深层卷积网络的演进与残差学习

机器学习 · 深度学习基础

学习目标

- 解释VGG的规则小核堆叠，连接卷积层、块、下采样与完整网络；在给定条件下比较感受野、参数和计算，读取原论文配置与实验。
- 说明AlexNet的背景作用，解释NiN与Inception各自的问题、设计和作用，区分通道变换、拼接与相加。
- 解释训练退化与残差学习，跟踪基本块、投影和相加时的形状条件。
- 依据链式法则解释两条梯度贡献，用原论文比较区分设计理由、机制与实验支持。
- 沿ResNet-18跟踪阶段、GAP与分类，辨认瓶颈的内部成本设计。

1 深层卷积网络的演进

规则加深、局部通道变换与多分支组合。

深层CNN的主要设计问题

深层CNN通过卷积、非线性激活、池化与分类层组织视觉表示。规则加深、局部通道变换和多分支组合分别处理不同的结构设计问题。

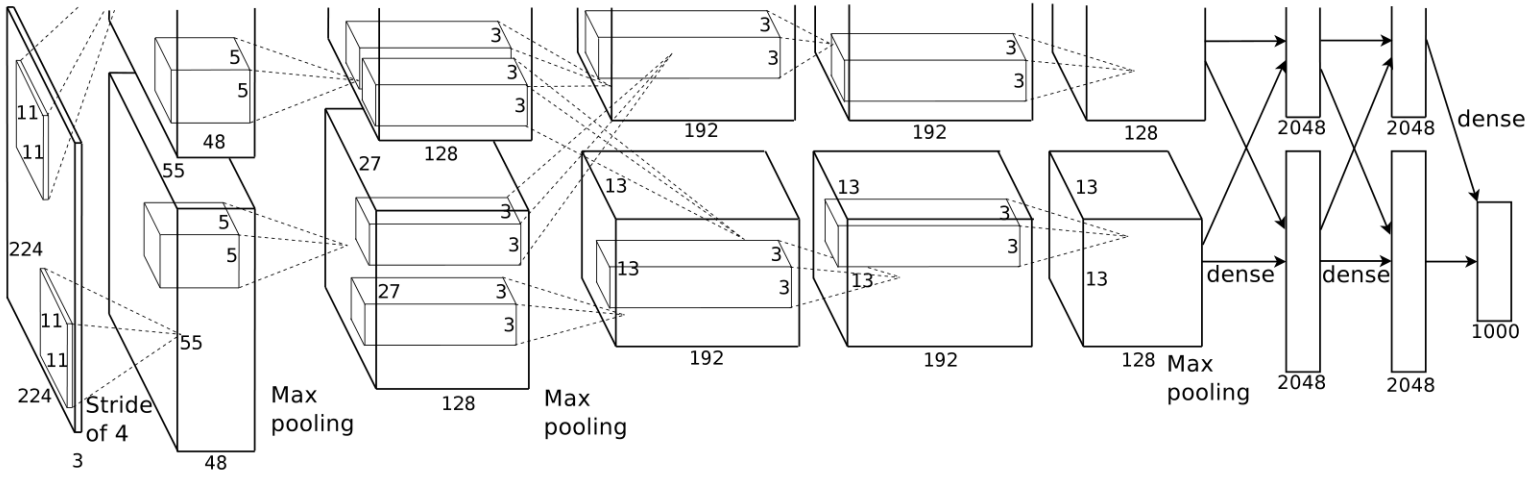
网络	结构设计问题	主要设计与作用
AlexNet	深层卷积怎样完成图像分类?	多层卷积、ReLU、池化与分类层，组合局部信息，形成整图预测
VGG：规则加深	怎样用统一结构研究网络深度?	堆叠小卷积核，组成卷积块与完整网络
NiN：局部变换	固定位置怎样增强通道表示?	1×1卷积配合ReLU；另用GAP形成分类出口
Inception：并行组织	怎样组合不同范围并控制成本?	多分支变换、1×1降维与通道拼接

VGG用规则的小卷积核堆叠构建深层网络。结构配置描述网络的组成，成本比较取决于通道与空间条件，实验检验给定设置下的识别效果。

VGG、NiN与GoogLeNet在时间与技术发展上存在交叠，三者并非简单的单线继承关系。

AlexNet：深层CNN的代表结构

AlexNet层配置各异，通过多层卷积、ReLU、池化与全连接分类逐级组合局部信息，形成整图预测，是大规模视觉识别的代表。



AlexNet，Fig.2（上下为GPU分组）

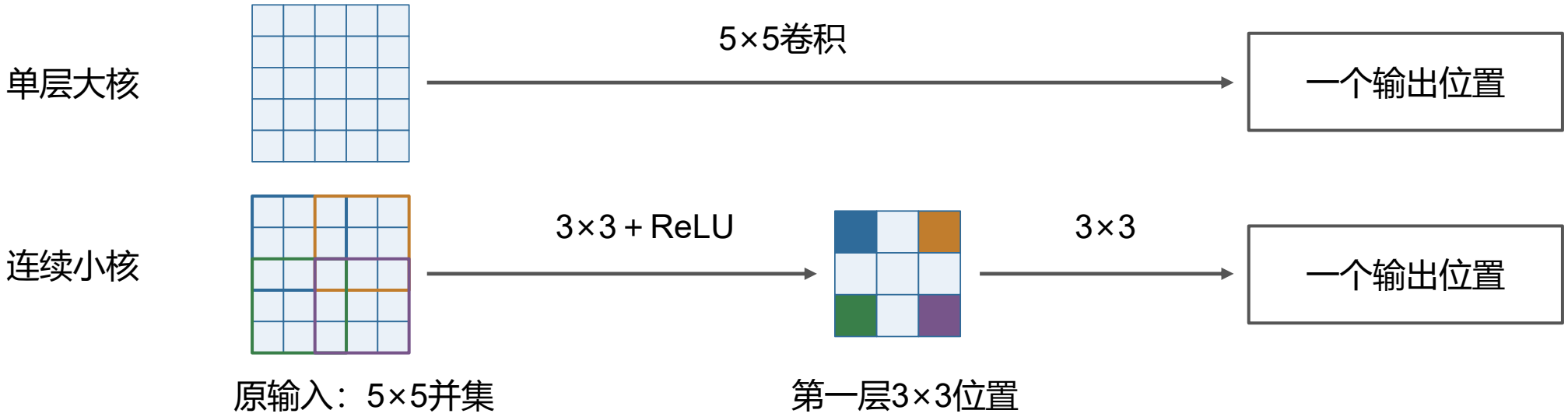
卷积以局部连接和参数共享提取响应；ReLU引入非线性，池化改变空间分辨率，分类层组合最终表示。

