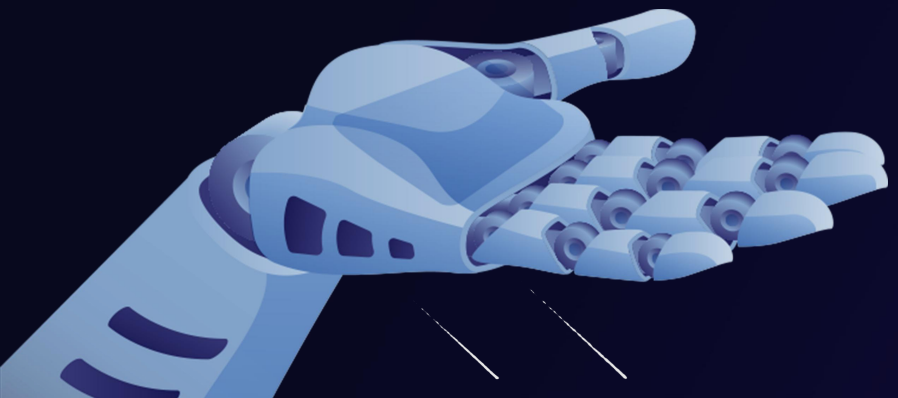
第三章：自然语言处理综述



CONTENTS

3.1 自然语言处理简介：概念、难点、任务和历史

自然语言通常指的是人类语言，是人类思维的载体和交流的基本工具，也是人类区别于动物的根本标志，更是人类智能发展的外在体现形式之一。

自然语言处理（Natural Language Processing，NLP）主要研究用计算机理解和生成自然语言的各种理论和方法，属于人工智能领域的一个重要甚至核心分支，是计算机科学与语言学的交叉学科，又常被称为计算语言学（Computational Linguistics，CL）。随着互联网的快速发展，网络文本呈爆炸性增长，为自然语言处理提出了巨大的应用需求。同时，自然语言处理研究也为人们更深刻地理解语言的机理和社会的机制提供了一条重要的途径，因此具有重要的科学意义。

