数据增广与迁移学习

机器学习 · 深度学习基础

学习目标

目标标注数据有限时，本讲依次回答：怎样利用原图的合理变化、怎样复用已有网络的表示，以及怎样检验这些做法是否有效。

- 解释增广所表达的任务条件，判断给定变换能否保留类别
- 区分预训练表示与任务输出，说明怎样把已有网络用于新类别
- 跟踪冻结和微调中的计算与更新，分清参数设置与训练 / 评价模式
- 沿一次完整示范连接数据划分、训练、验证选择与独立评价

1 数据增广与任务条件

先判断标签保持条件，再把允许变化写入训练。

有限照片难以覆盖同一类别的变化

设目标是把照片分为“猫”或“狗”。有限照片难以覆盖位置、尺度和光照变化；哪些变化后仍有充分理由沿用“猫”标签?

原图



标签：猫

水平翻转



标签：猫

局部裁剪



标签：猫

亮度×0.65



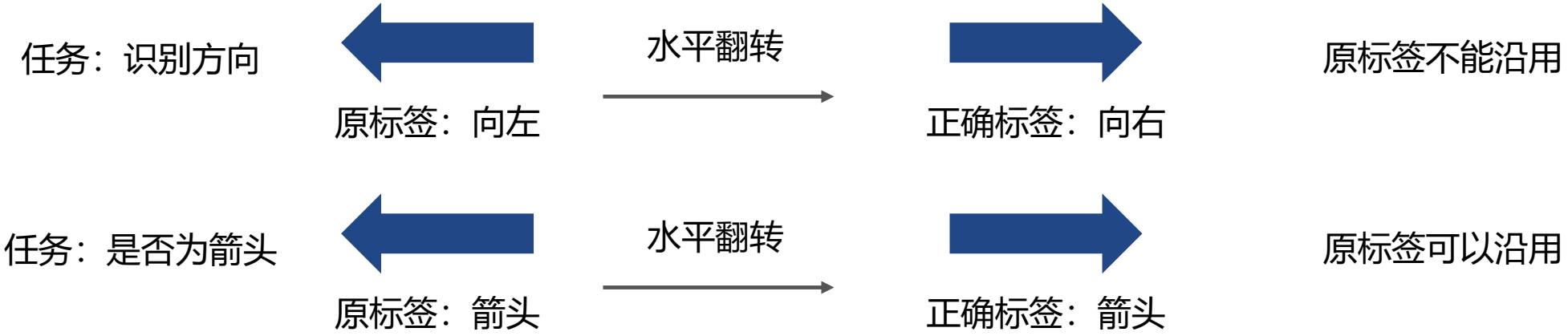
标签：猫

数据增广在保留当前任务标注含义等条件下，对训练输入施加选定变换，形成训练变体。

图中位置、尺度或亮度改变，支持“猫”判断的主要信息仍在。用同一标签训练这些变体，可以减少模型对偶然拍摄条件的依赖；效果仍需验证。

标签是否保持取决于任务

猫图水平翻转后仍可沿用“猫”标签，但水平翻转是否适用于所有分类任务？下面用同一个箭头、两种识别任务来比较。

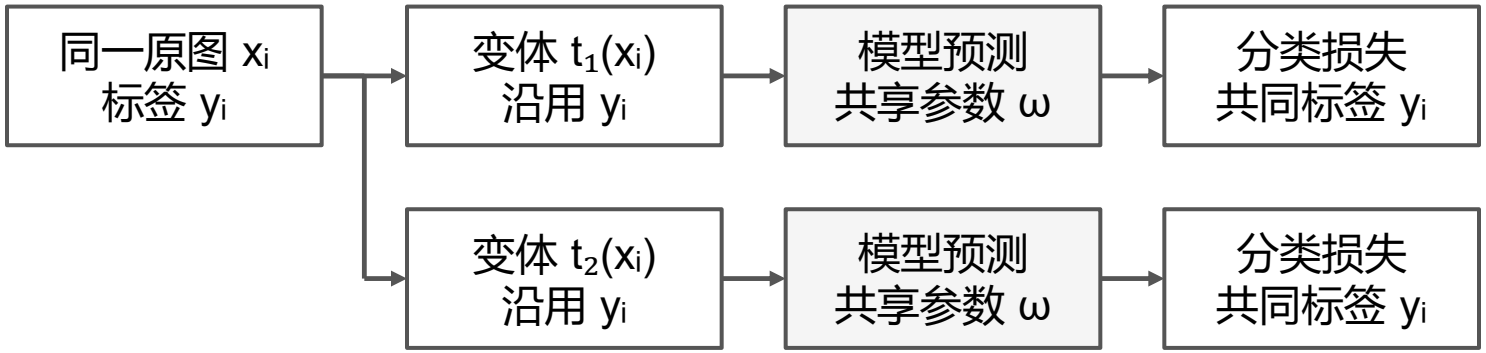


选择增广先说明**任务依赖的类别信息**及允许变化范围。同一翻转对方向识别会改变标签，对箭头识别则可以保持标签。