较大标注数据、可训练的非线性网络和GPU共同支持端到端学习，识别结果不能只归因于层数。

VGG采用简单、统一的模块构建更深的网络，集中研究网络深度的作用。

VGG：小卷积核的规则堆叠

VGG研究网络深度对视觉识别的影响，采用统一的小卷积核组织较深网络。简单、重复的小核结构用于研究网络深度的作用。



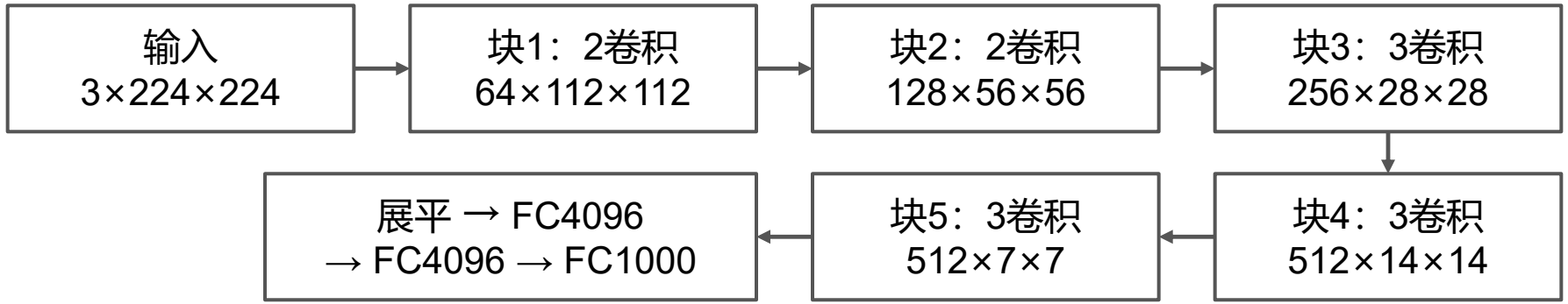
追踪：一个输出→第一层9个位置→输入5×5；四色仅示四角，9个邻域并集5×5

两个连续3×3卷积的内部**理论感受野**为5×5：第二层把第一层相邻位置的结果组合，第一层各位置又依赖原输入邻域。多一层激活还增加一次非线性变换。

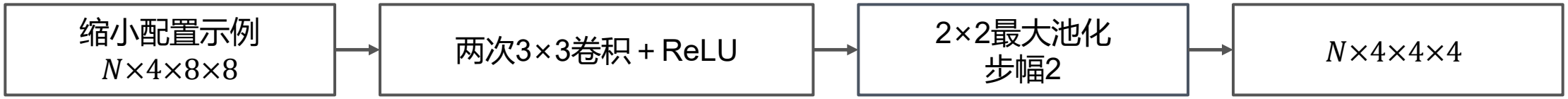
扩大输入范围可以通过多层小核组合实现；相同理论感受野并不表示两个网络计算相同函数。成本是否减少，还要核对通道、空间和具体层配置。

VGG：卷积块与完整分类网络

把若干次“3×3卷积 + ReLU”与一次下采样组合，可作为重复使用的**卷积块**。块内卷积保持空间尺度，块末池化缩小空间；重复的是结构形式，各层参数分别学习。



VGG-16配置D；省略N；各块末端2×2池化s2；所有卷积3×3 s1 p1 + ReLU



阶段组织连续的同尺度计算及其向下一尺度的转换。VGG的五组卷积及各组末端池化体现了这种组织；ResNet阶段的边界由其自身结构确定。

VGG-16有13个卷积层和3个全连接层，共16个带权重层；ReLU与池化不计入名称。VGG-19保留五组通道及分类结构，将后三组各增为4个卷积，共16卷积 + 3全连接。

VGG：小卷积核的参数与计算成本

在相同通道与空间尺寸下，比较单个5×5与连续两个3×3卷积。所有卷积均4输入 / 输出通道、无偏置、步幅1、膨胀1，使用填充使输出均为8×8；两个3×3之间有ReLU。

结构	卷积权重数	单样本主要卷积MAC
单个5×5	$4\times4\times25=400$	$8\times8\times400=25600$
两个3×3	$2\times4\times4\times9=288$	$8\times8\times288=18432$

一个**MAC**是一次乘积累加。本表只统计卷积，不含激活、池化等运算；不直接表示实际运行时间。

比较成立需要说明中间通道。一般两层3×3从C_i经C_m到C_o，权重为$9C_iC_m+9C_mC_o$；单5×5为$25C_iC_o$。只有所有通道均为C时，才得到$18C^2<25C^2$。

这里的小核组合以较少卷积权重与主要MAC覆盖同样的理论范围，但非线性及权重分解改变了函数。不能据此断言“两个3×3总比一个5×5更省”或性能一定更好。

VGG-16与VGG-19的网络配置

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

VGG原论文Table1
D：蓝框；E：橙框（图外标注）

原论文Table1列出A—E配置，卷积层按池化分为五组，末端接三个全连接层。D对应VGG-16，E对应VGG-19。

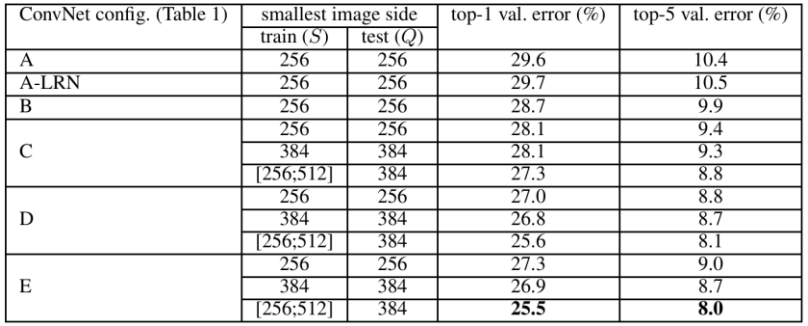
D的五组卷积数为2 / 2 / 3 / 3 / 3，E为2 / 2 / 4 / 4 / 4；后者在第3—5组各增加一个3×3卷积，通道序列和分类尾部保持。D有13卷积 + 3全连接，E有16卷积 + 3全连接；激活与池化不计入名称层数。

这些配置给出了不同深度的具体结构。识别效果还需在明确的训练与测试条件下比较，层数本身不提供效果结论。

VGG：网络深度与识别错误率

原论文Table3比较不同配置的验证错误率。下表选取固定训练尺度与单测试尺度均为256的A / B / D / E结果；配置与带权重层数对应原论文Table1。

配置	带权重层数	Top-1错误率（%）	Top-5错误率（%）
A	11	29.6	10.4
B	13	28.7	9.9
D	16	27.0	8.8
E	19	27.3	9.0



ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	28.7	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

VGG，Table3原表
含不同尺度；摘表仅选S = Q = 256

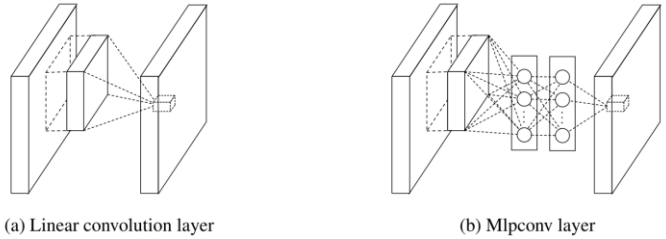
数据来自ILSVRC-2012验证集，固定训练最短边$S = 256$、单测试尺度$Q = 256$，采用论文的单网络稠密评价方式，错误率越低越好。

在这组条件下，A到D的结果改善，但E并未在两项指标上优于D。因此，论文支持这些配置下深度设计的价值，不证明网络越深总越好，也不能仅凭本表把差异归结为某一个因素。

VGG着重规则空间卷积。固定空间位置的通道表示，还可通过多层非线性变换增强。

NiN：局部通道的非线性变换

VGG的3×3已组合空间与通道。**Network in Network（NiN）**在空间卷积后增加多层**1×1卷积**与ReLU，形成局部多层感知机**mlpconv**，增强逐位置的非线性表示。



