

1 问题要求

结合生活中案例，利用物体检测类神经网络（如 YOLO 系列），实现物体的检测和识别。

- 要求 1：类别数量不得低于 4 个；
- 要求 2：每个类别的样本图片数量不得多于 20 个。

形成一个专题报告，详细描述选择该网络的理由、数据来源、训练环境、训练过程、验证和测试情况。

2 实现和说明

2.1 技术路线

2.1.1 网络构架

针对物体检测，我们备选的构架是 CNN 和 Transformer。两种构架的备选模型分别是 YOLO^[1] 系列和 Paligemma^[2]。针对问题和网络特性，最终选择 YOLO，理由如下：

- **数据量极小**：根据要求，每个类别仅 20 最多张图片。ViT 论文^[3] 指出，由于 CNN 带有归纳偏置，相比起 Transformers，小数据集下 CNN 很有可能表现更好。此外，对于每类仅 20 张图片的训练集大小，无法有效微调 3B 参数的 Paligemma 模型；相比之下，训练一个数据量 10M 左右的 CNN 更为可行。
- **资源受限**：我们使用 RTX 4060 8GB RAM 显卡，而 Paligemma 微调需要 12GB 显存才可能微调全部 Attnetion Layers。尽管通过 LoRA 或者 QLoRA，可能通过较为有限的资源对 Transformer 模型进行微调，但是能够对模型产生的影响也相对应地更小。相比之下，YOLOv8 更加轻量适配，完全可以对全部权重进行快速有效的微调。

2.1.2 YOLO 具体路线

具体方案选择：

- **YOLOv8s (small)**：11.2M 参数，适合细粒度分类且不易过拟合，适配已有资源的 8GB 显存，训练快，对于较少种类的进行视觉检测已经足够。
- **迁移学习**：在进行模型 fine-tuning 和从头训练两个路线之间，我们选择使用前者，使用基于 COCO 数据集预训练的模型进行权重微调。这是因为预训练的模型的较前的神经网络层数已经学习到了一些有效的 patterns，而使用 80 张图片从头训练一个较大的模型可能导致效果不佳。
- **强数据增强**：实验允许的训练数据极小，为了进行有效的学习，大量使用复制粘贴、旋转缩放、颜色变换等方法进行数据增强，以补偿数据不足。

2.2 任务和数据

任务设定：为了充分验证 YOLO 在少样本学习 (few-shot learning) 场景下的能力，我们选择了比较具有挑战性的任务，对猫的品种进行目标识别和检测。与猫狗分类等粗粒度任务不同，不同猫品种之间的视觉差异相对更小，需要模型学习更加精细和有效的特征表示以进行分类。

数据来源：使用 Cat-Breeds-Detection-Dataset¹，该数据集已按 YOLO 格式组织，包含标注完整的边界框信息，边界框标注了猫的头部。原始数据集涵盖 5 个猫品种，每个品种包含约 300 张图片。

数据集构建：为满足实验要求（每类不多于 20 张训练图片），我们挑选了 4 个较有视觉区分度的品种：

- 波斯猫 (Persian): 长毛，扁脸特征明显
- 斯芬克斯猫/加拿大无毛猫 (Sphynx): 无毛品种，皮肤褶皱特殊
- 布偶猫 (Ragdoll): 长毛，面部色块分布较为明显
- 苏格兰折耳猫 (Scottish Fold): 深色较多，有独特的折耳特征

