INITIALIZING MODELS WITH LARGER ONES

Zhiqiu Xu[1], **Yanjie Chen**[2], **Kirill Vishniakov**[3], **Yida Yin**[2], **Zhiqiang Shen**[3], **Trevor Darrell**[2], **Lingjie Liu**[1], **Zhuang Liu**[4]

[1]University of Pennsylvania [2]UC Berkeley [3]MBZUAI [4]Meta AI Research

ICLR2024 (Under Review-8686)

Paper Reading by Zhiying Lu

2023.12.05

智能多媒体内容计算实验室

Intelligent Multimedia Content Computing Lab

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结

作者介绍

Zhiqiu Xu

FOLLOW

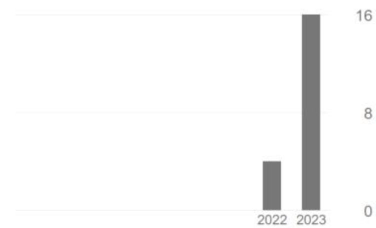
PhD student, University of Pennsylvania
Verified email at berkeley.edu - Homepage

Deep Learning   Representation Learning   Computer Vision   Natural Language Processing

TITLE	CITED BY	YEAR
Dropout Reduces Underfitting Z Liu*, Z Xu*, J Jin, Z Shen, T Darrell International Conference on Machine Learning (ICML), 2023	11	2023
Anytime dense prediction with confidence adaptivity Z Liu, Z Xu, HJ Wang, T Darrell, E Shelhamer International Conference on Learning Representations (ICLR), 2022	9	2021
A Coefficient Makes SVRG Effective Y Yin, Z Xu, Z Li, T Darrell, Z Liu arXiv preprint arXiv:2311.05589		2023

Cited by

	All	Since 2018
Citations	20	20
h-index	2	2
i10-index	1	1



Co-authors

Zhiqiang Shen

FOLLOW

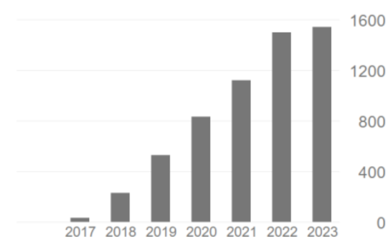
Assistant Professor at Mohamed bin Zayed University of Artificial Intelligence
Verified email at mbzuai.ac.ae - Homepage

Machine Learning   Computer Vision   Efficient Networks   Knowledge Distillation

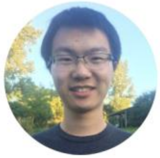
TITLE	CITED BY	YEAR
Learning efficient convolutional networks through network slimming Z Liu, J Li, Z Shen, G Huang, S Yan, C Zhang IEEE International Conference on Computer Vision (ICCV) 2017	2473	2017
Dsod: Learning deeply supervised object detectors from scratch Z Shen, Z Liu, J Li, YG Jiang, Y Chen, X Xue IEEE International Conference on Computer Vision (ICCV) 2017	810 *	2017
ReActNet: Towards Precise Binary Neural Network with Generalized Activation Functions Z Liu, Z Shen, M Savvides, KT Cheng European Conference on Computer Vision (ECCV), 2020	277	2020

Cited by

	All	Since 2018
Citations	5871	5795
h-index	26	26
i10-index	39	39



作者介绍



Zhuang Liu

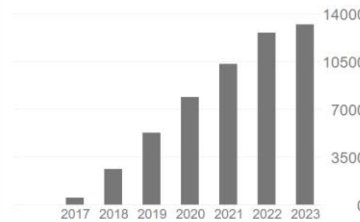
FOLLOW

Research Scientist, Meta AI Research
Verified email at berkeley.edu - Homepage
Deep Learning AI Computer Vision Machine Learning Neural Networks

TITLE	CITED BY	YEAR
Densely Connected Convolutional Networks G Huang*, Z Liu*, L Maaten, KQ Weinberger, *equal contribution Computer Vision and Pattern Recognition (CVPR), 2017	41068	2017
A ConvNet for the 2020s Z Liu, H Mao, CY Wu, C Feichtenhofer, T Darrell, S Xie Computer Vision and Pattern Recognition (CVPR), 2022	2712	2022
Learning Efficient Convolutional Networks through Network Slimming Z Liu, J Li, Z Shen, G Huang, S Yan, C Zhang International Conference on Computer Vision (ICCV), 2017	2473	2017
Deep Networks with Stochastic Depth G Huang*, Y Sun*, Z Liu, D Sedra, KQ Weinberger European Conference on Computer Vision (ECCV), 2016	2398	2016