NiN，Fig.1（原论文）



示例，步幅1、无偏置：一组权重(1,0,0,1)得到1 + 4 = 5
同一组权重用于各空间位置，其余7组输出不在本例求值

1×1仅指空间核大小，每组仍读取全部输入通道；激活引入非线性，权重跨位置共享，其本层空间感受野不扩大。

局部模块输出仍是空间特征图；图像分类还需要将各位置的响应汇总为整图类别得分。

NiN：全局平均池化与类别得分

局部特征仍分布在多个空间位置。NiN在末端**主动构造K个类别通道**，再把每个通道的空间响应汇总成一个类别得分；隐藏通道并非天然都有类别含义。

全局平均池化（Global Average Pooling，GAP）对每个通道的全部空间位置取平均：只汇总空间，不混合通道，也不引入可学习权重。

独立示例：K = 3，每张类别响应图为2×2；表中分号分隔两行。

类别通道	2×2空间响应	逐通道取平均	类别得分
类别1	[1,3;5,7]	(1 + 3 + 5 + 7)/4	4
类别2	[0,2;0,2]	(0 + 2 + 0 + 2)/4	1
类别3	[0,1;3,4]	(0 + 1 + 3 + 4)/4	2

三个通道分别得到一个得分，组成**(4,1,2)→softmax**。这些数字只解释平均操作，不是论文配置或训练结果。

NiN与ResNet的GAP分类结构

GAP只决定怎样汇总空间，**不决定通道代表什么**。类别得分怎样产生，还取决于GAP前后的网络结构。

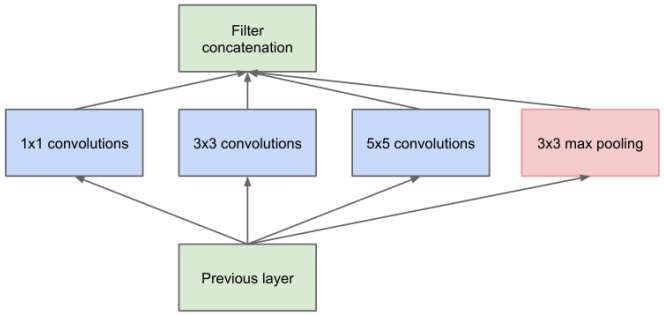
比较问题	NiN的类别通道出口	ResNet（10类配置）
GAP前是什么?	末端主动构造K个类别通道	512个一般特征通道
GAP得到什么?	K个类别得分	512个图像级特征
还需线性分类层吗?	此出口直接使用类别得分	线性层512→10，产生10类得分

NiN的隐藏通道并不天然对应类别；类别对应是末端架构的明确选择。三类示例中的(4,1,2)可直接作为三类得分送入softmax；不能据此把任意网络的GAP输出都当成类别得分。

NiN的两项设计各负其责：**局部模块增强非线性表示，GAP汇总空间响应**。局部通道变换之外，不同空间范围的变换还可以通过并行分支组合。

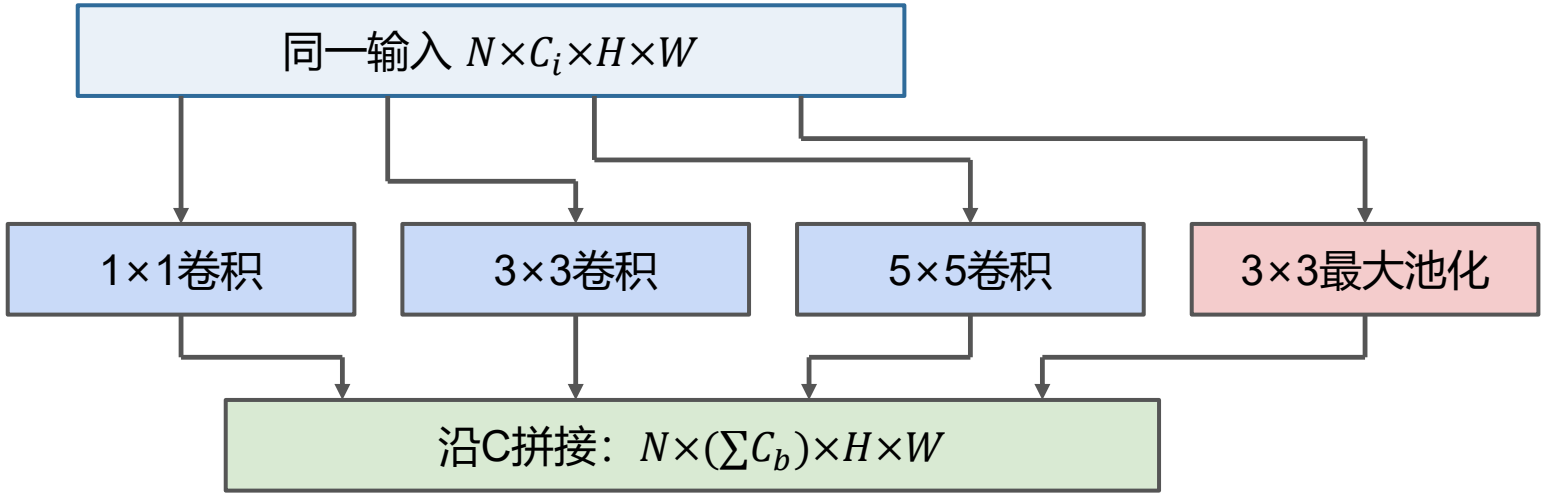
Inception：多分支变换与通道拼接

单一卷积配置只采用一种局部变换。Inception在同一输入上并行使用不同范围的卷积及池化，再把各支结果保留到下一层。



(a) Inception module, naïve version

Inception，Fig.2(a)
原论文朴素模块



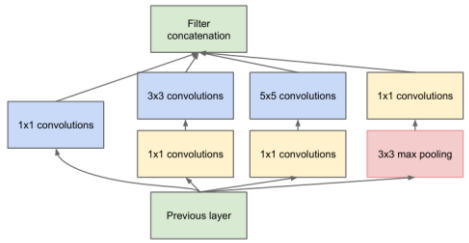
各支s1，填充保H/W；沿C拼接，不做相加

通道拼接把各支通道接在一起，输出通道数是各支之和；不同于逐元素相加，它不要求各支通道数相同，但批次与空间索引必须对齐。

不同范围的变换提供不同局部组合，拼接保留它们供后续使用；这是一种结构设计，不保证每一支都更有效。若高通道输入直接进入3×3、5×5，大核成本会很高；池化分支还可能使输出通道快速增长。

