激光与光电子学进展
先进成像

深度自适应动态神经网络进展综述

孙毅，李健*，徐昕**，王宇茹
国防科技大学智能科学学院，湖南 长沙 410000

摘要 深度神经网络极大地推动了视觉感知与自然语言处理任务的发展，但当前大多数深度模型执行的是静态推理图，即网络结构（深度）在推理阶段固定不变。这种静态推理模式使模型无法根据输入样本的特性（如复杂度）自适应地调节深度，无法平衡模型推理效率和预测精度。因此以深度自适应神经网络为代表的动态神经网络近年来引起了广泛的关注，该类神经网络能够根据输入样本的复杂度自动地调节推理深度。从三个角度对现有工作进行系统性的归纳总结：深度自适应网络的结构设计；样本复杂度估计方法的研究；深度自适应神经网络的训练方法。最后讨论了该方向未来有价值的问题。

关键词 动态神经网络；深度神经网络；深度自适应；样本复杂度；视觉感知
中图分类号 O436 **文献标志码** A **DOI**：10.3788/LOP202259.1415008

Depth-Adaptive Dynamic Neural Networks: A Survey

Sun Yi, Li Jian*, Xu Xin**, Wang Yuru
College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410000, *Hunan, China*

Abstract With the advancement of deep neural networks, research in visual perception and natural language processing has made significant progress. However, almost all current state-of-the-art deep neural models use static inference graphs, with the inference depth remaining constant throughout the inference stage. Because of this static inference mode, the model cannot adapt its depth to the complexity of the input data. Hence, static models cannot achieve a good trade-off between efficiency and accuracy. Conversely, depth-adaptive dynamic neural networks can decide the inference depth adaptively based on the complexity of the input data, indicating a promising research field for achieving efficient and robust deep models. We comprehensively review the works in this field and summarize the current literatures in three areas: depth-adaptive neural network structure design, data complexity estimation approaches, and depth-adaptive neural network training methods. Finally, we discuss the important future research problems in this field.

Key words dynamic neural network; deep neural network; depth adaptation; data complexity; visual perception

1 引 言

深度神经网络（DNN）极大地推动了人工智能技术的进步。DNN依靠级联非线性计算单元[1]和过参数化[2-3]等先验获得了比传统机器学习算法更强的非线性特征表征能力和基于隐式约束的可优化能力。以AlexNet[4]、ResNet[5]、VGG[6]为代表的神经网络结构极大地推动了视觉感知任务的发展。在目标识别与检测[7-12]、语义分割[13-16]和单目深度估计[17-18]等领域，深度神经网络模型持续刷新着最优性能。近年来，以注意力机制为核心模块的Transformer模型[19-21]进一步提升了深度神经网络模型的表达能力。虽然如此，深度神经网络模型仍然面临一系列挑战。一方面，从最初的LeNet[22]到AlexNet[4]，再到ResNet[5]以及现在的Transformer，在模型表达性能变得越来越强大时，网络深度也在不断增加，模型对计算资源和存储资源的需求也随之变大。另一方面，网络深度的增加虽然提升了网络的特征表达能力，但并非所有输入都需要同样深的网络结构来处理，使用特别深的网络结构去处理简单数据时（如背景十分简单的图像）会造成计算资

收稿日期：2022-04-12；**修回日期**：2022-05-20；**录用日期**：2022-05-23
基金项目：国家自然科学基金（61973311，61825305）
通信作者：*lijian@nudt.edu.cn；**xinxu@nudt.edu.cn

源的浪费。当前高效模型的解决方案有两种，一类是以Mobilenet[23]为代表的轻量化网络设计方法[23-26]，另外一类是以网络剪枝(network pruning)[27]和知识蒸馏(knowledge distillition)[28-30]等为代表的大模型裁剪和小模型训练方法。但无论是参数量巨大的深度模型，还是轻量化算法处理之后得到的小模型，它们的性能都受制于静态的推理方式，无法根据样本的复杂度来自动调整推理深度。具体来说，轻量化模型虽然提高了模型的计算效率，但是容量有限，当应对更加复杂的输入时，无法动态地扩增网络深度以提高网络的表达能力；大的深度模型虽然拥有了较为强大的特征表达能力，但是无论输入简单或复杂，针对所有输入都采用了同一个推理图(即推理深度)，这大大降低了模型的运算效率，不利于模型部署到对实时性和经济性要求高的平台中，比如边缘计算平台[31]。

不同于传统的静态神经网络，深度自适应网络致力于探索网络深度的自适应调整机制，即根据输入或者任务的复杂度动态调节网络的推理深度。对于简单的样本使用浅层子模型，而对于复杂的样本使用更深的模型以满足复杂特征的提取需求。人脑的视觉计算系统中也存在这种动态调节机制，人脑处理简单输入(低频的信息多，高频细节少)的速度比复杂输入(高频细节多)要快许多[32]。因此相比静态深度神经网络，深度自适应网络这种高效而具备可扩充性的计算模型更类似人类的感知系统，其不仅同时兼顾推理效率和模型的表达能力，也更具有可解释性。近年来许多关于深度自适应网络的工作从不同角度对深度自适应的机理进行了研究。本文通过对已有工作的抽象归纳，将从深度自适应网络的结构设计、样本复杂度估计方法、深度自适应神经网络的训练方法3个方面对现有工作进行全面的归纳总结，并进行相应的实验对比，以期为后面的研究提供更好的研究框架。

1) 深度自适应网络的结构设计(network structure)。深度自适应网络需要根据输入动态地调节推理深度，因此该网络的结构需要满足在不同深度进行输出的要求同时要保证浅层网络具备充足的感受野。目前主要有多退出结构(multi-exit)和基于残差学习的跳层连接结构(skip-connection)。

2) 样本复杂度估计方法的研究(input-complexity estimation--depth-adaptive policy)。深度自适应机制的关键是模型能够对样本复杂度进行衡量，进而获取自适应深度调整的策略依据。现有方法主要从置信度估计、数据驱动策略和互信息估计三个方面对样本复杂度估计进行研究。

3) 深度自适应神经网络的训练方法(training method)。深度自适应网络的训练方法研究主要围绕自适应网络训练所遇到的问题展开。一类是针对多退出结构的训练方法研究，主要探索如何有效地训练不同深度的输出模块，防止各个输出之间的相互干扰。另外一类研究则围绕基于跳层连接结构的自适应网络展开，主要关注如何训练要求稀疏化输出的跳层策略，正确地反映样本复杂度。

表1概括了深度自适应神经网络的主要研究内容。

表1 深度自适应网络相关研究概述

Table 1 Overview about the depth-adaptive neural networks

Method	Network structure	Depth-adaptive policy (input-complexity estimation)	Training method
Multi-exit	Independent output branches[33-36]; Additive/geometric ensemble[37-38]; Multi-scale feature fusion[39]; Multi-scale receptive field[34,38]	Confidence-based early exiting[33-37,40-42]; Mutual information estimation early exiting[43-44]; Learning policy networks for early exiting[45-47]	Weighted gradient descent[28,33,48]; Knowledge distillation[49-51]; Gradient adjustment[38,52]
Skip-style	Centralized gate module[53]; Distributed gate module[54-56]; Randomly block dropout[57-58]	Skipping non-linear blocks[53-58]	Sparse regularization[56]; Reinforcement-learning based[53]

2 深度自适应网络的结构设计

不同于静态深度的神经网络，如图1所示，深度自适应神经网络需要根据输入数据的复杂度动态调整深度。

为了实现动态调整网络深度的能力，已有工作[33-39]专门设计了不同于静态网络的网络结构，这些现有网络结构包括多退出分支结构和跳层自适应结构两类，并对现有的深度自适应结构进行总结和分析。多退出分支网络在不同深度处设置输出分支，在推理时根据预先设置的退出策略和输入样本，自动调节推理深度；跳层连接网络通过设置门控单元(Gate Module)，根据输入样本选择性地跳过非线性层，从而降低推理深度。两类自适应深度网络的架构如图2所示。