Cited by

	All	Since 2018
Citations	53379	52337
h-index	18	18
i10-index	23	23



Public access VIEW ALL

0 articles 11 articles

not available available

Trevor Darrell

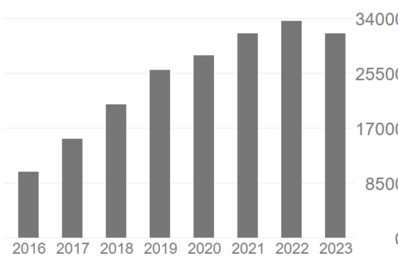
FOLLOW

Professor of Computer Science, U.C. Berkeley
Verified email at eecs.berkeley.edu - Homepage
Computer Vision Artificial Intelligence AI Machine Learning Deep Learning

TITLE	CITED BY	YEAR
Fully convolutional networks for semantic segmentation J Long, E Shelhamer, T Darrell Proceedings of the IEEE conference on computer vision and pattern …	44920	2015
Rich feature hierarchies for accurate object detection and semantic segmentation R Girshick, J Donahue, T Darrell, J Malik Proceedings of the IEEE conference on computer vision and pattern …	33944	2014
Caffe: Convolutional architecture for fast feature embedding Y Jia, E Shelhamer, J Donahue, S Karayev, J Long, R Girshick, … Proceedings of the 22nd ACM international conference on Multimedia, 675-678	17878	2014

Cited by VIEW ALL

	All	Since 2018
Citations	235638	172548
h-index	160	120
i10-index	458	348



室
Lab

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结

研究背景

- 模型训练的初始化可以帮助更好地训练
- 很多方法考虑random init模型并进行train from scratch
- 现有的大量预训练模型提供了网络初始化的另一种可能性
- 本文考虑利用大的预训练模型来初始化小模型

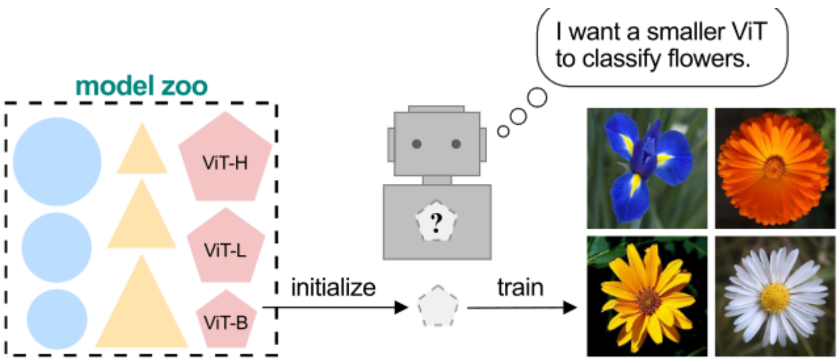


Figure 1: Large pretrained models offer new opportunities for initializing small models.

从零初始化—Xavier init

Understanding the difficulty of training deep feedforward neural networks

Xavier Glorot **Yoshua Bengio**
DIRO, Université de Montréal, Montréal, Québec, Canada

- 旨在保持激活函数的方差前向和反向传播过程中大致相同
- 避免梯度消失或爆炸的问题

2. 数学原理

考虑一个简单的全连接层，该层接受 n_{in} 个输入并产生 n_{out} 个输出。每个输出 o_i 可以表示为：

$$o_i = \text{activation}(\sum_{j=1}^{n_{in}} w_{ij} x_j + b_i)$$

其中 w_{ij} 是输入 x_j 到输出 o_i 的权重，b_i 是输出 o_i 的偏置，activation 是激活函数。

如果输入 x 的方差为 $Var(x)$，则线性函数 $\sum_{j=1}^{n_{in}} w_{ij} x_j$ 的方差将是 $n_{in} \times Var(w) \times Var(x)$（忽略偏执和激活函数）。