Inception：大核卷积前的通道降维

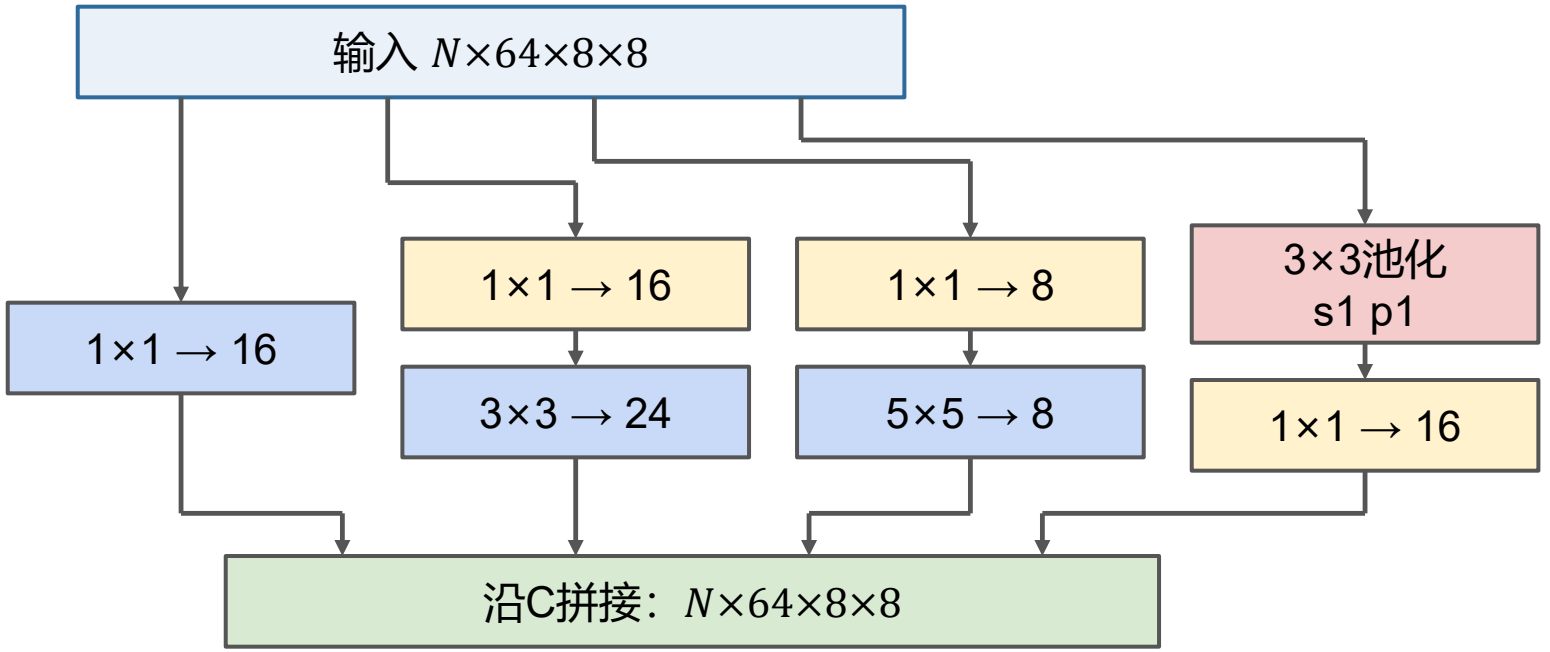
用1×1先压缩通道，再进行大核卷积；池化后也可用1×1控制送入拼接的通道数。



(b) Inception module with dimension reductions

Inception，Fig.2(b)
原论文降维模块

$k\times k$卷积需$k^2C_iC_o$个权重；先降到d通道则需$C_id + k^2dC_o$。同空间、无偏置时，是否更省取决于d及完整分支配置。



各支s1、填充保8×8、卷积后ReLU；16 + 24 + 8 + 16 = 64（示例）

3×3支路权重（无偏置）：64×24×9 = 13824
先64→16→24：64×16 + 16×24×9 = 4480（仅该支路）

拼接保留各支通道，残差相加要求同形状。结构选择仍不保证加深更易训练；普通网络与ResNet的对照实验检验加深时的优化表现。

2 残差学习与优化机制

训练退化、残差学习、梯度机制与实验证据。

深度增加与普通网络的训练退化

增加卷积层，是希望新增计算学到有助于任务的表示变化。但在原论文的 CIFAR-10 实验中，56 层普通网络的训练误差高于 20 层。

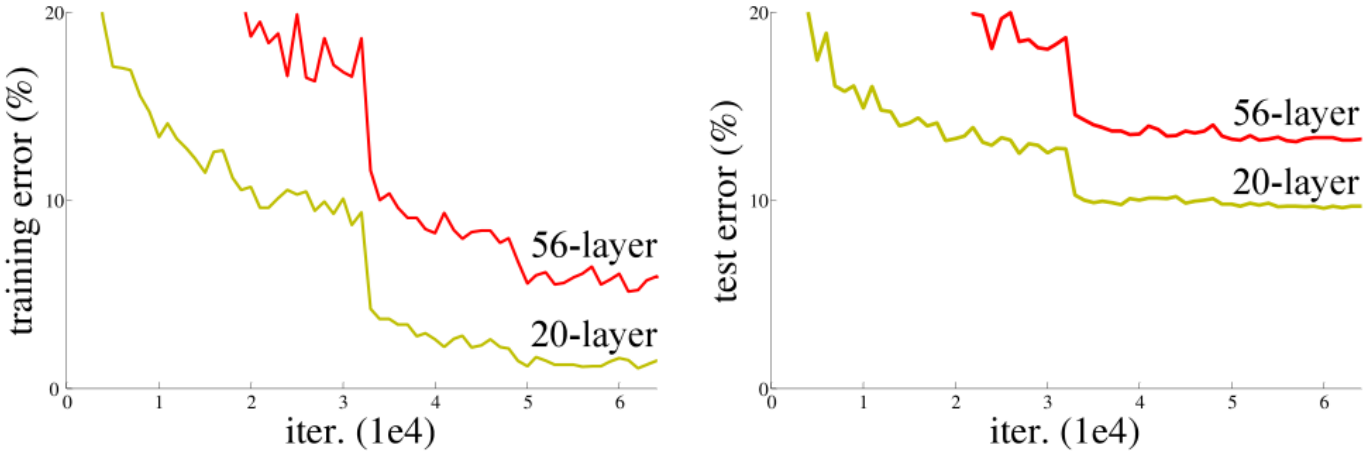


Fig.1：左训练，右测试；原论文实验

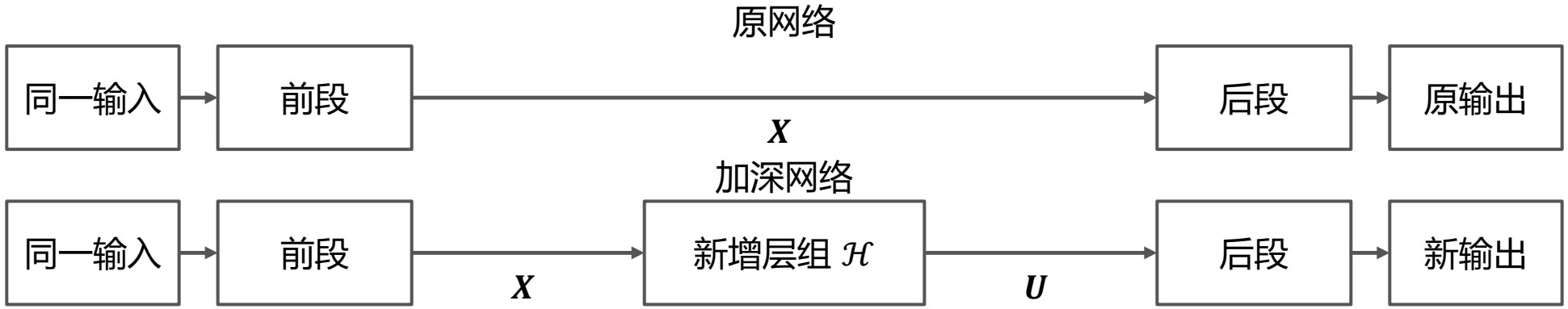
训练退化指这里随深度增加而出现的训练表现恶化。训练集上的拟合也变差，仅用过拟合不足以解释这一现象。

能否为新增层设置参数，使加深后仍得到原网络的输出？若能，加深就不必损失原有的拟合能力；训练是否找到这种参数，则是另一个问题。

He et al., Deep Residual Learning for Image Recognition, CVPR 2016.

新增计算怎样保留原来的结果?