其中，波斯猫和布偶猫由于都具有长毛特征，较容易混淆；无毛猫特征明显，容易分辨。在构建数据集时，每个品种选取 20 张训练图片和 5 张验证图片。样本展示如图1。

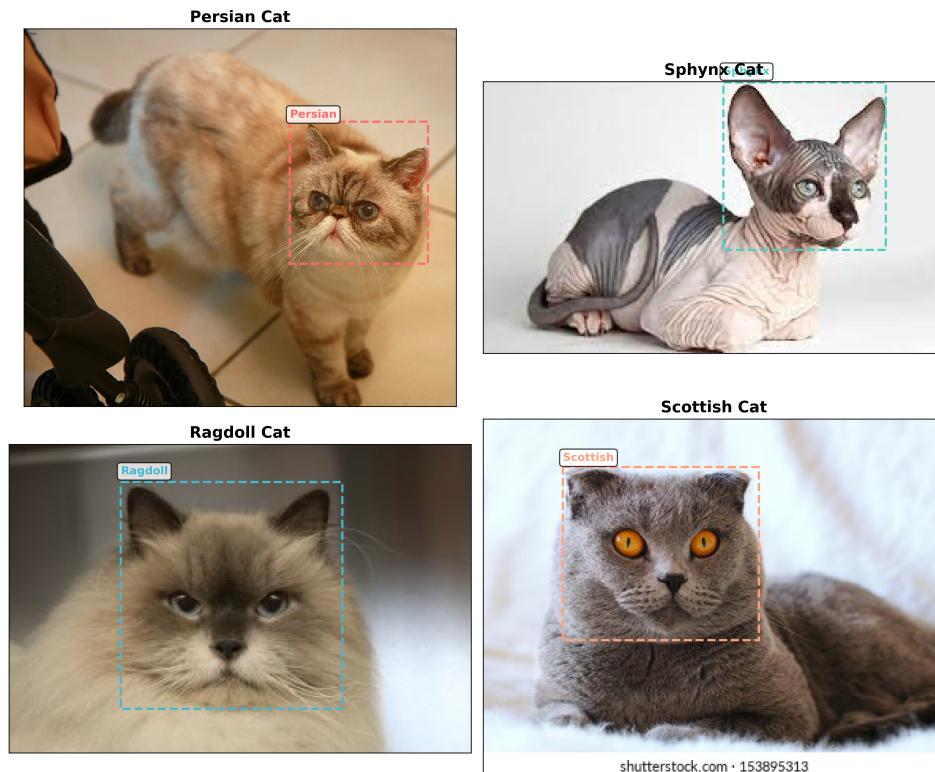


图 1：数据集样本展示

¹<https://github.com/AWREDD/Cat-Breeds-Detection-Dataset>

2.3 训练环境

实验在以下硬件和软件环境中进行：

- GPU: NVIDIA RTX 4060 (8GB 显存, 约 7GB 可用)
- CPU: Intel Core i7-14650HX
- 内存: 24GB
- 操作系统: Ubuntu 20.04 LTS
- Python 版本: 3.8

关键 Python 包详见附录表1。训练超参数配置详见附录表2，其中包含数据增强强度、优化器设置等信息。

2.4 训练过程

迁移学习: 训练采用在 COCO 数据集上预训练的 YOLOv8s 模型作为起点。由于 COCO 数据集中不包含猫的具体品种标注，预训练模型仅能识别“猫”这一通用类别（见图2）。通过微调，模型可以学习到区分不同品种的细节特征，实现特殊领域的物体识别。

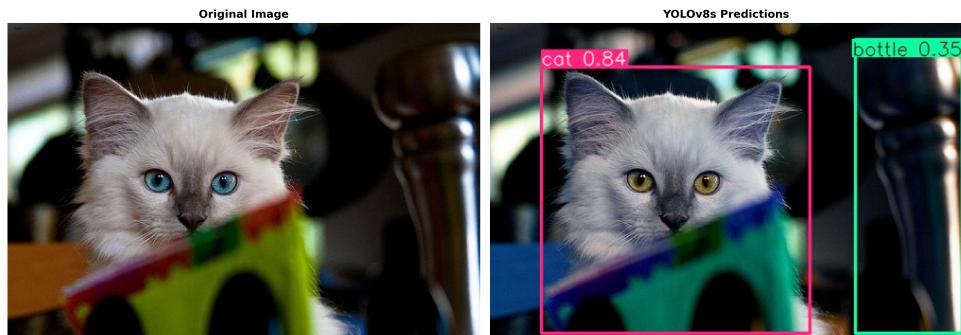


图 2: 预训练模型检测效果

数据增强: 针对仅 80 张训练图片的数据稀缺问题，我们使用了大量的数据增强。图3展示了经马赛克、颜色抖动、旋转缩放等增强后的效果。每张图片在每个 epoch 都经历不同的随机变换，有效扩充数据多样性，提高了模型的泛化能力，避免了模型过拟合。