即使任务允许某种变换，幅度也需检查：裁剪后只剩背景，或破坏了分类所需的关键颜色，原类别也可能失去依据。

增广把允许的变化写入训练损失

确定变换能保留当前任务的标签后，就可以把变体送入训练。对一个带标签样本$(\boldsymbol{x}_i, y_i)$，从约定变换集合中随机选取t，把$t(\boldsymbol{x}_i)$送入模型，仍以y_i监督。



图中两分支是不同抽样的示例；下式每个样本只使用一次变换。

$$L_{\text{aug}} = \frac{1}{B}\sum_{i=1}^{B} \ell\left(q_{\boldsymbol{\omega}}(t_i(\boldsymbol{x}_i)), y_i\right)$$

这里B为批量大小，$q_{\boldsymbol{\omega}}$为参数为$\boldsymbol{\omega}$的分类模型，ℓ为分类损失。同源变体沿用共同标签，鼓励模型减少对这些变化的敏感性。

数据增广通过标签保持的训练变体，将任务允许的输入变化转化为训练约束。变体并非独立观测；不变性及泛化改善均不获保证。

训练增广与评价预处理承担不同职责

- **先保留独立评价条件**
  - 先按原始样本划分训练、验证与测试数据；同一对象的近重复照片应按任务需要分组划分
  - 只从训练部分生成训练变体，避免同一原图的变体跨入评价集
- **再分别处理输入**
  - 训练时：在允许范围内随机变换，再按当前模型的输入约定处理尺寸和像素数值
  - 验证 / 测试时：使用预先约定、可重复的处理，以便比较方案