比较原网络与加深网络：输入相同，前段、后段的结构和参数不变，仅在两段之间增加一组层。前段产生表示 $\boldsymbol{X}$；新增层组的映射记为 $\mathcal{H}$，输出 $\boldsymbol{U} = \mathcal{H}(\boldsymbol{X})$与 $\boldsymbol{X}$形状相同。



- 若对所有相关 $\boldsymbol{X}$都有 $\mathcal{H}(\boldsymbol{X}) = \boldsymbol{X}$，新增层组就实现了**恒等映射**。后段输入相同、计算相同，因此整网输出相同。
- 新增结构需能表示恒等映射。参数存在，不等于普通深层网络能训练到同样好或更好的结果。

恒等映射只是保留原结果的一种选择，并非新增层的统一目标。加深网络还希望学到有用的变化：能否让新增部分既能保留输入，也能改变表示?

直接传递输入，再学习对表示的修正

残差结构用**恒等捷径**传递 $\boldsymbol{X}$，由可学习的**残差分支**计算变化 $\mathcal{F}(\boldsymbol{X})$。两路同形状结果逐元素相加，得到末端激活前输出 $\boldsymbol{U}$：

$$\boldsymbol{U} = \boldsymbol{X} + \mathcal{F}(\boldsymbol{X}) = \mathcal{H}(\boldsymbol{X})$$

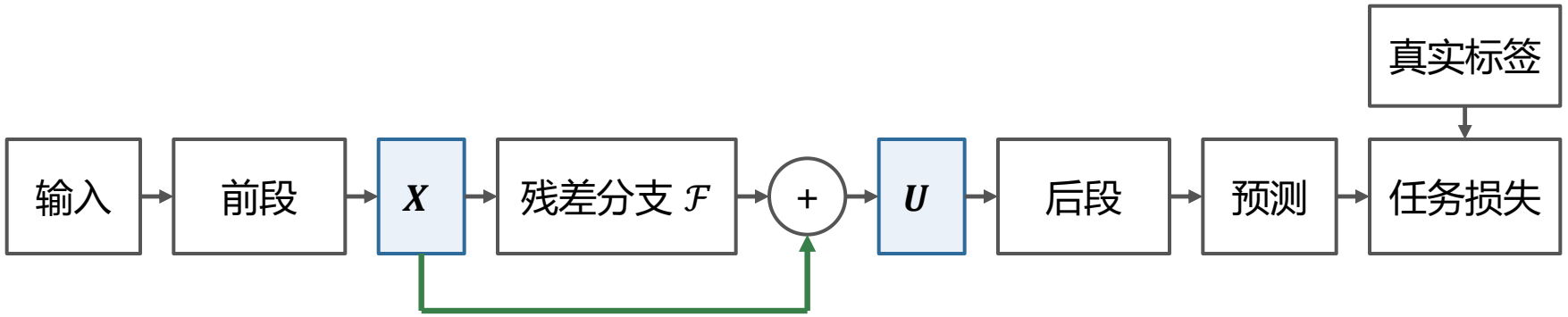


- 普通堆叠产生整个输出；残差结构直接传递输入。若任务适合保留或小幅调整表示，恒等捷径提供基准。
- $\mathcal{F}$用共享参数为不同输入计算变化 $\mathcal{F}(\boldsymbol{X})$：零变化保留输入，非零变化改变表示。
- 分支由最终任务损失引导学习，输出不预设为输入。训练优势还需传播分析与实验检验。

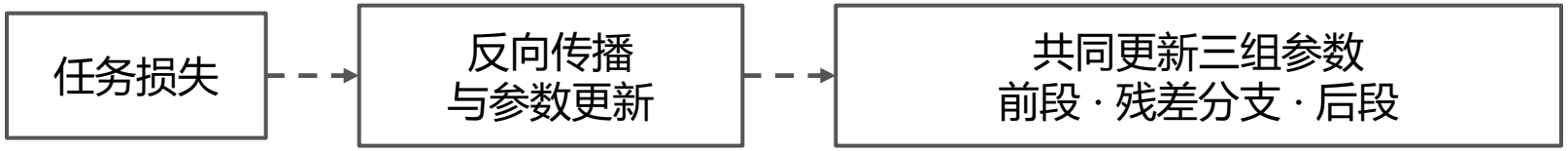
残差分支怎样与前后段一起学习?

采用残差结构时，前段、分支和后段共同训练；恒等构造只说明参数存在，不是先训练恒等网络再修改。

前向：从输入到任务损失（暂略块末端激活）



训练：三部分共同更新（参数更新关系示意）



反向传播计算梯度，优化器据此共同更新三组参数；标签评价最终预测，不给残差分支另设目标。

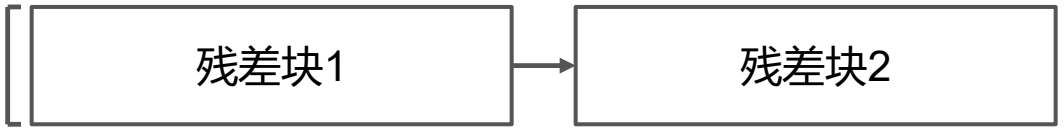
网络变深：含残差分支的路径增加层数，捷径仍短，表示形状不变。论文对照实验展示所比较设置下的训练改善。

ResNet的网络、阶段与残差块

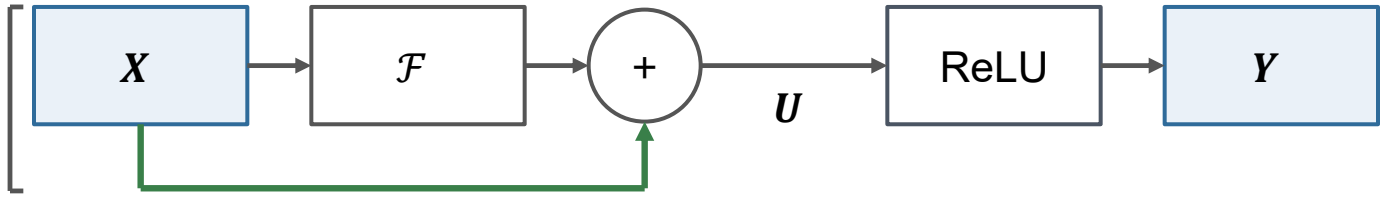
重复堆叠残差块可组成残差网络（ResNet）。一组层及其连接作为整体复用，称为**网络块**；多个残差块组成一个阶段，同一阶段通常保持空间尺度。