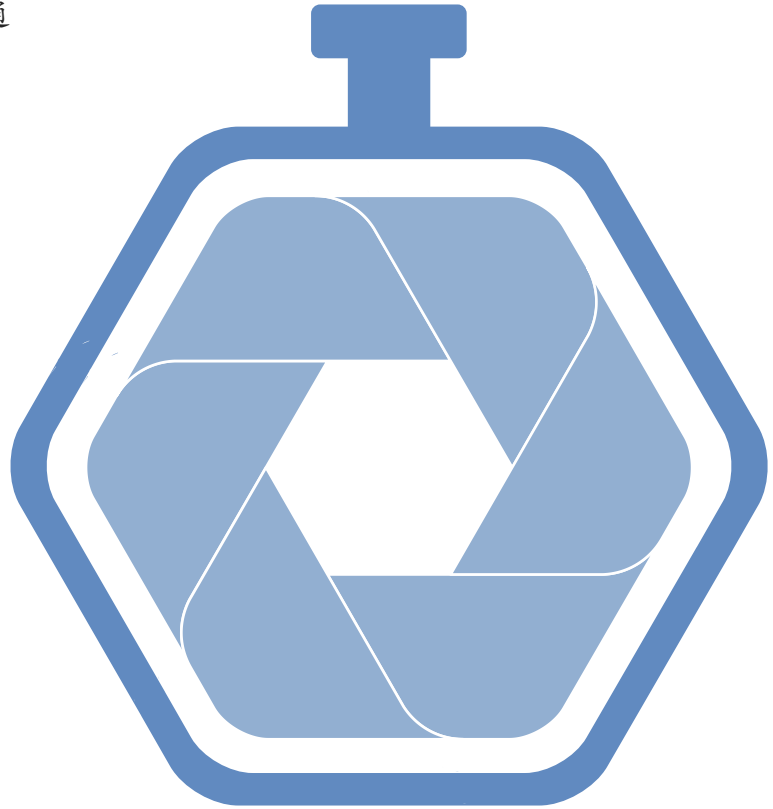
3.1 自然语言处理简介：概念、难点、任务和历史

1、抽象性
语言由抽象符号组成，如“车”表示各种交通工具-汽车、火车、自行车等，它们有共同的属性，有轮子、能载人。

2、组合性
有限的基本符号组成无穷的语义。

3、歧义性
“苹果”代表吃的水果，也指电脑设备。

4、进化性
新词层出不穷，旧词赋予新义。



5、非规范性
音近词：为什么-》为森么，单词简写或变形：please->pls，新造词：喜大普奔

6、主观性
主观性提高了数据标注的难度，也为准确评价系统的表现带来了一定的困难。

7、知识性
理解语言通常需要背景知识以及基于这些知识的推理能力。

8、难移植性
不同任务间移植困难，比如抽取任务难移植到对话系统。

3.1 自然语言处理简介：概念、难点、任务和历史

任务层级

任务类别

1、回归问题

即将输入文本映射为一个连续的数值，如对作文的打分，对案件刑期或罚款金额的预测等。

2、分类问题

又称为文本分类，即判断一个输入的文本所属的类别，如垃圾邮件识别和情感分类。

3、匹配问题

判断两个输入文本之间的关系，如：复述或非复述两类关系；蕴含、矛盾和无关三类关系。文本之间的相似性（0到1的数值）

4、解析问题

指对文本中的词语进行标注或识别词语之间的关系，包括词性标注、句法分析、分词、命名实体识别。

5、生成问题

特指根据输入（可以是文本，也可以是图片、表格等其他类型数据）生成一段自然语言，如机器翻译、文本摘要、图像描述生成等都是典型的文本生成类任务。

3.1 自然语言处理简介：概念、难点、任务和历史

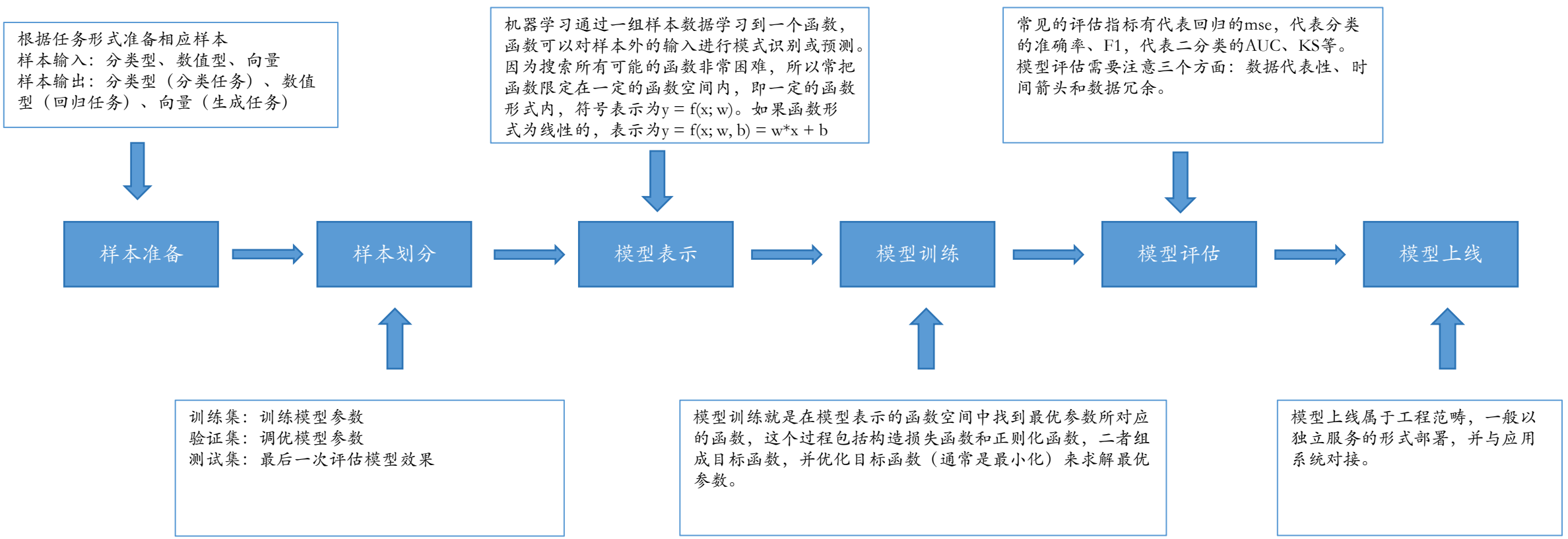
1）**小规模专家知识**：早期的自然语言处理主要采用基于理性主义的规则方法，通过专家总结的符号逻辑知识处理通用的自然语言现象。然而，由于自然语言的复杂性，基于理性主义的规则方法在面对实际应用场景中的问题时显得力不从心。