划分先保留独立评价条件，输入处理再分别服务训练与比较；不能把评价集上的表现用于参数训练。

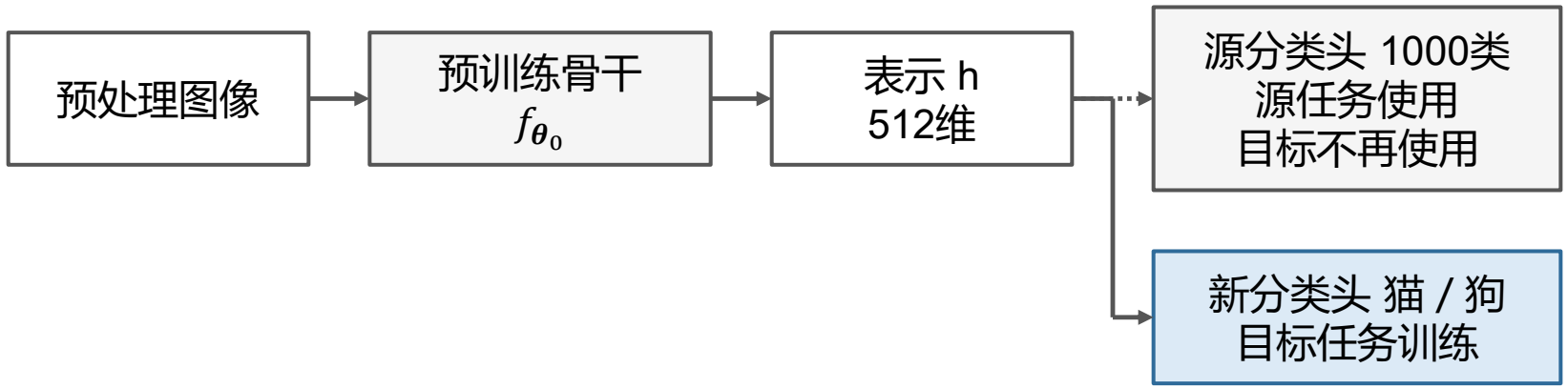
2 预训练表示与适应

增广利用原图变体，迁移复用已有表示；二者可共同用于目标训练。

目标类别改变，不必重学整个网络

目标只有少量猫 / 狗照片。从头训练需同时学习表示和判别规则；预训练骨干可能已保留目标所需的纹理、形状等信息。

预训练先在源数据与学习目标上训练模型；本例用于新的猫 / 狗任务和标签体系，称为**下游任务**。ImageNet-1K源分类已含猫犬相关类别。



先尝试保留表示、替换分类头。若已有表示不能有效区分猫与狗，仅更换线性分类头可能不够，需要进一步检验是否调整骨干。

分类头将表示映为类别得分

骨干网络$f_{\boldsymbol{\theta}}$把图像映为d维表示$\boldsymbol{h}$。**分类头**位于骨干之后，利用表示计算目标类别得分；本讲采用一个全连接线性层：

$$\boldsymbol{h} = f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{z} = \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}$$

- $\boldsymbol{h} \in \mathbb{R}^d$；$\boldsymbol{W} \in \mathbb{R}^{K\times d}$，$\boldsymbol{b} \in \mathbb{R}^K$，$K$为目标类别数
- $\boldsymbol{\theta}$为骨干参数，预训练值记$\boldsymbol{\theta}_0$；头参数合记$\boldsymbol{\phi} = (\boldsymbol{W}, \boldsymbol{b})$
- $\boldsymbol{z}$经softmax得到类别概率，再与目标标签计算交叉熵

$\boldsymbol{W}$的每一行对各特征加权求和，$\boldsymbol{b}$提供各类偏置。交叉熵的梯度更新新头参数，改变类别得分及判别边界，学会按目标标签组合已有特征。

把1000类输出改为猫 / 狗两类

目标分类只需猫 / 狗两类输出，因此保留512维表示，换用512→2的新线性头：

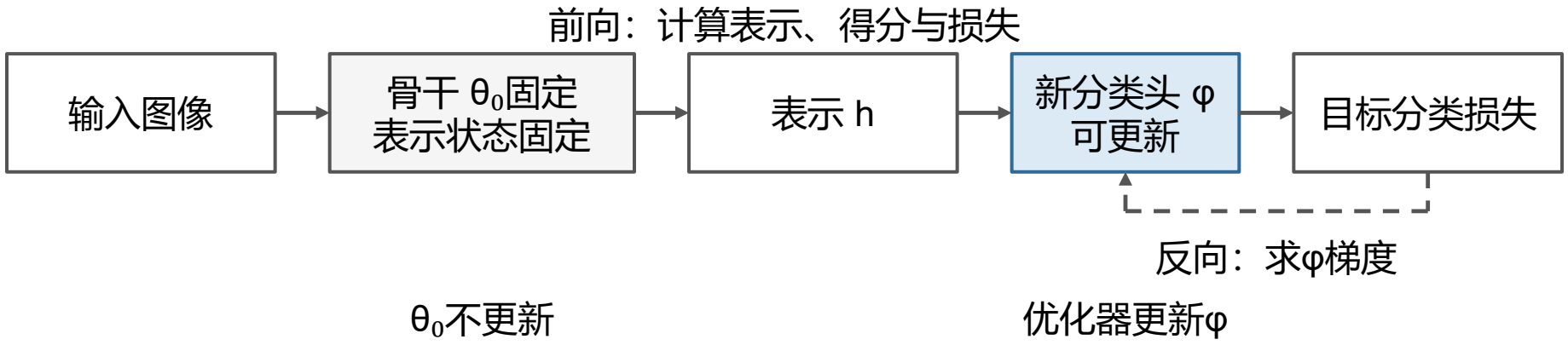
部分	输入 / 输出	参数初始化与使用
预训练骨干	图像→512维表示	读取源任务权重
原分类头	512→1000类得分	不用于目标两类输出
新分类头	512→2类得分	初始化后，用目标标签训练