阶段1的块结构



残差块1的内部结构

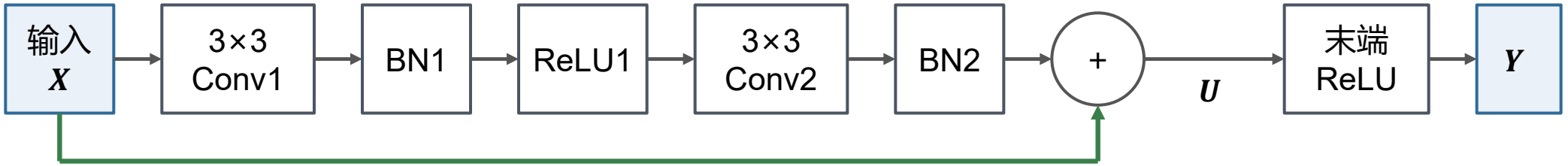


侧边括号表示展开范围；带箭头的实线表示计算

阶段1包含两个残差块。每个块由残差分支与捷径分支组成，两支相加后经过ReLU；残差分支可由卷积层实现。

卷积实现修正，两支相加后再激活

同一输入 $\boldsymbol{X}$进入两支。所示**基本残差块**采用后激活结构：两次 3×3 卷积形成残差分支，恒等捷径直接传递输入。



修正 $\mathcal{F}(\boldsymbol{X})$；在 $\boldsymbol{U}$ 处逐元素相加

恒等捷径 $\mathcal{S}(\boldsymbol{X}) = \boldsymbol{X}$

灰框：运算；浅蓝：表示；绿线：捷径

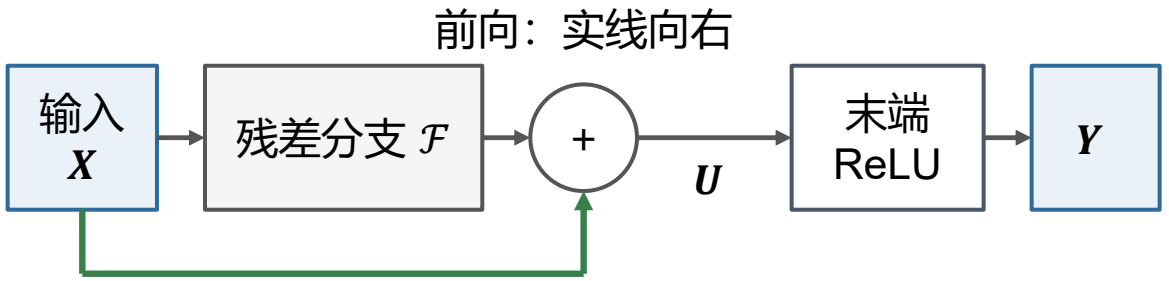
$$\boldsymbol{U} = \mathcal{F}(\boldsymbol{X}) + \boldsymbol{X}, \boldsymbol{Y} = ReLU(\boldsymbol{U})$$

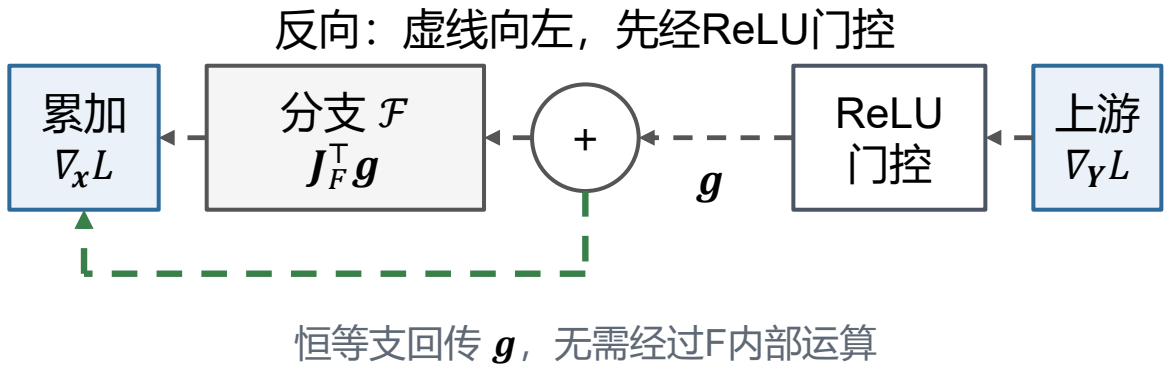
两支按相同 N, C, H, W及对应索引逐元素相加，通道数保持。CNN 中 BN 对每个通道汇总 N, H, W上的值并保持形状；启用运行统计时，训练用批次统计，推理用保存统计。

零残差使 $\boldsymbol{U} = \boldsymbol{X}$；整个块输出为 $ReLU(\boldsymbol{X})$，输入非负时才恢复该输入。