2.1 多退出分支网络

多退出分支网络(multi-exit network)是一种典型的动态深度推理网络结构，主要特点是在网络的不同深度处添加输出模块，比如分类分支或者目标检测分支，如图2(a)所示。定义网络输入为一个张量$\boldsymbol{X}$，

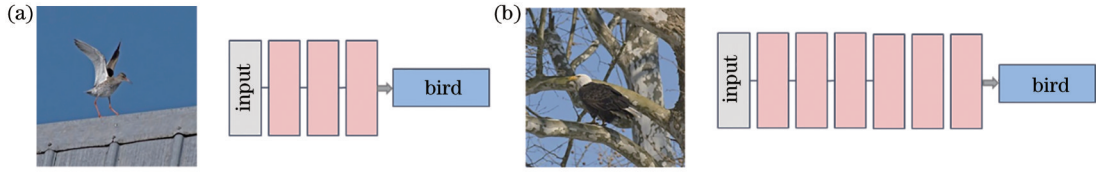


图 1 根据输入的复杂程度自动调节推理深度的深度自适应神经网络。(a)处理简单输入的网络结构;(b)处理复杂输入的网络结构

Fig. 1 Depth adaptive neural networks for automatically adjusting inference depth based on the input complexity. (a) Network structure for processing simple input; (b) network structure for processing complex input

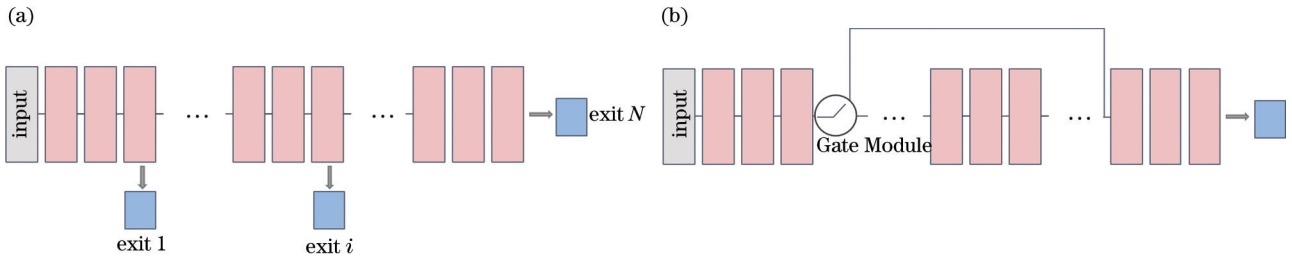


图 2 典型的深度自适应神经网络的结构。(a)多退出分支网络;(b)跳层连接网络

Fig. 2 Typical structures of depth-adaptive neural networks. (a) Multi-exit neural network; (b) skip-connection network

$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$,其中 C 代表输入的特征维度,(H, W) 分别表示输入的长和宽,$\boldsymbol{X}$ 可以是可见光图像、红外图像或者其他常见的传感器数据。由 N 个模块 $\{\boldsymbol{f}_1,\cdots,\boldsymbol{f}_N\}$ 级联组成一个深度为 N 的神经网络,其中 $\boldsymbol{f}_i$ 表示第 i 个非线性模块,$\boldsymbol{f}$ 可以由卷积[59]、注意力模块或者全连接层构成。为了方便表示,使用“∘”表示不同非线性模块间的连接关系,即 $f_N\circ f_{N-1}(\boldsymbol{X})=f_N\left[f_{N-1}(\boldsymbol{X})\right]$,网络最终的输出为 $\boldsymbol{Y}$,则一个 N 层级联神经网络可以表示为

$$\boldsymbol{Y}=\boldsymbol{f}_N\circ\boldsymbol{f}_{N-1}\circ\cdots\circ f_1(\boldsymbol{X})。\tag{1}$$

定义 multi-exit 网络在每一个深度处设置的输出模块为 $\boldsymbol{O}$,则 multi-exit 网络可以表示为

$$\begin{cases}\boldsymbol{Y}_1=\boldsymbol{O}_1\circ f_1(\boldsymbol{X})\\ \vdots\\ \boldsymbol{Y}_N=\boldsymbol{O}_N\circ\boldsymbol{f}_N\circ\boldsymbol{f}_{N-1}\circ\cdots\circ f_1(\boldsymbol{X})\end{cases}。\tag{2}$$

为了设计性能良好的多退出神经网络结构,已有工作主要研究模块之间的信息融合和网络感受野(receptive field)设计。不同深度模块之间的信息融合能有效地提升模型的表达性能,而感受野则与网络捕捉的大尺度模式和抽象特征紧密相关[60-61]。

2.1.1 模块之间的信息融合

构造多退出输出网络最直接的方式是,直接附加输出模块到网络不同深度处(independent output branches)[33,35-36],这些模块间除了共享部分网络参数外,输出相互独立,没有信息交流。文献[40,45,62]分别在语义分割和分类、超分辨率恢复任务中采用了这种结构。融合不同输出分支之间的信息可以有效地提高多退出网络的性能[37-39],图 3 概括表达了这种方案。不同输出模块信息交流方案主要包含加性集成(additive ensemble)[38]、几何集成(geometric ensemble)[37]和多尺度特征融合(multi-scale feature fusion)[39]三种

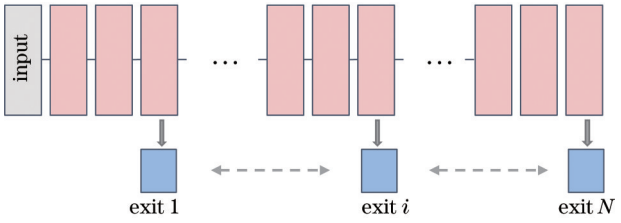


图 3 输出模块的信息交流方案

Fig. 3 Information exchange scheme of output module

思路。以分类任务为例,定义各个分支的预测结果为 $\{\boldsymbol{P}_i|i\in[1,N]\}$,$\boldsymbol{P}_i\in\mathbf{R}^C$,$C$ 为数据集的类别个数。

基于加性集成的思路是把每一层预测的概率分布 $\{\boldsymbol{P}_i|i\in[1,N]\}$ 作为特征,通过学习特征变换 $\boldsymbol{T}$ 或者直接用恒等变换[38],对不同层的预测结果进行加性组合,得到融合后的预测结果。特征变换 $\boldsymbol{T}$ 可以是浅层神经网络,也可以是用于对各个分支结果进行加权的权重:

$$\hat{\boldsymbol{P}}_m=\sum_{i<m}\boldsymbol{T}_i\circ\boldsymbol{P}_i。\tag{3}$$

基于几何集成的融合则是基于加权联合概率的形式。相比于加性集成的思路,基于连乘形式的几何集成方法对各个输出分支的预测一致性要更加严格,即如果网络需要以较高概率在某个输出层预测输入的类别,要求在此层之前的所有预测结果具有较高的一致性。该类思路的计算范式为

$$\hat{\boldsymbol{P}}_m=\frac{1}{Z}\prod_{j=1}^{m}(\boldsymbol{P}_j)^{\alpha_j}=\frac{1}{Z}\exp\left(\sum_{j=1}^{m}\alpha_j\ln\boldsymbol{P}_j\right),m\in[1,N],\tag{4}$$

式中:Z 为规范化因子,使得分布 $\{\hat{\boldsymbol{P}}_m|m=1,\cdots,N\}$ 能够成为概率分布;α 为可学习的加权网络参数。文献[37]基于几何集成的思路,使用各个预测结果的联合概率来提升单个输出模块的预测结果。

