深度学习2025年考试原题

选择部分(4 × 10)

1、下列哪家AI公司是中国的？

2-9、知识点比较基础，需要大家根据老师画的资料进行分析，每年都不太一样，具体而言今年考到了鞍点，交叉熵，激活函数等知识点。第一题往往是较为热点的问题，也非常简单，比如24年考到了诺奖Hinton，今年考了Deepseek。

简答题(60)

1、ReLU激活函数

（1）给出 ReLU 函数的定义，并写出其导数。

（2）已知 $\sigma(x)$ 为 ReLU 函数，计算 $\sigma(2)$ 和 $\sigma'(2)$。

（3）比较 Sigmoid 与 ReLU 在反向传播中的梯度特性，说明为什么 ReLU 可以缓解梯度消失问题。

2、给定线性回归模型$\hat{y} = \omega x + b$；和平方损失函数$L = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2$；

（1）推导损失函数对参数ω和b的梯度$\frac{\delta L}{\delta \omega}$ $\frac{\delta L}{\delta b}$。

（2）在给定一组样本数据$(x_1, y_1) = (1, 10)$ $(x_2, y_2) = (2, 5)$以及当前$\omega = 0.5, b = 0.5$的情况下，计算梯度数值。给定学习率$\alpha = 0.01$，写出一次梯度下降后的参数更新结果。

3、梯度消失与梯度爆炸

（1）简述反向传播中梯度消失和梯度爆炸产生的原因。

（2）梯度消失和梯度爆炸会带来什么问题？请各给出至少一种缓解方法。

4、RNN 基本计算公式

简述RNN的计算流程，需要写出循环神经网络（RNN）中隐状态h_t和y_t的计算公式。

5、Adam 优化器的核心思想

Adam 优化算法结合了哪些关键思想？请简要说明。

6、Transformer 结构

叙述 Transformer 编码器的基本结构，并按顺序列出。

7、Encoder - Decoder 架构

简述编码器解码器架构的基本结构和计算流程。

8、BERT 与 GPT 的训练目标对比

分别说明 BERT 和 GPT 的预训练目标，并指出二者的主要区别。

参考答案

第 2 题（激活函数）

ReLU 函数定义：

$$\sigma(x) = \max(0, x)$$

ReLU 的导数：

$$\sigma'(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases}$$

因此有：

$$\sigma(2) = 2, \quad \sigma'(2) = 1$$

Sigmoid 与 ReLU 的比较：

Sigmoid 函数的导数最大值小于 1，在反向传播过程中多层连乘后容易导致梯度消失；而 ReLU 在正区间导数恒为 1，可以在一定程度上缓解梯度消失问题。

第 3 题（线性回归与梯度下降）

损失函数关于参数的梯度：

$$\frac{\partial L}{\partial w} = -\sum_{i=1}^{n} (y_i - (wx_i + b))x_i$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} (y_i - (wx_i + b))$$

梯度计算结果：

$$\frac{\partial L}{\partial w} = -16$$

$$\frac{\partial L}{\partial b} = -12.5$$

参数更新（学习率为 α）：

$$w = w - \alpha \frac{\partial L}{\partial w} = 0.66$$

$$b = b - \alpha \frac{\partial L}{\partial b} = 0.625$$

第 4 题（梯度消失与梯度爆炸）

在反向传播过程中，由于链式法则的连乘形式，当各层梯度项小于 1 时，梯度会不断衰减，导致**梯度消失**；当各层梯度项大于 1 时，梯度会迅速增大，导致**梯度爆炸**。

梯度消失的影响与缓解方法：

梯度消失会导致模型训练缓慢甚至无法有效更新参数。常见缓解方法包括使用 ReLU 激活函数、减少网络深度、批归一化（Batch Normalization）以及残差连接。

梯度爆炸的影响与缓解方法：

梯度爆炸会使训练过程不稳定。常见缓解方法包括减小学习率、使用正则化或权重衰减、梯度裁剪（Gradient Clipping）以及批归一化。

第 5 题（循环神经网络）

隐状态更新公式：

$$h_t = f(W_x x_t + W_h h_{t-1} + b)$$

输出计算公式：

$$y_t = g(W_y h_t)$$

其中，$f(\cdot)$ 和 $g(\cdot)$ 为非线性激活函数。

第 6 题（Adam 优化算法）

Adam 优化算法结合了以下思想：

- **动量（Momentum）**：利用历史梯度信息，引入参数更新的惯性；
- **自适应学习率**：为每个参数分别调整学习率；
- **偏差修正（Bias Correction）**：修正一阶和二阶矩在训练初期的偏置。

答出其中任意两点即可。

第 7 题（Transformer 结构）

Transformer 中一个典型的编码器模块由以下部分组成：

1. 多头自注意力机制
2. 残差连接
3. 层归一化（Layer Normalization）
4. 逐位置前馈网络
5. 残差连接
6. 层归一化（Layer Normalization）

第 8 题（Encoder–Decoder 架构）

编码器（Encoder）：

编码器接收一个长度可变的输入序列，并对其进行编码与特征提取，生成上下文表示向量，通常记为 C

解码器（Decoder）：

解码器利用上下文向量 C，逐步生成长度可变的目标输出序列。

第 9 题（BERT 与 GPT）

BERT：

- 使用掩蔽语言模型（Masked Language Model），随机掩蔽词元并利用上下文预测被掩蔽词元；
- 使用下一句预测（Next Sentence Prediction）任务，判断两个句子是否在原文中相邻。

GPT：

- 使用自回归语言模型，通过预测下一个词元进行训练。

上述答案表述不完全，具体信息请详细阅读复习资料。