Deep Learning (WJ64007N): Homework #2

姓名 Name: _____ 学号 Student ID: _____ Due: 2026-12-08 23:59

本次作业覆盖第 5–10 周课程，共 6 题，总分 100 分。请在回答中说明理由，涉及计算时应给出必要步骤。所有对数均为自然对数。

题目 1（卷积网络，15 分）

(a) 卷积层为什么采用局部连接，并在不同空间位置共享同一个滤波器的参数？使用多个滤波器有什么作用？

(b) 输入图像为 $32 \times 32 \times 3$（高 × 宽 × 通道）。考虑下列网络。卷积层均采用标准二维卷积（不分组，膨胀率为 1），卷积层和线性层均包含偏置。计算输出形状时不计批次维度。

层	卷积核/窗口	步长与每侧填充	输出通道/维数
Conv1	3×3	$s = 1,\ p = 1$	16
ReLU	—	—	不变
MaxPool	2×2	$s = 2,\ p = 0$	不变
Conv2	3×3	$s = 1,\ p = 1$	32
ReLU	—	—	不变
全局平均池化（GAP）	对每通道取空间平均	—	32
线性分类器	—	—	10

写出各层的输出形状与可训练参数量，并计算网络的总参数量。

(c) 以 (b) 中第二个 ReLU 的输出作为输入，比较以下三种分类头。三种方案均输出 10 个类别得分，所有线性层和卷积层均包含偏置。方案 A：GAP 后接线性分类器；方案 B：展平（flatten）后接线性分类器；方案 C：先用 1×1 卷积将通道数变为 10，再做 GAP。分别计算三种分类头的参数量，以及方案 B 的参数量与方案 A、方案 C 的参数量之比。结果可保留分数形式。

解答

题目 2（目标检测与语义分割，15 分）

(a) 目标检测与语义分割分别需要输出什么？当各类别的像素数量差异很大时，为什么仅用像素准确率评价分割效果可能具有误导性？

(b) 课件介绍的 FCN 语义分割网络为什么需要对下采样后的特征图进行上采样？跳跃连接（skip connection）融合浅层特征和深层特征时，分别利用了两者的哪些信息？

(c) 一个二分类语义分割模型在 100 个像素上的预测结果如下，行表示真实类别，列表示预测类别：

	预测为背景	预测为目标
真实为背景	90	5
真实为目标	3	2

计算像素准确率（PA）、平均像素准确率（MPA）、背景与目标各自的交并比（IoU），以及平均交并比（mIoU）。MPA 与 mIoU 均对背景、目标两类取平均，可保留分数形式。

解答

题目 3（自注意力、位置编码与 ViT，15 分） 设输入矩阵 X 的每一行表示一个 token，d_k 为查询和键的维度。单头自注意力定义为

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V, \qquad A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \qquad S(X) = AV,$$

其中 softmax 按行计算。以下不考虑 Dropout 和输出投影。

(a) 查询（Query）、键（Key）和值（Value）分别如何由输入得到？哪两个矩阵用于计算注意力权重？归一化后的权重用于对哪个矩阵进行加权求和？

(b) 如果自注意力只使用词的内容表示，而不引入位置信息，为什么无法充分利用词序？位置编码如何补充位置信息？请结合一个语言序列的例子，说明词序为什么重要。

(c) 将一张 $224 \times 224 \times 3$ 的图像切分为不重叠的 16×16 图像块。每个图像块展平后，通过**同一个**带偏置的线性层映射到 $D = 192$ 维，形成一个 token。不添加分类 token，不计位置编码的参数。计算 token 数、图像块嵌入层的参数量，以及单张图像对应的单头注意力分数矩阵的元素个数。若改用 8×8 图像块，且嵌入维度不变，计算新的 token 数、注意力分数矩阵的元素个数，以及后者与原方案的比值。

解答

题目 4（有监督微调与 LoRA，15 分）

(a) 比较自回归预训练和有监督微调（SFT）的训练数据与学习目标。在只对回答 token 计算损失的 SFT 中，为什么仍需把提示送入模型？如果不计算提示部分的预测损失，回答部分的损失能否通过提示的中间表示，反向传播到相关的可训练参数？为什么？