2）**大规模语料库统计模型**：基于语料库的统计自然语言处理方法能够更加客观、准确和细致地捕获语言规律。在这一时期，词法分析、句法分析、信息抽取、机器翻译和自动问答等领域的研究均取得了一定程度的成功。但它也有明显的局限性，也就是需要事先利用经验性规则将原始的自然语言输入转化为机器能够处理的向量形式，也被称为特征工程。

3）**大规模语料库深度学习**：2010年之后，随着基于深度神经网络的表示学习方法（也称深度学习）的兴起，该方法直接端到端地学习各种自然语言处理任务，不再依赖人工设计的特征。另一个好处是打通了不同任务之间的壁垒。完全革新了生成类模型，比如机器翻译、文本摘要和人机对话等任务。但是基于深度学习的算法有一个致命的缺点，就是过度依赖于大规模有标注数据。

4）**大规模预训练语言模型**：早期的静态词向量预训练模型，以及后来的动态词向量预训练模型，特别是2018年以来，以BERT、GPT为代表的超大规模预训练语言模型恰好弥补了自然语言处理标注数据不足的缺点，帮助自然语言处理取得了一系列的突破，使得包括阅读理解在内的所有自然语言处理任务的性能都得到了大幅提高，在有些数据集上达到或甚至超过了人类水平。

