视觉Transformer（ViT）

机器学习 · 深度学习基础

学习目标

- 跟踪图像分块、共享投影与序列形状
- 解释位置表示和分类词元的作用
- 跟踪Pre-LN编码器到类别得分与监督学习
- 从空间联系和计算规模比较CNN与ViT
- 结合预训练、输入规格与微调条件理解使用

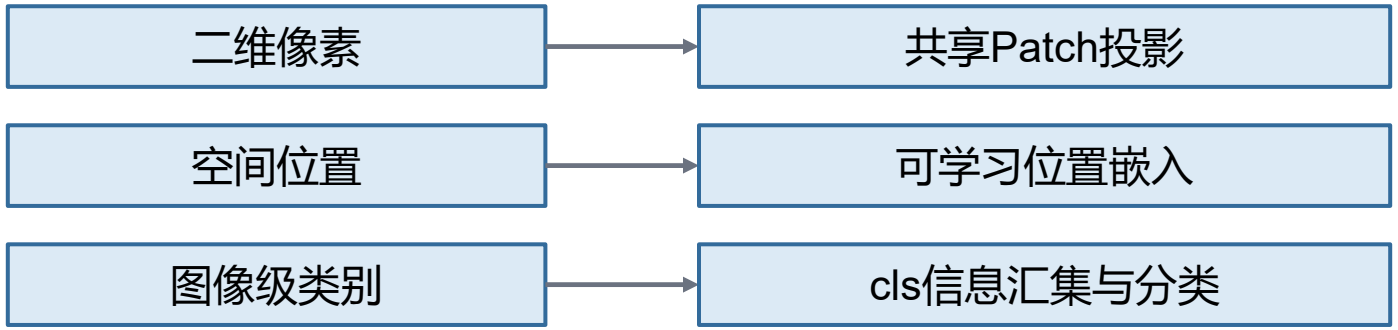
1 图像成为序列

把局部像素与位置组合为编码器输入。

视觉Transformer的建模问题

CNN已经能处理图像，Transformer已经能处理向量序列。能否让Transformer直接完成图像分类?

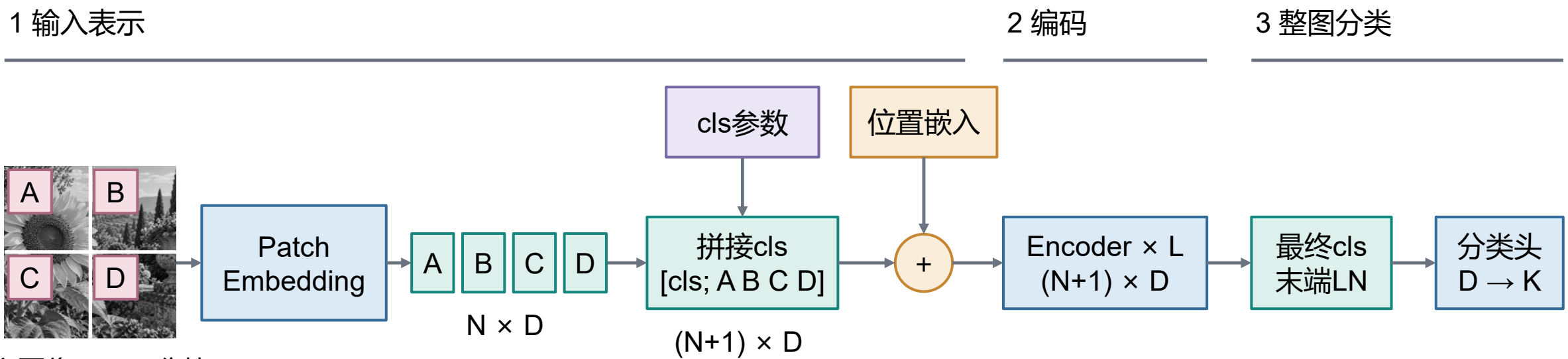
图像是二维空间信号。需要解决三个接口问题：局部像素怎样变成向量、区域位置怎样进入计算、多个区域怎样形成一个图像级输出。



ViT把图像转成带位置的词元序列，在Transformer编码器上学习视觉表示。分块建立输入接口，自注意力学习内容相关的区域联系，分类词元提供整图读出。

ViT图像分类架构

完整模型分为输入表示、Transformer编码、整图分类三个阶段。图中A—D在二维网格与序列中保持相同身份。



真实图像：2×2分块

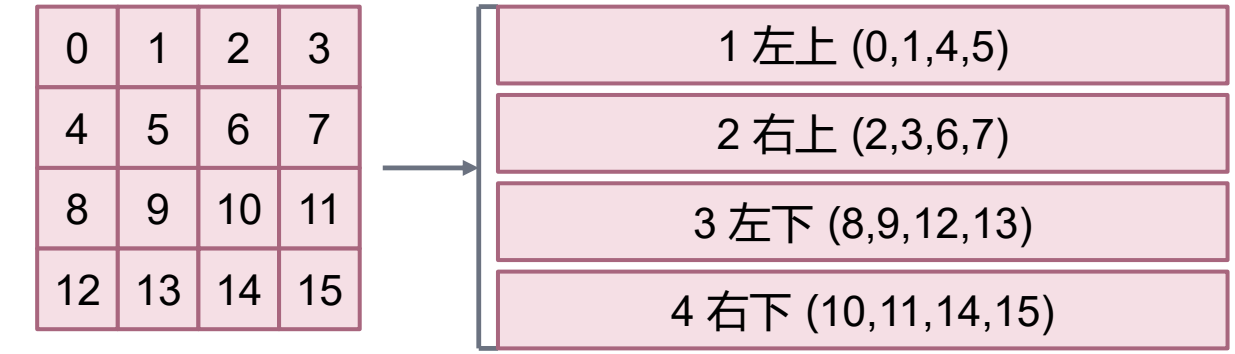
N个图像块共享投影为D维词元；先沿序列轴拼接cls，再与位置矩阵逐元素相加，得到$(N+1)\times D$输入。经过L个编码块，最终cls经LN和分类头输出K个得分。

