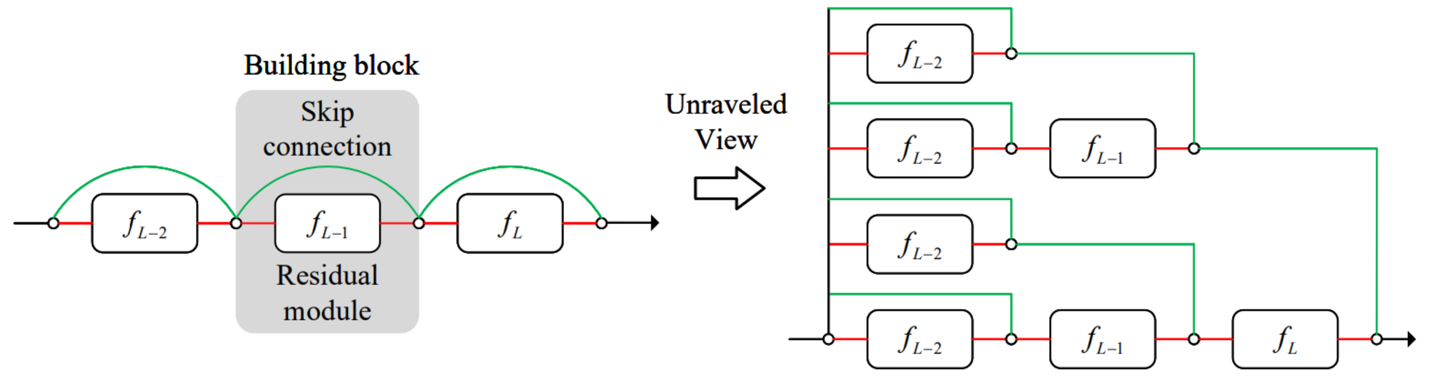
Stimulative Training of Residual Networks: A Social Psychology Perspective of Loafing

Peng Ye[1], Shengji Tang[1], Baopu Li[2], Tao Chen[1], Wanli Ouyang[3]
[1]School of Information Science and Technology, Fudan University [2]Oracle Health and AI, USA,
[3]The University of Sydney, SenseTime Computer Vision Group, Australia, and Shanghai AI Laboratory