3.2 机器学习一般过程



四大机器学习任务

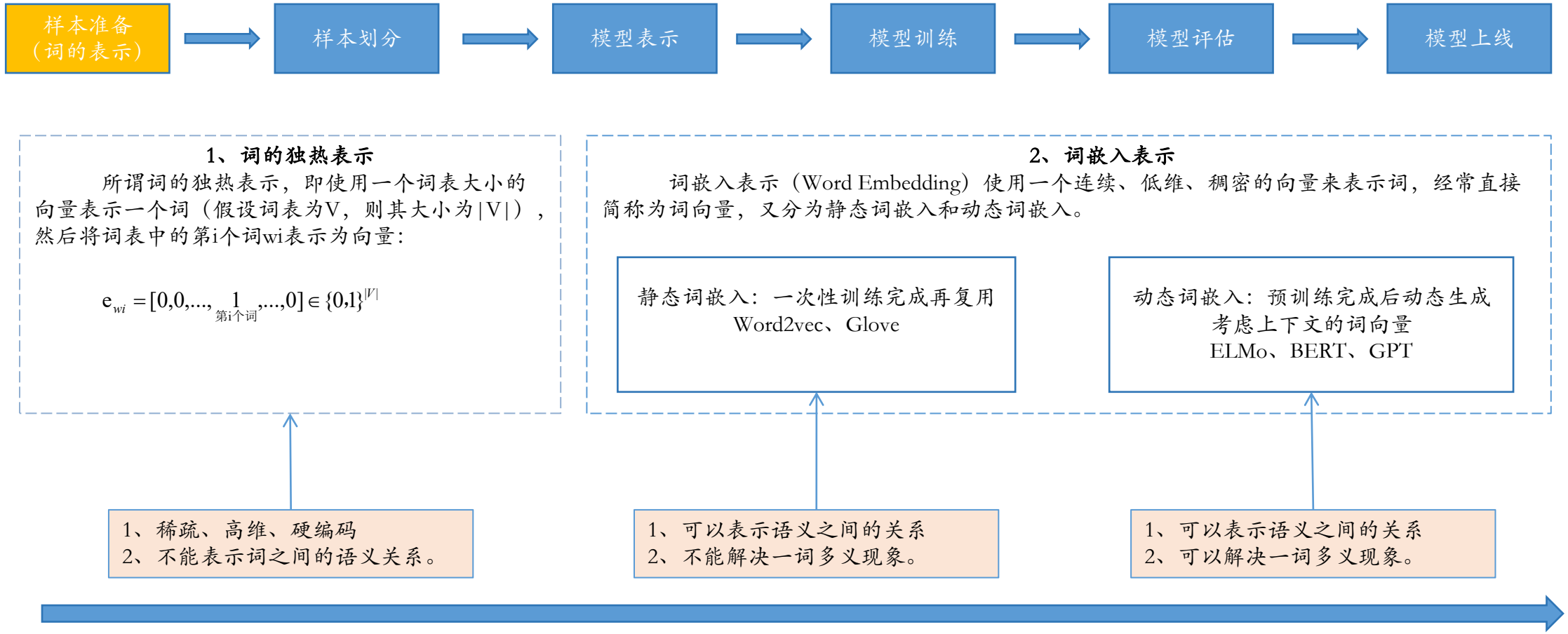
1、监督学习

2、无监督学习

3、自监督学习

4、强化学习

3.3 词的表示：独热表示和词嵌入表示

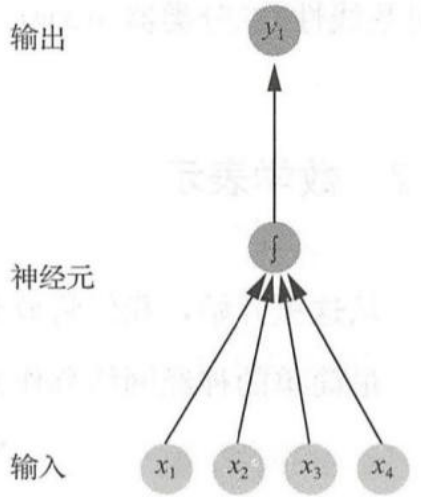


3.4 模型表示：MLP、RNN和LSTM三种神经网络结构

1、神经网络：一个大脑的比喻

顾名思义，神经网络的灵感来源于大脑的计算机制，它由被称为神经元的计算单元组成。在这个比喻中，神经元是具有标量输入和输出的计算单元，每个输入都有与其相关联的权重，神经元将每个输入乘以其权重并将它们相加，然后使结果通过一个非线性函数，最终传递给其输出。

神经元彼此连接，形成网络，神经元的输出可能会提供给一个或多个神经元作为输入。这样的网络被证明是功能强大的计算工具。如果权重设置正确，具有足够多神经元和非线性激活函数的神经网络可以近似模拟种类非常广泛的数学函数。