Xavier 初始化试图使得每一层的输出的方差接近于其输入的方差。具体地，它设置权重 w 的初始方差为：

$$Var(w) = \frac{2}{n_{in} + n_{out}}$$

这样，无论 n_{in} 和 n_{out} 的大小如何，这一层的输出方差都接近于其输入方差。

`torch.nn.init.xavier_uniform_` 函数从均匀分布 $U(-v, v)$ 中抽取权重，其中

$$v = \sqrt{3 \times Var(w)} = \sqrt{\frac{6}{n_{in} + n_{out}}}$$

`torch.nn.init.xavier_normal_` 函数从正态分布 $N(0, \sigma^2)$ 中抽取权重，其中

$$\sigma = \sqrt{Var(w)} = \sqrt{\frac{2}{n_{in} + n_{out}}}$$

n_{in} 和 n_{out} 分别是权重的输入节点数和输出节点数。

从零初始化—Kaiming init

Delving Deep into Rectifiers:
Surpassing Human-Level Performance on ImageNet Classification

Kaiming He　Xiangyu Zhang　Shaoqing Ren　Jian Sun

Microsoft Research

- 使网络每一层的输入输出方差尽可能相等，避免梯度消失或爆炸的问题
- Xavier是针对tanh和sigmoid激活函数设置的，不满足ReLU情况，此时需要用到Kaiming init

三、均匀分布

设参数w服从均匀在[-a, a]区间内均匀分布，则w的方差为：

$$D(w)=\frac{(a+a)^2}{12}=\frac{4a^2}{12}=\frac{a^2}{3}=\frac{2}{n_{in}}$$

所以

$$a=\sqrt{\frac{6}{n_{in}}}$$

即w的是均匀分布在 $(-\sqrt{\frac{6}{n_{in}}},\sqrt{\frac{6}{n_{in}}})$ 上的随机变量。

四、正态分布

如果我们假设w是服从正态分布的，则w服从

$$w\sim N(0,\sqrt{\frac{2}{n_{in}}})$$

五、Pyotrch实现

```
nn.init.kaiming_uniform_
nn.init.kaiming_normal_
```

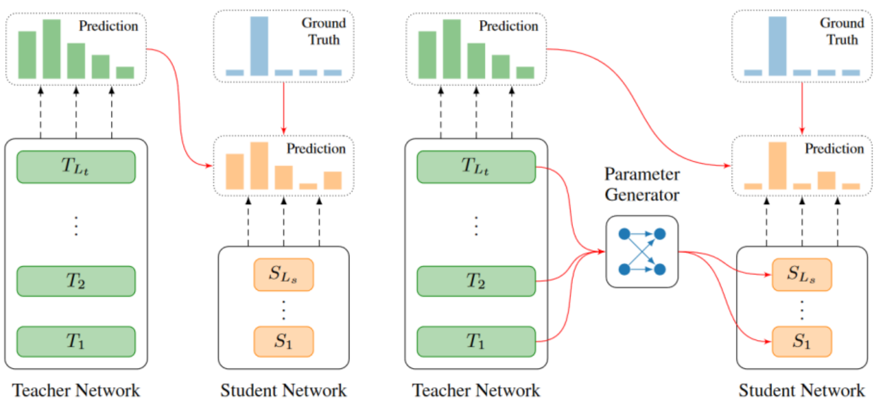
值得注意的是，kaiming方法并没有gain增益系数，只有a的一个修正系数，实际公式如下：