多尺度特征融合是对网络不同深度的特征进行融合的方案。定义网络在各个深度提取的特征图为$\{\boldsymbol{F}_i|i\in[1,N]\}$，为了综合利用各层提取的特征，现有工作主要通过特征合并(Concatenation)或者加权相加[39]的方式来实现特征融合，然后通过特征变换$\boldsymbol{T}$对融合的特征进行转换，得到融合提升的结果。特征变换$\boldsymbol{T}$通常由恒等映射和全连接层表示：

$$\hat{\boldsymbol{F}}_m=\boldsymbol{T}\circ\text{Concatenation}(\{\boldsymbol{F}_i|i=1,\cdots,m\}),m\in[1,N], \tag{5}$$

$$\hat{\boldsymbol{F}}_m=\boldsymbol{T}\circ\sum_{i=1}^{m}\alpha_i\boldsymbol{F}_i, \tag{6}$$

式中：α_i为加权系数。由于当前研究缺乏统一的对照比较(统一的基础网络结构和数据集)，为了更加直观地体现以上方法的有效性，基于MSDNet[34]中设计的多退出网络结构，在CIFAR100数据集上分别训练了基于三类信息融合方式的多退出网络模型。训练的迭代次数为100，其余学习策略采用文献[34]中的设置。共设置4个中间输出模块，MSDNet作为baseline模型进行对照，输出模块间相互独立，没有信息融合。加性集成采用文献[38]中的方案、几何集成采用文献[37]中的方案、多尺度特征融合则采用文献[39]中的加权相加方案。表2中的结果显示相比于baseline模型，输出模块间的信息融合方案能够有效地提升部分或者整体的性能，其中Exit-n表示第n个分支输出。

表2 各种特征融合方法在CIFAR100数据集上的效果对比

Table 2 Performance comparison of different information fusion approaches on CIFAR100 dataset

Method	Exit-1	Exit-2	Exit-3	Exit-4
Baseline	66.77	70.31	71.93	73.0
Additive-ensemble	66.04	70.70	72.49	73.23
Geometric-ensemble	63.91	70.35	72.67	73.01
Multi-scale feature fusion	66.60	70.53	72.75	73.05

2.1.2 浅层网络多尺度感受野设计

网络感受野表示的是网络计算操作所包含的空间范围。卷积神经网络的感受野与卷积层的深度密切相关，深层网络拥有相对于浅层网络更大的感受野。多尺度感受野对于模型提取多尺度特征至关重要[61]。定义每一层的卷积核的大小为k，则网络感受野S随着深度d的变化公式为

$$S_d=S_{d-1}+(k-1)\times\prod_{j'=1}^{d-1}s_{j'}, \tag{7}$$

式中：$s_{j'}$表示第j'层的卷积核移动步长(stride)或者是下采样倍率。神经网络通过级联结构获得了不同尺度的感受野[60]。在其他任务中[63-64]，相关实验结果表明增大卷积感受野能提升算法的检测性能。但深度自适应网络的浅层模块受限于级联深度，其感受野相比于深层模块要小。

在不增加网络深度的情况下，提升浅层网络的感受野，以捕捉更大尺度的特征。Huang等[34]设计了一种基于多尺度下采样的网络结构MSDNet，如图4所示，该网络依靠多尺度下采样在浅层获得了不同尺度的感受野，通过融合不同尺度的感知结果，提升了浅层网络的计算性能。图5为MSDNet与Ensemble-ResNets的性能对比，相比Ensemble-ResNets未在浅层网络配置多尺度感受野的网络，MSDNet以更少计算量取得了在ImageNet数据集上更高的分类精度。对于基于Transformer构造的自适应网络[44]，由于自注意力机制的全局计算属性，其感受野与网络深度无关，但是也大大增加了计算的复杂度。

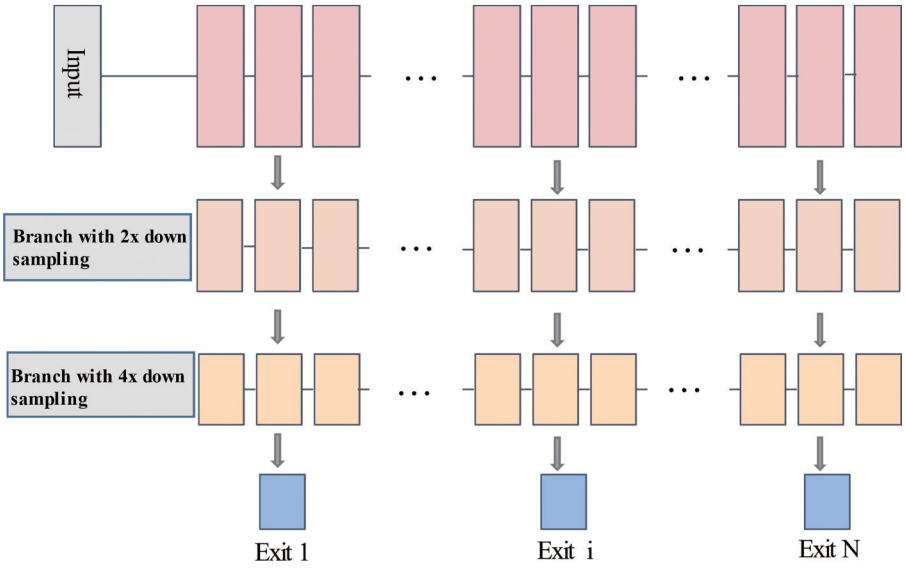


图4 基于多尺度下采样的网络结构MSDNet[34]

Fig. 4 Network structure MSDNet based on multi-scale down sampling[34]

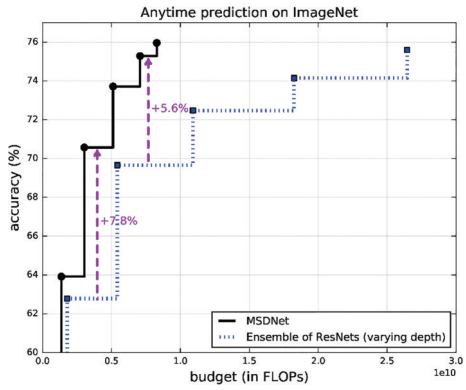


图 5 MSDNet 和 Ensemble-ResNets 在 ImageNet 数据集上的分类精度[34]

Fig. 5 Classification accuracy of MSDNet and Ensemble-ResNets on ImageNet dataset[34]

2.2 跳层自适应结构

当残差单元输出忽略不计时，ResNet[5]的 skip-connection 结构可以由恒等映射替换，等价于降低了网络的深度。受这种特性启发，设计了基于跳层结构的深度自适应网络，该类自适应深度神经网络依靠 Gate Module 调节推理深度，Gate Module 的基本结构如图 6 所示。Gate Module 根据输入的特征图来选择执行恒等映射或者非线性层，跳过非线性计算模块可以等效地降低网络的推理深度。

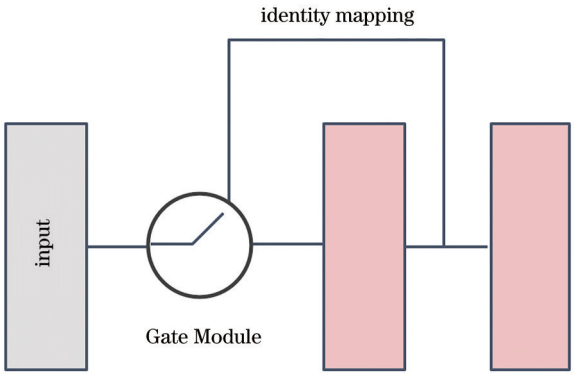


图 6 Gate Module 的基本结构

Fig. 6 Basic structure of Gate Module

定义第 l 层和第 $l+1$ 层的特征变换映射为 $T_{l\to l+1}: F_l(\boldsymbol{x})\to F_{l+1}(\boldsymbol{x})$，则跳层自适应结构可以表述为

$$T_{l\to l+1}=\begin{cases} I(\cdot), & G_l(\boldsymbol{x})<\tau \\ f_{l\to l+1}(\cdot), & G_l(\boldsymbol{x})\geqslant\tau \end{cases}, \tag{8}$$

