论文领读与课堂测试

机器学习导引 • 第一章绪论（配套材料）

Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115–118.

一、读懂论文所需的基础概念（对照绪论）

概念	论文中的体现	对应绪论
监督学习	输入图像 → 输出疾病标签，属于分类任务	三大范式 • 监督学习
输入空间 X / 输出空间 Y	X= 皮肤病变图像（像素），Y= 疾病类别（离散集合）	学习问题的形式化
训练/验证/测试集	127,463 训练 + 验证 / 1,942 活检证实测试	数据集划分
标签与金标准	医生标注；测试集用活检证实作金标准	标注 / 数据质量
模型 vs 学习算法	Inception v3 = 模型；反向传播 = 学习算法	模型与学习算法
参数	CNN 数以百万计的权重	参数 θ
Softmax	CNN 输出各类别的概率分布	Softmax / Logits
迁移学习	ImageNet 预训练 + 皮肤图像微调	（绪论自然引出）
评估指标	准确率、灵敏度、特异度、ROC	模型评估
泛化	对未见活检图像仍达专家水平	泛化
数据量	比之前数据集大两个数量级	数据 = 燃料 / 经验

二、课堂测试题

（一）选择题

Q1 这篇论文的任务属于哪种学习范式？

A. 无监督学习　B. 监督学习　C. 强化学习　D. 半监督学习

Q2 论文里“输入 x”和“输出 y”分别是什么？

A. x= 疾病名，y= 图像　B. x= 皮肤病变图像，y= 疾病类别概率分布　C. x= 像素标签，y=CNN 权重

Q3 CNN 先在 ImageNet（128 万张自然图像、1000 类）预训练，再在皮肤图像上微调，这种做法的学名是？

A. 强化学习　B. 迁移学习　C. 降维　D. 聚类

Q4 为什么测试集要用“活检证实（biopsy-proven）”的图像？

A. 为了更清晰　B. 测试集标签必须是金标准，否则无法客观评价泛化能力　C. 因为数量不够

Q5 CNN 输出的是 757 个细粒度类别的概率；要得到“黑色素瘤 vs 良性痣”的二分类，作者做了什么？

A. 重新训练　B. 把子类的概率相加　C. 取最大概率那一类　D. 随机选

Q6 论文用 sensitivity（灵敏度）评估。sensitivity = 真阳性 / 阳性总数，它等价于哪个指标？

A. 准确率　B. 召回率 recall　C. 精确率 precision　D. F1

Q7 “CNN 在两类任务上达到与皮肤科医生相当的水平”，这主要体现模型的什么能力？
A. 记忆　B. 泛化　C. 过拟合　D. 降维

Q8 论文强调数据集“比之前大两个数量级”，结合绪论这说明什么？
A. 数据是机器学习的燃料，数据规模是性能的关键　B. 算法不重要　C. 标签可以没有

（二）简答题

Q9 用 Mitchell 定义的三件套，写出这篇论文的 E（经验）、T（任务）、P（性能度量）。

Q10 迁移学习为什么有效？试用“表示 / 特征”的角度解释一句。

三、论文领读讨论题

1. **三要素拆解**：用绪论的“数据、任务、模型”三要素各用一句话描述这篇论文；再指出“学习算法”在这里对应什么、“最终模型”又是什么。
2. **测试集的讲究**：论文为何特意用“活检证实”图像做测试？若作者在调参时反复看这 1,942 张图，会发生什么？（联系“测试集只用一次”“数据泄露”）
3. **从概率到决策**：CNN 输出概率分布，但临床要下“良性/恶性”判断。为什么需要一个阈值把概率变决策？阈值变化如何影响 sensitivity 与 specificity？（提示：ROC）
4. **泛化的证据**：论文凭什么说模型“达到专家水平”？与皮肤科医生的对比是在“训练过的图”还是“未见的活检图”上做的？为什么这是“泛化”而非“记忆”？
5. **局限与改进**：论文承认了哪些局限？（数据来源、多样性、仅图像等）若让你改进，你会先在“数据”还是“任务定义”上动手？
6. **迁移到表示学习**：把“预训练 + 微调”对应到绪论的“表示学习”——为什么在 ImageNet（没有皮肤癌类别）上学到的表示，能帮助识别皮肤病变？

四、参考答案与解析

选择题

Q1 B——监督学习：图像 $\rightarrow$ 疾病标签，有标注的分类任务。

Q2 B——x 是皮肤病变图像（原始像素），y 是疾病类别的概率分布。

Q3 B——迁移学习：预训练于大数据，再在新任务上微调。

Q4 B——测试集标签必须是病理金标准，否则评估不可信（对应“测试集只用一次、标签要可靠”）。

Q5 B——按疾病分类树，把细粒度子类的概率求和得到粗粒度类别的概率。

Q6 B——sensitivity = recall = 查全率。

Q7 B——泛化：在未见图像上的表现，而非死记训练样本。

Q8 A——数据是机器的“燃料/经验”，规模是性能关键（呼应绪论“数据、学习、泛化”）。

简答题

Q9 E = 127,463 张训练图像（经验）；T = 皮肤病变分类（两类关键二分类 + 九分类）；P = 灵敏度/特异度（或准确率）。

Q10 底层的卷积层学到的边缘、纹理等通用视觉特征可迁移复用，只需微调顶层来适配新任务，从而在小数据集上也能学得好。

讨论题要点

- **1 三要素**：Data = 129,450 张皮肤图像；Task = 良/恶性分类；Model = 训练好的 Inception v3。学习算法 = 反向传播/梯度下降；模型 = 训练后权重确定的 CNN。
- **2 测试集**：活检是病理金标准，保证标签可靠；反复偷看测试集会数据泄露、评估失真。
- **3 阈值**：设阈值 τ，P(恶性) $> \tau$ 判恶性；τ 降低则 sensitivity 升、specificity 降，ROC 刻画这一权衡。
- **4 泛化**：对比在未见活检图像上进行，而非训练图，证明学到的是规律而非死记。
- **5 局限**：数据来源单一、缺乏多样性、仅图像不结合病史等；可先扩充数据多样性或重新定义任务。
- **6 迁移**：ImageNet 学到的底层特征（边缘、纹理）是通用视觉表示，可迁移到皮肤病变，顶层再适配具体任务。