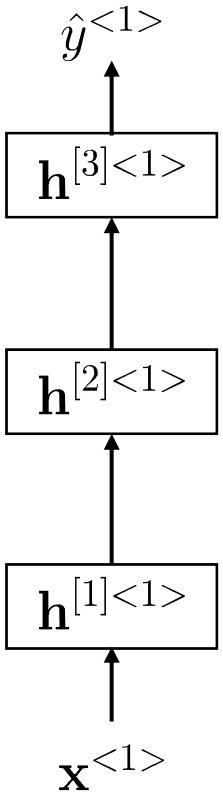


更多结构的RNN

大连理工大学 Dalian University of Technology

- 深层循环神经网络 **(Deep RNN)**

  前馈神经网络

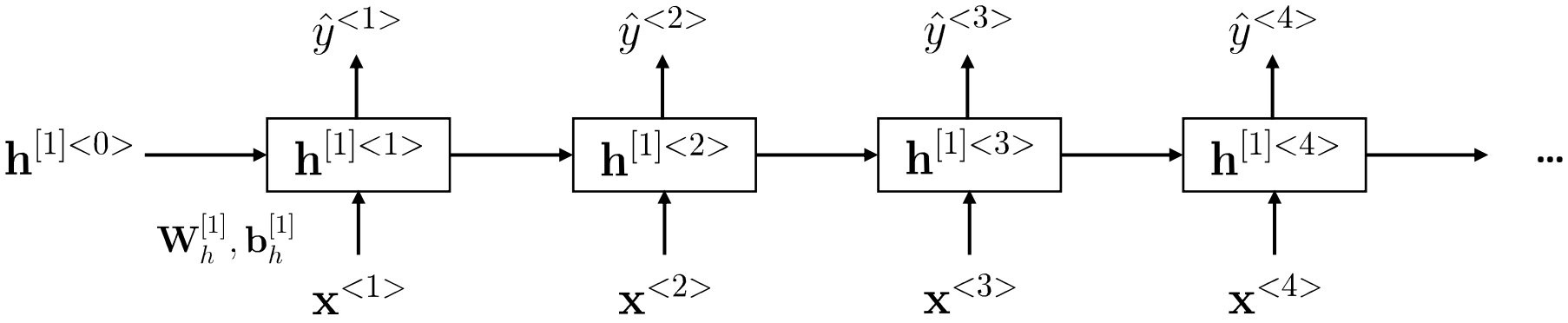


更多结构的RNN

大连理工大学 Dalian University of Technology

- 深层循环神经网络 **(Deep RNN)**

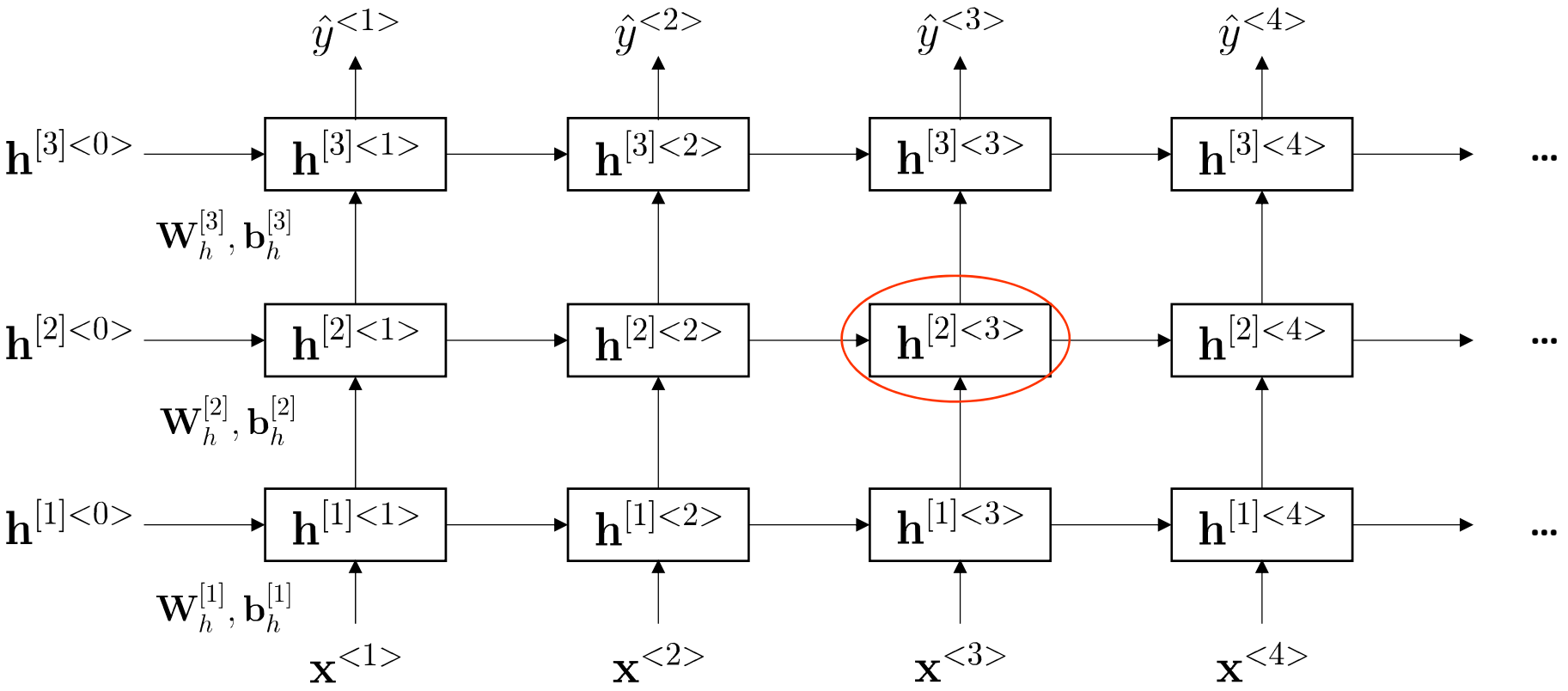
  循环神经网络



更多结构的RNN

大连理工大学 Dalian University of Technology

■ 深层循环神经网络 (Deep RNN)