$$bound=\sqrt{\frac{6}{(1+a^2)n_{in}}}$$

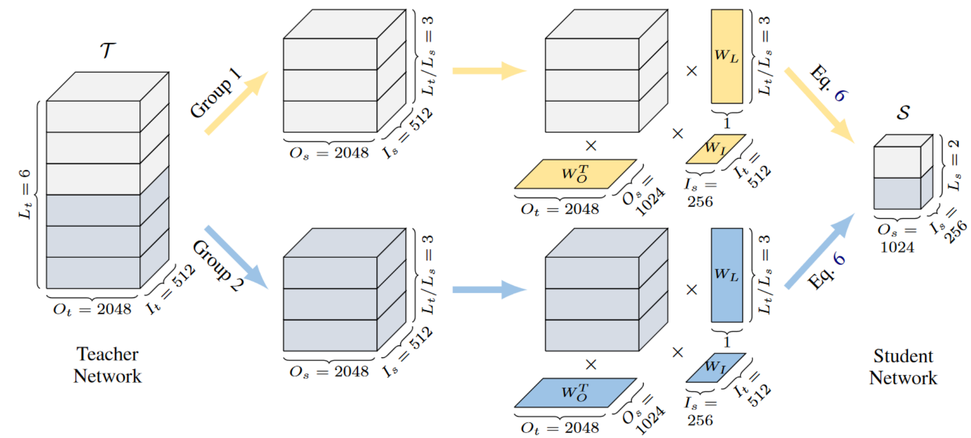
权重蒸馏

Weight Distillation: Transferring the Knowledge in Neural Network Parameters

Ye Lin[1]*, Yanyang Li[2]*, Ziyang Wang[1], Bei Li[1], Quan Du[1], Tong Xiao[1,3], Jingbo Zhu[1,3]†
[1]NLP Lab, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
[2]The Chinese University of Hong Kong, Hong Kong, China
[3]NiuTrans Research, Shenyang, China



(a) Knowledge Distillation (b) Weight Distillation



- 可学习的权重变换矩阵

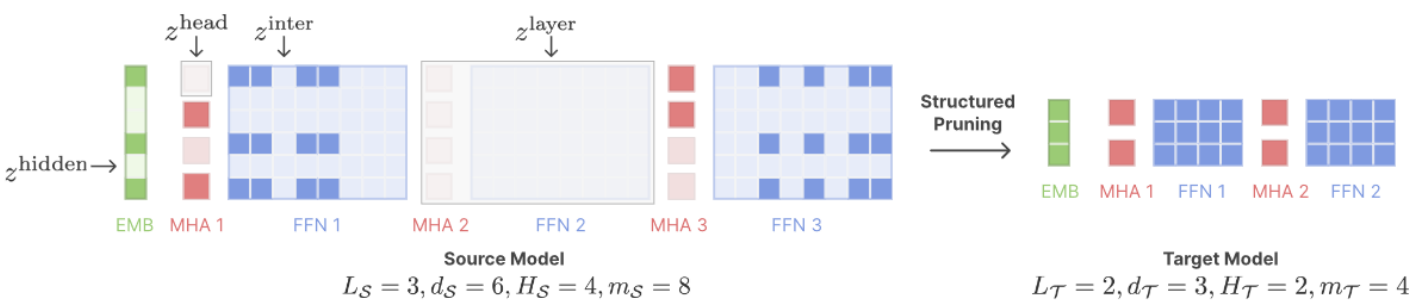
$$\mathcal{S} = \tanh(\hat{\mathcal{T}}) \odot W + B$$

$$\bar{\mathcal{S}} = \arg\min_{\mathcal{S}} [(1-\alpha)\mathcal{L}(y_{\mathcal{T}}, y_{\mathcal{S}}) + \alpha\mathcal{L}(y, y_{\mathcal{S}})]$$

权重裁剪

SHEARED LLAMA: ACCELERATING LANGUAGE MODEL PRE-TRAINING VIA STRUCTURED PRUNING

Mengzhou Xia[1], Tianyu Gao[1], Zhiyuan Zeng[2]*, Danqi Chen[1]
[1]Department of Computer Science & Princeton Language and Intelligence, Princeton University
[2]Department of Computer Science and Technology, Tsinghua University



- 学习应该mask哪些权重

Granularity	Layer	Hidden dimension	Head	Intermediate dimension
Pruning masks	$z^{\text{layer}} \in \mathbb{R}^{L_{\mathcal{S}}}$	$z^{\text{hidden}} \in \mathbb{R}^{d_{\mathcal{S}}}$	$z^{\text{head}} \in \mathbb{R}^{H_{\mathcal{S}}}$ $(\times L_{\mathcal{S}})$	$z^{\text{int}} \in \mathbb{R}^{m_{\mathcal{S}}}$ $(\times L_{\mathcal{S}})$

$$\tilde{\mathcal{L}}^{\text{head}}(\lambda, \phi, z) = \lambda^{\text{head}} \cdot \left(\sum z^{\text{head}} - H_{\mathcal{T}}\right) + \phi^{\text{head}} \cdot \left(\sum z^{\text{head}} - H_{\mathcal{T}}\right)^2.$$