新线性头的权重为2×512，偏置为2维，共有$2\times512+2=1026$个参数。

训练时用固定的猫 / 狗类别索引解释两维输出。512维是表示的维度，不是类别数；其分量也无需各自对应可命名属性。

冻结骨干时，完整前向仍然经过骨干

冻结特征提取保持骨干参数不变，利用其输出训练新的分类头。本讲也固定影响表示计算的状态，使相同输入对应固定表示。



$min_{\boldsymbol{\phi}} L(\boldsymbol{\theta}_0, \boldsymbol{\phi})$

新头学习“已有表示怎样对应新标签”；表示缺失的类别信息，单靠新头也可能难以恢复。参数与状态怎样分别固定，将在完整训练流程中说明。

固定表示与表示适应

同一组猫 / 狗样本，分别从相同的预训练表示与新头初值出发；比较训练时允许改变什么。

冻结骨干：点位固定

微调骨干：点位可改变

猫 ● 狗 ■；虚线：原边界，实线：更新后；右图空心→实心：同一样本

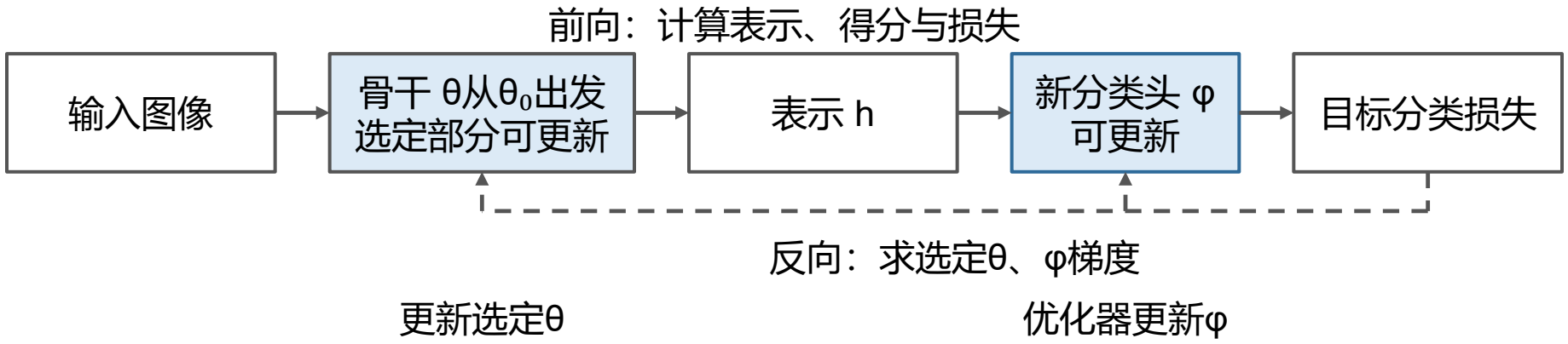


二维概念示意，非真实512维特征投影；变化方向不代表改善

冻结利用固定表示训练新头，已有表示可分时可能足够。**微调**允许表示与判别规则共同适应，固定表示不足时可尝试；收益仍需验证。

微调让目标损失继续改变表示

微调从预训练参数出发，用目标任务损失继续更新部分或全部骨干参数，并训练目标分类头。



$min_{\boldsymbol{\theta}_{\mathrm{sel}},\boldsymbol{\phi}} L(\boldsymbol{\theta},\boldsymbol{\phi})$