式中：$G(\boldsymbol{x})$表示门控(Gate)单元，常由浅层的多层感知机(MLP)和 Sigmoid 激活函数或者 Softmax 激活函数组成。Gate 单元根据输入的特征图 $\boldsymbol{x}$ 来输出决策值。当决策值小于某一个阈值 τ 时，该映射 $T_{l\to l+1}$ 成为恒等映射 $I(\cdot)$，等价于降低了网络的推理深度，否则执行非线性层 $f_{l\to l+1}(\cdot)$。根据门控单元的构造形式，将目前的跳层结构分为三类：集中式跳层选择机制(centralized gate module)[53]、分布式跳层选择机制(distributed gate module)[54-56]、基于 Dropout 机制的随机跳层(randomly block Dropout)[57-58]。

1) 集中式跳层选择机制。BlockDrop[53]中设计了一个独立于任务网络的策略网络 Φ(policy network)，$\boldsymbol{\pi}=\Phi(\boldsymbol{x})$，一次性输出任务网络所有 N_G 个 Gate 单元 $\{G_1,\cdots,G_{N_G}\}$的决策 $\boldsymbol{\pi}\in\mathbf{R}^N$。

2) 分布式跳层选择机制。文献[54-56]中使用分布式结构，将各个 Gate 单元 $G_l=\Phi_l(\cdot)$分布在任务网络的各个跳层连接处，分别进行计算，如图 2(b)所示。

3) 基于 Dropout 机制的随机跳层。Dropout 常被用作网络训练时的正则化机制，以防止网络过拟合。文献[57-58]中利用这种思想，在训练时以某一个概率 p 随机去除部分网络层，使得网络经过训练后获得了在不同深度结构下的推理能力；在部署时也以概率 p 进行网络层去除，实现动态随机深度推理，其 Gate 单元等价于一个服从伯努利分布的随机采样模块。

由于 Gate 单元的输出是稀疏的，因此文献[53-56]中采用的 skip 结构对 Gate 单元的训练要求比较高，且该类自适应策略仅适用于具有 skip-connection 结构的网络，因此相关的研究工作相比于多退出神经网络少。使用 Dropout 机制的网络依靠概率随机决定推理深度，虽然可以通过调控 Dropout 概率 p 获取不同的期望推理深度，但是推理过程仍然过于随机，无法有效把控。

3 样本复杂度估计方法的研究

如图 7 所示，神经网络的输入数据常常拥有不同的复杂度(如复杂图片常常拥有丰富的纹理信息)，深

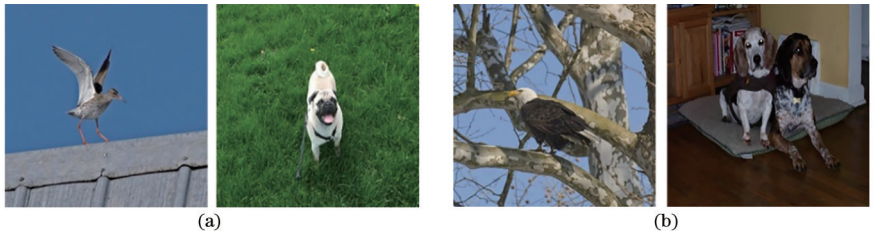


图 7 具有不同复杂度的样本。(a)纹理和背景相对简单的样本；(b)复杂样本

Fig.7 Samples with different complexity. (a) Samples with relatively simple texture and background; (b) complex samples

度自适应机制的关键是模型能够对(输入)样本的复杂度进行衡量。样本复杂与否是相对于模型而言的,当样本复杂度与模型深度不相匹配时,模型将不能有充足的非线性特征表达能力去对样本特征进行编码,进而影响预测性能。由N个非线性模块$\{\boldsymbol{f}_1,\cdots,\boldsymbol{f}_N\}$级联组成一个深度为$N$的神经网络,样本复杂度的定义:使得网络对数据$\boldsymbol{x}$的预测误差$L$小于一个预定的范围$\varepsilon$的最小网络深度$d$。在分类任务中,$L$常常为网络预测结果$\boldsymbol{f}_d\circ\cdots\circ f_1(\boldsymbol{x})$与真值$\boldsymbol{y}$的交叉熵(cross-entropy),样本复杂度估计形式化描述为

$$D=\arg\min_{d}\{L\left[\boldsymbol{f}_d\circ\cdots\circ f_1(\boldsymbol{x}),\boldsymbol{y}\right]<\varepsilon\}。\tag{9}$$

但实际上无法在推理过程中获取真值,因此需要一个不依赖真值的样本复杂度衡量机制M,其能够仅根据输入便能预测任务损失或者能够衡量任务损失的相关变量,

$$D=\arg\min_{d}\{M\left[\boldsymbol{f}_d\circ\cdots\circ f_1(\boldsymbol{x})\right]<\varepsilon\}。\tag{10}$$

样本复杂度估计是深度网络执行自适应推理所需的基本功能,是模型自适应调整推理深度的依据。传统计算机视觉中的图像复杂度主要是通过统计图像的底层特征信息来进行度量的,如边缘、独立成分或者压缩后的像素误差(Kolmogorov complexity)[65-66],这种方式独立于深度网络本身,无法有效地融合进网络自适应推理过程中。为了解决以上问题,现有方法主要从置信度估计、数据驱动策略和互信息估计三个方面结合网络的输出对样本复杂度估计进行研究。将现有方法归纳为三类:基于预测置信度的策略(confidence-based early exiting policy)、基于数据驱动学习的策略(learning policy networks early exiting policy)和基于互信息估计的策略(mutual information estimation early exiting policy)。

3.1 基于预测置信度的样本复杂度估计策略

基于预测置信度的样本复杂度估计策略主要应用于深度自适应的分类或者分割网络,以及其他以概率形式进行输出的网络。定义网络的预测输出为$\{\boldsymbol{p}_i|i\in[1,C]\}$,$C$为概率分布的维度,在分类任务中对应类别个数。根据香农信息熵,网络关于样本$\boldsymbol{x}$的预测不确定性可以由熵$H(\boldsymbol{x})$来表示:

$$H(\boldsymbol{x})=-\sum_{c=1}^{C}p_c(\boldsymbol{x})\log_{\mathrm{e}}p_c(\boldsymbol{x})。\tag{11}$$

在实际应用中,文献[33-34,36-37,40-42]中的工作对式(11)做了部分修改,使用网络输出的最大置信度$p^{\max}$来估计样本复杂度$\tilde{H}(\boldsymbol{x})$,即

$$\tilde{H}(\boldsymbol{x})=-p^{\max}(\boldsymbol{x})\log_{\mathrm{e}}\left[p^{\max}(\boldsymbol{x})\right]\leqslant H(\boldsymbol{x})=-\sum_{c=1}^{C}p_c(\boldsymbol{x})\log_{\mathrm{e}}p_c(\boldsymbol{x})。\tag{12}$$

当最大置信度$p^{\max}$越低时,网络关于样本类别的不确定性越高,代表样本越复杂。这种方式需要提前设置一系列的阈值,并面临over-confidence[67]的问题,即样本有很高的类别预测概率却被错误分类。此外,对于检测或者回归等其他任务来说,基于置信度的方法适用性也不强。除了基于信息熵的衡量标准外,文献[35]中以网络在不同深度预测结果的变化量为依据,判断是否早退,当网络连续t个输出分支的结果不发生较大变化时就退出,不再执行更深的网络。