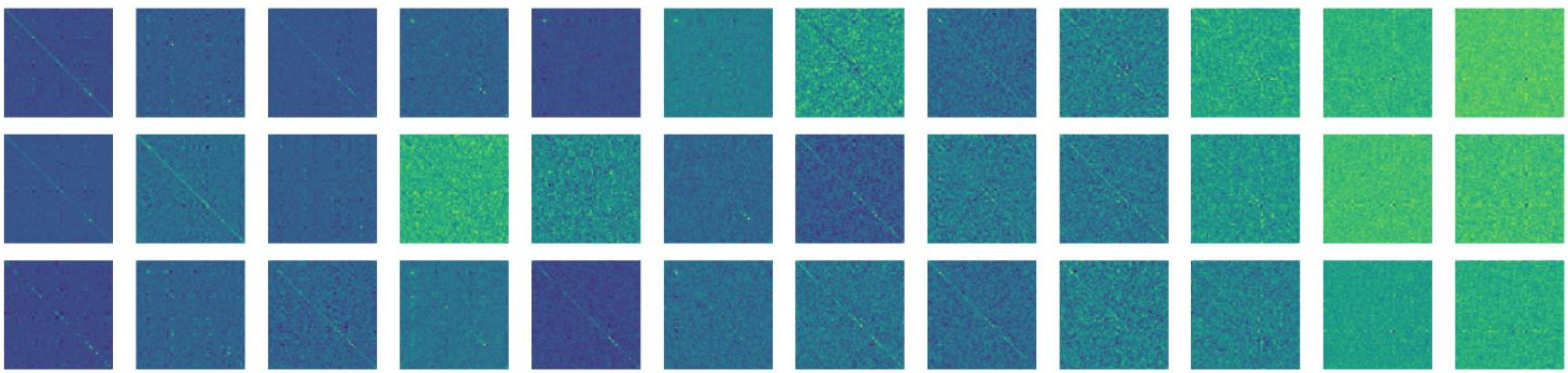
$$\mathcal{L}_{\text{prune}}(\theta, z, \lambda, \phi) = \mathcal{L}(\theta, z) + \sum_{j=1}^{L_{\mathcal{S}}} \tilde{\mathcal{L}}_j^{\text{head}} + \sum_{j=1}^{L_{\mathcal{S}}} \tilde{\mathcal{L}}_j^{\text{int}} + \tilde{\mathcal{L}}^{\text{layer}} + \tilde{\mathcal{L}}^{\text{hidden}},$$

权重初始化

Mimetic Initialization of Self-Attention Layers

Asher Trockman [1] J. Zico Kolter [1 2]

(a) $W_Q W_K^T$ often has a noticeable positive diagonal. → Layers 1-12, ↓ Attention Heads 1-3



(b) $W_V W_{proj}$ often has a prominent negative diagonal. Here, we sum over heads.

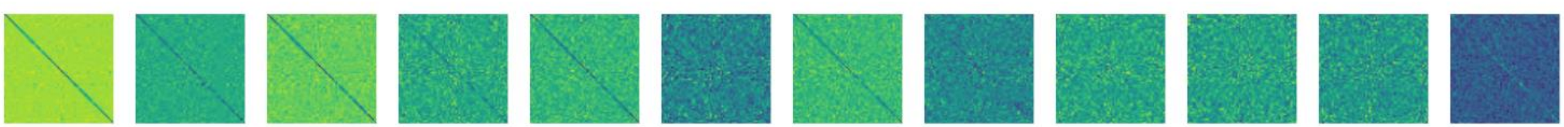


Figure 1. Self-attention weights of an ImageNet-pretrained ViT-Tiny. Pictured are 3 heads for each of the 12 layers. Clipped to 64x64.

