为了提升在小数据集上的性能，有学者让神经网络像生物一样“进化”了 | CVPR2021 Oral

原创 CV开发者都爱看的 极市平台 2021-03-27 22:01:00 手机阅读

收录于话题

#神经网络结构设计 25 #小样本学习 1 #模型训练 1

作者 | 二玖

审稿 | 邓富城

编辑 | 极市平台

极市导读

如何在较小的数据集上训练神经网络？马里兰大学的学者提出了一种以“进化”为灵感的训练方法：Knowledge Evolution(KE)。这种名为KE的方法能像生物进化一样，巧妙地提升神经网络在小数据集上的性能。该方法已开源。 >>加入极市CV技术交流群，走在计算机视觉的最前沿

如果将深度学习比作汽车，那数据集就是石油。深度学习对于数据集的依赖是不言而喻的。那么一个深度学习领域悬而未决的挑战则是：**如何在较小的数据集上训练神经网络？**

针对这一问题，马里兰大学的学者提出了一种以“**进化**”为灵感的训练方法：**Knowledge Evolution(KE)**。这种名为KE的方法能**像生物进化一样，巧妙地提升神经网络在小数据集上的性能**。目前作者已经开源了这一方法。

Knowledge Evolution in Neural Networks

Ahmed Taha Abhinav Shrivastava Larry Davis
University of Maryland, College Park

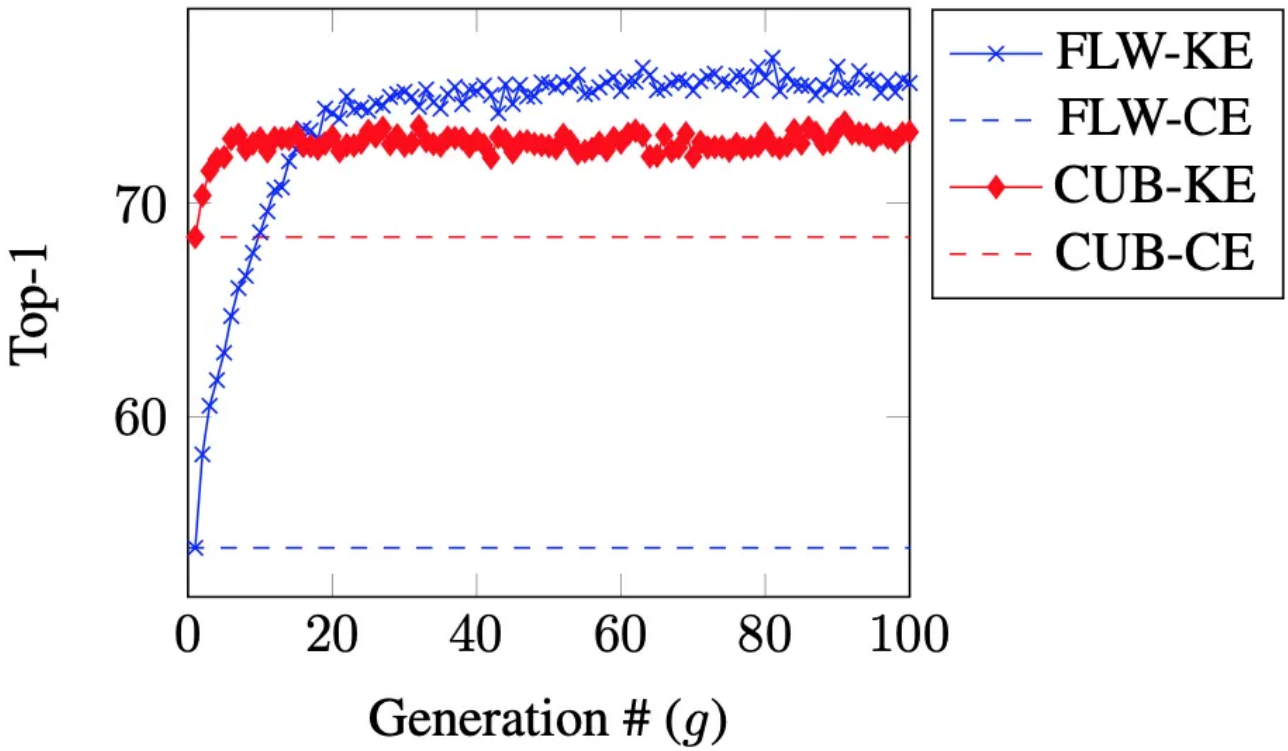
KE将神经网络用两种假设进行拆分：拟合假设和重置假设，并通过多次扰动重置假设来迭代拟合假设中知识。这种方法不仅能提升性能，还能以较低的推理成本来获取一个精简的神经网络。同时，KE能减少过拟合和数据收集的负担。不但如此，KE支持各种网络结构和损失函数，还能与残差卷积网络，以及其他正则化技术（如标签平滑）等无缝集成。