Main Problems

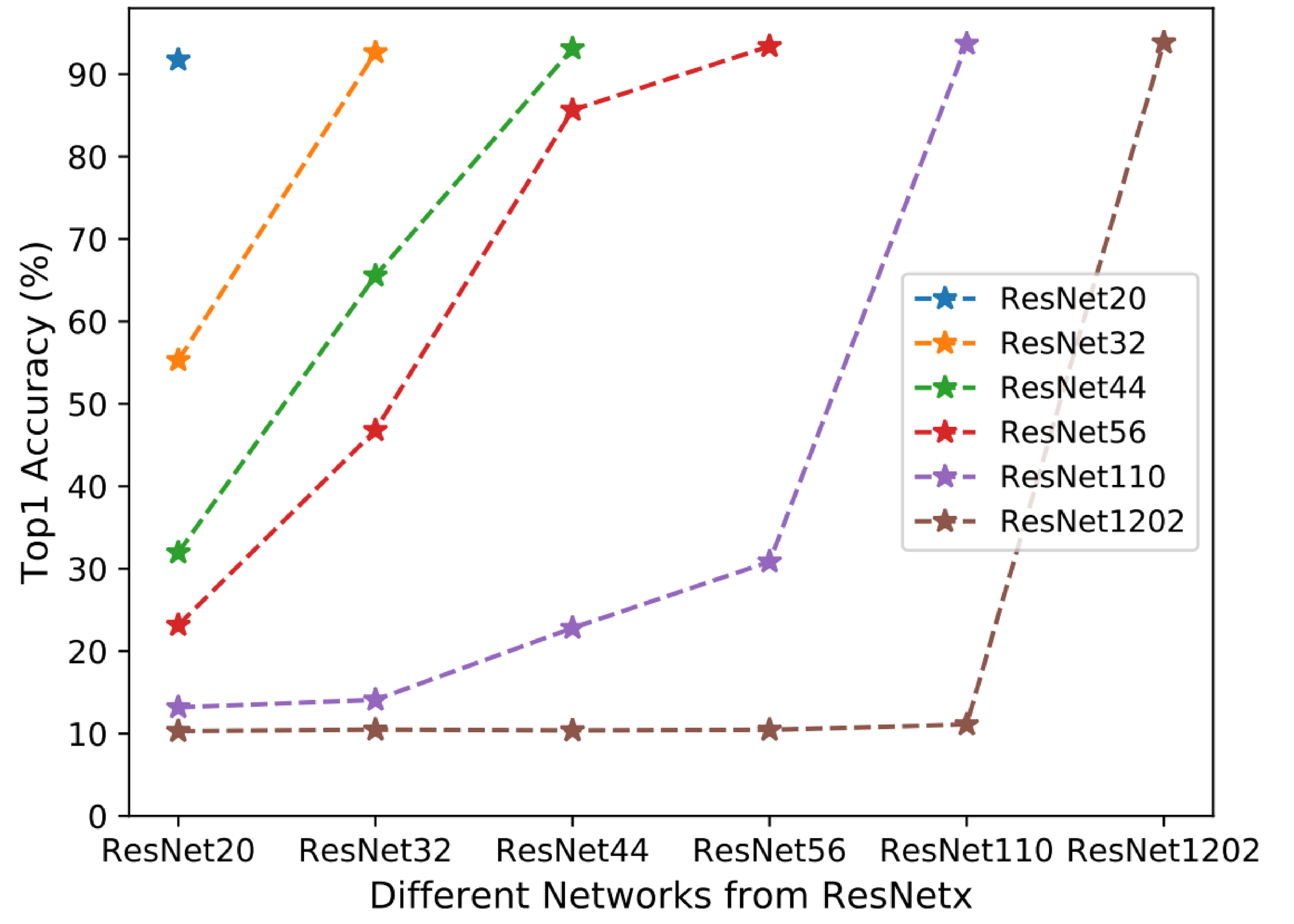
➢Unraveled view of residual network



➢Figure 1: The unraveled view of residual network.

➢As residual networks can be viewed as ensembles of relatively shallow networks (i.e., **unraveled view**) in prior works, we also start from such view and consider that the final performance of a residual network is co-determined by a group of sub-networks.