- 初始化模型使其注意力具有对角线性质 $\text{Softmax}\left(\frac{1}{\sqrt{k}}(\beta_1 dI + \beta_1 PP^T)\right).$

$$W_Q W_K^T = \alpha Z + \beta I \quad \mathbb{E}[(X + P)(\alpha Z + \beta I)(X + P)^T] = \beta dI + \beta PP^T.$$

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结

Weight Selection

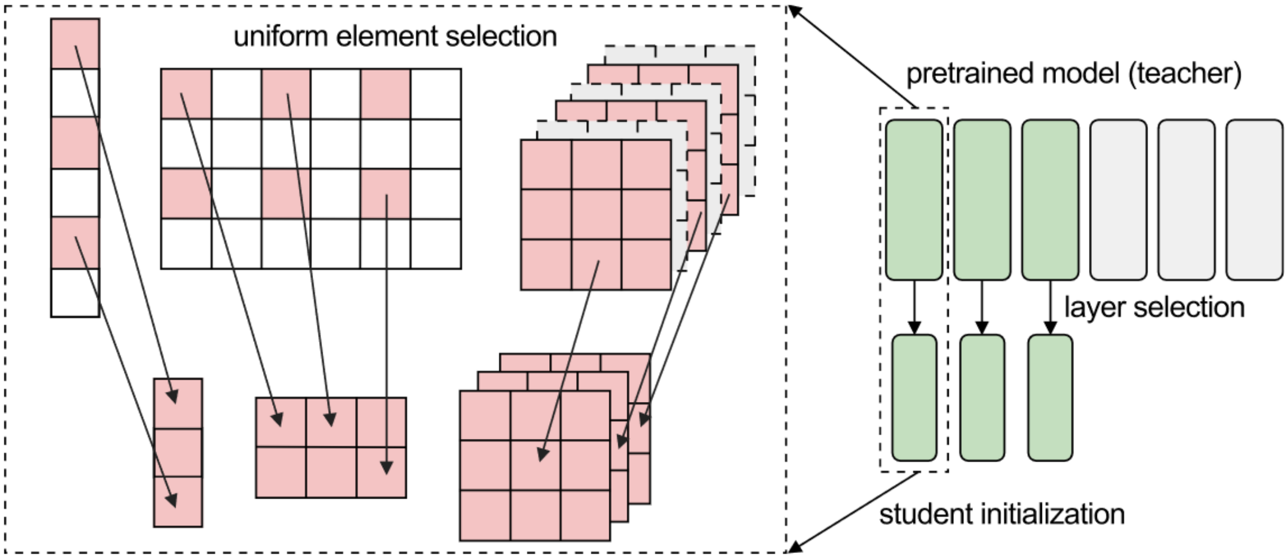


Figure 2: **Weight selection.** To initialize a smaller variant of a pretrained model, we uniformly select parameters from the corresponding component of the pretrained model.