$\boldsymbol{\theta}_{\mathrm{sel}}$为选定骨干参数，其余固定。目标损失的梯度经新头传回骨干，再由链式法则求出选定参数的更新方向。

骨干中的运算支持梯度反向传播，可从预训练参数继续优化；更新参数会改变同一输入的表示，使特征随目标损失调整。是否改善仍由验证决定。

更新范围需要结合数据条件验证

方案	可更新对象	主要权衡
固定特征提取	新分类头	成本较低；受已有表示限制
部分微调	新头与选定骨干层	允许部分表示适应，需管理冻结范围
全部微调	新头与整个骨干	适应范围大，训练与过拟合风险需检查
从头训练	随机初始化的全网	作为参照时明确数据与训练预算

目标数据少、已有表示较相关时，可先建立冻结基线。需要适应时，再用验证集比较微调范围；预训练层可采用比新头更小的学习率，以控制其每步变化。

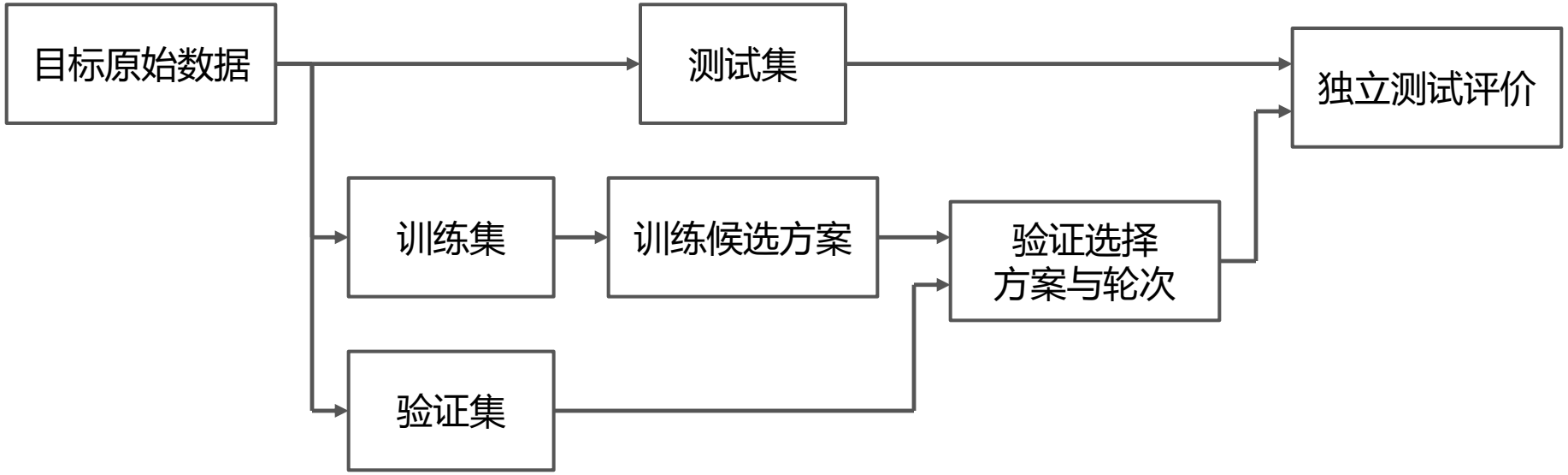
这些是试验起点。数据量、任务差异与验证结果共同决定选择，微调和预训练都不保证优于其他方案。

3 迁移学习的完整使用

由更新范围进入共同的训练、验证选择与独立评价流程。

迁移流程先固定数据与比较条件

前面已经明确两种更新范围，现在把它们放入同一使用流程。本例比较猫 / 狗分类的冻结与微调方案。沿用前面的独立划分规则，在相同划分、评价处理与指标下训练候选方案，再用验证集选择。