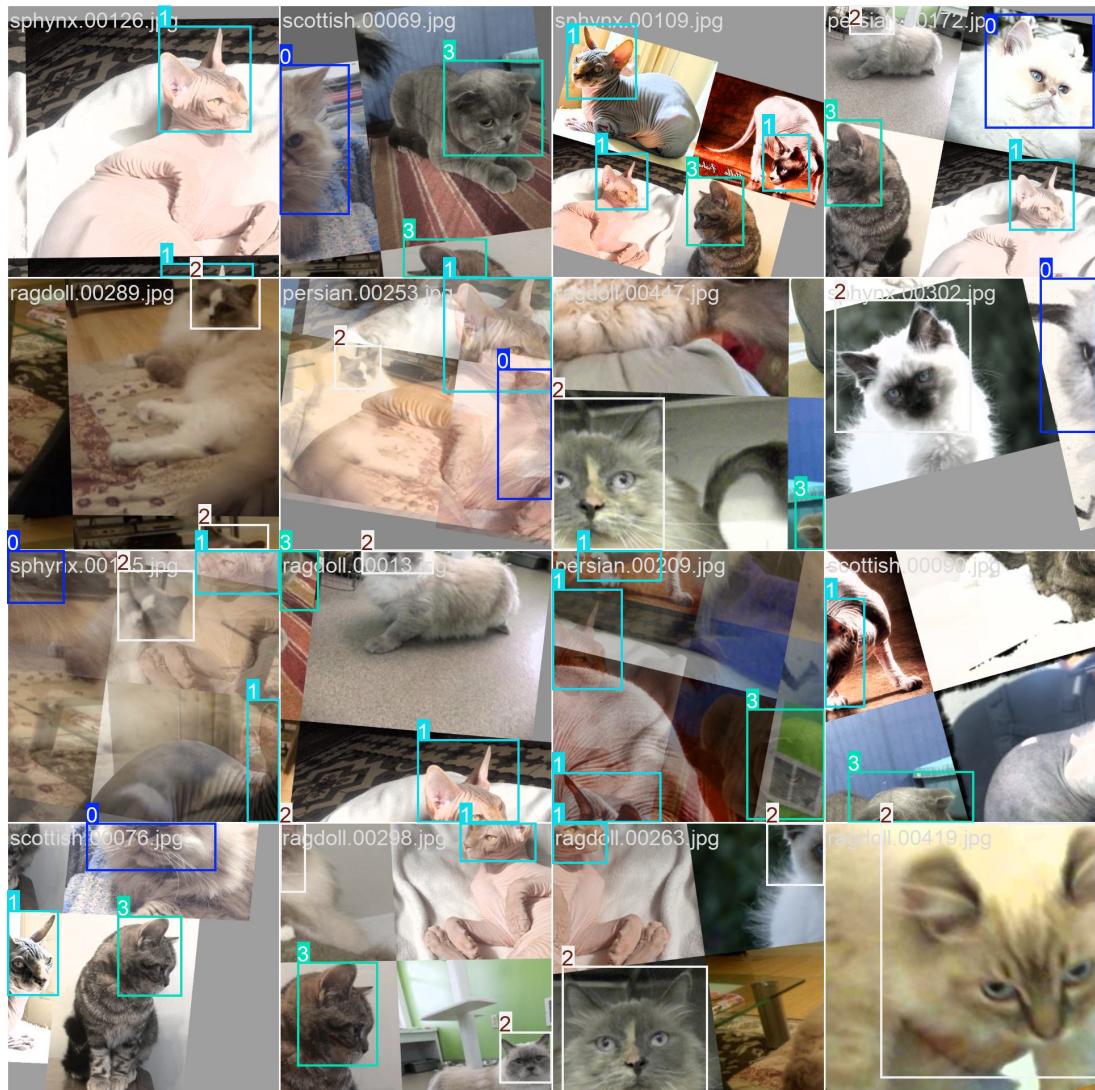


图 3: 训练数据增强效果

训练过程和指标: 训练中, loss 呈较为稳定下降的趋势 (见图4)。分类 loss 和标注框 loss 同时下降, 说明模型在品种分类和定位能力上都有提升。尽管设置 150 轮训练, Early Stopping 机制在 43 轮时触发, 表明模型已充分收敛。最终验证集准确率和召回率均达 60% 以上, mAP50 和 mAP50-95 指标也有所上升, 表明了训练的有效性。