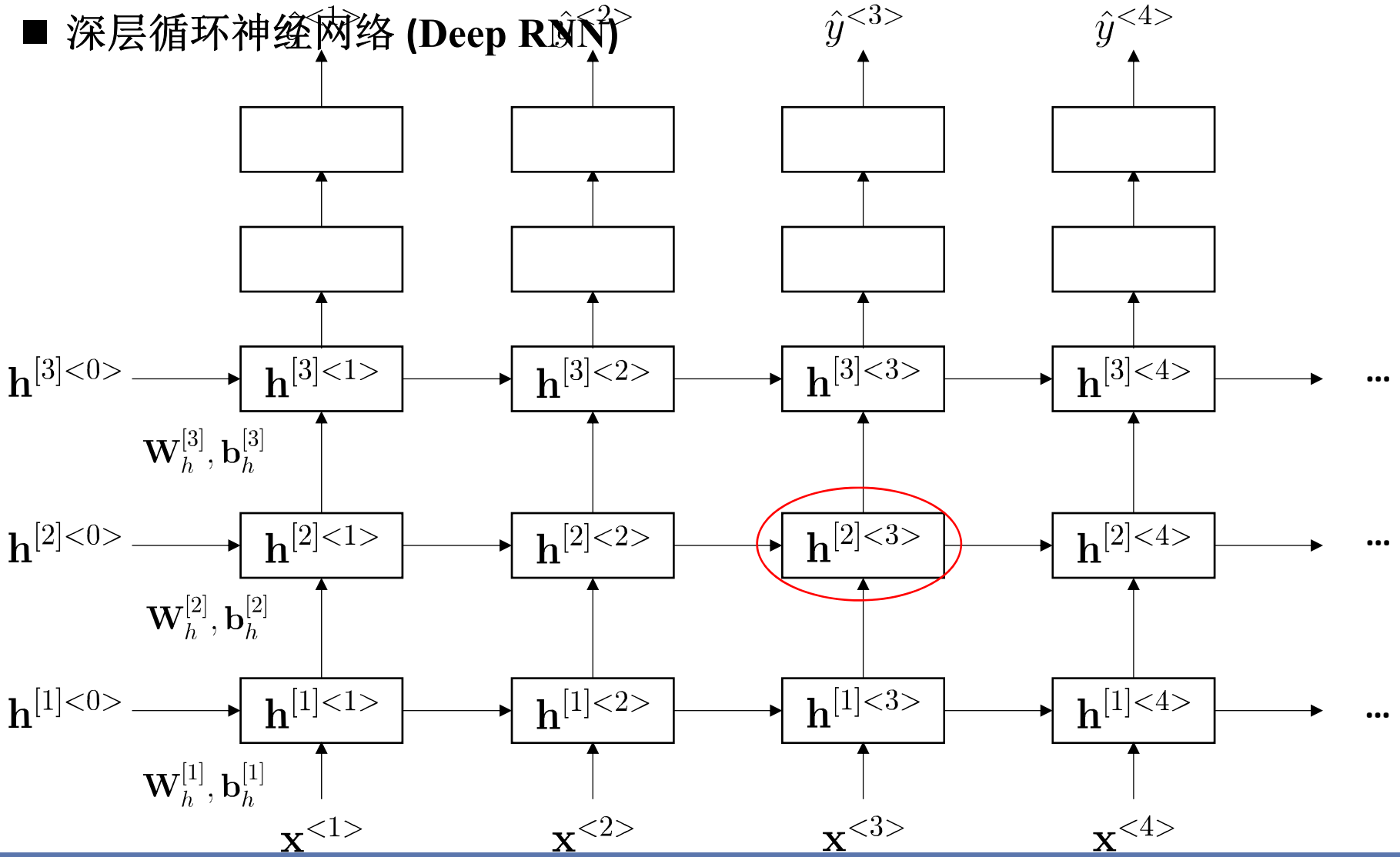


$$\mathbf{h}^{[2]<3>} = g(\mathbf{W}_h^{[2]}[\mathbf{h}^{[2]<2>}, \mathbf{h}^{[1]<3>}] + \mathbf{b}_h^{[2]})$$

$$\mathbf{h}^{[l]<t>} = g(\mathbf{W}_h^{[l]}[\mathbf{h}^{[l]<t-1>}, \mathbf{h}^{[l-1]<t>}] + \mathbf{b}_h^{[l]})$$

更多结构的RNN

- 深层循环神经网络 (Deep RNN)