下面我们将更详细地介绍这篇论文。

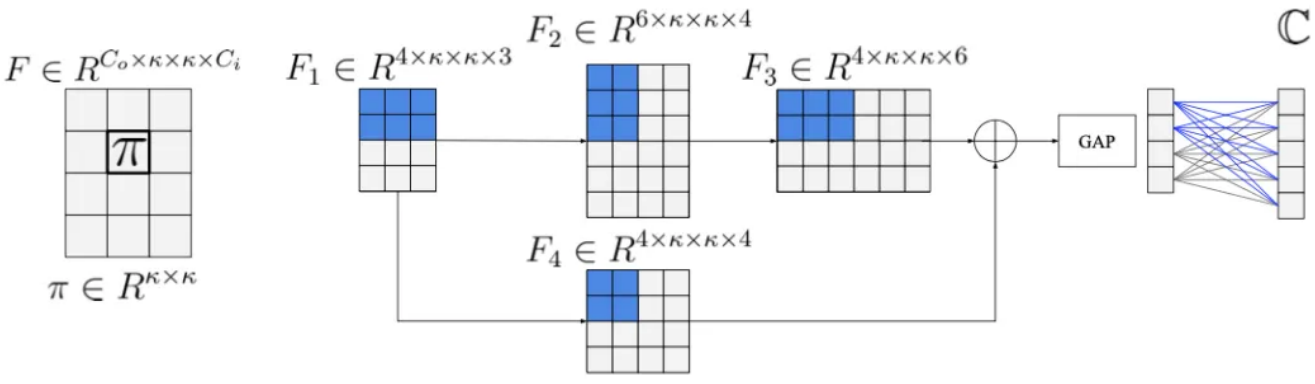
KE的主要贡献

在本文中，作者尝试用神经网络复制一个这样的生物过程：基因编码了从祖先到后代的遗传信息（知识），而基因传递将遗传信息从父母传递至其后代。虽然祖先并不一定具有更好的知识，但是遗传信息（知识）在几代人之间的发展将会促进后代更好的学习曲线。

因此，作者将深度神经网络的知识封装在一个名为拟合假设的子网络$H^{\triangle}$ 中，然后将拟合假设的知识从**父母网络**传递至其后代，即下一代神经网络。并反复重复此过程，在后代网络中证明了其性能的显著提升：



如下图所示，KE将神经网络分为两个假设（子网络）：拟合假设$H^{\triangle}$和重置假设H^{∇}。通过重新训练多代网络来进化$H^{\triangle}$中的知识。而每一代都会通过扰乱H^{∇}内部的权重以鼓励$H^{\triangle}$学习独立的表达形式。这种知识进化方法能够提高神经网络在小数据集上的性能。



此外，为了降低推理成本，作者提出了一种为CNN量身定制的一种拆分技术，即内核级卷积感知拆分（**ke**rnel-**l**evel-convolution al-aware **s**plitting，KELS）。KELS同时支持CNN和残差网络，且既不引入超参数也不引入正则项。

知识进化方法详解

假设一个具有L层的深度网络 N。 网络 N 具有卷积滤波器F，批范数Z以及权重为W，偏置项为B的完全连接层。

知识进化（KE）首先从概念上将深度网络N分为两个互斥假设（子网络）：拟合假设$H^{\triangle}$和重置假设H^{∇}。这些假设由二进制掩码M概述。 $H^{\triangle}$为1，H^{∇}为0，即$H^{\triangle}=MN$和$H^{\nabla}=(1-M)N$。 随后，网络N被随机初始化，即$H^{\triangle}$和H^{∇}都被随机初始化。训练e期N，并将已训练的网络称为第一代N_1，其中$H_1^{\triangle}=MN_1$，$H_1^{\nabla}=(1-M)N_1$。

为了学习更好的网络（下一代），作者使用$H_1^{\triangle}$**重新初始化**网络N，然后**重新训练**N以学习N_2。网络N使用卷积滤波器F和来自N_1 的拟合假设$H_1^{\triangle}$中的权重W进行**重新初始化**，而网络的剩余部分$\left(H^{\nabla}\right)$则被随机初始化。 作者使用哈达玛积重新初始化每层l，如下所示：

$$F_l = M_l F_l + (1 - M_l) F_l^r$$

类似地，作者通过它们相应的二进制掩码重新初始化权重W_l和偏置B_l。 网络架构仅在最后一个完全连接层中具有偏差项$\left(B \in R^{\mathbb{C}}\right)$。 因此，对于这些架构，所有偏差项都属于拟合假设，即$B \subset H^{\triangle}$。并将学习的批规范Z进行无随机化的跨代传递。