3.2 基于数据驱动学习的策略

如图6所示,定义某个跳跃连接处的输入特征为$\boldsymbol{x}$,文献[54-56]中的方法在网络的各个跳跃连接(skip connection)处设置Gate单元$G_l=\Phi_l(\cdot)$,来完成网络深度的自适应选择。Gate单元G_l常常由浅层的MLP构成,最后用非线性函数Softmax或者Sigmoid对结构进行归一化,使预测结果成为一个分布:

$$\begin{cases}G_l:\Phi_l(\boldsymbol{x})=\mathrm{Sigmoid}\left\{\mathrm{MLP}\left[\mathrm{GAP}(\boldsymbol{x})\right]\right\}\\G_l:\Phi_l(\boldsymbol{x})=\mathrm{Softmax}\left\{\mathrm{MLP}\left[\mathrm{GAP}(\boldsymbol{x})\right]\right\}\end{cases}。\tag{13}$$

Gate单元的作用是预测网络在跳层连接处选择恒等映射或者非线性映射的概率。多退出网络也可以通过训练策略网络来实现样本复杂度的估计[45]。

3.3 基于互信息估计的策略

样本复杂度估计本身是一个无监督问题,因此可以通过重构任务估计输入和隐变量的互信息,这里的隐变量为样本复杂度分布。Faster-Transformer[44]通过估计自然语言中关键字与样本训练损失的相互关系,实现了通过关键字来判定样本复杂度的目的。文献[43]将策略网络在每一层退出的概率建模成随输入分布相关的隐变量,使用变分信息瓶颈(variational information bottleneck)的思路来优化策略网络。定义每一层执行退出的概率为π,则深度网络执行到第l层退出的概率φ_l为一个多项式分布:

$$\begin{cases}\varphi_l=\pi_l\prod_{d=1}^{l-1}(1-\pi_d), & l<N'\\\varphi_l=\prod_{d=1}^{N'-1}(1-\pi_d), & l=N'\end{cases},\tag{14}$$

式中:N'是网络总的深度。如上所述,一开始并不知道样本复杂度的实际分布(早退的概率),因此文献[43]中使用变分估计模型$q(\boldsymbol{x})$拟合这种未知分布,同时以训练集中获得的最优退出分布$p(\boldsymbol{x})$(获得准确分类的最浅深度)作为目标分布,通过最大化$q(\boldsymbol{x})$与$p(\boldsymbol{x})$之间的互信息来获得复杂度的变分估计模型$q(\boldsymbol{x})$,采用Kullback-Leibler(KL)散度:

$$\mathrm{KL}(p\|q)=\int p(\boldsymbol{x})\log_{\mathrm{e}}\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}。\tag{15}$$

已有方法主要从两方面来评价样本复杂度估计的精度:在指定计算量的情况下,模型的平均精度;在指定数据集上的平均计算量与平均精度。目前仍亟需进一步探索深度自适应神经网络的样本复杂度估计算

法，并进一步统一样本复杂度估计的评价框架。

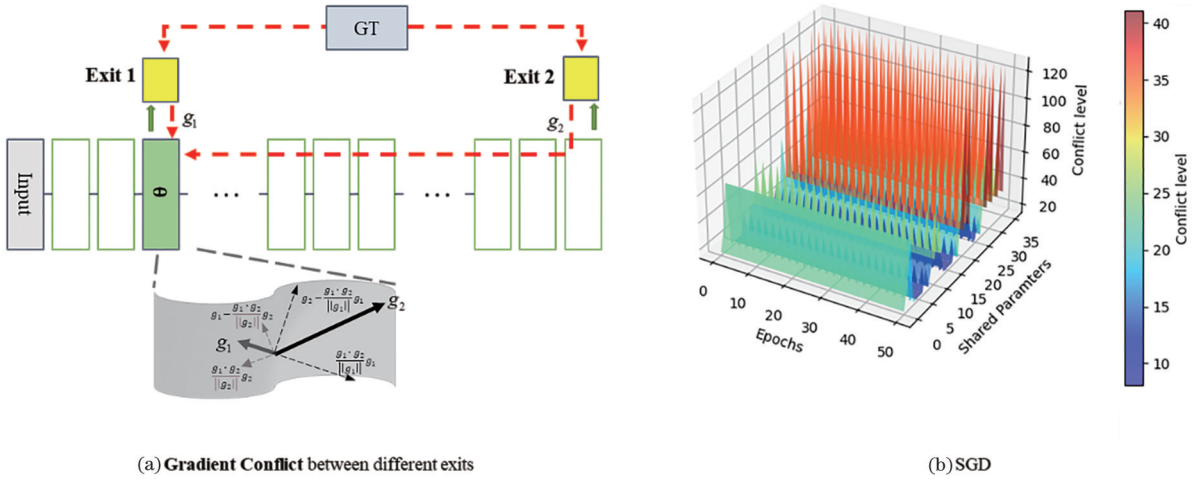
4 深度自适应神经网络的训练方法

深度自适应神经网络面临不同深度输出之间冲突、Gate单元稀疏化训练的问题，因此将从多退出网络的训练和跳层自适应网络的Gate单元训练两方面进行归纳总结。

4.1 多退出网络的训练

多退出自适应深度网络（multi-exit networks）的训练主要会遇到“cross-talk”的问题，即不同输出层之间在训练过程中相互干扰，进而影响网络整体性能。该问题宏观上是不同输出分支在学习目标上产生了冲突，即虽然各层的训练真值一样，但是深层网络对于数据的拟合迭代方向与浅层网络的方向并不一致，可以理解为对知识的表示不一致。这种冲突微观上体现在网络共享参数收到了相互冲突的梯度，即共享参数上来自不同任务梯度在更新方向上相反。在文献[68]中，梯度冲突被定义为两个梯度的余弦夹角小于0。不失一般性，分析具有两个输出的网络，如图8(a)所示。定义输出1和输出2关于共享参数的梯度为$\boldsymbol{g}_1=\nabla L_1$和$\boldsymbol{g}_2=\nabla L_2$，当二者冲突时有$\boldsymbol{g}_1\cdot\boldsymbol{g}_2<0$（内积）。定义学习率为$\eta>0$，只用梯度$\boldsymbol{g}_1$对网络进行更新时，可以通过一阶Taylor展开近似估计ΔL_1，表达式为

$$\Delta L_1=L_1\left[f(\boldsymbol{x}_1,\boldsymbol{\theta}-\eta\boldsymbol{g}_1),\boldsymbol{y}_1\right]-L_1\left[f(\boldsymbol{x}_1,\boldsymbol{\theta}),\boldsymbol{y}_1\right]=-\eta\boldsymbol{g}_1\cdot\boldsymbol{g}_1+o(\eta^2)\quad。(16)$$



(a) **Gradient Conflict** between different exits　　(b) SGD

图8 不同输出关于共享参数$\boldsymbol{\theta}$的梯度相互冲突。(a) 两个冲突梯度的余弦夹角小于零；(b) 训练过程中统计的不同共享参数上的冲突程度

Fig. 8 Shared parameters $\boldsymbol{\theta}$ receives conflict gradients from different exits. (a) Conflicted gradients have negative cosine similarity value; (b) level of gradient conflict in the training stage

由于$\boldsymbol{g}_1\cdot\boldsymbol{g}_1\geqslant 0$，所以输出1的损失呈现下降趋势。如果同时使用输出2的梯度对权重进行更新，则

$$\Delta L_1=L_1\left\{f\left[\boldsymbol{x}_1,\boldsymbol{\theta}-\eta(\boldsymbol{g}_1+\boldsymbol{g}_2)\right],\boldsymbol{y}_1\right\}-L_1\left[f(\boldsymbol{x}_1,\boldsymbol{\theta}),\boldsymbol{y}_1\right]=-\eta(\boldsymbol{g}_1\cdot\boldsymbol{g}_1+\boldsymbol{g}_1\cdot\boldsymbol{g}_2)+o(\eta^2)。(17)$$

由于$\boldsymbol{g}_1\cdot\boldsymbol{g}_2<0$，输出1的损失变化$\Delta L_1$相比于只用$\boldsymbol{g}_1$，受到了输出2的负面影响。为了验证梯度冲突的存在，基于MSDNet在CIFAR100上训练了具有2个退出模块的多退出神经网络。图8(b)为各个共享参数上的冲突值（余弦相似度的负相关函数，当余弦相似度为1时，冲突值为0）随训练迭代次数的变化情况，可以很直观地看到在部分网络共享参数上，不同输出之间的相互干扰一直存在。