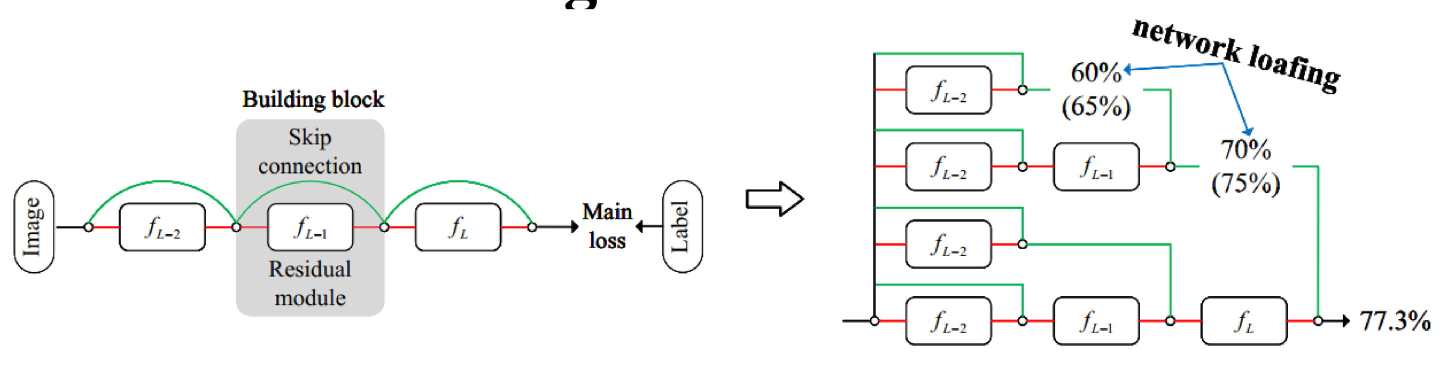
➢The problem of network loafing



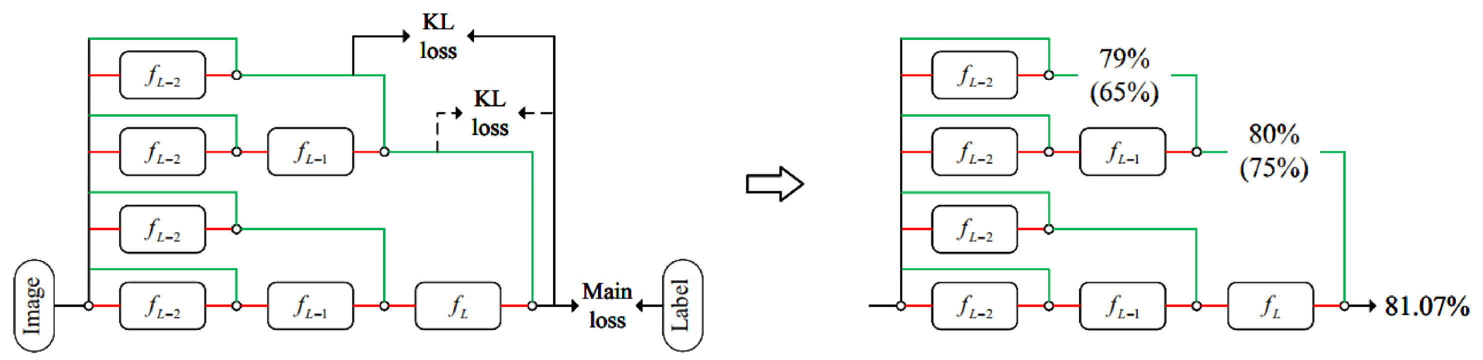
➢Figure 2: Different residual networks invariably suffer from the problem of network loafing (i.e., **the subnets of given residual network perform significantly worse than the same architecture trained alone**), and deeper residual network tends to have more serious loafing problem. The horizontal axis means the sampled different sub-networks from ResNetx.

Approach

➢Stimulative training framework



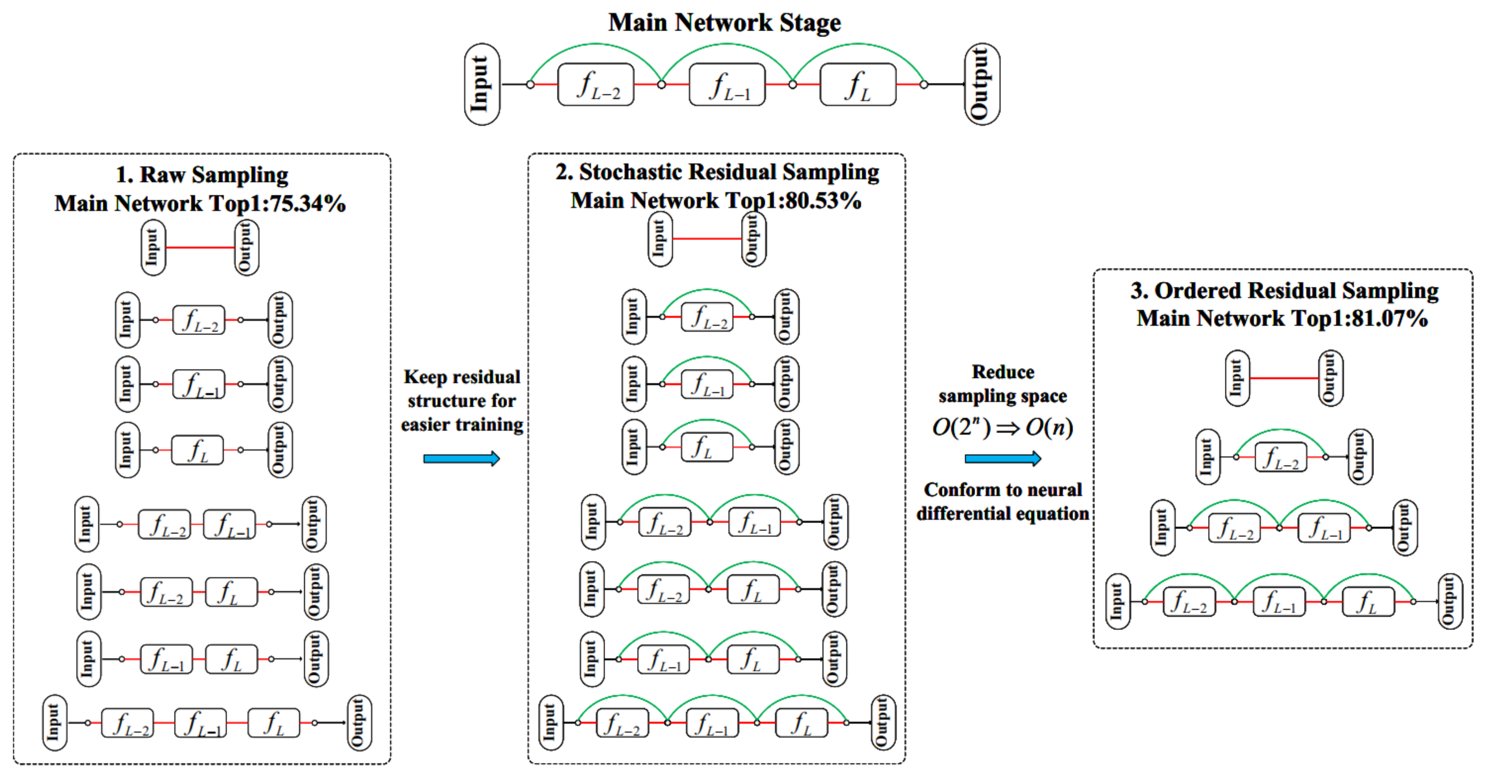
(a) **Common training** scheme suffers from severe **network loafing** problem