3.4 模型表示：MLP、RNN和LSTM三种神经网络结构

2、多层感知机（MLP）

1）多层感知机（MLP），也称为前馈神经网络（Feedforward Neural Network），由输入层、隐藏层和输出层组成。

2）隐藏层通常包括一层或多层，每层由多个神经元组成。

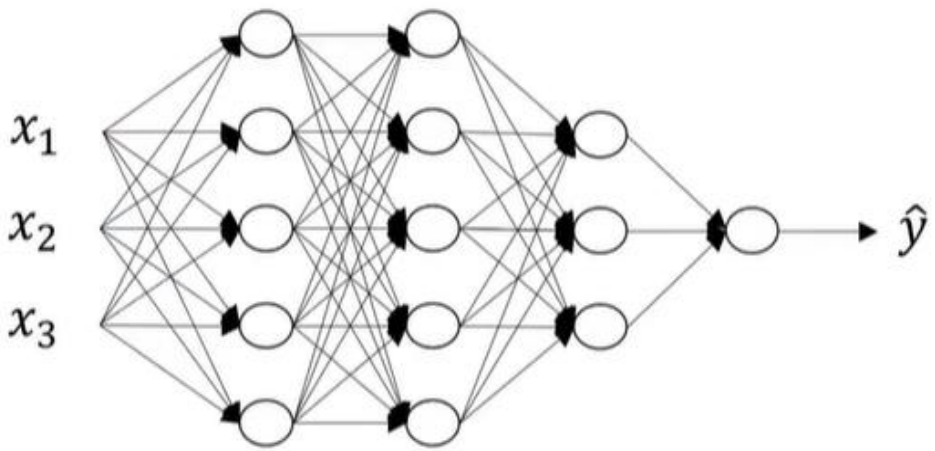
3）右侧的MLP网络结构有三个隐藏层组成，符号表示如下：

$$a^{[0]} = x$$

$$z^{[l]} = W^{[l]} \cdot a^{[l-1]} + b^{[l]}$$

$$a^{[l]} = g^{[l]}(z^{[l]})$$

$$\hat{y} = a^{[L]}$$

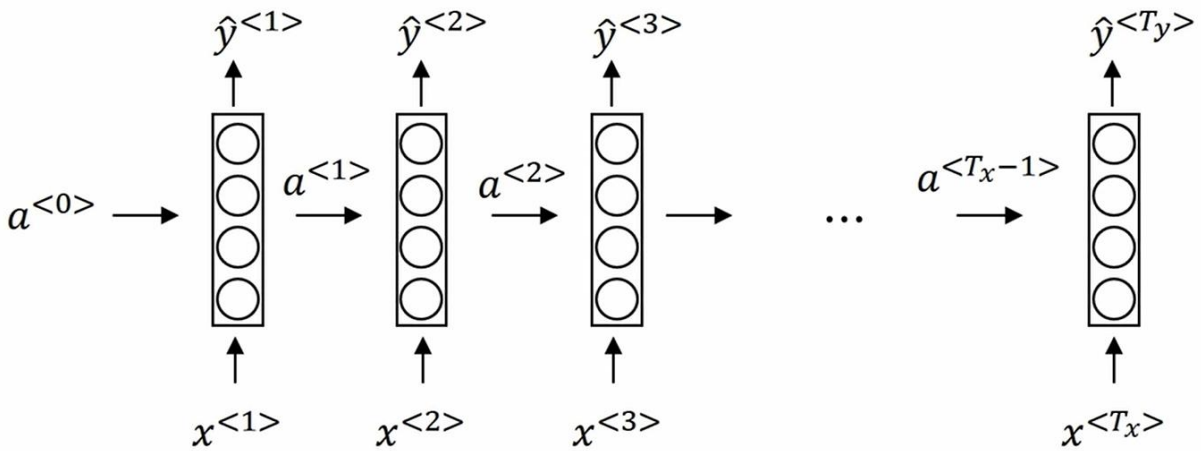


3.4 模型表示：MLP、RNN和LSTM三种神经网络结构

3、循环神经网络（RNN）

循环神经网络（Recurrent Neural Network，RNN）是一类具有短期记忆能力的神经网络，适合用于处理视频、语音、文本等与时序相关的问题。

在 RNN 中，神经元不仅可以接收其他神经元的信息，还可以接收自身的信息，形成具有环路的网络结构。相比于传统的前馈神经网络，RNN 能够处理任意长度的序列，对于处理顺序或时间问题（如语言翻译、语音识别）非常有效。



$$a^{<t>} = g(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

$$\hat{y}^{<t>} = g(W_{ya}a^{<t>} + b_y)$$

3.4 模型表示：MLP、RNN和LSTM三种神经网络结构

3、循环神经网络(LSTM)

在原始的循环神经网络中，信息是通过多个隐含层逐层传递到输出层的。当层数非常深时，可能会引起梯度爆炸和梯度消失，对于梯度爆炸可以采用梯度等比例缩减技术来解决，长短时记忆网络（LSTM）可以较好地解决梯度消失问题。在LSTM中，通过三个门（更新门、遗忘门、输出门）来决定保留上一节点的信息，还是使用更新后的信息。