为了解决不同输出层相互干扰的问题，一种缓解方法是设计特殊的网络结构：如图9所示，文献[34, 45]通过在网络不同层之间添加密集连接（dense connection），使得深层的输出模块也能够直接接触到浅层的特征提取模块，降低浅层输出对深层输出模块

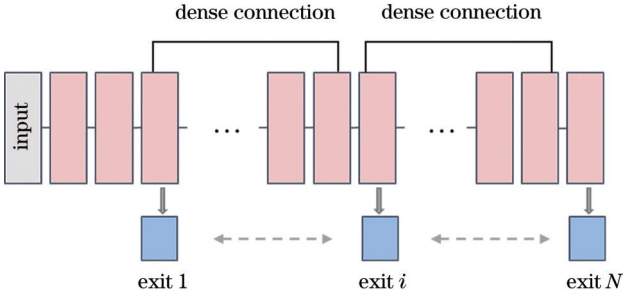


图9 添加密集连接的网络结构

Fig. 9 Network structure with dense connection

的影响，这种密集连接提升了multi-exit网络的综合性能。除此之外，已有的专门研究消除输出层相互干扰的方法可以归纳为三类：加权梯度下降（weighted gradient descent）、知识蒸馏（knowledge distillation）和梯度调整（gradient adjustment）。

1）加权梯度下降。主要对各个输出模块的损失函数进行加权[28,33,48]，通过权重系数α调整不同输出模

块在反向传播过程中的梯度幅值，以此调整各个模块的损失函数对模型更新的贡献，平衡不同模块之间的训练，缓解冲突问题，表达式为

$$\nabla L=\sum_{i=1}^{N}\alpha_i\nabla L_i。\tag{18}$$

2）知识蒸馏。根据网络过参数化的相关研究[69-70]，过参数化（比如深度）会对网络的学习过程起到隐式正则化的效果，使得模型的 Loss Landscape 更加平坦，即深层网络拥有更好的学习能力和泛化能力。因此以知识蒸馏为代表的小模型训练方法[28-29]用大网络的预测结果（soft-target）训练小网络，提升小网络的泛化能力。这个过程可以看作是知识从大网络传递到小网络的过程，这种思路被用到了深度自适应网络的训练过程中。文献[49-51]使用深层网络的输出来训练网络浅层的输出，保证其在拟合目标上的一致性：

$$L_d=\mathrm{KL}\left[f_N(\boldsymbol{x}),f_d(\boldsymbol{x})|d<N\right],\tag{19}$$

式中：$f_N(x)$表示网络最后一层的预测结果。对于回归问题，可以使用 Hint-Loss[29]替代 KL 散度，实现知识传递。Meta-Distiller[51]中的工作则进一步对知识蒸馏过程进行研究，专门设计了一个对蒸馏出来的知识进行调整的网络，使之能够更好地引导浅层网络的训练。表 3 对比了不同知识蒸馏方法 DBT[50]和 H-DBT[49]的效果，相比于未使用知识蒸馏的方法 MSDNet[34]和 IMPR[38]，知识蒸馏有效地提升了网络的训练效果。

表 3 基于知识蒸馏训练的多退出网络性能对比

Table 3 Performance comparison of multi-exit networks trained by knowledge distillation

Method	Exit-1	Exit-2	Exit-3	Exit-4	Exit-5
MSDNet[34]	79.25	86.46	89.15	89.83	90.75
IMPR[38]	80.15	87.89	90.52	91.33	91.74
DBT[50]	80.80	86.92	88.82	89.15	89.73
H-DBT[49]	83.06	87.12	90.85	91.9	92.04

3）梯度调整。另外一类缓解“cross-talk”的方法是对各层的梯度进行调整，文献[38]中提出了 Gradient Equilibrium（GE）的解决方案，针对某个共享参数 $\boldsymbol{\theta}$，假设其被 k 个输出分支共享，则 GE 方案对所有关于 $\boldsymbol{\theta}$ 的梯度 $\{\nabla_\theta L_d|d\in[1,k],\ k\leqslant N\}$ 进行平均，以控制参数 $\boldsymbol{\theta}$ 更新量的散布方差，最终 $\boldsymbol{\theta}$ 的更新梯度为

$$\boldsymbol{g}_\theta=\frac{1}{k}\sum_{d=1}^{k}\nabla_\theta L_d。\tag{20}$$

式（20）对深层传递到浅层的原始梯度进行了拉伸。这种简单的梯度拉伸思路对某个节点上接收到的所有任务梯度进行平均，取得了比原始梯度更好的性能。与加权梯度下降方法不同的是，这种调节过程更精细地考虑到了参数被共享的次数。

除了平衡梯度的幅值以外，梯度之间的方向冲突也影响了多退出网络的训练效果。为了平衡各个任务间的冲突问题，文献[52]中使用梯度正交投影的方式（PCgrad）来消除梯度之间的冲突成分。类似的梯度调整工作在多任务学习领域也有体现，文献[68，71-72]中的工作从正交投影和 Pareto 优化的角度对共享参数上的梯度进行调整，尽可能减少不同输出间的相互干扰。表 4 为使用梯度调整策略后模型性能的变化对比，可以明显看到无论是 GE 还是 PCgrad 都能有效地提升模型的综合性能。

表 4 使用不同梯度调整方法后多退出网络在 ImageNet 数据集上的效果对比

Table 4 Performance comparison of multi-exit networks after using different gradient adjustment approaches on ImageNet dataset

Method	Exit-1	Exit-2	Exit-3	Exit-4	Exit-5
MSDNet[34]	58.48	65.96	68.66	69.48	71.03
IMPR-GE[38]	57.75	65.54	69.24	70.27	71.89
PCgrad+GE[52]	57.62	64.87	68.93	71.05	72.45

值得一提的是，加权梯度下降、知识蒸馏和梯度调整是从不同的角度来解决多退出神经网络的训练问题的，相互之间可以互补，通过综合使用可以进一步提升网络的性能[52]。

4.2 跳层自适应网络的 Gate 单元训练

跳层自适应网络通过 Gate 单元来选择执行的深度，为了让 Gate 单元能根据样本的难易程度自适应地调整输出，现有训练方法都将计算量作为损失函数的一部分对 Gate 单元进行监督。由于 Gate 单元要做出离散决策，即 skip or not，所以要解决稀疏训练的问题。目前有两类方法来确保 Gate 单元输出的稀疏性。

1）基于 Gumble-Softmax 的训练策略[56]。Gumble-Softmax 训练策略对数据添加 Gumble 噪声后然后使用 Softmax 得到概率分布，Gumble 噪声的添加促进了概率分布的二值化。Gumble 分布的概率密度函数为

$$\sigma\sim-\log\left[-\log(u)\right],\quad u\sim U(0,1),\tag{21}$$

式中：$U(0,1)$为标准正态分布。定义 Gate 单元的输出为$\{\pi_1,\pi_2\}$，1 和 2 分布对应执行恒等映射和非线性映射的未归一化概率值，则加入 Gumble 噪声之后的归一化预测概率值$\{p_1,p_2\}$的计算方式为

$$p_i=\frac{\mathrm{e}^{\frac{\pi_i+\sigma_i}{\tau}}}{\mathrm{e}^{\frac{\pi_1+\sigma_1}{\tau}}+\mathrm{e}^{\frac{\pi_2+\sigma_2}{\tau}}},\ i=1,2。\tag{22}$$

类似的稀疏化训练函数还有 Improved semi-hash[73]。

2）基于强化学习的训练。文献[45-47，53]中单独设计了一个独立于任务网络的策略网络 Q，并采用强化学习的方法，如 Q-learning[45-47]、curriculum learning[53]训练策略网络。通过合理地设置 Reward 函数，强化学习策略能够启发式地引导网络在数据集中