(b) 低秩适配（LoRA）将线性层的权重写为 $W = W_0 + BA$，其中 W_0 保持冻结，只训练低秩因子 A 和 B。当所选秩远小于原权重矩阵的两个维度时，这种方法为什么能减少可训练参数量及其对应的优化器状态？通常将 A 随机初始化、B 初始化为零，这样做有何作用？为什么不把两者都初始化为零？

(c) 训练序列为 $[\mathrm{BOS}, p_1, p_2, y_1, y_2]$，其中 BOS 为序列起始符，$(p_1, p_2)$ 为提示，(y_1, y_2) 为回答。使用真实前缀预测下一个 token，且只计算回答部分的损失。按预测目标 $[p_1, p_2, y_1, y_2]$ 的顺序写出损失掩码，其中 1 表示计入损失，0 表示不计入损失。已知

$$p_\theta(y_1 \mid \mathrm{BOS}, p_1, p_2) = \tfrac{1}{4}, \qquad p_\theta(y_2 \mid \mathrm{BOS}, p_1, p_2, y_1) = \tfrac{1}{2}.$$

计算给定起始符和提示时，整个回答序列的条件概率、按回答 token 数取平均的交叉熵损失，以及相应的困惑度（perplexity）。

(d) 对一个 2048×2048 权重矩阵使用 LoRA，$B \in \mathbb{R}^{2048 \times r}$、$A \in \mathbb{R}^{r \times 2048}$，忽略偏置与缩放系数。当 $r = 8$ 时，计算全量微调与 LoRA 的可训练参数量、LoRA 参数量占全量微调的比例，以及 $\Delta W = BA$ 的秩上界。若该层 LoRA 的可训练参数量不超过 65,536，求秩参数 r 可取的最大整数值。

解答

题目 5（VAE 与 GAN，20 分）

(a) VAE 如何利用编码器的输出表示给定输入时隐变量的分布？以对角高斯分布为例，说明均值和标准差分别描述什么。对同一个输入，即使编码器给出的分布参数不变，多次独立采样得到的隐变量是否一定相同？为什么？

(b) 比较 VAE 重构已有图像与进行无条件生成的流程：两种情况下，隐变量分别从哪里获得？生成新图像时，是否必须先提供一张真实图像并运行编码器？为什么训练图像重构得好，并不一定意味着从先验采样也能生成质量较高的新图像？

(c) GAN 中的生成器与判别器分别学习什么？训练初期，当判别器很容易区分真假样本时，原始极小极大目标中的生成器损失可能出现什么梯度问题？可以采用什么替代损失来缓解这一问题？

(d) 假设训练判别器时，样本来自真实数据或生成器的概率各为 1/2。如果生成分布与真实数据分布完全一致，达到最优的判别器应如何判断样本来源？为什么？反过来，如果实际训练中的判别器始终输出 1/2，能否仅凭这一现象认定两个分布已经一致？说明理由。

解答

题目 6（DDPM 与流匹配，20 分） 以下讨论无条件生成：DDPM 使用噪声预测器，流匹配采用课件中的线性路径。

(a) DDPM 的前向加噪过程与反向生成过程分别以什么为起点，目标是什么？哪个过程遵循预设的加噪规则，哪个过程依赖训练得到的神经网络？生成新样本时，是否需要知道某张训练图像及其加噪记录？为什么？

(b) 训练 DDPM 的噪声预测器时，网络的输入和预测目标分别是什么？训练所需的噪声标签如何获得，为什么不需要人工标注？为什么训练时可以随机抽取一个时间步来构造训练样本，而生成时通常需要反复调用网络？

(c) 在线性路径的条件流匹配中，独立采样噪声端点 z_0 和数据端点 z_1。如何利用这两个端点构造时刻 $t \in [0,1]$ 的中间位置和目标速度？速度预测网络的输入是这两个端点，还是时间与当前位置？如果给定时间与位置可能对应多组端点和不同的速度标签，平方误差回归的理想最优解应是什么？

(d) 在上述线性路径设定下，为什么流匹配训练时可以直接构造随机时刻的监督样本，而无需先求解完整轨迹？生成时，为什么通常需要从噪声出发，沿学到的速度场逐步更新样本？即使网络准确学到了目标速度场，训练所用的直线路径是否也能保证生成轨迹为直线？说明理由。

解答