重新初始化后，训练e期N以学习第二代N_2。 为了学习更好的网络，反复为g代**重新初始化**以及**重新训练**N。 通过拟合假设$H^{\triangle}$将知识（卷积滤波器和权重）从一代传递至下一代。需要注意的是：

- 一代网络的贡献在初始化下一代后立刻结束，即每一代的训练都是独立的；
- 在训练了新一代后，两个假设中的权重都会发生变化 ，即 $H_1^{\triangle} \neq H_2^{\triangle}$ ， $H_1^{\nabla} \neq H_2^{\nabla}$,；
- 所有网络代都使用精确的超参数，即相同的期数、优化器、学习率调度器等）进行训练。

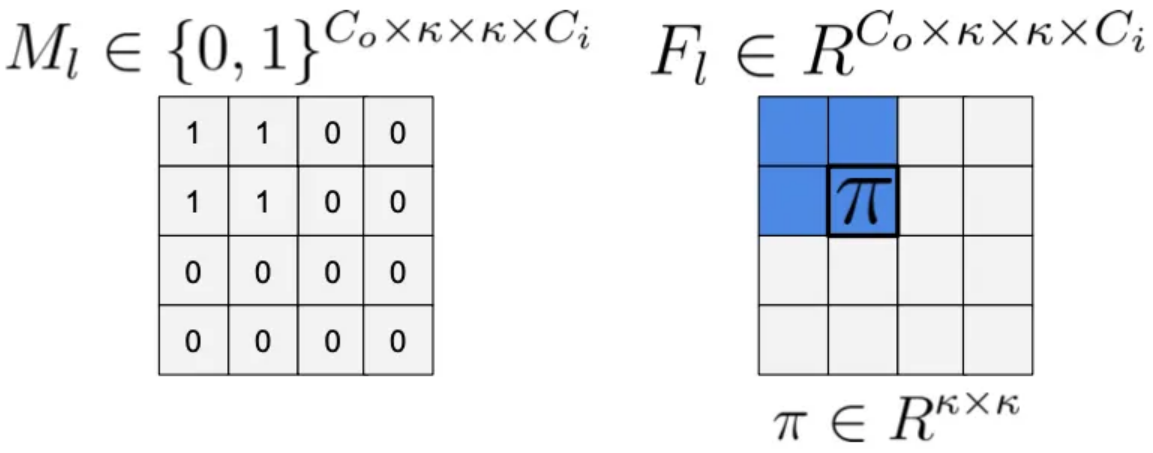
拆分网络

KE需要进行网络拆分。 作者主要采用了两种拆分技术：

1. **we**ight-**l**everl **s**plitting（WELS），能用于突出KE通用性；
2. **ke**rnel-**l**evel convolutional-aware **s**plitting（KELS），一种有效的CNN技术。

WELS技术十分简单：对每一层 l，二进制掩码 M_l 将 l 分为两个专有部分：拟合假设 $H^{\triangle}$ 和重置假设 H^{∇} 。给定拆分率 $0<s_r<1$，使用掩码$M_l \in \{0,1\}^{|W_l|}$随机分配权重$W_l \in R^{|W_l|}$，其中$|W_l|$是内部权重的数量，$\operatorname{sum}(M_l)=s_r \times |W_l|$。WELS技术支持完全连接，卷积，递归以及图卷积，这也印证了KE的通用性。

虽然KE通过WELS跨代提升了网络性能，但是WELS不能从CNN的连接中受益。因此，作者提出了一种既能提高性能，又能减少在小数据集上的推理成本的拆分技术，即KELS。利用CNN的连接性并引出拟合假设$H^{\triangle}$，对网络进行修剪。用内核遮罩替代加权遮罩，因此为内核级卷积感知分割（ KELS）技术。给定一个拆分率s_r和一个卷积滤波器$F_l \in R^{\hat{C}_o \times \kappa \times \kappa \times C_i}$，KELS引出拟合假设，以将第一个$\lceil s_r \times C_i \rceil$内核包括在第一个$\lceil s_r \times C_o \rceil$滤波器中，如下图所示。KELS保证了结果卷积滤波器之间的维度匹配。因此，KELS可以无缝集成在经典CNN（AlexNet和VGG）和具有残差连接的最新网络架构中。