$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$

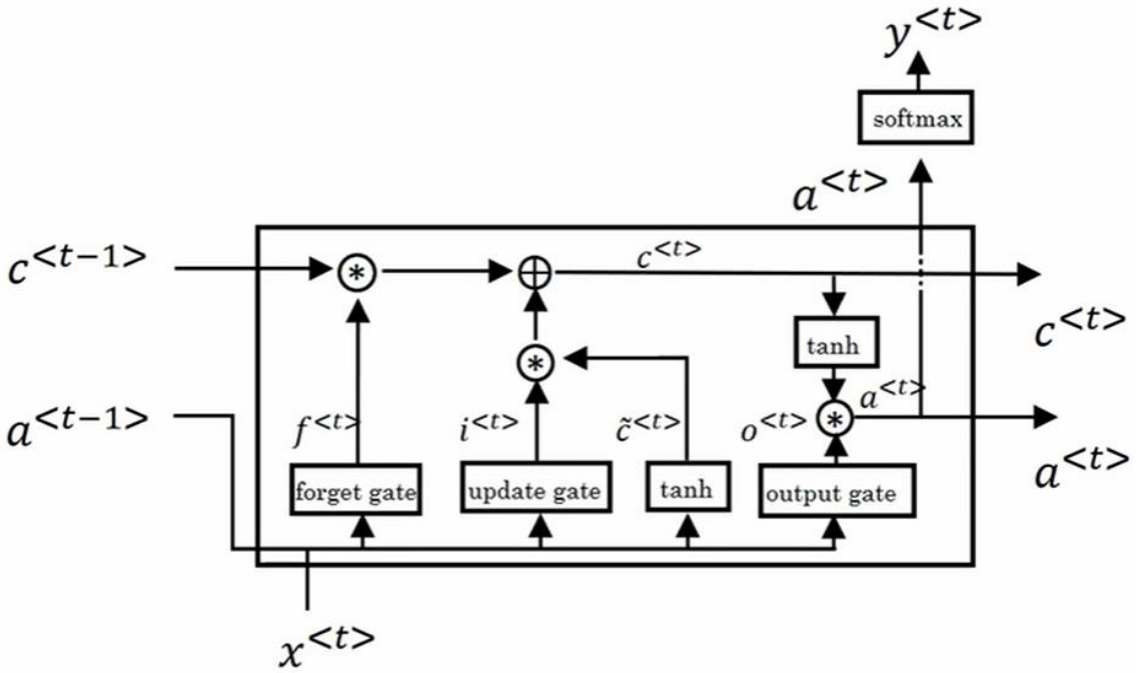
$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$$

$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

$$a^{<t>} = \Gamma_o * \tanh c^{<t>}$$



3.5 模型训练：损失函数、正则化和参数优化

1、损失函数

二元交叉熵：$l(y,\hat{y}) = -y\log\hat{y} - (1-y)\log(1-\hat{y})$

多元交叉熵：$\mathrm{l}(y,\hat{y}) = -\sum_{\mathrm{i}=1}^{M} y_i \log(\hat{y}_i)$

差平方：$\mathrm{l}(y,\hat{y}) = (y-\hat{y})^2$

2、正则化函数

*L*1正则：$R_{L1}(w) = \|w\|_1 = \sum_i |w_i|$

L2正则：$\mathrm{R}_{\mathrm{L2}}(w) = \|w\|_2^2 = \sum_i w_i^{\ 2}$

弹性正则：$\mathrm{R}_{elastic-net}(w) = \alpha R_{L1}(w) + (1-\alpha)R_{L2}(w)$

dropout 正则

3、目标函数&参数优化

$$\text{objective} = \sum l(y_{\mathrm{i}}, \hat{y}_i) + \lambda R(w)$$

$$\mathrm{w}^* = \arg\min_w objective = \arg\min_w(\sum l(y_{\mathrm{i}}, \hat{y}_i) + \lambda R(w))$$

3.5 模型训练：损失函数、正则化和参数优化

梯度下降法（一）

前面提到的深度学习参数求解在数学上称为无约束凸优化，梯度下降法是无约束优化问题的常用求解方法，求解过程如下公式，先从任意一个点出发，沿着函数值下降最快的一个方向走一小段，然后在新的起点，继续沿着函数值下降的一个方向走一小段，直至走到最小值所在的点，即满足收敛条件。下图中p代表方向，x代表每一步的点,λ代表步幅。

$$x^{(1)} = x^{(0)} + \lambda p^{(0)}$$

...