- 本文方法只利用large pretrained的权重做初始化，不会在训练过程中使用pretrain，没有额外可学习参数，也无需损失函数和蒸馏来监督训练过程

Weight Selection

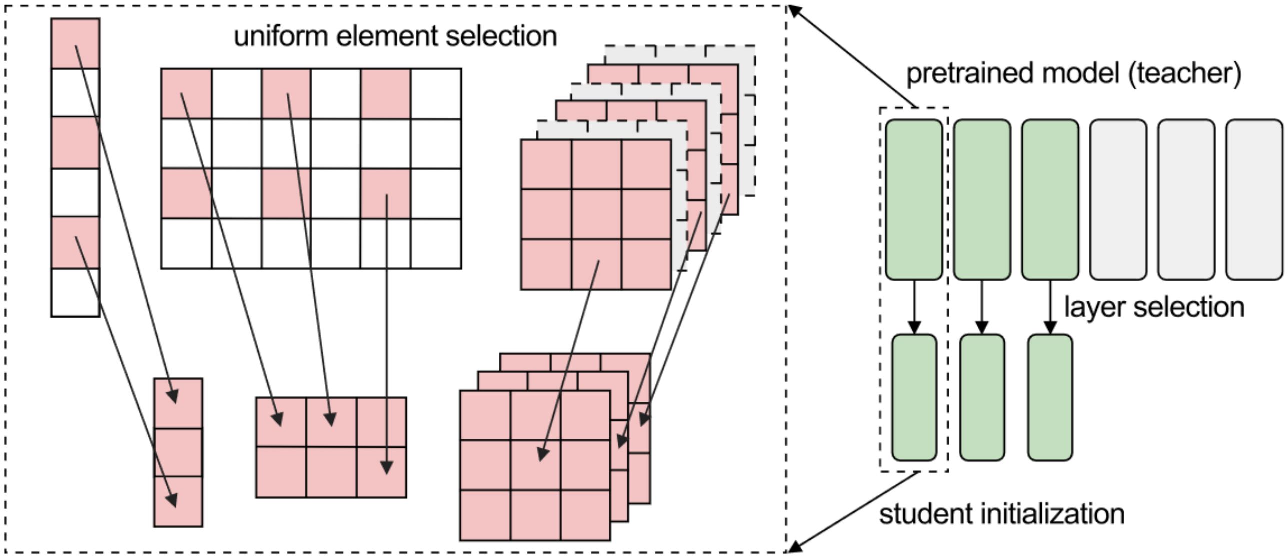


Figure 2: **Weight selection.** To initialize a smaller variant of a pretrained model, we uniformly select parameters from the corresponding component of the pretrained model.

- 包含三个步骤：
- Layer Selection, Component Mapping, Element Selection
- 选择层数，模块对应，选择元素

Weight Selection

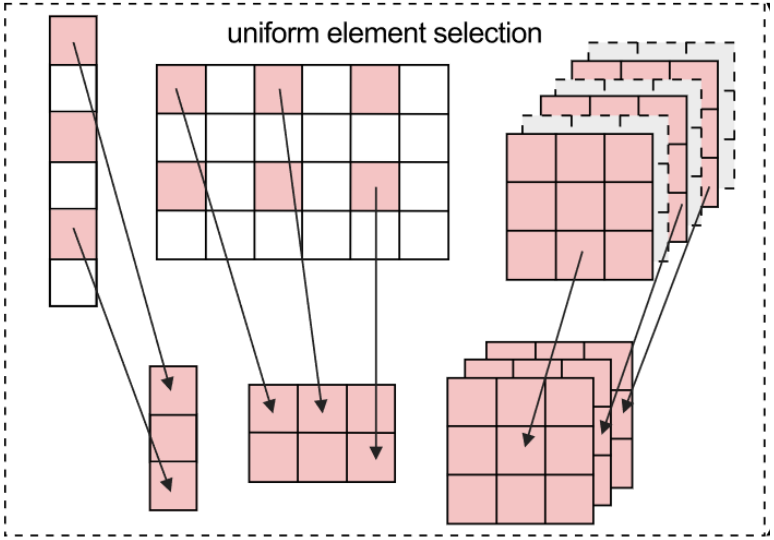
Layer Selection

- 默认采用first-N layer selection，即选择teacher网络连续N层
- 对于isotropic架构，直接选择前N层
- 对于hierarchical架构，每个stage选择前N层

Component Mapping

- 模块化的设计，一一对应
- Conv to Conv, Linear to Linear, Attn to Attn

Element Selection



Uniform Selection (Default)

- 均匀采样pretrain的参数，按照index来进行选择
- 例如dim=6的Linear层选择1, 3, 5维度
- 支持任意维度变换，可以利用线性插值

Consecutive Selection

- 连续选择一定区域的参数
- 例如9*9卷积选择左上角的3*3区域

Element Selection

Uniform Selection (Default)

- 均匀采样pretrain的参数，按照index来进行选择
- 例如dim=6的Linear层选择1, 3, 5维度
- 支持任意维度变换，可以利用线性插值

Algorithm 1 Uniform element selection from teacher's weight tensor

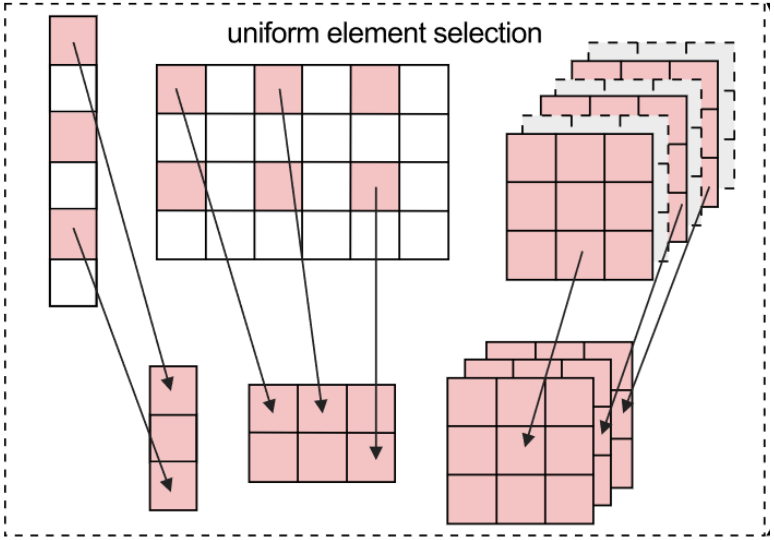
Input: W_t ▷ teacher's weight tensor
Input: s ▷ desired dimension for student's weight tensor
Output: W_s with shape s

```
procedure UNIFORMELEMENTSELECTION(W_t, student_shape)
    W_s ← Copy of W_t                                   ▷ student's weight tensor
    n ← length of W_t.shape
    for i = 1 → n do
        d_t ← W_t.shape[i]
        d_s ← s[i]
        indices ← Select d_s evenly-spaced numbers from 1 to d_t
        W_s ← Select indices along W_s's i^th dimension
    end for
    return W_s
end procedure
```