接下来P6—9只展开输入表示；编码与整图读出在第二部分展开。

规则网格将图像划成固定大小的块

图像块（patch）是按固定网格划出的局部像素区域。它不必对应一个物体或部件。

给定4×4单通道图像，按2×2不重叠分块。块网格先行后列，每块内逐行读取像素：



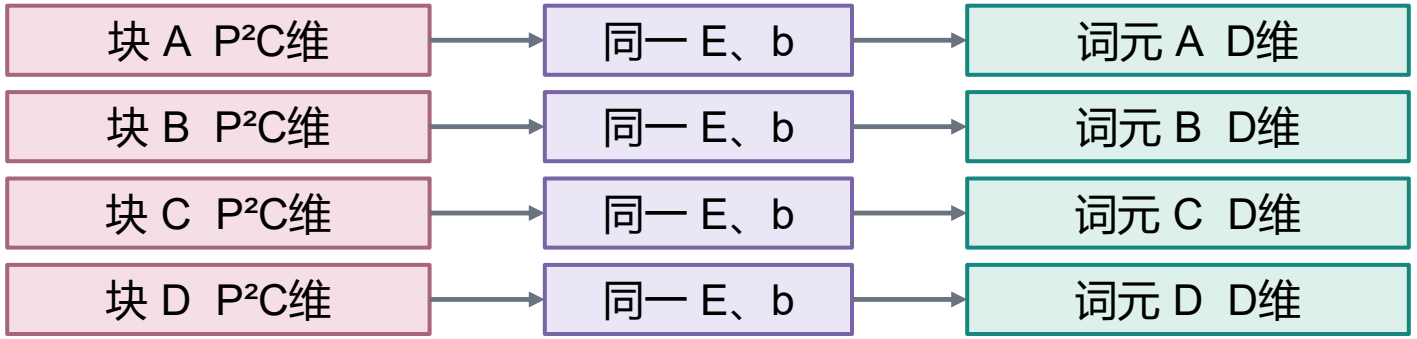
展平改变数值的排列形状，不改变这些像素的数值。下一步才用可学习参数变换每个块。

共享线性投影把每个块映成同宽表示

图像高H、宽W、通道C，块边长P；H、W均可整除P。块数$N = (H/P)(W/P)$，每块展平宽度P^2C。

$$\boldsymbol{T} = \boldsymbol{X}_{\text{patch}}\boldsymbol{E} + \boldsymbol{b}$$

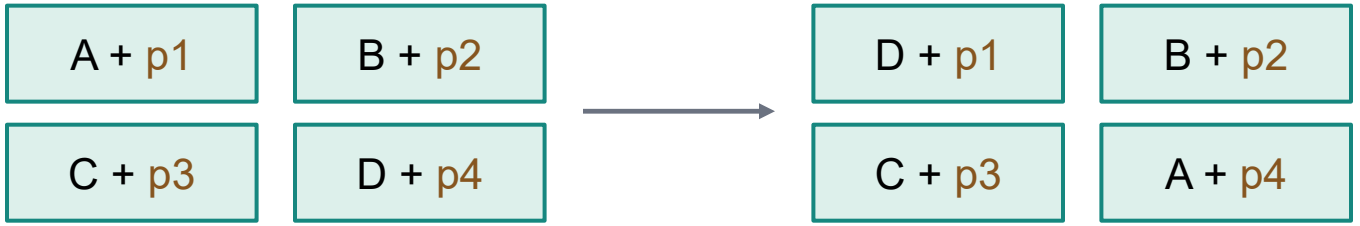
$\boldsymbol{X}_{\text{patch}}$为$N \times P^2C$，共享矩阵E为$P^2C \times D$，偏置b沿行广播；输出T为$N \times D$。



四块小例中N=4、块向量宽4。D是模型特征宽度，可不同于块内像素数。各块使用同一组参数，既不把块编号当类别，也不假设每块已经对应一个物体。

位置表示使区域所在位置参与计算

共享投影变换块内容，位置嵌入补充“内容位于哪个槽位”。本讲为每个固定序列位置保存一个可学习D维向量。



原始内容　　交换A、D；位置槽位不变

交换左上A与右下D的内容，位置向量p_1, p_4仍留在原槽位：输入由$t_A + p_1, t_D + p_4$变成$t_D + p_1, t_A + p_4$。