$$x^{(k+1)} = x^{(k)} + \lambda p^{(k)}$$

...

$$x^{(e)} = x^{(e-1)} + \lambda p^{(e-1)}$$

梯度下降法（二）

其中方向p(k)为函数负梯度方向，推理过程如下。首先f(x(k+1))的泰勒公式见下，从x(k)走一小段，是希望f(k+1)下降的，下降的越多越好，这个等式中除了p(k)其它都是确定的值，所以能让f(k+1)下降的只能靠p(k) 了，可以发现当p(k)的方向为x(k)的梯度方向的反方向时，取值最小，且为负值，这时候f(k+1)下降的最多。因此，沿着函数负梯度方向前进可以最快逼近最小值。

$$\mathrm{f}(x^{\mathrm{k+1}}) = f(x^k + \lambda p^k)$$

$$= f(x^k) + \lambda \nabla f(x^k) p^k + \mathrm{R}_n$$

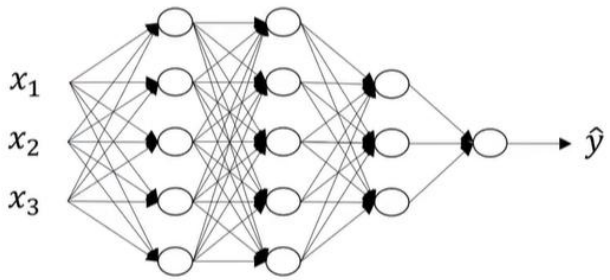
$$\approx f(x^k) + \lambda \nabla f(x^k) p^k$$

3.5 模型训练：损失函数、正则化和参数优化

样本准备（词的表示） → 样本划分 → 模型表示（神经网络） → **模型训练** → 模型评估 → 模型上线

梯度下降求解最优参数在神经网络中的具体表现为前向和后向传播算法。下图以一个三层神经网络为例来说明，其符号表示为：

$$a^{[0]} = x$$
$$z^{[l]} = W^{[l]} \cdot a^{[l-1]} + b^{[l]}$$
$$a^{[l]} = g^{[l]}(z^{[l]})$$
$$\hat{y} = a^{[L]}$$



前向传播

作用：更新最新参数的函数值

输入：$a^{[l-1]}$

输出：$a^{[l]}$

过程：$z^{[l]} = W^{[l]} \cdot a^{[l-1]} + b^{[l]}$

$a^{[l]} = g^{[l]}(z^{[l]})$

后向传播

作用：更新参数W

输入：$da^{[l]}$

输出：$dw^{[l]}$ $db^{[l]}$ $da^{[l-1]}$

过程：$dz^{[l]} = da^{[l]} * g^{[l]\prime}(z^{[l]})$

$dw^{[l]} = dz^{[l]} \cdot a^{[l-1]}$

$db^{[l]} = dz^{[l]}$

$da^{[l-1]} = w^{[i]T} \cdot dz^{[l]}$

3.6 模型评估：分类、解析、生成、回归四大问题评估方法

1、分类问题\匹配问题

某一类别准确率 = 某一类别正确识别的数量/某一类别识别的总数量
某一类别召回率 = 某一类别正确识别的数量/某一类别样本中的总数量
某一类别F1 = 2*某一类别准确率*某一类别召回率/(某一类别准确率+某一类别召回率)
整体准确率 = 整体正确识别的数量/整体识别的总数量
整体F1 = 某一类别F1的简单平均或加权平均（类别样本量比例作为权重）

2、解析问题

准确率 = 正确识别的数量/识别的总数量
召回率 = 正确识别的数量/样本中的总数量
F1 = 2*准确率*召回率/(准确率+召回率)

3、生成问题

BLEU = 预测文本和参考文本的N-gram匹配数/参考文本的N-gram总数

4、回归问题

$$R^2 = SSR / SST = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = 1 - SSE / SST = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| * 100$$

感谢您的观看