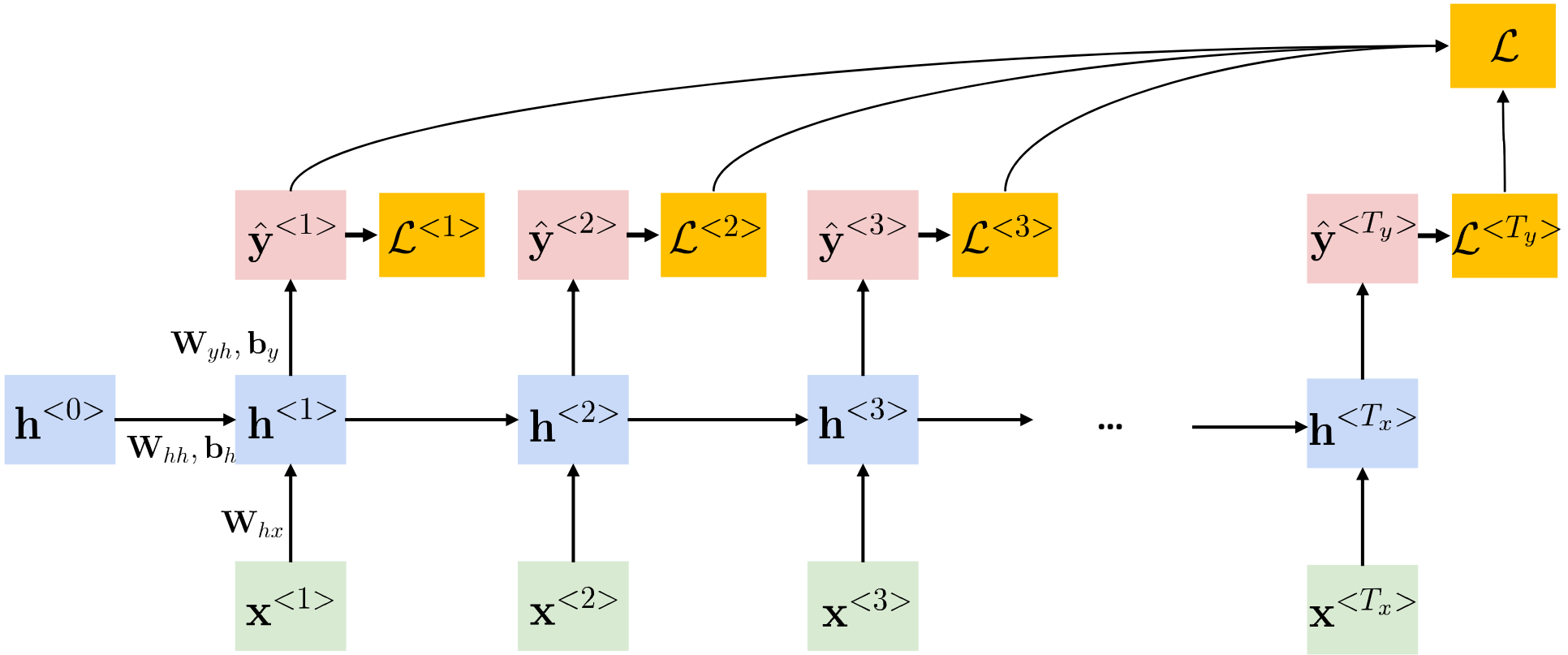
大连理工大学
Dalian University Of Technology

02

PART TWO

循环神经网络训练
RNN Training

循环神经网络训练



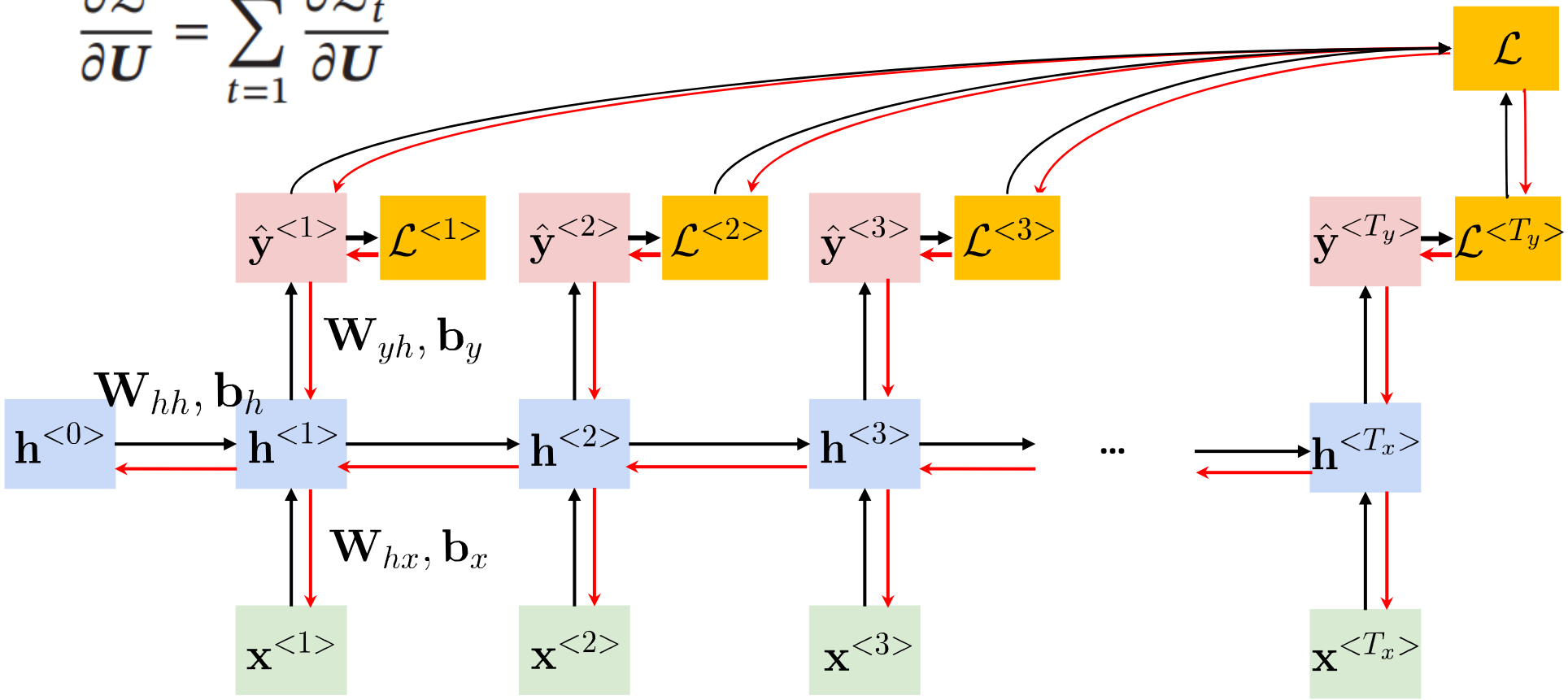
$$\mathcal{L}_t = \mathcal{L}(y_t, g(\boldsymbol{h}_t))$$

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t$$

循环神经网络训练

■ **BPTT (Back-propagation through time)：随时间反向传播**

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{U}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{U}}$$