(b) **Stimulative training** scheme can relieve **network loafing** and improve the performance

➢Figure 3: Illustration of common and stimulative training schemes. **Stimulative training** can relieve the network loafing problem, and improve the performance of a given residual network and all of its sub-networks.

➢Ordered Residual Sampling

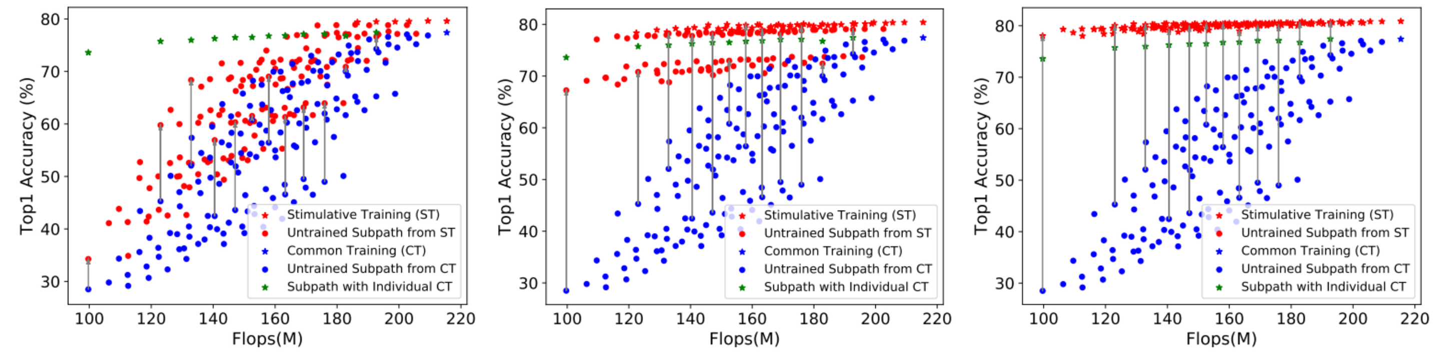


➢Figure 4: The **ordered residual sampling** rule, consisting of keeping residual structure and ordered sampling, can facilitate stimulative training effectively and efficiently.