进行策略探索，使得模型获得判别样本复杂度的能力。

5 未来的研究课题

深度自适应神经网络的研究虽然取得了一定的发展，但是其在网络结构、训练方法以及自适应策略上仍旧面临挑战：当前的自适应网络结构面向的任务单一，主要以分类任务为主，需要进一步探索在其他任务上的自适应网络结构；自适应网络的训练理论需要进一步的优化，尤其要解决多退出网络面临的梯度冲突问题和跳层连接网络中 Gate 单元稀疏化训练问题；网络推理深度的自适应调整策略与输入复杂度的估计算法紧密相关，而此过程是一个无监督学习的过程，需要探索更加高效且扩展性强的样本复杂度估计算法。

将未来值得研究的课题详细阐述如下。

1）探索更加通用的深度自适应网络结构。当前的深度自适应网络结构主要是为分类任务设计的，需要探索能同时适用于目标检测任务、语义分割等其他任务的自适应结构。此外，不同图片的不同位置的复杂度是不一样的，而目前的深度自适应是以整张图为单位进行自适应推理的。因此，未来需要同时结合深度自适应与空间自适应机制，即网络应该有能力针对图像的不同区域使用不同深度的网络。基于 Transformer 的研究近年来在视觉领域发展迅速，如何设计更有效的基于 Transformer 的深度自适应网络也是一个值得研究的课题，因为 Transformer 在数据处理上比 CNN 更加灵活（例如 Token 的数目是可以变化的）。

2）探索有效的自适应网络训练方法。在训练多输出自适应神经网络时，为了解决不同输出之间的冲突，现有解决方案是基于梯度调整。但单纯基于梯度调整的方法忽略了共享参数之间的差异性，有的共享参数可能有明显的偏好性，即该参数相比于输出 n，可能对于另外一个输出 m 更为重要。如何在梯度调整的时候考虑到这种差异性是一个值得研究的课题。此外，如何在训练的过程中就引入样本复杂度估计，也是需要研究的问题，一方面如果让过浅的子网络接触过于复杂的训练数据，其学习过程势必受到负面影响。另外一方面如果能让复杂度判断网络和主干任务网络实现有效地协同训练，会极大地提升训练效率（相当于在训练阶段就引入了自适应深度推理）。

3）探索更加有效的、通用的复杂度估计方法。基于置信度的样本复杂度评价方法不仅会面临 over-confidence 的问题，而且这种评价机制使用范围受限，仅能够用于分类任务。用基于策略网络的数据驱动方式替代人工设计规则是一个更好的选择，不过样本复杂度估计本身是一个无监督问题，当前基于强化学习的训练策略虽然有效，但是其训练结果对于 reward 的设计、采样过程的设计等因素比较敏感，训练周期较长。利用样本的预测损失和样本复杂度之间的互信息，进行策略网络的训练是一条较好的思路，不过需要注意的是训练集上的预测损失可能会因为过拟合而变得很小，从而无法客观反映样本的复杂度。

参 考 文 献

[1] An S, Boussaid F, Bennamoun M. How can deep rectifier networks achieve linear separability and preserve distances?[C]//International Conference on Machine Learning, July 6-11, 2015, Lille, France. Cambridge: PMLR, 2015: 514-523.

[2] Arora S, Cohen N, Hazan E. On the optimization of deep networks: implicit acceleration by overparameterization[C]// Proceedings of the 35th International Conference on Machine Learning, July 10-15, 2018, Stockholm, Sweden. Cambridge: PMLR, 2018.

[3] Guo S, Alvarez J M, Salzmann M. Expandnets: linear over-parameterization to train compact convolutional networks[C]//Advances in Neural Information Processing Systems 2020, December 6-12, 2020, Virtual. New York: Curran Associates, 2020, 33: 1298-1310.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[5] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2021-05-04]. https://arxiv.org/abs/1409.1556.

[7] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2021-05-04]. https://arxiv.org/abs/2004.10934.

[8] Feng C J, Zhong Y J, Gao Y, et al. TOOD: task-aligned one-stage object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 3490-3499.

[9] Duan K W, Bai S, Xie L X, et al. CenterNet: keypoint triplets for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 6568-6577.

[10] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[11] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[12] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern

Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[13] Wang X L, Zhang R F, Kong T, et al. Solov2: dynamic and fast instance segmentation[C]//Advances in Neural Information Processing Systems, December 6-12, 2020, Virtual. New York: Curran Associates, 2020, 33: 17721-17732.

[14] Yu C Q, Xiao B, Gao C X, et al. Lite-HRNet: a lightweight high-resolution network[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 10435-10445.

[15] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 5686-5696.

[16] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[17] Godard C, Aodha O M, Brostow G J. Unsupervised monocular depth estimation with left-right consistency [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6602-6611.

[18] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 12159-12168.

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA. New York: Curran Associates, 2017: 5998-6008.

[20] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12346: 213-229.

[21] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 9992-10002.

[22] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[23] Howard A G, Zhu M, Chen B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2021-05-06]. https://arxiv.org/abs/1704.04861.

[24] Goyal A, Bochkovskiy A, Deng J, et al. Non-deep networks[EB/OL]. (2021-10-14) [2022-01-04]. https://arxiv.org/abs/2110.07641.

[25] Han K, Wang Y H, Tian Q, et al. GhostNet: more features from cheap operations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1577-1586.

[26] Tan M X, Le Q V. Efficientnet: rethinking model scaling for convolutional neural networks[C]//36th International Conference on Machine Learning, June 9-15, 2019, Long Beach, California, USA. Cambridge: PMLR, 2019: 6105-6114..

[27] Molchanov P, Tyree S, Karras T, et al. Pruning convolutional neural networks for resource efficient inference[C]//5th International Conference on Learning Representations, April 24-26, 2017, Toulon, France. Cambridge: ICLR, 2017.

[28] Zhang L F, Song J B, Gao A N, et al. Be your own teacher: improve the performance of convolutional neural networks via self distillation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 3712-3721.

[29] Chen G, Choi W, Yu X, et al. Learning efficient object detection models with knowledge distillation[C]//Advances in Neural Information Processing Systems 30, December 4-9, 2017, Long Beach, CA, USA. New York: Curran Associates, 2017, 30: 743-752.

[30] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[EB/OL]. (2015-05-09)[2021-01-01]. https://arxiv.org/abs/1503.02531.

[31] Scardapane S, Scarpiniti M, Baccarelli E, et al. Why should we add early exits to neural networks?[J]. Cognitive Computation, 2020, 12(5): 954-966.

[32] Kauffmann L, Ramanoël S, Peyrin C. The neural bases of spatial frequency processing during scene perception [J]. Frontiers in Integrative Neuroscience, 2014, 8: 37.

[33] Kaya Y, Hong S, Dumitras T. Shallow-deep networks: understanding and mitigating network overthinking[C]//36th International Conference on Machine Learning, June 9-15, 2019, Long Beach, California, USA. Cambridge: PMLR, 2019: 3301-3310.

[34] Huang G, Chen D, Li T, et al. Multi-scale dense networks for resource efficient image classification[C]//International Conference on Learning Representations, April 30-May 3, 2018, Vancouver, BC, Canada. New York: Curran Associates, 2018.

[35] Zhou W, Xu C, Ge T, et al. Bert loses patience: fast and robust inference with early exit[C]//Advances in Neural Information Processing Systems, December 6-12, 2020, Virtual. New York: Curran Associates, 2020, 33: 18330-18341.

[36] Teerapittayanon S, McDanel B, Kung H T. BranchyNet: Fast inference via early exiting from deep neural networks[C]//2016 23rd International Conference on Pattern Recognition (ICPR), December 4-8, 2016, Cancun, Mexico. New York: IEEE Press, 2016: 2464-2469.