比较冻结与微调时，保持目标任务、划分、评价处理和指标一致，并记录更新范围、训练预算及预训练来源。若源数据可能包含目标评价样本，还需核对这种重合。

预训练配置包含权重、输入和类别接口

本例选用ResNet-18的IMAGENET1K_V1权重；实际使用时将以下配置一起记录：

配置对象	本例约定
源权重	ImageNet-1K预训练参数，记录确切权重版本
训练输入	合理随机裁剪 / 翻转，保留主体；输出224×224 RGB
评价输入	短边缩放至256，再中心裁剪224×224
像素处理	缩放到[0,1]，按权重约定逐通道标准化
目标输出	新建512→2线性头，固定猫 / 狗索引

该权重按RGB顺序使用均值(0.485, 0.456, 0.406)、标准差(0.229, 0.224, 0.225)。实际输入处理应以所选权重配置为准。

输入逐通道标准化处理像素，网络内的**BatchNorm**处理特征统计量；两者职责不同，不能互相替代。

冻结参数与评价模式控制不同对象

固定参数还不够：表示计算还可能受模块状态影响。默认跟踪运行统计的BN即使参数已冻结，在训练模式下仍可能更新运行均值与方差。例如运行均值初始为0，当前批次均值为4，momentum为0.1，更新后为0.4。

设置	直接作用	仍需单独处理
requires_grad=False	该参数不累计本轮自动微分梯度	已有梯度、优化器范围及模块状态
eval()	切换BN、Dropout等模块的行为	不会自动关闭参数梯度
no_grad()	区域内不记录反向计算图	不会自动切换BN / Dropout模式

本例采用固定特征训练：**将骨干设为评价模式、冻结骨干参数，优化器只包含新分类头参数**；分类头保持训练模式。每轮设置模式时均保持这些设置。

固定特征训练仍包含损失、反向与更新

```
backbone.eval()
head.train()
optimizer.zero_grad(set_to_none=True)
with torch.no_grad():
    features = backbone(images)
logits = head(features)
loss = criterion(logits, labels)
loss.backward()
optimizer.step()
```