恒等捷径增加一项直接梯度贡献

恒等捷径使输入梯度增加一项直接贡献，它不经过残差分支内部运算。反向图中的两项在输入处累加。





$$\nabla_x L = g + J_F^{\top} g, g = \nabla_u L$$

将同形状张量 X, U按同一顺序展开为列向量 x, u。**雅可比矩阵**J_F记录分支各输出对各输入的局部导数；公式两项对应恒等支与残差支。

标量示例（无末端激活）：$F(x) = 0.2x$、$g = 1.4$，恒等支 1.4 + 残差支 0.28 = 输入梯度 1.68。

g已包含末端 ReLU 门控，两支仍可能抵消。直接通路不保证收敛，训练退化也不等同于梯度消失。深层训练还需适当的初始化和归一化，控制信号及更新的尺度。

ResNet：训练退化的对照实验

ImageNet 上，18 / 34 层普通与残差网络采用相同主干配置比较；残差捷径为 Option A（恒等 / 增维补零），没有额外可学习参数。

①普通网络：18→34层　　②ResNet：18→34层

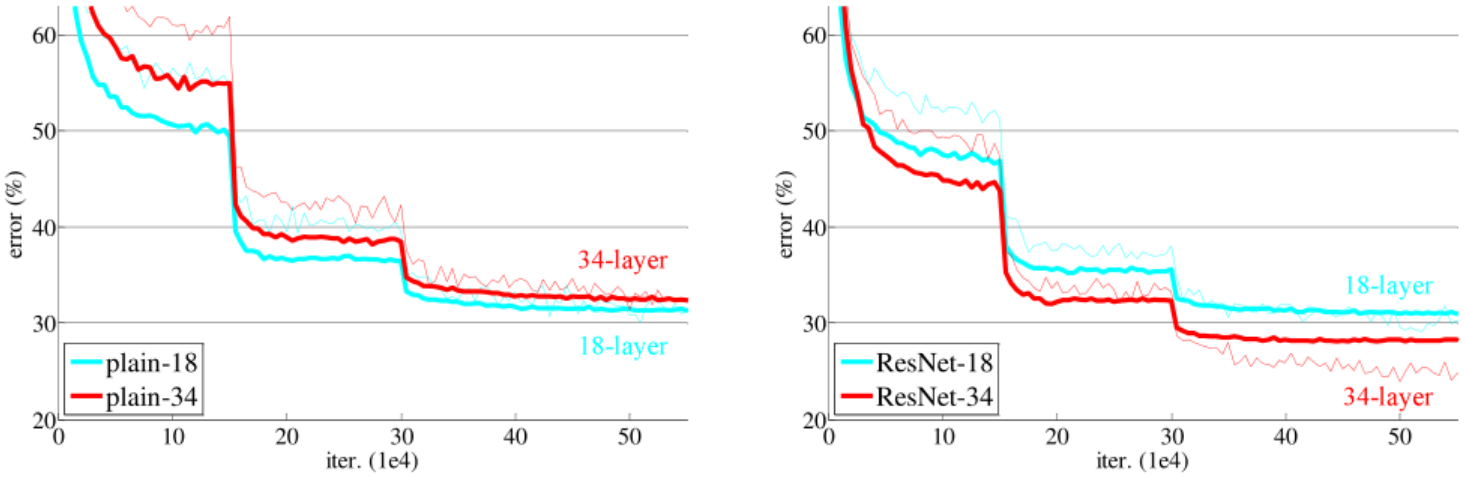


Fig.4　细线：训练；粗线：验证（中心裁剪）

深度	普通网络	ResNet
18 层	27.94%	27.88%
34 层	28.54%	**25.03%**

表为论文 Table 2：ImageNet 验证集 Top-1 错误率，10-crop，越低越好；图中验证曲线采用中心裁剪。

细线表示训练误差：①左图18→34层，普通网络加深后升高；②右图18→34层，残差网络加深后降低；③同为34层，残差网络低于普通网络。表中列出对应的验证结果。

原论文 §3.3—4.1、Fig.4、Table 2

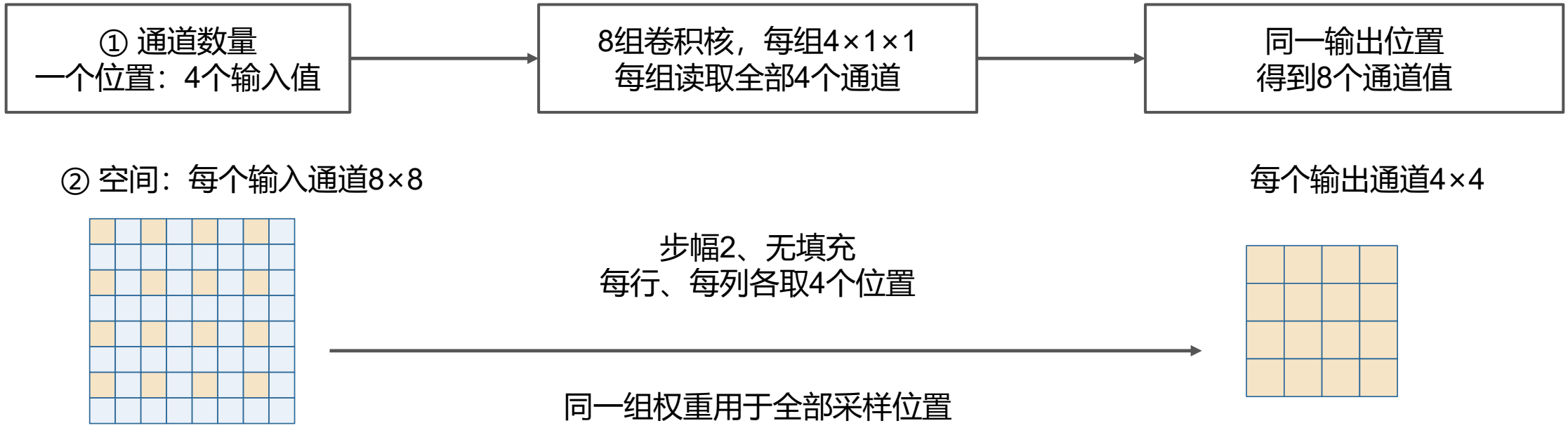
18层两者的验证结果接近；34层 ResNet 比普通网络低3.51个百分点。训练曲线支持本实验设置下的退化缓解。

3 从残差块到完整ResNet

阶段变换、投影捷径与完整图像分类网络。

1×1卷积怎样同时匹配通道与空间?

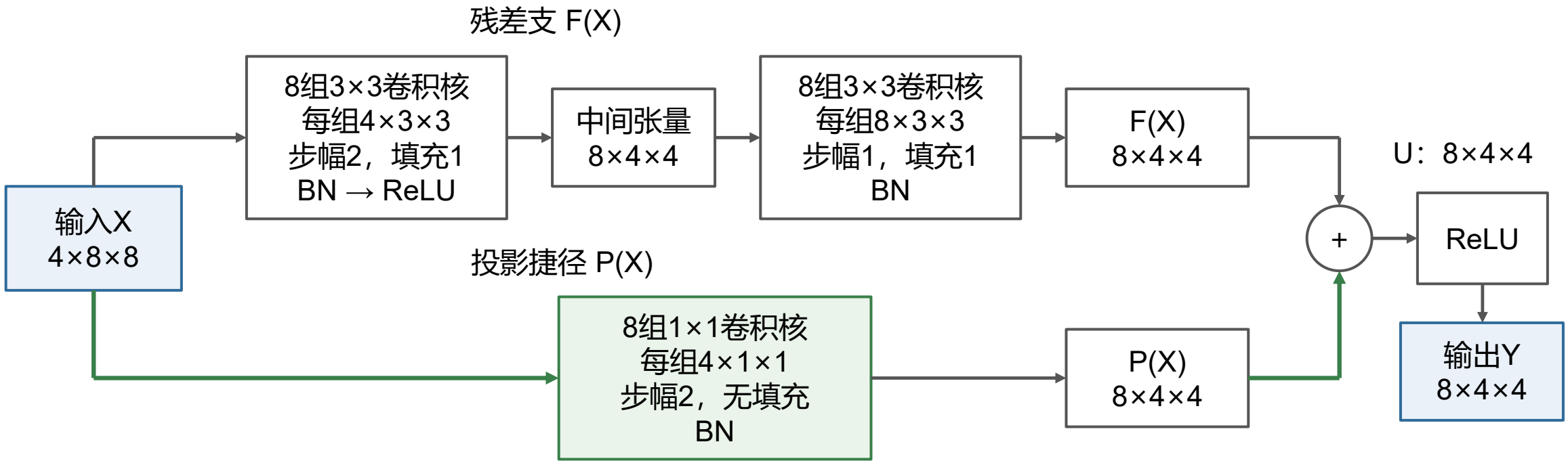
所示ResNet跨阶段缩小空间、增加通道。若残差支把 $N\times4\times8\times8$变为 $N\times8\times4\times4$，输入就不能直接相加。1×1卷积可变换捷径的通道数量，并通过步幅调整空间尺寸。