[37] Wołczyk M, Wójcik B, Bałazy K, et al. Zero time waste: recycling predictions in early exit neural networks [C]//Advances in Neural Information Processing Systems 34, December 6-14, 2021, Virtual. New York: Curran Associates, 2021, 34: 2516-2528.

[38] Li H, Zhang H, Qi X J, et al. Improved techniques for training adaptive deep networks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1891-1900.

[39] Passalis N, Raitoharju J, Tefas A, et al. Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits[J]. Pattern Recognition, 2020, 105: 107346.

[40] Kouris A, Venieris S I, Laskaridis S, et al. Multi-exit semantic segmentation networks[EB/OL]. (2021-06-07) [2021-10-03]. https://arxiv.org/abs/2106.03527v1.

[41] Xin J, Tang R, Lee J, et al. DeeBERT: dynamic early exiting for accelerating BERT inference[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 5-10, 2020, Online. Stroudsburg: Association for Computational Linguistics, 2020: 2246-2251.

[42] Schwartz R, Stanovsky G, Swayamdipta S, et al. The right tool for the job: matching model and instance complexities[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 5-10, 2020, Online. Stroudsburg: Association for Computational Linguistics, 2020: 6640-6651.

[43] Chen X, Dai H, Li Y, et al. Learning to stop while learning to predict[C]//37th International Conference on Machine Learning, July 13-18, 2020, Virtual Event. Cambridge: PMLR, 2020: 1520-1530.

[44] Liu Y J, Meng F D, Zhou J, et al. Faster depth-adaptive transformers[C]//Proceedings of the AAAI Conference on Artificial Intelligence, February 2-9, 2021, Virtual Event. Virigina: AAAI Press, 2021, 35(15): 13424-13432.

[45] Jie Z Q, Sun P, Li X, et al. Anytime recognition with routing convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(6): 1875-1886.

[46] Huang C, Lucey S, Ramanan D. Learning policies for adaptive tracking with deep feature cascades[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 105-114.

[47] Dai X, Kong X N, Guo T. EPNet: learning to exit with flexible multi-branch network[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management, October 19-23, 2020, Virtual Event, Ireland. New York: ACM Press, 2020: 235-244.

[48] Duggal R, Freitas S, Dhamnani S, et al. Elf: an early-exiting framework for long-tailed classification[EB/OL]. (2020-06-22)[2021-06-05]. https://arxiv.org/abs/2006.11979.

[49] Wang X, Li Y. Harmonized dense knowledge distillation training for multi-exit architectures[C]//Proceedings of the AAAI Conference on Artificial Intelligence, February 2-9, 2021, Virtual Event. Virigina: AAAI Press, 2021, 35(11): 10218-10226.

[50] Phuong M, Lampert C. Distillation-based training for multi-exit architectures[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1355-1364.

[51] Liu B L, Rao Y M, Lu J W, et al. MetaDistiller: network self-boosting via meta-learned top-down distillation[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12359: 694-709.

[52] Wang X L, Li Y M. Gradient deconfliction-based training for multi-exit architectures[C]//2020 IEEE International Conference on Image Processing, October 25-28, 2020, Abu Dhabi, United Arab Emirates. New York: IEEE Press, 2020: 1866-1870.

[53] Wu Z X, Nagarajan T, Kumar A, et al. BlockDrop: dynamic inference paths in residual networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8817-8826.

[54] Wang Y, Shen J H, Hu T K, et al. Dual dynamic inference: enabling more efficient, adaptive, and controllable deep inference[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(4): 623-633.

[55] Wang X, Yu F, Dou Z Y, et al. SkipNet: learning dynamic routing in convolutional networks[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11217: 420-436.

[56] Veit A, Belongie S. Convolutional networks with adaptive inference graphs[J]. International Journal of Computer Vision, 2020, 128(3): 730-741.

[57] Fan A, Grave E, Joulin A. Reducing transformer depth on demand with structured dropout[C]//8th International Conference on Learning Representations, April 26-30, 2020, Addis Ababa, Ethiopia. [S.l.: s.n.], 2020.

[58] Huang G, Sun Y, Liu Z, et al. Deep networks with stochastic depth[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 646-661.

[59] 邸江磊, 唐雎, 吴计, 等. 卷积神经网络在光学信息处理中的应用研究进展[J]. 激光与光电子学进展, 2021, 58(16): 1600001.
Di J L, Tang J, Wu J, et al. Research progress in the applications of convolutional neural networks in optical information processing[J]. Laser & Optoelectronics Progress, 2021, 58(16): 1600001.

[60] Luo W, Li Y, Urtasun R, et al. Understanding the effective receptive field in deep convolutional neural networks[EB/OL]. (2017-01-15) [2021-05-06]. https://arxiv.org/abs/1701.04128.

[61] 肖万新, 李华锋, 张亚飞, 等. 多尺度特征学习和边缘增强的医学图像融合[J]. 激光与光电子学进展, 2022, 59(6): 0617029.

Xiao W X, Li H F, Zhang Y F, et al. Medical image fusion based on multi-scale feature learning and edge enhancement[J]. Laser & Optoelectronics Progress, 2022, 59(6): 0617029.

[62] Jeon G W, Choi J H, Kim J H, et al. LarvaNet: hierarchical super-resolution via multi-exit architecture [M]//Bartoli A, Fusiello A. Computer vision-ECCV 2020 workshops. Lecture notes in computer science. Cham: Springer, 2020, 12537: 73-86.

[63] 马天浩, 谭海, 李天琪, 等. 多尺度特征融合的膨胀卷积残差网络高分一号影像道路提取[J]. 激光与光电子学进展, 2021, 58(2): 0228001.
Ma T H, Tan H, Li T Q, et al. Road extraction from GF-1 remote sensing images based on dilated convolution residual network with multi-scale feature fusion[J]. Laser & Optoelectronics Progress, 2021, 58(2): 0228001.

[64] Peng C, Zhang X Y, Yu G, et al. Large kernel matters: improve semantic segmentation by global convolutional network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1743-1751.

[65] Yu H H, Winkler S. Image complexity and spatial information[C]//2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), July 3-5, 2013, Klagenfurt, Austria. New York: IEEE Press, 2013: 12-17.

[66] Perkiö J, Hyvärinen A. Modelling image complexity by independent component analysis, with application to content-based image retrieval[M]//Alippi C, Polycarpou M, Panayiotou C, et al. Artificial neural networks-ICANN 2009. Lecture notes in computer science. Heidelberg: Springer, 2009, 5769: 704-714.

[67] Han Y, Huang G, Song S, et al. Dynamic neural networks: a survey[EB/OL]. (2021-02-09)[2021-05-08]. https://arxiv.org/abs/2102.04906.

[68] Yu T, Kumar S, Gupta A, et al. Gradient surgery for multi-task learning[C]//Advances in Neural Information Processing Systems, December 6-12, 2020, Virtual. New York: Curran Associates, 2020, 33: 5824-5836.

[69] Li H, Xu Z, Taylor G, et al. Visualizing the loss landscape of neural nets[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems, December 3-8, 2018, Montreal, Canada. New York: Curran Associates, 2018: 6391-6401.

[70] Nguyen Q, Hein M. The loss surface and expressivity of deep convolutional neural networks[C]//International Conference on Learning Representations Workshop, April 30-May 3, 2018, Vancouver, BC, Canada. [S.l.: s.n.], 2018.

[71] Liu B, Liu X, Jin X, et al. Conflict-averse gradient descent for multi-task learning[C]//Advances in Neural Information Processing Systems, December 6-14, 2021, Virtual. New York: Curran Associates, 2021: 18878-18890.

[72] Sener O, Koltun V. Multi-task learning as multi-objective optimization[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems, December 3-8, 2018, Montreal, Canada. New York: Curran Associates, 2018: 525-536.

[73] Li Y, Ji R, Lin S, et al. Interpretable neural network decoupling[EB/OL]. (2019-06-04)[2021-04-05]. https://arxiv.org/abs/1906.01166v2.