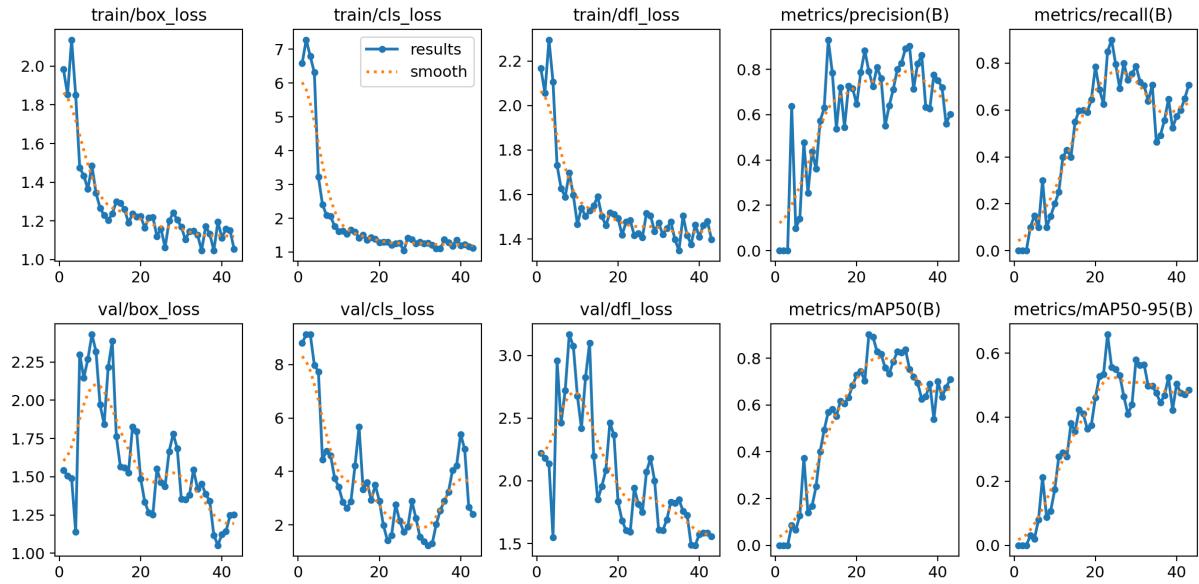


图 4: 训练和表现指标变化曲线

2.5 验证和测试

除去使用图4中的指标进行训练结果的验证，我们还对训练结果进行了可视化。图5, 6展示了预训练模型和微调后模型在同一张验证图片上的检测效果对比。可以明显看出，微调后的模型能够更有效地识别出具体的猫品种，而预训练模型只能标注通用的“cat”类别。在图5中，微调后的模型正确辨认出了所有猫的品种，而微调前的模型甚至将无毛猫辨认为“dog”。然而，在小样本上学习的模型仍然可能出现错误。在图6中，微调后的模型错误地将波斯猫辨认为布偶猫。