位置表示使模型有条件学习内容与位置的关联，不保证每次交换都改变预测。参数按块网格的固定扫描顺序与二维位置对应。

分类词元为整图输出设置一个读取位置

图像分类需要整幅图像的一个类别输出。标准ViT在块序列开头加入可学习的**分类词元（cls token）**，其最终表示供分类头读取。

$$\boldsymbol{Z}_0 = [\boldsymbol{c}; \boldsymbol{T}] + \boldsymbol{P}_{\text{pos}}$$

分号表示沿序列行拼接。分类词元c是一行D维参数；位置矩阵$\boldsymbol{P}_{\text{pos}}$为(N+1)×D，含cls位置。

四个块加cls，序列共有五行；每行仍为D维。

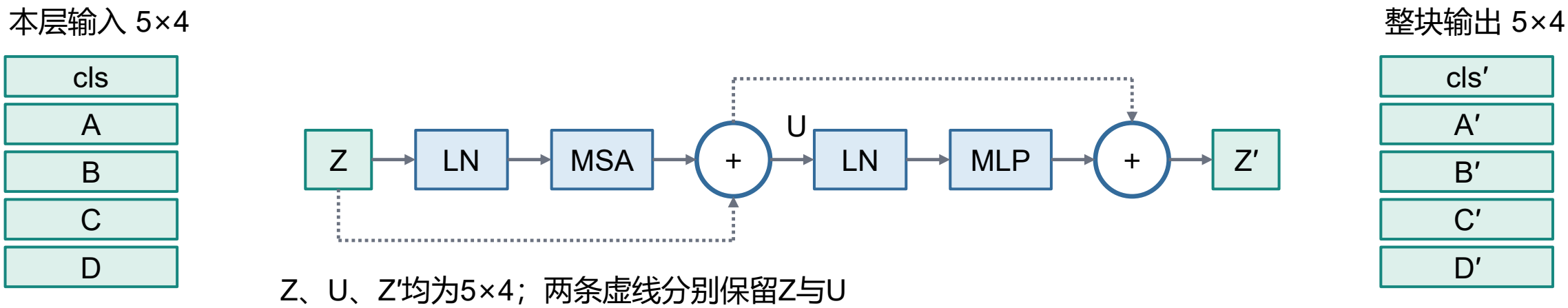
同一组参数用于不同图像，初始cls输入相同。它经过编码器后读取图像信息，输出才依赖当前图像；cls本身不保存正确类别标签。

2 编码器与分类

建立Pre-LN、信息汇集和标签监督的连续计算。

完整Pre-LN编码块

一个ViT编码块有两个残差子层：自注意力混合位置之间的信息，MLP逐位置变换特征。此编码块重复堆叠L次，各层通常使用各自参数。



Z、U、Z'均为5×4；两条虚线分别保留Z与U

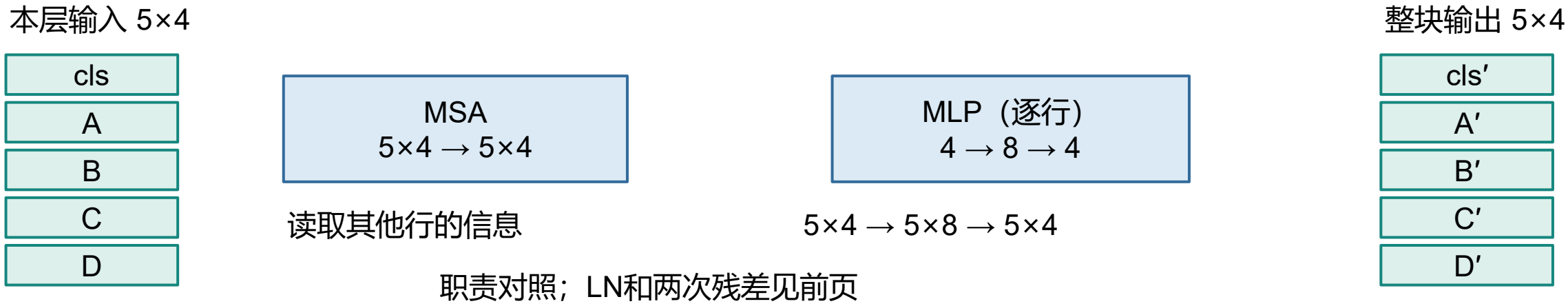
$$\boldsymbol{U} = \boldsymbol{Z} + MSA(LN(\boldsymbol{Z}))$$