Code: https://github.com/Sunshine-Ye/NIPS22-ST

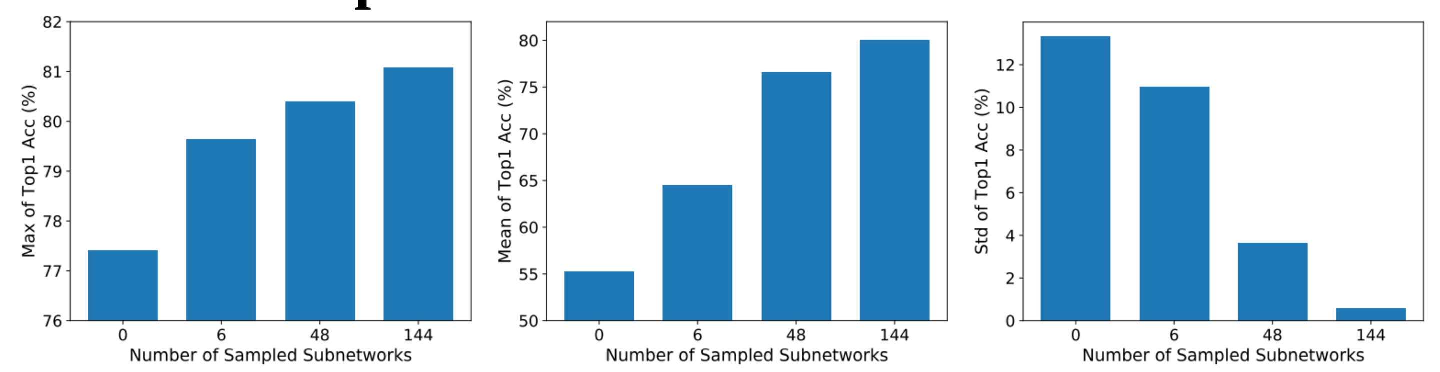
Experimental Results

➢Stimulative training can alleviate network loafing



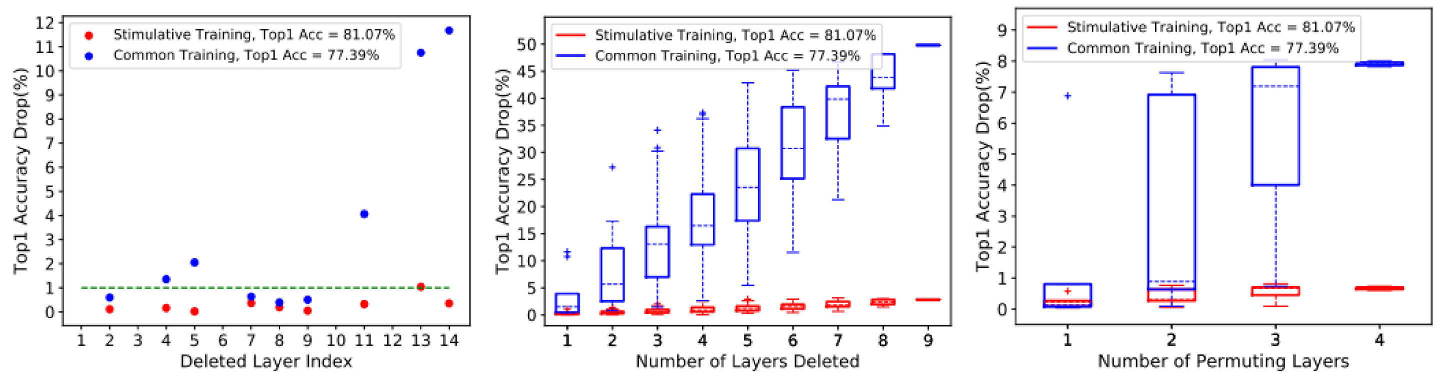
(a) 6 sub-networks (b) 48 sub-networks (c) 144 sub-networks

➢Stimulative training show the best statistical characteristics of the performance of all residual sub-networks