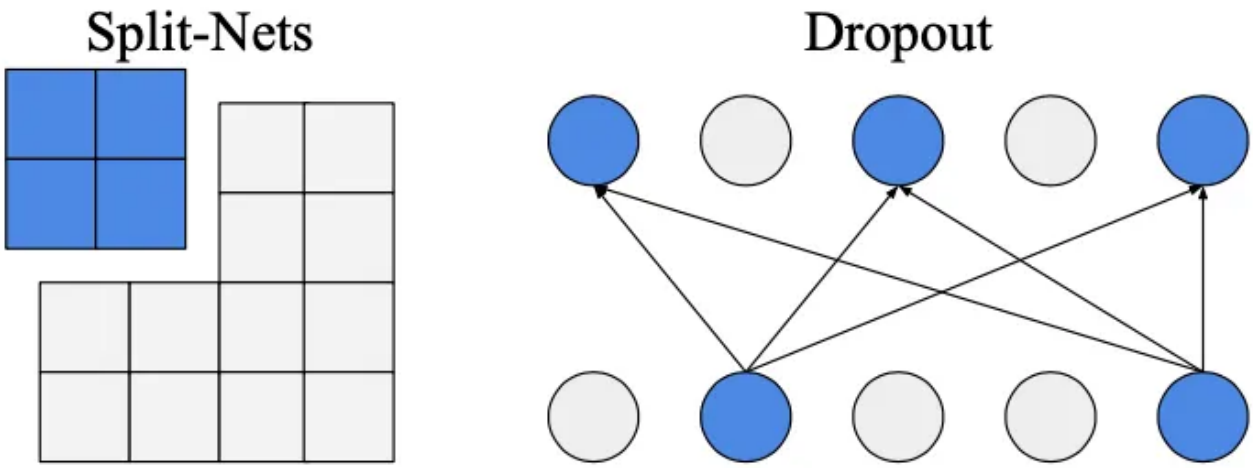


KELS

知识进化的直觉

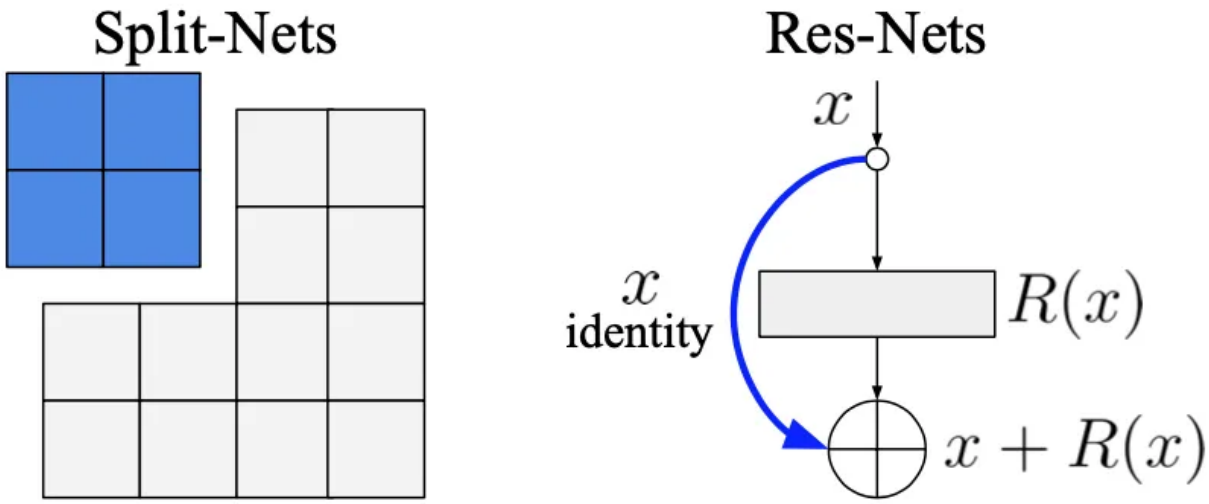
为了理解KE，作者给出了不需要KELS技术的两个互补直觉：Dropout和残差网络。

Dropout在训练过程中随机丢弃神经网络单元，使得神经网络单元减少相互依赖并学习独立表示。 类似的，在KE中通过在每一代之前随机初始化重置假设H^{∇}，我们可以在重新初始化期间丢弃H^{∇}。这同样让$H^{\triangle}$减少对H^{∇}的依赖，并学习独立表示。 通过评估各代$H^{\triangle}$的性能验证了这种直觉，且$H^{\triangle}$的性能随着代数的增加而增加。如下图所示：