计算与更新

输入经增广进入固定骨干；新头将表示映为两类得分，与类别索引计算交叉熵。反向后，优化器只更新头参数。

设置与评价

骨干参数已冻结，输出为特征；骨干设为评价模式，分类头设为训练模式。criterion为交叉熵，optimizer只含分类头参数。

这段代码展示一个训练批次。逐批重复完成一轮，再以评价模式和无梯度计算验证损失，保存该轮参数；下一页把同一流程改为微调。

建立微调候选，需要改变哪些设置

与前页的冻结候选使用相同的预训练权重和新头初值，另建允许选定骨干参数更新的微调候选。

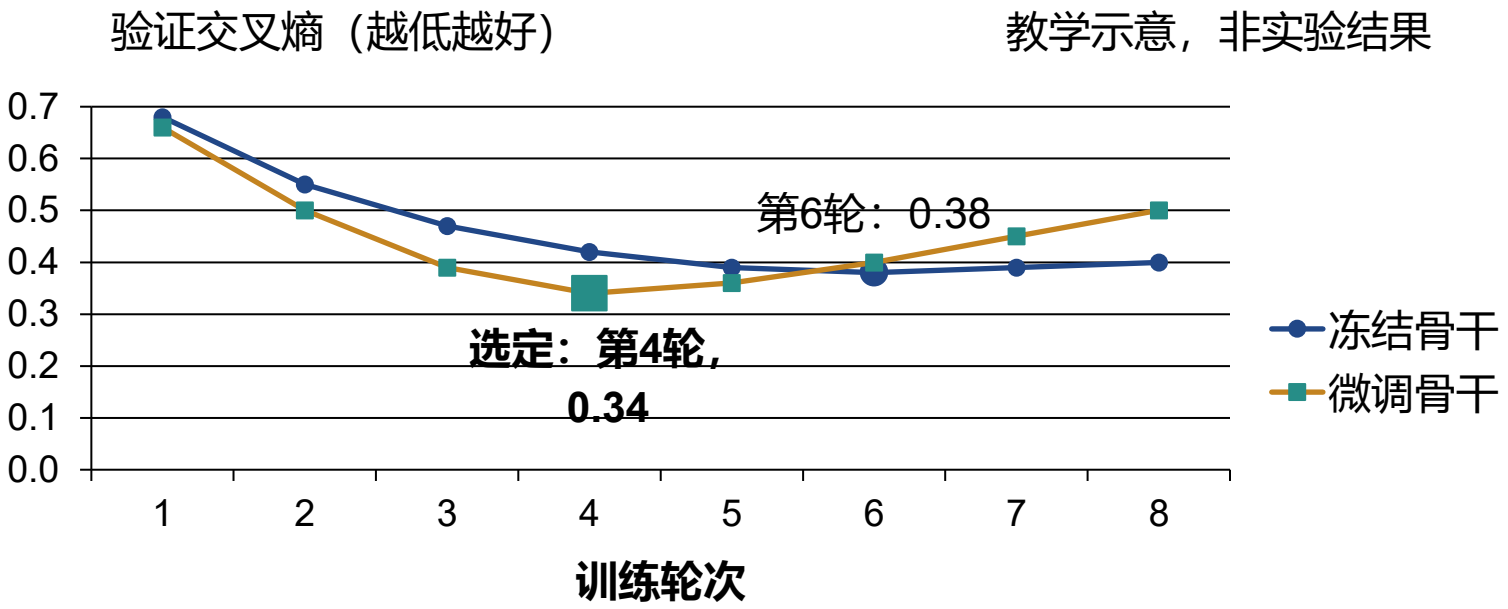
设置	冻结候选	微调候选
骨干参数	全部requires_grad=False	选定参数改为True，其余仍为False
优化器参数	仅新分类头	新头 + 选定骨干参数；清除旧梯度后重建
骨干前向	在no_grad()内计算	移出no_grad()，让梯度传回可训练部分
本例模块模式	骨干eval()，新头train()	保持同样模式，固定骨干BN运行统计