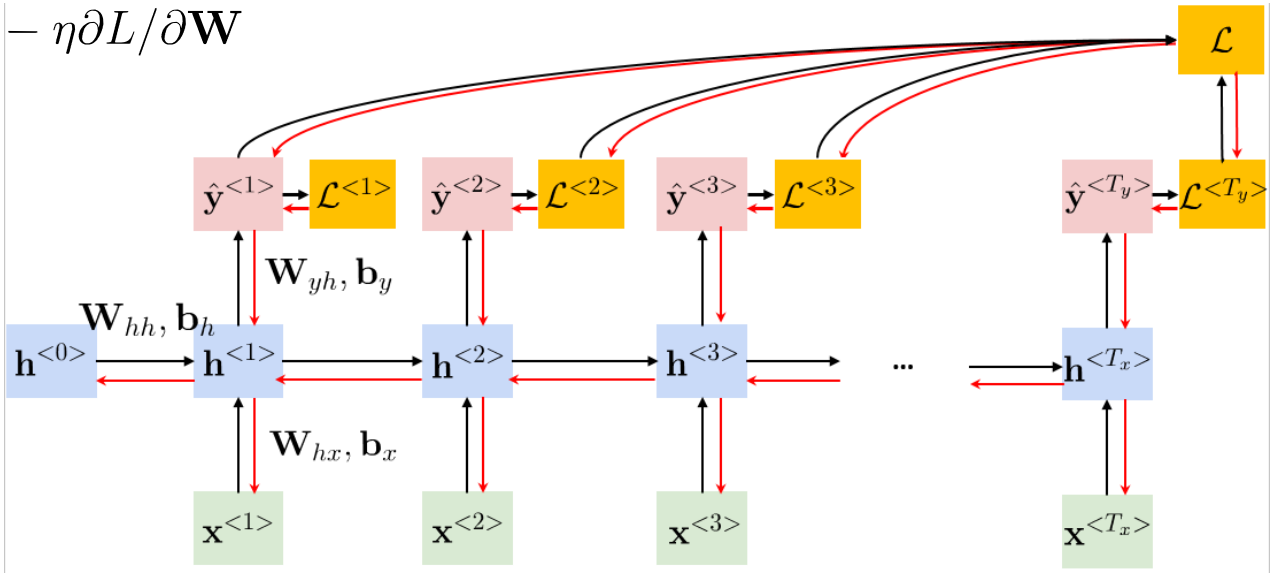


循环神经网络训练

- **BPTT (Back-propagation through time)：随时间反向传播**

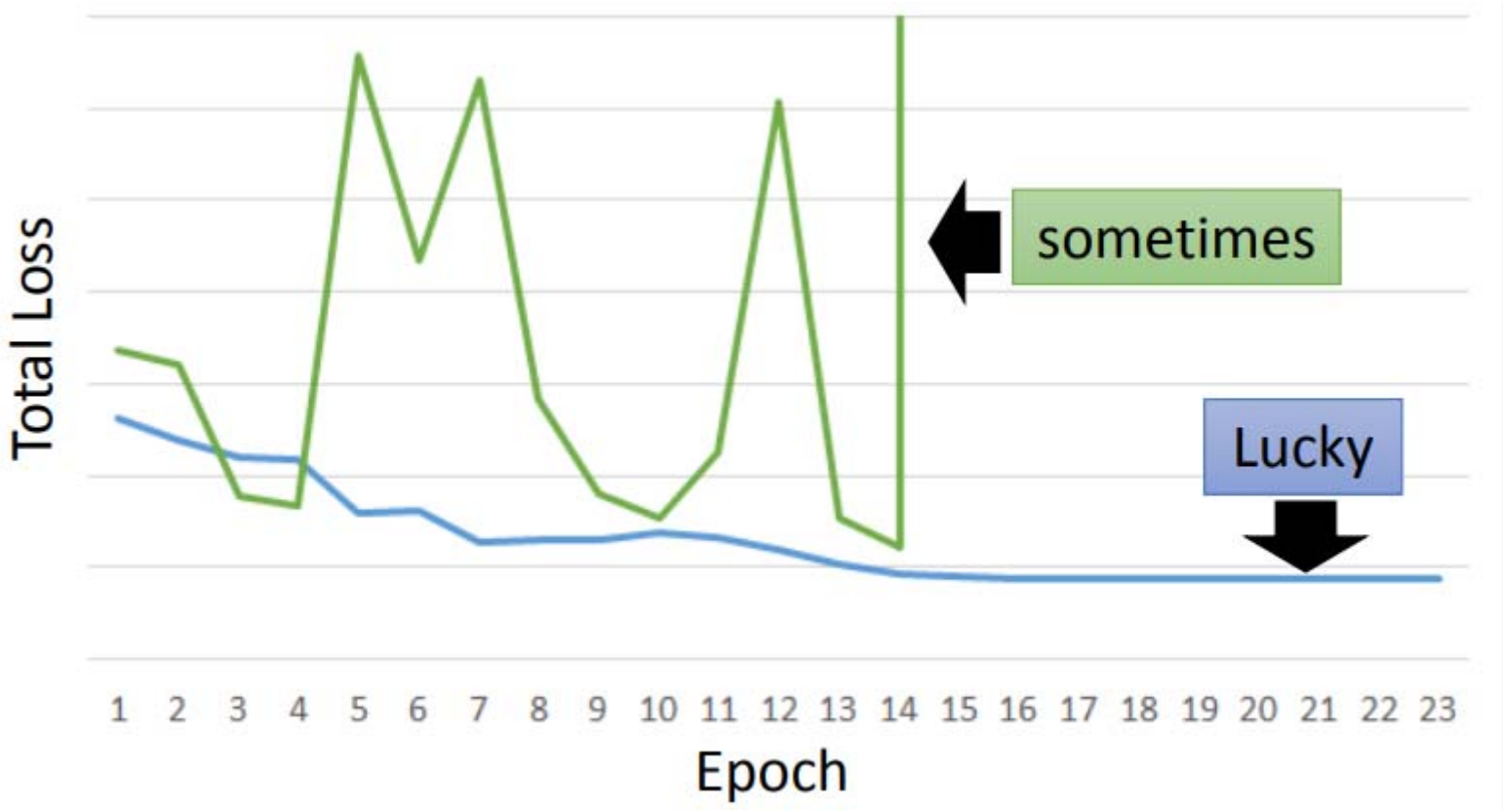
BPTT的中心思想和BP算法相同，本质是梯度下降法。需要寻优的参数有三个， $\mathbf{W}_{hh}, \mathbf{W}_{hx}, \mathbf{W}_{yh}$ **.与BP算法不同的是，** $\mathbf{W}_{hh}, \mathbf{W}_{hx}$ **参数在寻优的过程中需要追溯之前历史数据，因此RNN有记忆功能。**