$$\boldsymbol{Z}' = \boldsymbol{U} + MLP(LN(\boldsymbol{U}))$$

两条捷径分别保留Z与U，形状均为$(N+1)\times D$。第15讲Post-LN在相加后归一化；这里的LN位于子层支路内。图中省略Dropout。

注意力联系区域，MLP分别变换特征

MSA跨词元组合信息，MLP独立变换每一行的特征。 以下小例取D=4。



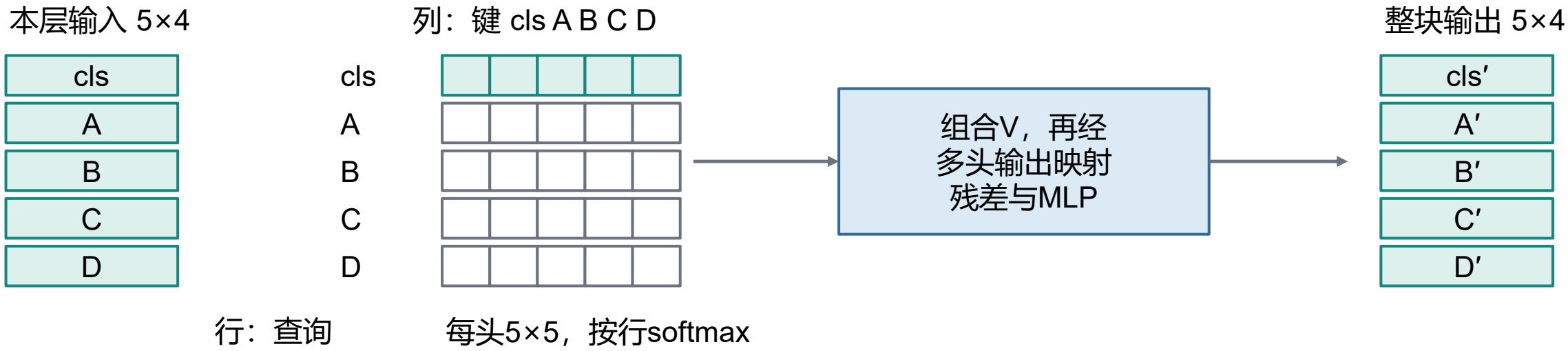
MSA的Q、K、V由同一份归一化输入分别投影。本例双头、每头2维；每头权重为5×5，多头合并并映射后恢复5×4。

MLP接收注意力残差结果U，经LN后逐行作4→8→4变换，位置数始终为5。各行共享参数，但不直接读取其他行。

GELU逐分量提供非线性；LN对每行的D个特征归一化。两个残差相加沿用前页。

同一层并行更新所有位置

一层中的全部Q、K、V都由同一份本层输入得到；五个查询各自形成结果，再一起交给下一层。

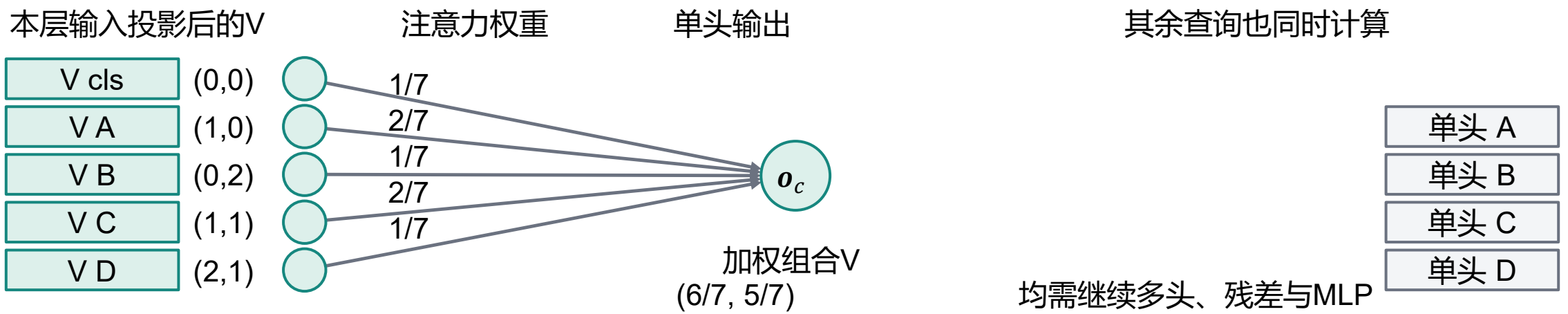


每头权重为5×5：行是查询，列是键。cls与四个块各有读取权重，均可读取五行输入的Value。

cls和各块同时从本层输入形成新表示，再经MLP逐行加工。下一层才接收全部更新结果；最后分类只读cls。下一页放大cls的单头读取。

分类词元的信息汇集

在每个注意力头中，归一化后的各行分别经线性投影得到Q、K、V。cls的查询与各位置的键形成读取权重，再对对应的Value作加权组合。



$$\alpha_{cj} = softmax_j(\boldsymbol{q}_c \boldsymbol{k}_j^{\top}/\sqrt{d_k}), \boldsymbol{o}_c = \sum_{j=0}^{N} \alpha_{cj}\, \boldsymbol{v}_j$$

图示为每头2维的局部算例，Value与权重逐行对应。单头输出还需多头合并、输出映射、残差与MLP，才完成cls更新。

权重随输入与参数变化。多层更新使最终cls依赖图像内容；权重图不是完整的决策因果解释。

分类头读取编码后的cls表示

编码器输出仍含cls与各块的表示。取最后一层的cls行，经末端LayerNorm得到图像表示h：

$$\boldsymbol{h} = LN(\boldsymbol{Z}_L[0,:]), \boldsymbol{z} = \boldsymbol{h}\boldsymbol{W}_{\text{head}} + \boldsymbol{b}_{\text{head}}$$