Split-Nets vs Dropout

Res-Nets将连续层之间的默认映射设置为身份。但是从一个不同的角度来看，没有限制网络容量，Res-Nets在某些子网络（残差连接）中实现零映射。 类似地，KE通过跨代重复使用拟合假设$H^{\triangle}$来实现重置假设H^{∇}的零映射。 在第一代N1之后，与包含随机值的H^{∇}相比，$H^{\triangle}$总是更趋于收敛。 因此，KE促进了新一代网络在$H^{\triangle}$和H^{∇}中对前代网络的知识进化。



Split-Nets vs Res-Nets

对KE进行评估

作者在两个分类和度量学习这两个监督任务上对KE进行了评估。分类任务已有研究对深度神经网络在小型数据集上的性能进行了广泛的探讨，因此它提供了严格的性能基准。而在度量学习上的评估则突出了KE的灵活性与普遍性。

在分类任务上对KE进行评估

下图为用带有KELS的ResNet18进行定量分类评估。 N_g表示g^{th} 网络代的性能。 第一代N_1既是KE的基准又是起点。 随着代数的增加，KE的性能提高。

Method	Flower	CUB	Aircraft	MIT	Dog
CE + AdaCos	**55.45**	**62.48**	**57.06**	56.25	**65.34**
CE + RePr	41.90	42.88	39.43	46.94	50.39
CE + DSD	51.39	53.00	57.24	53.21	63.58
CE + BANs-N_{10}	48.53	53.71	53.19	55.65	64.16
CE (N_1)	48.48	53.57	51.28	55.28	63.83
CE + KE-N_3 **(ours)**	52.53	56.73	52.53	57.44	64.28
CE + KE-N_{10} **(ours)**	56.15	58.11	53.21	**58.33**	64.56
Smth (N_1)	50.97	59.75	55.00	57.74	65.95
Smth + KE-N_3 **(ours)**	56.87	62.88	57.47	58.78	66.91
Smth + KE-N_{10} **(ours)**	**62.56**	**66.85**	**60.03**	**60.42**	**67.06**
CS-KD (N_1)	55.10	67.71	58.15	57.37	69.60
CS-KD + KE-N_3 **(ours)**	61.74	71.63	**59.97**	**58.41**	70.62
CS-KD + KE-N_{10} **(ours)**	**69.88**	**73.39**	59.08	57.96	**70.81**

下图为用带有WELS的DenseNet169进行的定量评估。

Method	Flower	CUB	Aircraft	MIT	Dog
CE + AdaCos	49.96	**62.20**	56.15	50.89	65.33
CE + RePr	39.75	47.01	36.04	49.77	55.63
CE + DSD	48.85	56.11	53.66	58.31	65.76
CE + BANs-N_{10}	44.92	57.30	52.56	57.66	65.49
CE (N_1)	45.85	55.16	51.73	56.62	64.82
CE + KE-N_3 **(ours)**	52.44	57.75	56.70	**59.67**	67.06
CE + KE-N_{10} **(ours)**	**60.15**	58.01	**59.73**	58.71	**67.75**
Smth (N_1)	46.34	59.93	57.74	57.81	65.12
Smth + KE-N_3 **(ours)**	55.46	**62.53**	62.86	**60.27**	**68.21**
Smth + KE-N_{10} **(ours)**	**64.18**	61.34	**65.86**	59.75	67.46
CS-KD (N_1)	46.97	67.32	58.87	56.62	69.83
CS-KD + KE-N_3 **(ours)**	59.36	69.77	59.91	**59.00**	**71.70**
CS-KD + KE-N_{10} **(ours)**	**65.27**	**70.36**	**61.22**	57.44	70.72