$$\mathbf{W} \longleftarrow \mathbf{W} - \eta \partial L / \partial \mathbf{W}$$



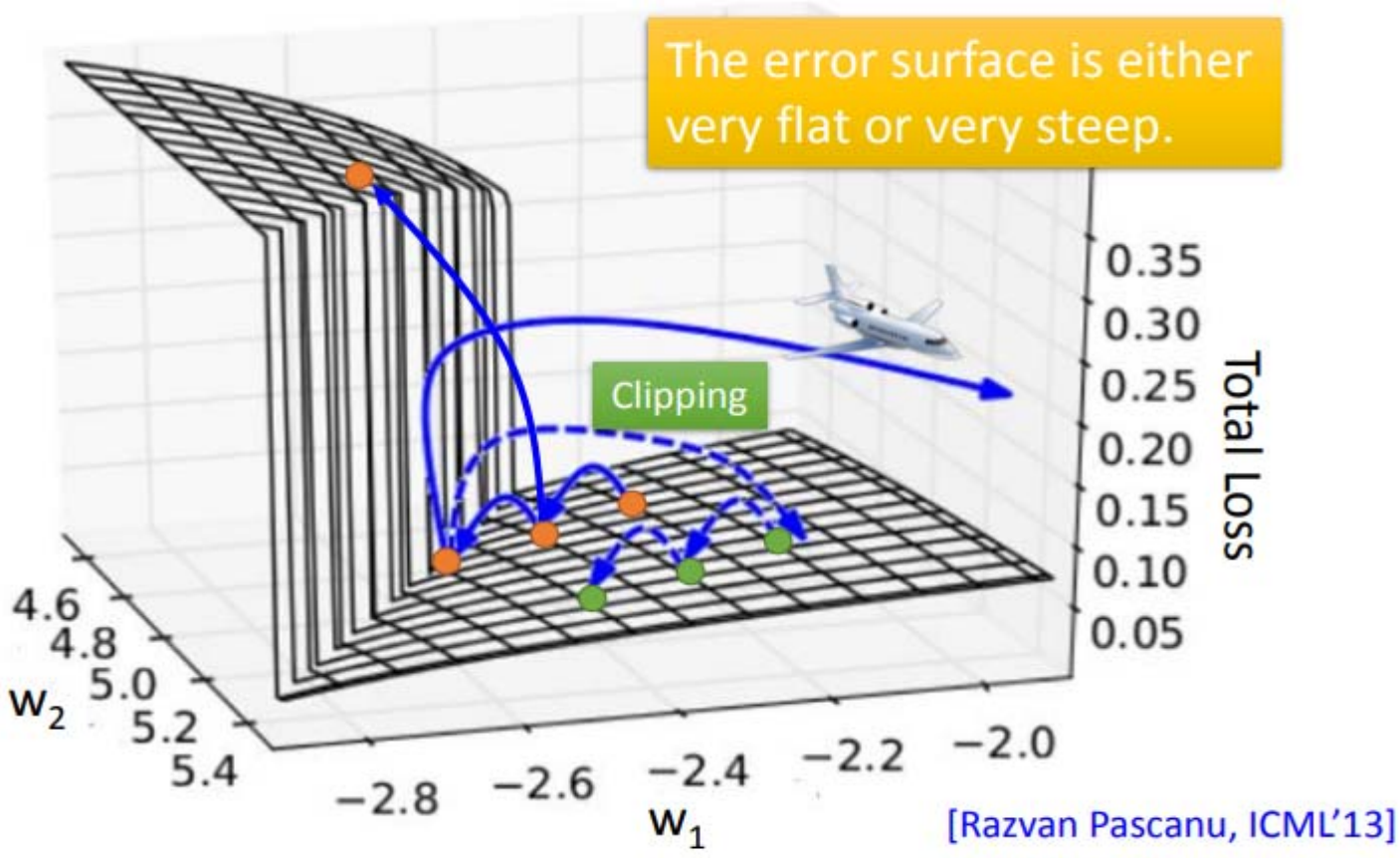
循环神经网络训练

- **BPTT (Back-propagation through time)：随时间反向传播**
  - **训练RNN通常比较困难**



循环神经网络训练

- **BPTT (Back-propagation through time)：随时间反向传播**
  - **训练RNN通常比较困难**



循环神经网络训练

- **BPTT (Back-propagation through time)：随时间反向传播**
  - 训练RNN通常比较困难

$w = 1 \rightarrow y^{1000} = 1$

$w = 1.01 \rightarrow y^{1000} \approx 20000$

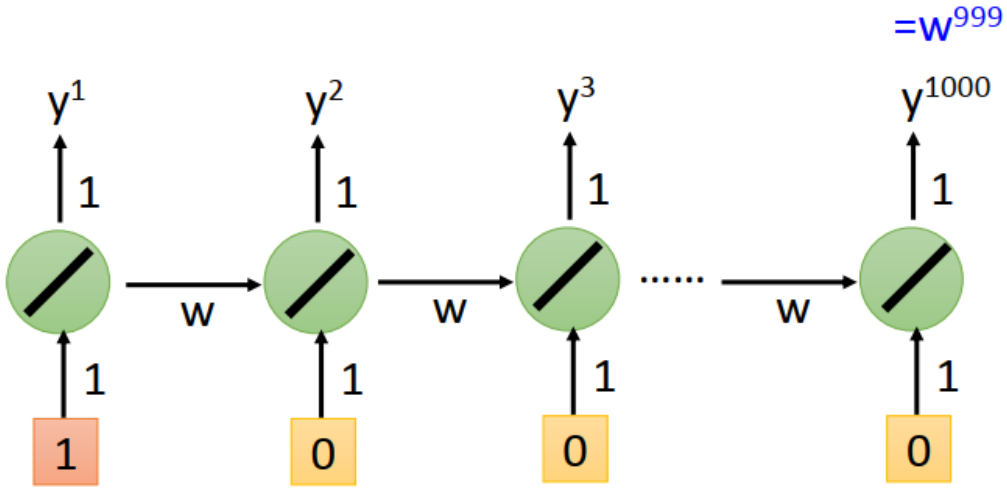
Large $\partial L/\partial w$ → Small Learning rate?

$w = 0.99 \rightarrow y^{1000} \approx 0$

$w = 0.01 \rightarrow y^{1000} \approx 0$

small $\partial L/\partial w$ → Large Learning rate?

Toy Example



循环神经网络训练

大连理工大学 Dalian University of Technology

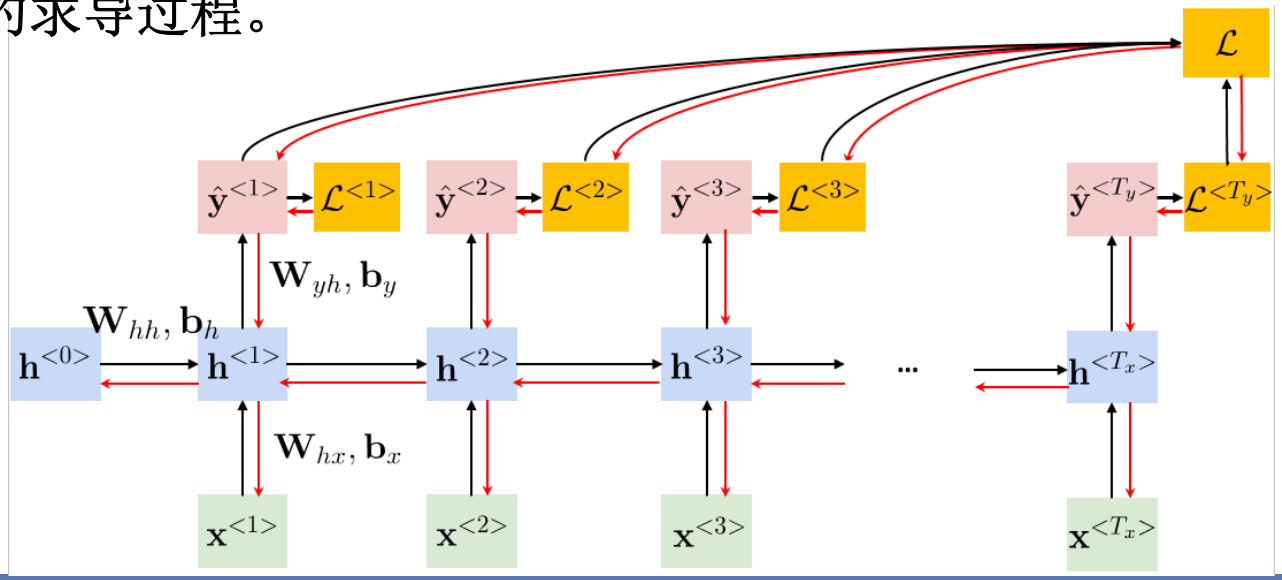
■ **BPTT (Back-propagation through time)：随时间反向传播**

□ 对 $\mathbf{W}_{yh}$ 求偏导:

$$\frac{\partial \mathcal{L}^{<t>}}{\partial \mathbf{W}_{yh}} = \frac{\partial \mathcal{L}^{<t>}}{\partial \hat{y}^{<t>}} \cdot \frac{\partial \hat{y}^{<t>}}{\partial \mathbf{W}_{yh}}$$

其中嵌套着激活函数，是复合函数的求导过程。

$$\mathbf{h}^{<t>} = g(\mathbf{W}_{hh}\mathbf{h}^{<t-1>} + \mathbf{W}_{hx}\mathbf{x}^{<t>} + \mathbf{b}_h)$$
$$\hat{y}^{<t>} = g_2(\mathbf{W}_{yh}\mathbf{h}^{<t>} + \mathbf{b}_y)$$
$$\mathcal{L}(\hat{\mathbf{y}}^{<t>}, \mathbf{y}^{<t>}) = -\sum_i y_i^{<t>} \log \hat{y}_i^{<t>}$$
$$\mathcal{L} = \sum_t \mathcal{L}^{<t>}(\hat{\mathbf{y}}^{<t>}, \mathbf{y}^{<t>})$$