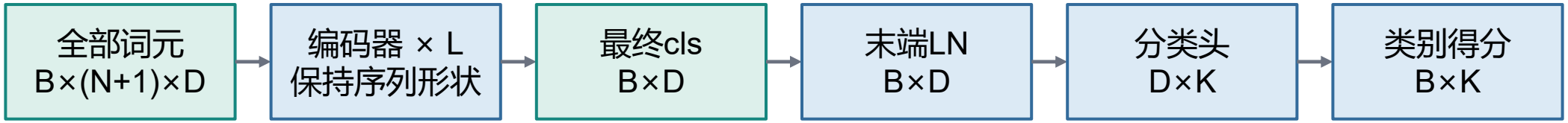
这里h按行表示、长D，L为编码层数，索引0是cls。$\boldsymbol{W}_{\text{head}}$为D×K，K是类别数；z含K个类别得分，softmax后得到类别概率。

本例D=4、K=3：五个位置的表示只取cls一行，形成三个类别得分。分类头的参数由标签监督学习，不把N个图像块当成N个类别。

本讲采用cls读出与线性分类头。原论文预训练时使用含一个隐藏层的MLP头，微调时使用线性头；也可设计块表示的池化读出，但需另定结构及训练。

完整前向把图像映成类别得分

多层注意力使cls汇集区域信息。分类头读取最终cls表示，形成整幅图像的类别得分。



标签损失通过可训练路径更新表示与分类映射

小例B=1、N=4、D=4、K=3；编码器保持1×5×4，最终图像表示为1×4，得分为1×3。每头权重为1×5×5，双头未平均权重为1×2×5×5。

全部有效图像词元均可见。这里没有目标移位、文本解码器或自回归因果掩码。图像标签指导注意力汇集与分类映射的学习。

图像标签通过交叉熵训练整条计算链

每幅图像提供一个类别标签。对正确类y的预测概率p_y，单样本交叉熵为：

$$L = -\log p_y$$

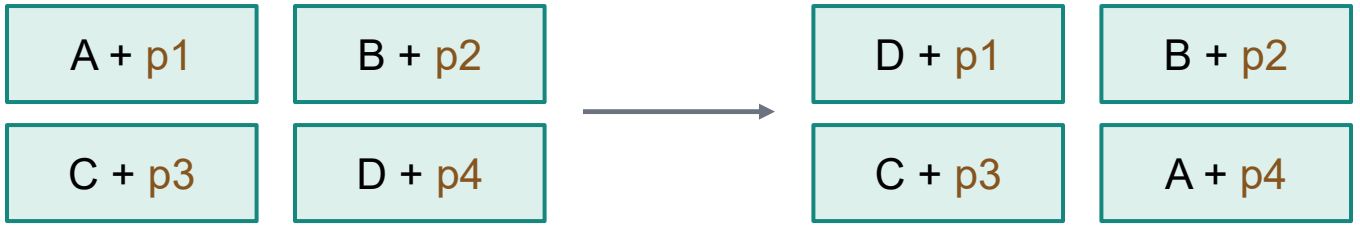
给定三个得分(0, ln2, 0)，标签索引为1，softmax概率为(1/4, 1/2, 1/4)，损失为ln2，约0.6931。这里是独立给定算例，不是未训练整网的预测。

阶段	参数使用	输出用途
训练	反向传播并更新可训练参数	标签损失指导分类头、编码器、cls、位置与块投影
评价 / 使用	本讲固定参数，只作前向	产生类别得分并预测；评价时与保留标签比较

logits经softmax得到概率，由交叉熵反向传播并更新可训练参数。分类效果需在独立数据上评价。

没有位置时，交换块顺序不会改变cls读出

基本：只交换A、D的内容，位置向量仍留在原网格槽位。



原始内容　　交换A、D；位置槽位不变

无位置时，交换两块内容的cls读出相同；有位置时，内容与位置的配对改变，cls读出可以改变，但不保证改变。

深化：交换完整输入行。 内容和各自位置一起移动，原配对不变；键值对及对应权重只是重排，cls加权和保持相同：

$$\alpha_{cA}\boldsymbol{v}_A + \alpha_{cD}\boldsymbol{v}_D = \alpha_{cD}\boldsymbol{v}_D + \alpha_{cA}\boldsymbol{v}_A$$

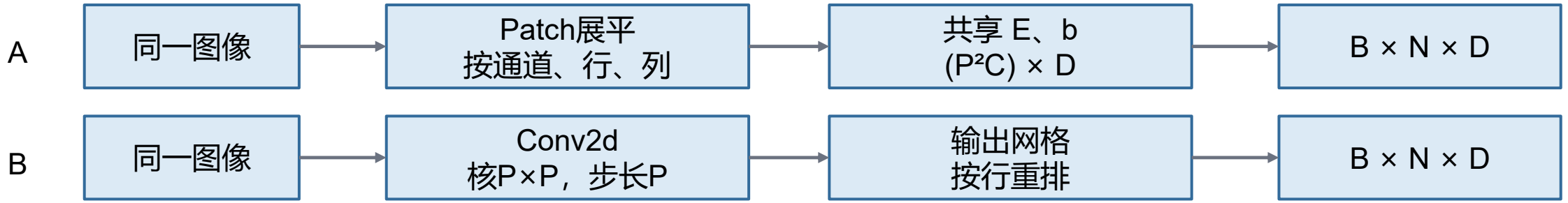
比较固定cls及模型参数，关闭Dropout，无因果掩码或其他位置约束，忽略数值舍入。逐行LN、共享MLP和多层注意力保持完整行重排关系。