零基索引下，输出 (i,j)读取输入 $(2i,2j)$处的4个通道值。本例无偏置；1×1是仿射变换，非线性由配套激活引入。

投影捷径与残差分支的形状匹配

同一输入 X分别经残差分支与投影捷径，得到同形状结果。

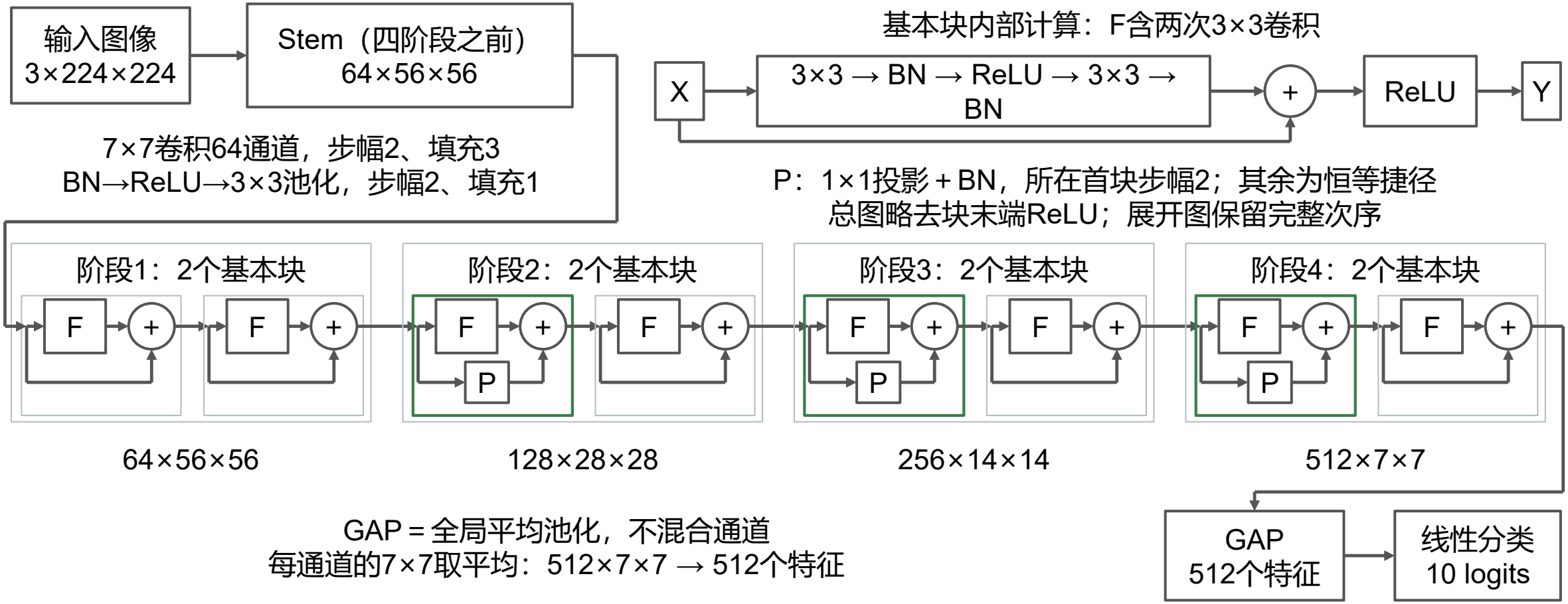


每组核：输入通道×核高×核宽；BN与ReLU保持形状（省略N；卷积无偏置）

逐元素相加后仍为8×4×4，通道数不相加。ResNet-18阶段2—4的首块采用这种变形结构。

原论文对照实验用Option A（恒等 / 增维补零）；所示块用可学习投影。

ResNet-18的四个主体阶段

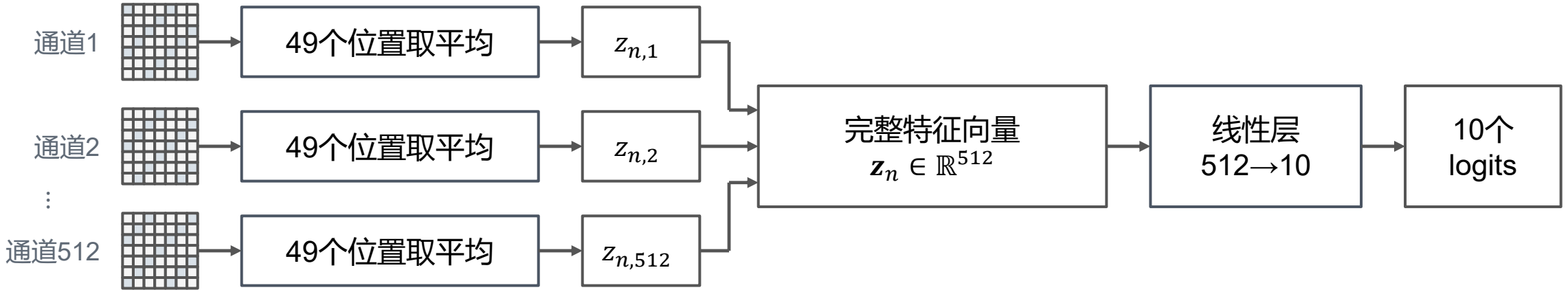


18 = Stem的1个卷积＋8块×2个主支卷积＋1个分类层；投影不计入命名

C×H×W表示通道×高×宽；省略N。含BN投影、10类分类头、无预训练，区别于Option A实验。

全局平均池化（GAP）：逐通道汇总，再分类

全局平均池化（GAP）对每个通道的空间数值取平均。例如一个2×2通道的四个值为1、3、5、7，平均后为4。512个通道分别平均，得到512个特征；线性层再跨通道组合成类别得分。



示意样本n；批量形状：$N\times512\times7\times7 \rightarrow N\times512 \rightarrow N\times10$；灰度非实测

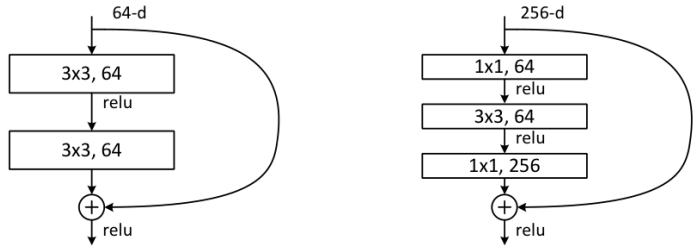
$$z_{n,c} = \frac{1}{HW}\sum_{h,w} A_{n,c,h,w}$$

$\boldsymbol{A}$为 $N\times C\times H\times W$激活，$n, c, h, w$依次索引样本、通道和空间位置。GAP不混合通道；隐藏通道无固定类别含义。

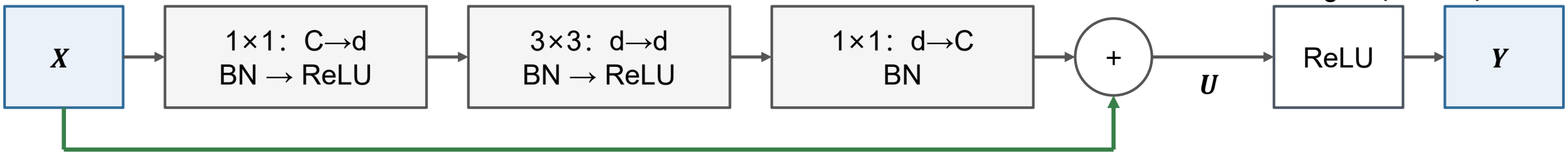
10类分类头输出 $N\times10$个 logits，与 N个 0—9 类别标签接入交叉熵及反向传播。残差连接改变网络结构，损失与反向传播接口保持；该配置与ImageNet论文实验不同。

瓶颈残差块的通道与成本设计

瓶颈残差块保留残差连接，以较窄通道承担主要空间卷积；分支先压缩通道，再恢复输出。这里两端为C个通道，中间为$d < C$个通道，空间尺寸不变。



ResNet Fig.5（原论文）



中间：$N \times d \times H \times W$，$d < C$；两端：$N \times C \times H \times W$；恒等捷径
三次卷积均s1，3×3填充1；前两次BN后ReLU，第三次BN后直接相加

3×3 卷积作用在较窄的 d个通道上，主要权重数为 $9d^2$；完整分支还包括两次 1×1 的 $Cd + dC$个权重。具体总成本须连同形状、BN及捷径配置计算。

基本块使用两次 3×3 主卷积；瓶颈块使用三次主卷积。内部成本设计与捷径连接是可以分别调整的两个方面。

原论文 §4.1、Fig.5

总结：深层CNN的结构设计与残差学习

深层CNN的演进包含不同的设计问题，不能只用网络层数或一个性能排名概括。

网络	主要结构设计	作用及必要区别
AlexNet	多层卷积、ReLU、池化与分类	建立深层视觉表示背景，数据与计算共同支持训练
VGG	规则小核、重复块与下采样	以统一结构研究深度；范围、非线性与成本分别判断
NiN	mlpconv、逐位置通道变换与GAP	增强局部非线性，区分空间计算与通道组合
Inception	并行变换、1×1降维与通道拼接	保留不同范围的结果并控制通道成本；拼接不等于相加
ResNet	恒等 / 投影捷径与残差分支	围绕输入基准学习变化，以阶段组成完整分类网络

残差相加使输入梯度含两条路径的贡献；两条贡献受实际激活、投影和抵消条件影响。优化作用由传播分析和原论文对照共同支持，不是任意深度的训练保证。

VGG与ResNet分别重点回答“怎样有规则地加深”和“怎样用残差结构改善深层网络的优化”。这些模型在时间与技术发展上存在交叠，并非严格的单线继承关系。