循环神经网络训练

大连理工大学 Dalian University of Technology

■ **BPTT (Back-propagation through time)：随时间反向传播**

□ 对 $\mathbf{W}_{hh}$ 求偏导:

$$\frac{\partial \mathcal{L}^{<3>}}{\partial \mathbf{W}_{hh}} = \frac{\partial \mathcal{L}^{<3>}}{\partial \hat{y}^{<3>}} \cdot \frac{\partial \hat{y}^{<3>}}{\partial \mathbf{h}^{<3>}} \cdot \frac{\partial \mathbf{h}^{<3>}}{\partial \mathbf{W}_{hh}}$$

$$+ \frac{\partial \mathcal{L}^{<3>}}{\partial \hat{y}^{<3>}} \cdot \frac{\partial \hat{y}^{<3>}}{\partial \mathbf{h}^{<3>}} \cdot \frac{\partial \mathbf{h}^{<3>}}{\partial \mathbf{h}^{<2>}} \cdot \frac{\partial \mathbf{h}^{<2>}}{\partial \mathbf{W}_{hh}}$$

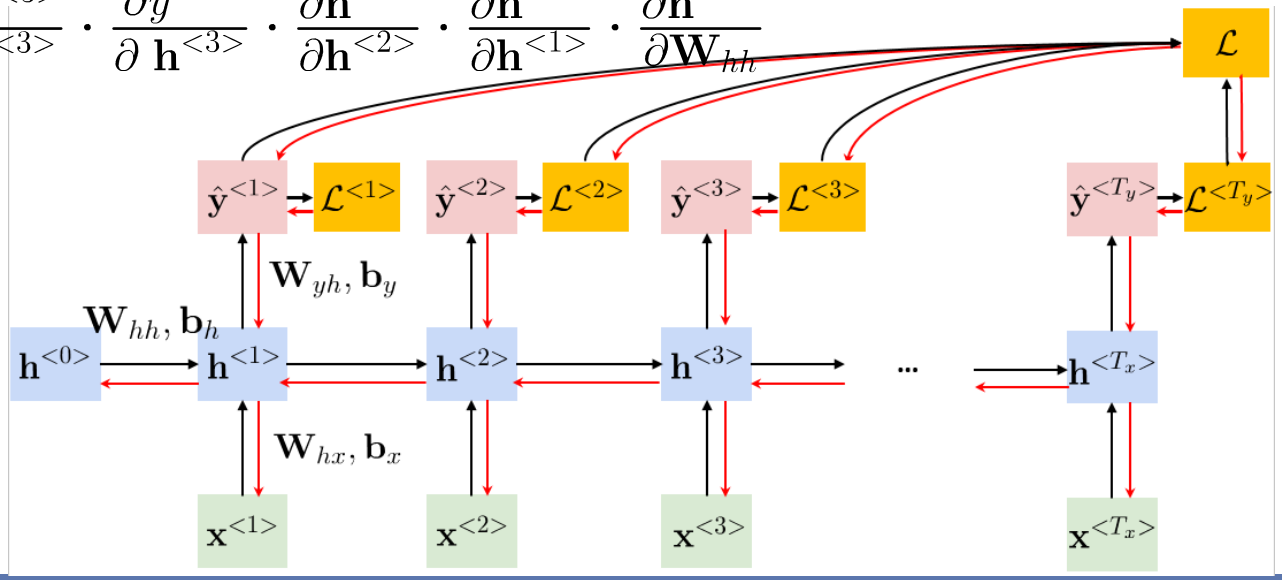
$$+ \frac{\partial \mathcal{L}^{<3>}}{\partial \hat{y}^{<3>}} \cdot \frac{\partial \hat{y}^{<3>}}{\partial \mathbf{h}^{<3>}} \cdot \frac{\partial \mathbf{h}^{<3>}}{\partial \mathbf{h}^{<2>}} \cdot \frac{\partial \mathbf{h}^{<2>}}{\partial \mathbf{h}^{<1>}} \cdot \frac{\partial \mathbf{h}^{<1>}}{\partial \mathbf{W}_{hh}}$$

$$\mathbf{h}^{<t>} = g(\mathbf{W}_{hh}\mathbf{h}^{<t-1>} + \mathbf{W}_{hx}\mathbf{x}^{<t>} + \mathbf{b}_h)$$

$$\hat{y}^{<t>} = g_2(\mathbf{W}_{yh}\mathbf{h}^{<t>} + \mathbf{b}_y)$$

$$\mathcal{L}(\hat{\mathbf{y}}^{<t>}, \mathbf{y}^{<t>}) = -\sum_i y_i^{<t>} \log \hat{y}_i^{<t>}$$

$$\mathcal{L} = \sum_t \mathcal{L}^{<t>}(\hat{\mathbf{y}}^{<t>}, \mathbf{y}^{<t>})$$



循环神经网络训练

大连理工大学 Dalian University of Technology

■ **BPTT (Back-propagation through time)：随时间反向传播**

□ 对 $\mathbf{W}_{hx}$ 求偏导:

$$\frac{\partial \mathcal{L}^{<3>}}{\partial \mathbf{W}_{hx}} = \frac{\partial \mathcal{L}^{<3>}}{\partial \hat{y}^{<3>}} \cdot \frac{\partial \hat{y}^{<3>}}{\partial \mathbf{h}^{<3>}} \cdot \frac{\partial \mathbf{h}^{<3>}}{\partial \mathbf{W}_{hx}}$$

$$+ \frac{\partial \mathcal{L}^{<3>}}{\partial \hat{y}^{<3>}} \cdot \frac{\partial \hat{y}^{<3>}}{\partial \mathbf{h}^{<3>}} \cdot \frac{\partial \mathbf{h}^{<3>}}{\partial \mathbf{h}^{<2>}} \cdot \frac{\partial \mathbf{h}^{<2>}}{\partial \mathbf{W}_{hx}}$$