3 结构与计算规模

空间关系怎样建模？输入规格变化带来什么代价？

卷积可以实现同样的分块投影

同一批图像可通过两条路径得到相同的块表示；这里比较输入映射。



E的第d列 ↔ 第d个卷积核；相同偏置，padding=0

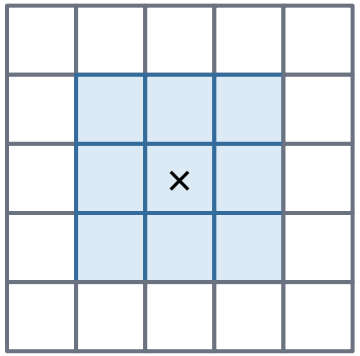
将E的每一输出列按通道、行、列重排为一个卷积核；偏置逐项相同，输出网格按行读取。这样每个输出分量都使用相同像素、权重及偏置。

本例1×1×4×4，经卷积得到1×4×2×2，重排后为1×4×4，与共享投影逐项相同。

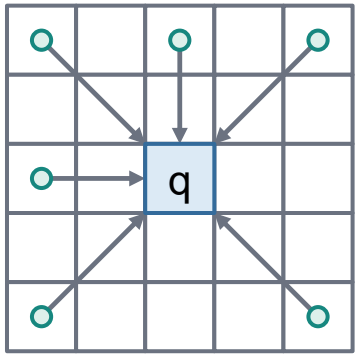
等价的是分块输入映射。 后续区域交互仍要分别比较全局注意力与局部卷积。

CNN与ViT的空间建模

归纳偏置是结构或学习方法预先规定的偏好。下面固定一个输出位置，比较它在一层中怎样联系输入区域。



CNN：3×3局部邻域



ViT：向q读取（代表联系）

CNN卷积在空间上限定局部邻域，同一卷积核跨位置共享；多层可扩大感受野并形成全局信息。

标准ViT单层可读取全部有效词元，权重由Q、K计算，位置表示补充空间信息。图中六条箭头仅示意读取方向，统一线宽，不表示实测权重或因果重要性。

结构规定可连接范围；有用联系与任务表现还取决于训练。

全局读取提供联系，也增加规模要求

全局可连接提供信息读取的机会；哪些联系有用依靠学习，读取规模也有代价。

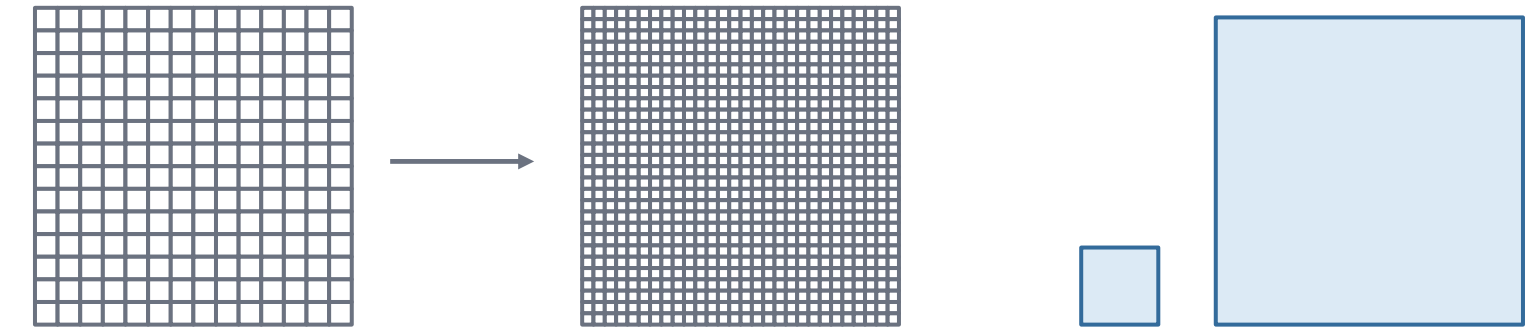
- **联系能力**：一个区域可在单次注意力中读取远处区域。例如判断局部纹理时，可同时利用图像其他区域的表示。
- **学习条件**：可连接不规定“哪个远处区域有用”。位置表示、训练数据和目标共同影响学习结果；标准ViT没有CNN中逐层局部邻域与常见分层下采样的同样约束。
- **计算代价**：标准ViT保持固定块网格。提高分辨率会增加词元数，也增加查询与键的评分配对；下一页具体计算矩阵规模。

Swin以局部窗口和分层表示调整这些结构选择，本讲只作定位。注意力权重的解释边界沿用前文：单张权重图不是完整的决策因果解释。

块数决定每头评分矩阵的大小

Patch边长减半 → 块数4倍 → 每头逻辑评分规模约16倍。

固定224×224图像。块数为N，含cls的序列长度S=N+1，每头评分矩阵为$S \times S$。



P=16：14×14=196块，序列长197，评分元素38,809。

P=8：28×28=784块，序列长785，评分元素616,225，约为前者15.88倍。

固定P=16，将图像长宽都加倍也得到784块。这里固定批量和头数，只比较逻辑评分规模；现代实现可避免完整存储评分矩阵，实际GPU显存和整网耗时不必按同一比例增长。

改变分辨率与改变块大小，适配对象不同

保持特征宽度D，分块投影参数形状为$P^2C\times D$。块数决定序列长度，块内像素数决定投影输入宽度；两者不能混为同一个尺寸。

改变条件	分块投影	位置嵌入	含cls长度
224→448，保持P=16、C=3	仍为768×D，可沿用	块位置196→784，需适配	197→785
P=16→8，保持224、C=3	768×D→192×D，不能原样载入	块位置196→784，需适配	197→785