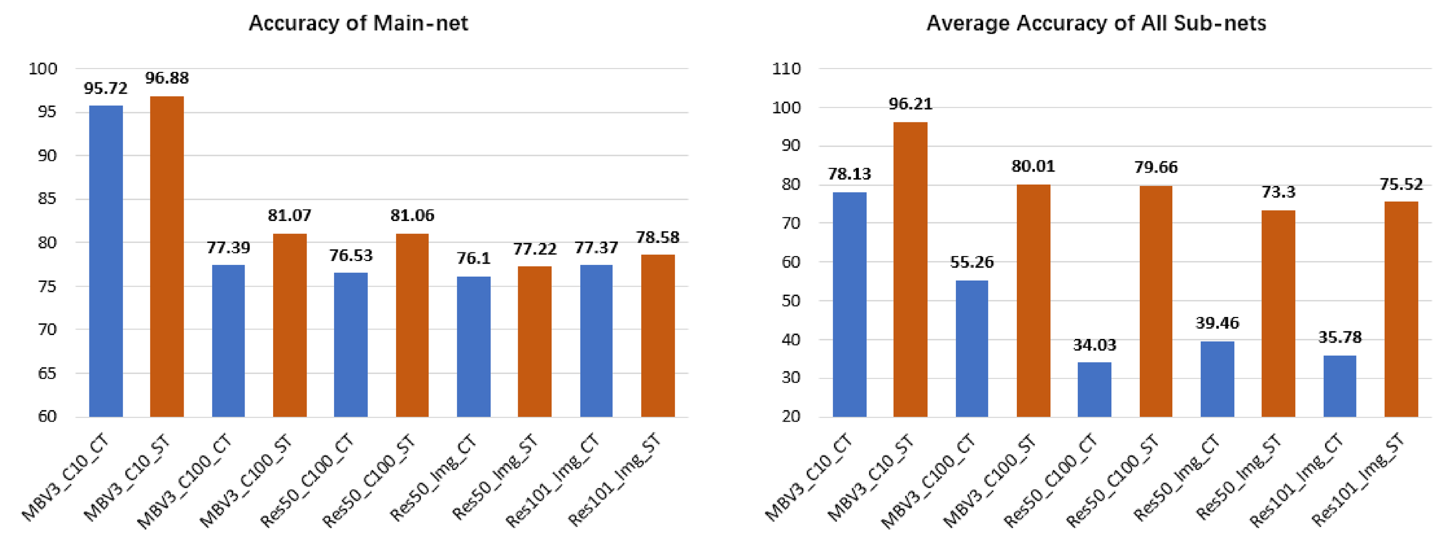
(a) Maximum Value (b) Mean Value (c) Standard Deviation

➢Stimulative training provides stronger robustness in resisting various network destruction operations



(a) deleting one (b) deleting more (c) permuting

➢Experimental results on various models and datasets



➢Please refer to our paper for more experimental and theoretical analysis

More Discussions

➢Comprehensive comparisons among different methods

➢ Table 1: Comprehensive comparisons among different methods, including common training (CT), stimulative training (ST), common training (CT) with layer/stage supervision, Self-Distillation and Stochastic Depth

Method	Time	Memory	Main(%)	All(%)
CT	16.91h	3291MiB	77.39	55.26±13.37
CT+layer supervision	23.3h	7193MiB	78.77	59.18±11.12
CT+stage supervision	19.3h	5197MiB	78.59	54.82±13.31
Self distillation	26.8h	3887MiB	79.59	50.39±14.22
Stochastic depth	**16.9h**	**3291MiB**	78.43	70.72±3.76
ST	24.08h	**3291MiB**	**81.07**	**80.01±0.59**

➢The loafing problem also exists on DenseNet networks

➢ Table 2: Different DenseNet networks trained on ImageNet invariably suffer from the problem of network loafing

Main-net\Sub-net	DenseNet121	DenseNet169	DenseNet201
DenseNet121	74.86	20.91	11.57
DenseNet169	-	76.46	51.18
DenseNet201	-	-	77.44

➢ Table 3: Different DenseNet networks trained on CIFAR100 invariably suffer from the problem of network loafing

Main-net\Sub-net	DenseNet121	DenseNet169	DenseNet201	DenseNet264
DenseNet121	78.84	43.64	31.51	10.01
DenseNet169	-	79.64	70.78	48.41
DenseNet201	-	-	79.77	62.29
DenseNet264	-	-	-	79.81

➢Discussions&Limitations

➢Residual structure is widely applied in numerous different types of models including **DenseNet** and **transformer**. It would be of vital value to study whether the loafing problem exists in these models and explore the proper method to solve this problem.

➢The proposed method suffers from about **1.4 times** of computation cost of the original training to get **better performance and robustness**. As the first research of network loafing problem, the proposed method is a positive pioneer-like exploration. We believe designing a more efficient method to solve the loafing problem is a worthy research direction in the future.