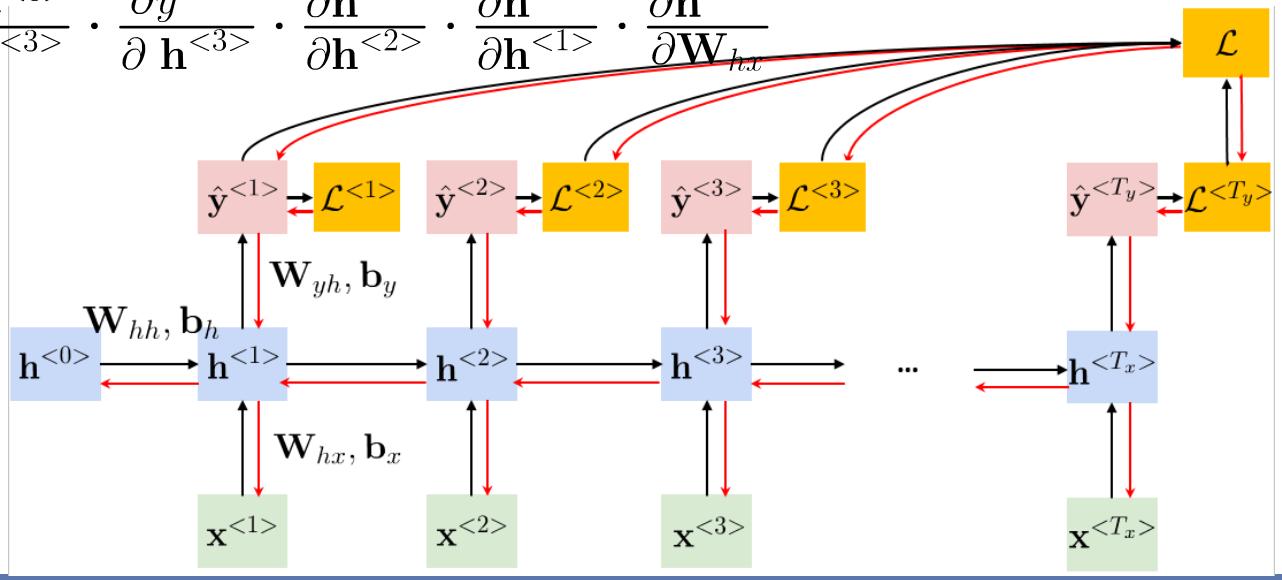
$$+ \frac{\partial \mathcal{L}^{<3>}}{\partial \hat{y}^{<3>}} \cdot \frac{\partial \hat{y}^{<3>}}{\partial \mathbf{h}^{<3>}} \cdot \frac{\partial \mathbf{h}^{<3>}}{\partial \mathbf{h}^{<2>}} \cdot \frac{\partial \mathbf{h}^{<2>}}{\partial \mathbf{h}^{<1>}} \cdot \frac{\partial \mathbf{h}^{<1>}}{\partial \mathbf{W}_{hx}}$$

$$\mathbf{h}^{<t>} = g(\mathbf{W}_{hh}\mathbf{h}^{<t-1>} + \mathbf{W}_{hx}\mathbf{x}^{<t>} + \mathbf{b}_h)$$

$$\hat{y}^{<t>} = g_2(\mathbf{W}_{yh}\mathbf{h}^{<t>} + \mathbf{b}_y)$$

$$\mathcal{L}(\hat{\mathbf{y}}^{<t>}, \mathbf{y}^{<t>}) = -\sum_i y_i^{<t>} \log \hat{y}_i^{<t>}$$

$$\mathcal{L} = \sum_t \mathcal{L}^{<t>}(\hat{\mathbf{y}}^{<t>}, \mathbf{y}^{<t>})$$



循环神经网络训练

大连理工大学 Dalian University of Technology

- **BPTT (Back-propagation through time)：随时间反向传播**

  - 对 $\mathbf{W}_{hh}$ 求偏导:

$$\frac{\partial \mathcal{L}^{<t>}}{\partial \mathbf{W}_{hh}} = \sum_{k=0}^{t} \frac{\partial \mathcal{L}^{<t>}}{\partial \hat{y}^{<t>}} \cdot \frac{\hat{y}^{<t>}}{\partial \mathbf{h}^{<t>}} \cdot \left(\prod_{j=k+1}^{t} \frac{\partial \mathbf{h}^{<j>}}{\partial \mathbf{h}^{<j-1>}} \right) \cdot \frac{\partial \mathbf{h}^{<k>}}{\partial \mathbf{W}_{hh}}$$

$$\prod_{j=k+1}^{t} (tanh)' \cdot \mathbf{W}_{hh}$$

  - 对 $\mathbf{W}_{hx}$ 求偏导:

$$\frac{\partial \mathcal{L}^{<t>}}{\partial \mathbf{W}_{hx}} = \sum_{k=0}^{t} \frac{\partial \mathcal{L}^{<t>}}{\partial \hat{y}^{<t>}} \cdot \frac{\hat{y}^{<t>}}{\partial \mathbf{h}^{<t>}} \cdot \left(\prod_{j=k+1}^{t} \frac{\partial \mathbf{h}^{<j>}}{\partial \mathbf{h}^{<j-1>}} \right) \cdot \frac{\partial \mathbf{h}^{<k>}}{\partial \mathbf{W}_{hx}}$$

$$\prod_{j=k+1}^{t} (sigmoid)' \cdot \mathbf{W}_{hh}$$

连乘会导致激活函数导数的连乘，进而会导致“梯度消失”和“梯度爆炸”现象的发生

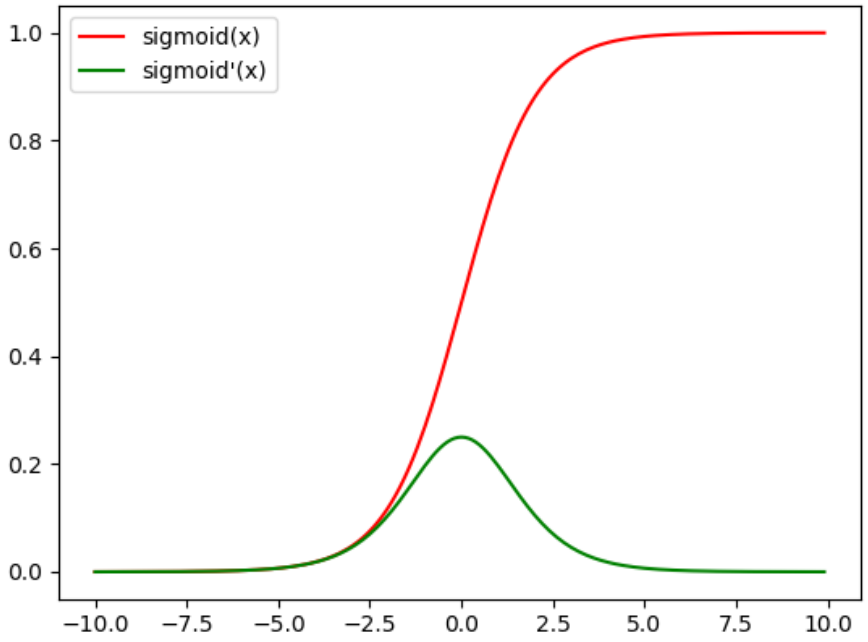
$$\mathbf{h}^{<t>} = g(\mathbf{W}_{hh}\mathbf{h}^{<t-1>} + \mathbf{W}_{hx}\mathbf{x}^{<t>} + \mathbf{b}_h)$$