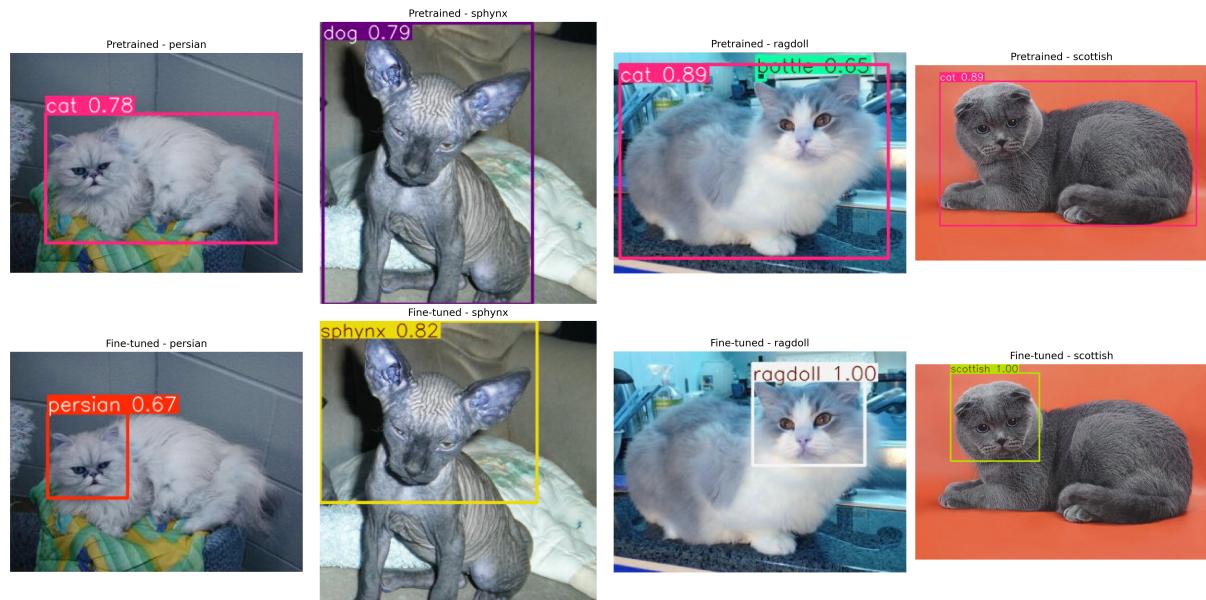


图 5: 模型对比结果 1

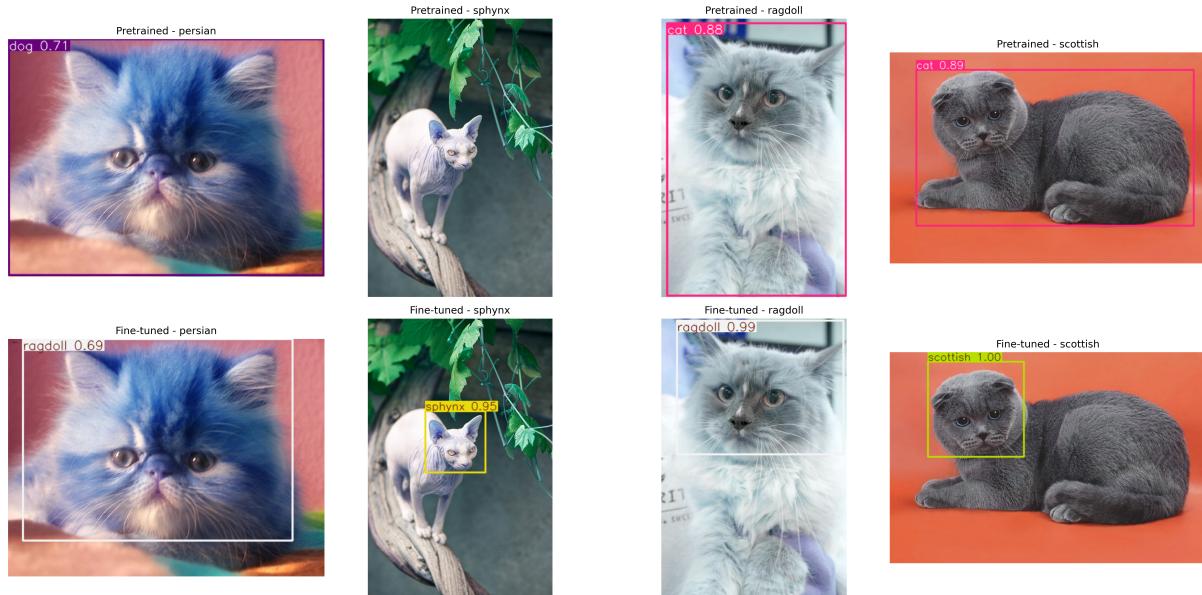


图 6: 模型对比结果 2

总体而言，针对猫的品种辨认，微调后的模型有效地学习到了猫的品种特征，也习得了对应品种猫面部位置的知识。如果训练样本量增加，微调结果将更加有效。

附录

表 1: 关键 Python 包版本信息

包名称	版本
torch	2.4.1+cu124
torchvision	0.19.1+cu124
ultralytics	8.3.43
opencv-python	4.10.0
numpy	1.24.1
matplotlib	3.7.2
pandas	2.0.3
PyYAML	6.0.2
tqdm	4.66.5

表 2: 训练超参数配置

参数	数值
epochs	150
batch size	32
image size	640
optimizer	AdamW
learning rate	0.001
weight decay	0.0005
saturation augmentation	0.7
value augmentation	0.4
rotation range	15.0°
scale range	0.5
mosaic probability	1.0
mixup probability	0.2
copy-paste probability	0.3

参考文献

- [1] HUSSAIN M. When, where, and which?: Navigating the intersection of computer vision and generative ai for strategic business integration[J/OL]. IEEE Access, 2023, 11:127202-127215. DOI: 10.1109/ACCESS.2023.3332468.
- [2] BEYER L, STEINER A, PINTO A S, et al. Paligemma: A versatile 3b vlm for transfer[EB/OL]. 2024. <https://arxiv.org/abs/2407.07726>.
- [3] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. 2021. <https://arxiv.org/abs/2010.11929>.