本例微调时保留骨干eval()，仍可更新选定参数：评价模式控制模块行为，不关闭梯度。BN状态策略是另一个决定；本例将其固定，以突出参数更新范围的差异。

两个候选分别训练，逐轮在同一验证集上记录损失并保存参数；下一页据此同时选择方案和轮次。

验证集选择方案与训练轮次

比较冻结与微调时，固定划分、评价处理及指标。以下构造数据用于说明方案与训练轮次的选择。



在这组示意数据中，冻结的最低值为第6轮0.38，微调为第4轮0.34；因此选择微调第4轮，恢复该轮保存的参数，而不是直接取训练结束时的参数。

选定后才在未参与训练与选择的测试集上评价，报告数据来源、覆盖和训练预算。真实方案的优劣需由实验检验；仅凭验证损失曲线上升不能确定原因。

预训练信号与下游使用方式分别选择

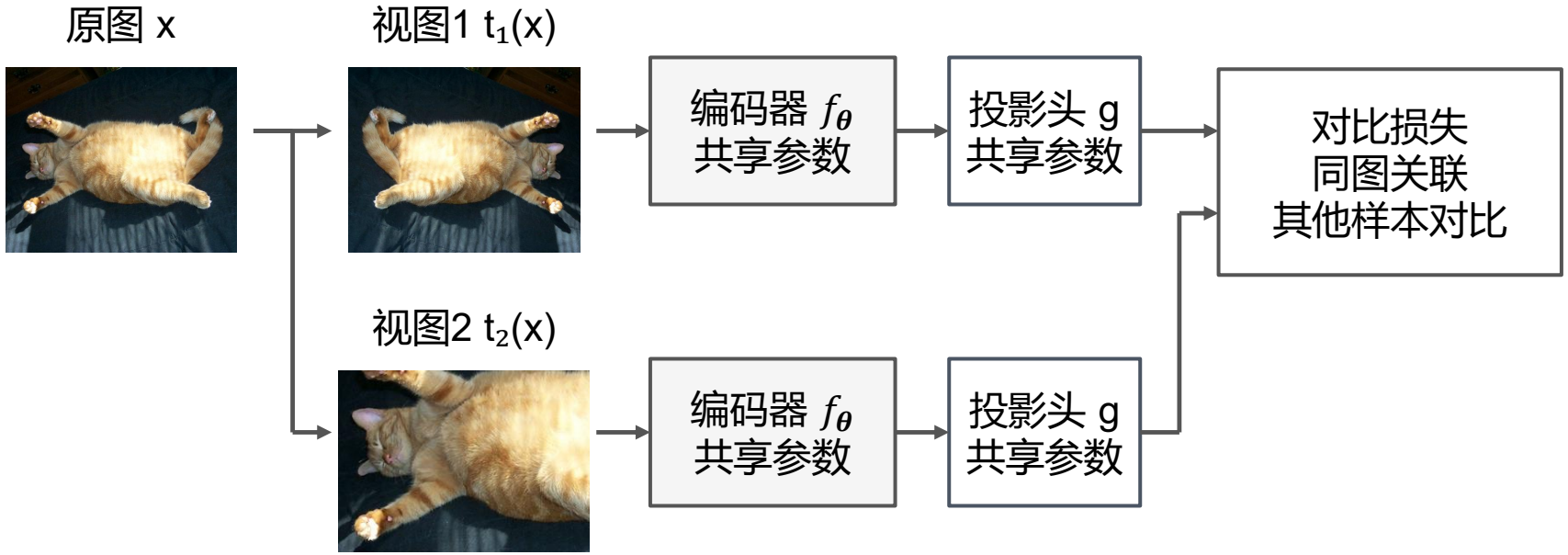
前面的完整流程采用ImageNet类别标签学得的预训练权重。现在回看预训练阶段：已有表示也可以从自监督信号学得；下游仍可选择冻结或微调。

- **预训练的学习信号**
  - 监督或自监督决定预训练所用信号。它们可能形成不同表示，是否适合目标任务需要检验。
- **下游任务的使用方式**
  - 固定特征提取：保留表示计算，训练新的任务输出
  - 微调：在已有参数基础上，用目标任务损失继续调整部分或全部表示

预训练信号回答“表示从哪里学来”，冻结或微调回答“目标阶段怎样使用”。下面用同图双视图说明一种自监督信号。

自监督从同图双视图构造学习关系

自监督学习利用数据本身可确定的关系构造训练目标。SimCLR根据增广操作记录确定同图视图，能自动形成“找出同图另一视图”的任务，无需人工类别标签。



这里的**编码器**承担前文骨干的表示提取职责，**投影头**把表示映入对比空间。损失要求同图投影相对批内其他样本更相似；梯度经投影头更新共享编码器，鼓励保留样本差异、减少对所选增广变化的依赖。

同图来源关系提供训练信号；增广须保留有用内容，表示是否适合下游仍需验证。用于猫 / 狗分类时，保留编码器、去掉投影头，以类别标签训练新头，再接入冻结或微调。

本讲总结

- 数据增广用共同标签监督任务允许的输入变体，将标签保持的先验转为训练约束；变体仍共享原图信息。
- 迁移学习复用预训练骨干及其表示，新分类头根据目标标签学习各特征怎样组合成类别得分。
- 冻结利用固定表示训练新头，微调还将目标损失的梯度经新头传回骨干，更新选定参数并改变表示；固定表示需兼顾参数与状态。
- 增广和迁移的收益都需检验。应在一致条件下用验证集选择方案与轮次，再由独立测试集评价。