$$\hat{y}^{<t>} = g_2(\mathbf{W}_{yh}\mathbf{h}^{<t>} + \mathbf{b}_y)$$

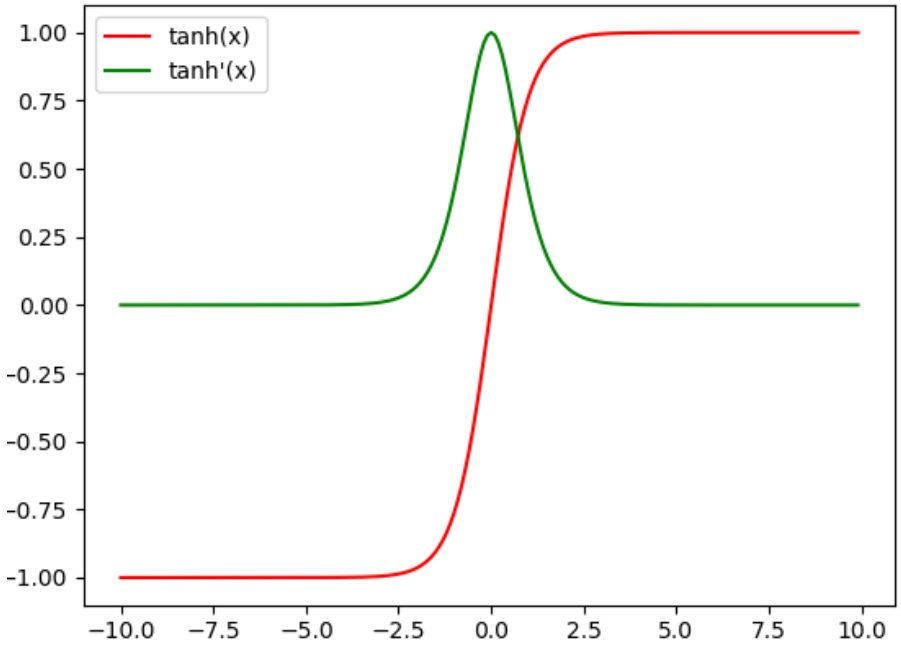
$$\mathcal{L}(\hat{\mathbf{y}}^{<t>}, \mathbf{y}^{<t>}) = -\sum_i y_i^{<t>} \log \hat{y}_i^{<t>}$$

$$\mathcal{L} = \sum_t \mathcal{L}^{<t>}(\hat{\mathbf{y}}^{<t>}, \mathbf{y}^{<t>})$$

循环神经网络训练



Sigmoid函数及其导数



Tanh函数及其导数

Sigmoid和tanh的导数都在[0,1]范围内，随着时间序列的不断深入，小数的累乘就会导致梯度越来越小直到接近于0，这就是“梯度消失”现象。梯度消失就意味消失那一层的参数再也不更新，那么那一层隐层就变成了单纯的映射层，毫无意义了。

循环神经网络训练

大连理工大学 Dalian University of Technology

- 长程依赖问题

-- 梯度爆炸问题：梯度削减

-- 梯度消失问题： 1. 选取更好的激活函数

2. 改变传播结构

-- 记忆容量问题：随着 $\boldsymbol{h}_t$ 不断累积存储新的输入信息，会发生饱和现象．假设 $g(\cdot)$ 为 Logistic 函数，则随着时间 t 的增长，$\mathbf{z}_t$会变得越来越大， 从而导致 $\boldsymbol{h}$ 变得饱和. 也就是说，隐状态 $\boldsymbol{h}_t$ 可以存储的信息是有限的，随着记忆单元存储的内容越来越多, 其丢失的信息也越来越多.

$$\mathbf{h}^{<t>} = g(\mathbf{W}_{hh}\mathbf{h}^{<t-1>} + \mathbf{W}_{hx}\mathbf{x}^{<t>} + \mathbf{b}_h)$$

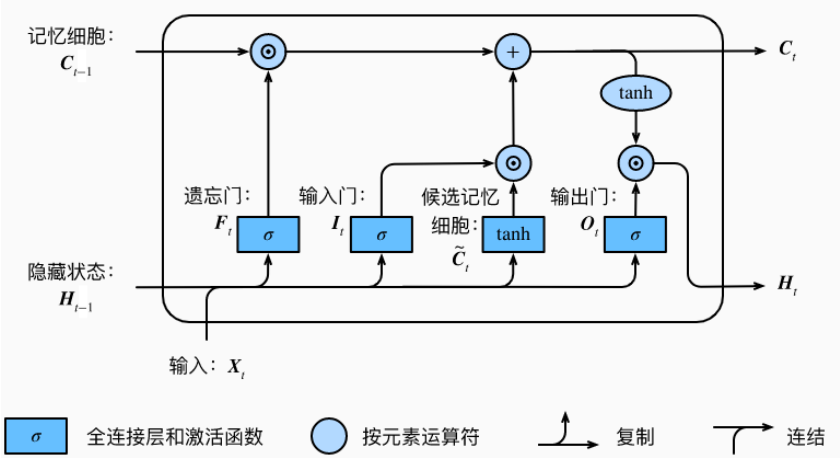
循环神经网络训练

- 长程依赖问题
  - -- 梯度爆炸问题
  - -- 梯度消失问题
  - -- 记忆容量问题

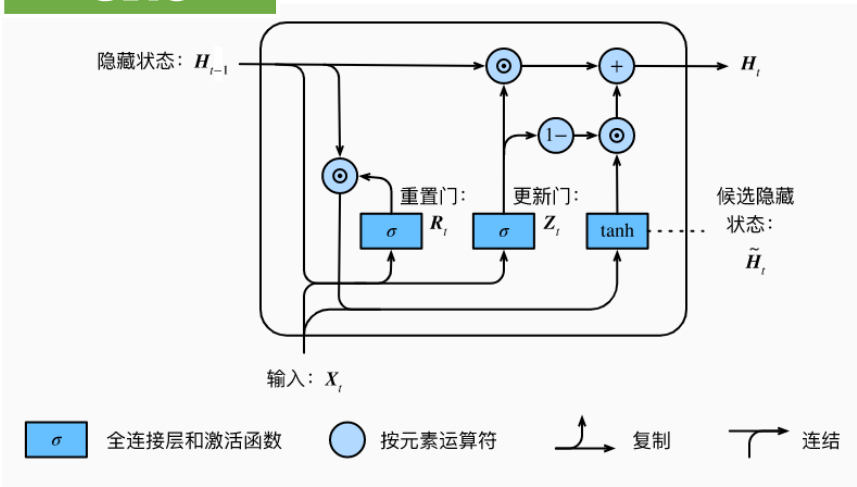
- 如何解决？

引入门控机制来控制信息的累积速度， 包括有选择地加入新的信息，并有选择地遗忘之前累积的信息.

LSTM



GRU



大连理工大学
Dalian University Of Technology

PART THREE

小结和资源

Take Home Message and Resource

小结

- **RNN的结构**
- 语言模型
- **Bi-RNN & Deep RNN**
- **RNN的训练**
  - 随时间反向传播
  - 梯度爆炸
  - 梯度消失