两种改变得到相同词元数，但第二种还改变每块输入宽度。只对位置向量插值，不能修复分块投影参数的形状不匹配。

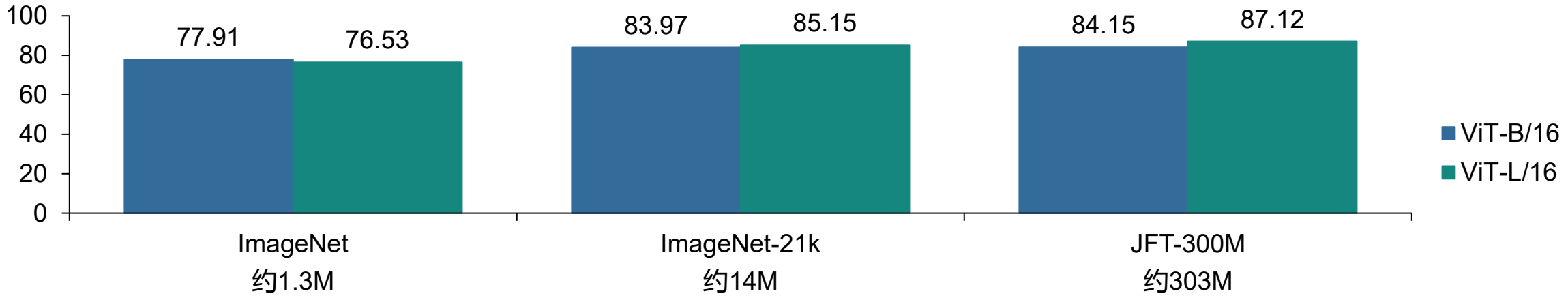
选较细分块时，同时核对局部输入表示、预训练权重和序列规模；更细的网格是否改善任务效果，需要验证，不能只由块数推出。

4 预训练与当前任务

说明输入适配、分类头及微调的使用条件。

预训练规模与模型表现

Dosovitskiy等，ICLR 2021：比较Base（B）与Large（L）两种容量；/16表示块边长16。



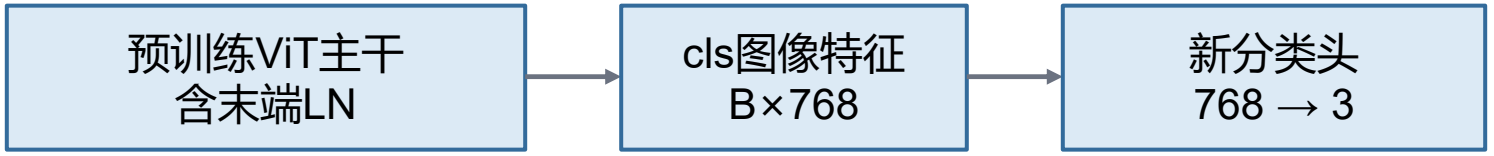
ImageNet Top-1 (%)；原论文表5；384分辨率微调

各数据源分别比较B与L：ImageNet下L低于B，另外两种数据源下L高于B。较大模型的相对优势随预训练条件改变。

Top-1是最高得分类别正确的样本比例。横轴是数据源类别，不是仅改变样本数的受控实验。

把预训练表示接到一个三分类任务

沿用第09讲迁移学习。假设RGB三分类任务，选择输入224×224、P=16、D=768的预训练模型，旧头1000类。



移除旧1000类头；新头按猫、狗、鸟输出

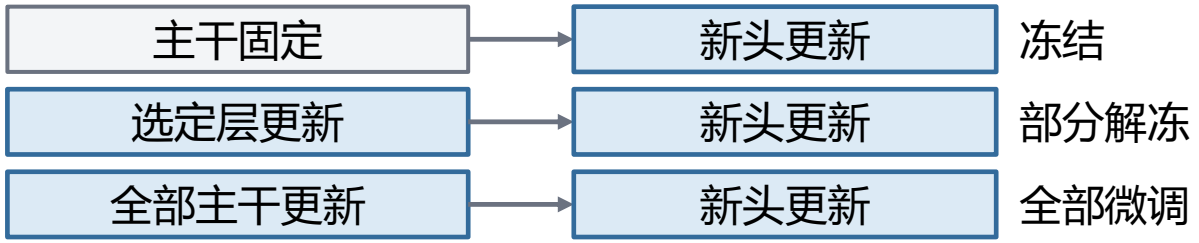
主干输出最终cls图像特征B×768。移除768→1000的旧头，接入待训练的768→3新头，对应固定顺序[猫、狗、鸟]。

输入须匹配该权重的通道、缩放、裁剪和归一化；此规格产生196个块与B×197×768序列。训练集更新、验证集选设置，测试集作最终评价。新头未训练时不能声称已完成三分类。