在度量学习任务上对KE进行评估

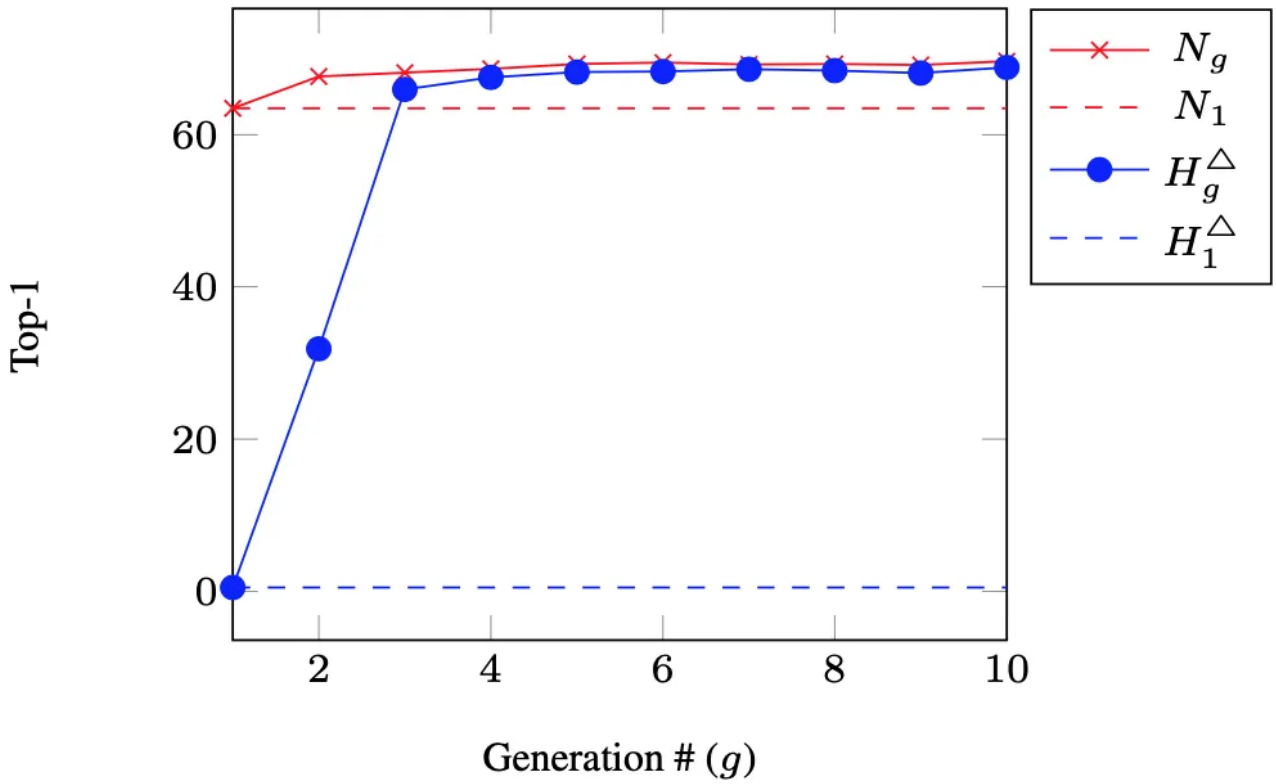
下图为使用标准度量学习数据集和架构进行的定量检索评估。

	ResNet50			GoogLeNet		
Datasets	NMI	R@1	R@4	NMI	R@1	R@4
CUB (N_1)	0.396	13.01	30.37	0.396	10.16	25.71
CUB + KE-N_3 **(ours)**	0.424	17.22	36.14	0.418	13.94	33.78
CUB + KE-N_{10} **(ours)**	**0.429**	**18.25**	**39.40**	**0.419**	**15.34**	**34.30**
Cars (N_1)	0.374	11.63	28.66	0.319	5.29	17.94
Cars + KE-N_3 **(ours)**	0.514	34.28	60.25	0.476	24.98	50.06
Cars + KE-N_{10} **(ours)**	**0.523**	**42.36**	**68.11**	**0.495**	**32.63**	**58.84**

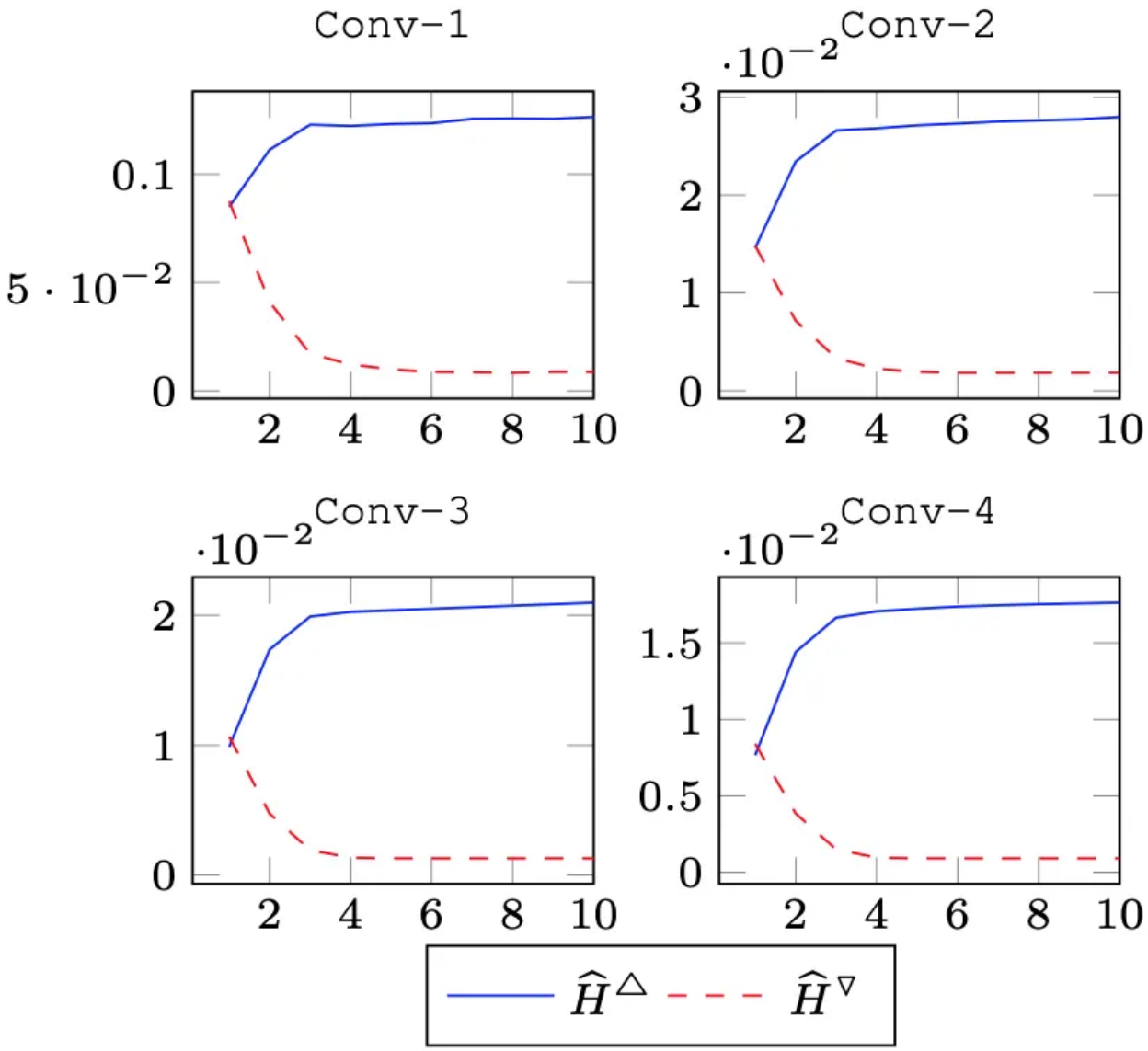
消融实验

（1）Dropout与Res-Net

在VGG11_bn上使用CUB-200进行定量分类评估，下图显示了10代密集网络N和拟合假设$H^{\triangle}$的性能。

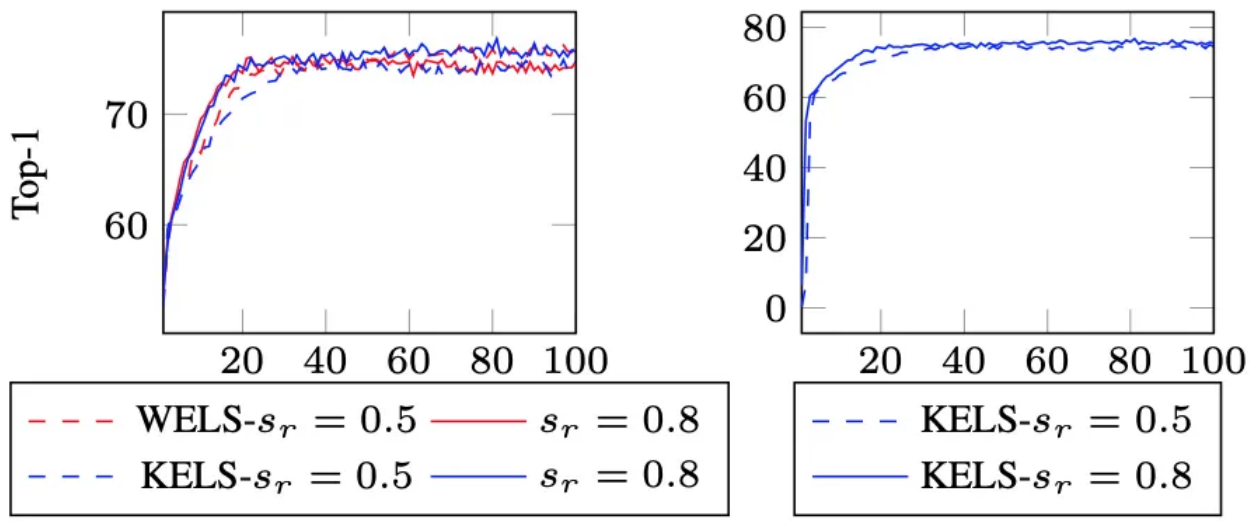


拟合假设$H^{\triangle}$在$g=1$处的性能较差，但随着代数的增加，其性能也会提高。 $\widehat{H}^{\triangle}$和$\widehat{H}^{\nabla}$表示$H^{\triangle}$和 H^{∇}内部的平均绝对值。



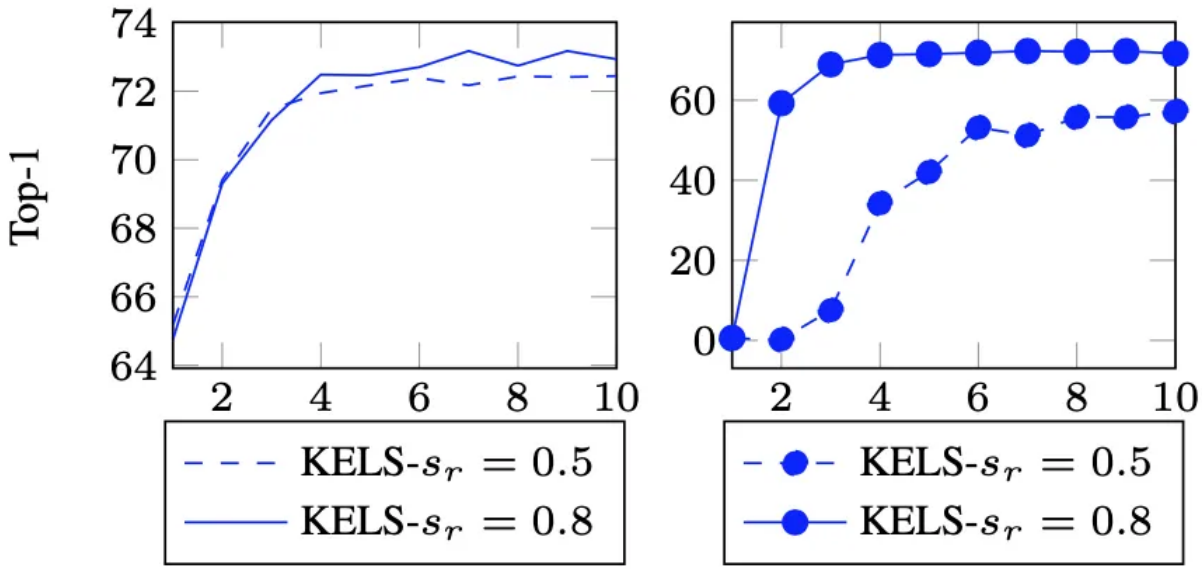
（2）WELS 与 KELS

在ResNet18上使用Flower-102对KELS和WELS进行100代定量评估。下图左为密集网络N的分类性能，右为拟合假设$H^{\triangle}$的性能。



（3）拆分率s_r的tradeoffs

拆分率s_r控制拟合假设的大小； 小的s_r会降低推理成本，但是小的s_r会减少$H^{\triangle}$的容量。下图左比较了使用CUB-200和GoogLeNet的两个拆分率($s_r = \{0.5, 0.8\}$)的10代。两种分割率都在密集网络N上显著提升了边际量。但是，下图右显示，较大的拆分率$s_r = 0.8$ 可以帮助拟合假设 $H^{\triangle}$更快地收敛并获得更好的性能。 因而对于大型数据集，需要大的拆分率才能使拟合假设$H^{\triangle}$的性能具有竞争力。



论文传递门

论文：

https://arxiv.org/abs/2103.05152

代码：

https://github.com/ahmdtaha/knowledge_evolution

推荐阅读

CVPR'21 | Involution：超越卷积和自注意力的神经网络新算子
2021-03-23

目标检测一卷到底之后，终于有人为它挖了个新坑 | CVPR2021 Oral
2021-03-12

CVPR 2021 Oral | Transformer再发力！华南理工和微信提出UP-DETR：无监督预训练检测器
2021-03-11

觉得有用麻烦给个在看啦~

阅读原文

喜欢此内容的人还喜欢

15个目标检测开源数据集汇总

极市平台