冻结与微调改变的是学习路径

表示主干包括分块投影、cls、位置、编码器及末端LN。记其输出与新头计算为：

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



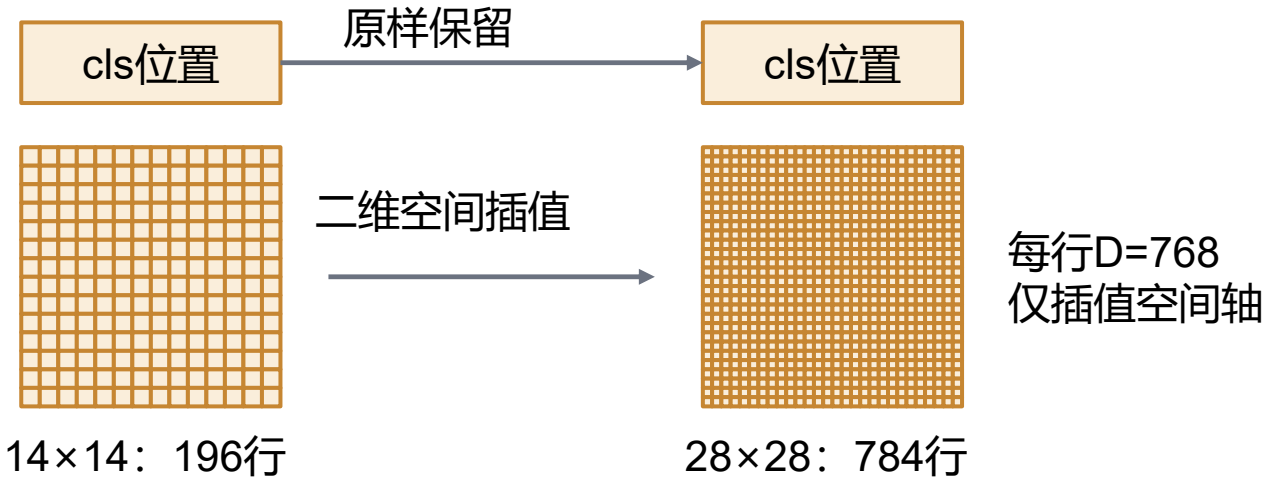
策略	可考虑的条件	取舍与限制
冻结	先检验预训练特征能否区分当前类别	反传范围较小；表示固定，仍需验证任务适用性
部分解冻	需要调整部分表示，并能选择解冻层	更新范围可控；层数与学习率需由验证集比较
全部微调	需要调整完整表示，且有相应数据与计算预算	反传与优化开销较大；需控制过拟合并验证泛化

冻结时要关闭主干参数的更新，优化器只更新选定参数。

训练 / 评价模式与参数冻结分别控制；固定特征方案还需固定主干的随机行为。可先比较冻结方案，再按验证表现决定是否微调。

提高分辨率时怎样适配位置参数

保持P=16、D=768，把输入从224×224提高到448×448。分块投影仍为768×768，块位置从196增为784。

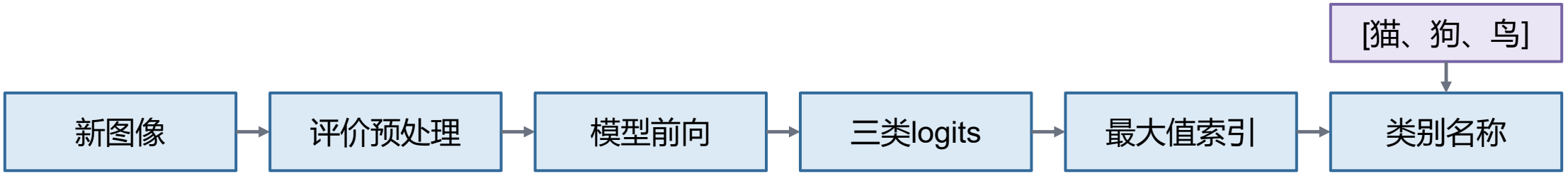


先单独保留cls的一行位置向量；按固定扫描顺序把块位置恢复为14×14×768，在两个空间轴上插值至28×28×768，再展平为784×768并接回cls行，得到785×768。

特征轴D不插值，cls不参与图像网格插值。注意力与MLP参数可复用，实现仍须支持新规格；还需匹配预处理并验证任务表现。改变Patch大小则另涉及投影参数形状，不能只插值位置。

使用模型时同时保留预处理和类别映射

应用深化：新头已训练并选定参数后，把一张新图像变成类别名称。



预处理、配置与权重、类别映射共同保存

评价预处理包含权重要求的缩放、裁剪、归一化，并形成批量轴。设置评价模式与关闭梯度记录分别控制，二者互不替代。

取logits最大值的索引，通过训练时保存的[猫、狗、鸟]顺序恢复名称。如需概率，另作softmax；其大小不能直接当作预测正确率。

共同保存模型配置与权重、评价预处理及类别映射，才能复现这条使用流程。

本讲总结

- **图像如何成为序列？** 分块与共享投影生成同宽词元；拼接cls，再加位置。
- **区域信息如何形成整图分类？** Pre-LN编码器更新全部词元，最终cls经末端LN与D→K分类头得到得分，由标签损失学习。
- **结构选择带来什么能力与代价？** CNN局部读取、ViT全局联系；多层CNN也能形成全局信息。细分块增加评分规模，表现仍受数据、容量与训练方案影响。
- **迁移应用闭环**：匹配预处理，替换新任务头并选择更新范围；分辨率变化与Patch变化分别核对位置及投影参数。