Element Selection

Random w/ consistency

- 随机采样，但对于每个参数，都选择固定位置的
- 例如随机一组index，对于所有卷积核均按照这一组index选择



Random w/o consistency

- 对于所有参数，所有index完全随机

- 作者介绍
- 研究背景
- Tip-Adapter
- 实验效果
- 总结

Experiment

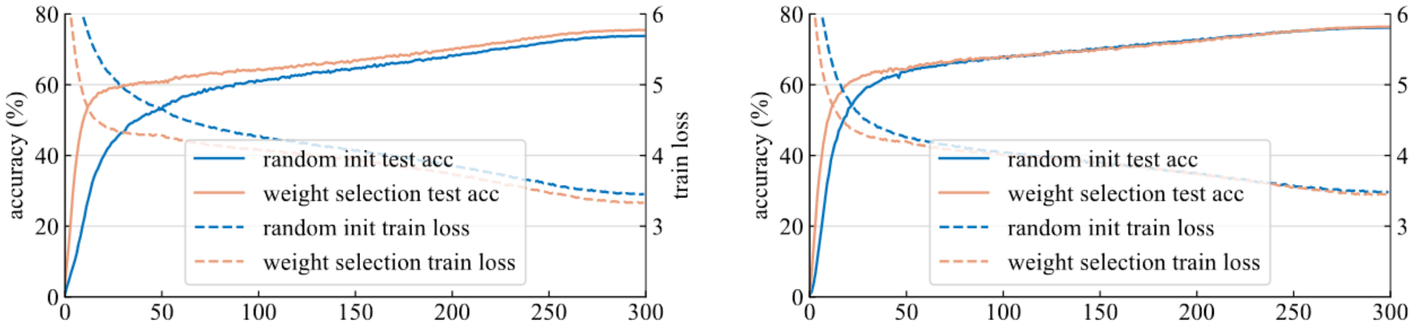
configuration	student		teacher	
model	ViT-T	ConvNeXt-F	ViT-S	ConvNeXt-T
depth	12	2 / 2 / 6 / 2	12	3 / 3 / 9 / 3
embedding dimension	192	96 / 192 / 384 / 768	384	48 / 96 / 192 / 384
number of heads	3	-	6	-
number of parameters	5M	5M	22M	28M

Table 1: **Model Configurations.** We perform main experiments on ConvNeXt and ViT, and use student that halve the embedding dimensions of their corresponding teacher.

dataset (scale ↓)	random init	weight selection	change	random init	weight selection	change
ImageNet-1K	73.9	75.6	↑1.6	76.1	76.4	↑0.3
SVHN	94.9	96.5	↑1.6	95.7	96.9	↑1.2
Food-101	79.6	86.9	↑7.3	86.9	89.0	↑2.1
EuroSAT	97.5	98.6	↑1.1	98.4	98.8	↑0.4
CIFAR-10	92.4	97.0	↑4.6	96.6	97.4	↑0.8
CIFAR-100	72.3	81.4	↑9.1	81.4	84.4	↑3.0
STL-10	61.5	83.4	↑21.9	81.4	92.3	↑10.9
Flowers	62.4	81.9	↑19.5	80.3	94.5	↑14.2
Pets	25.0	68.6	↑43.6	72.9	87.3	↑14.4
DTD	49.4	62.5	↑13.1	63.7	68.8	↑5.1

(a) ViT-T (b) ConvNeXt-F

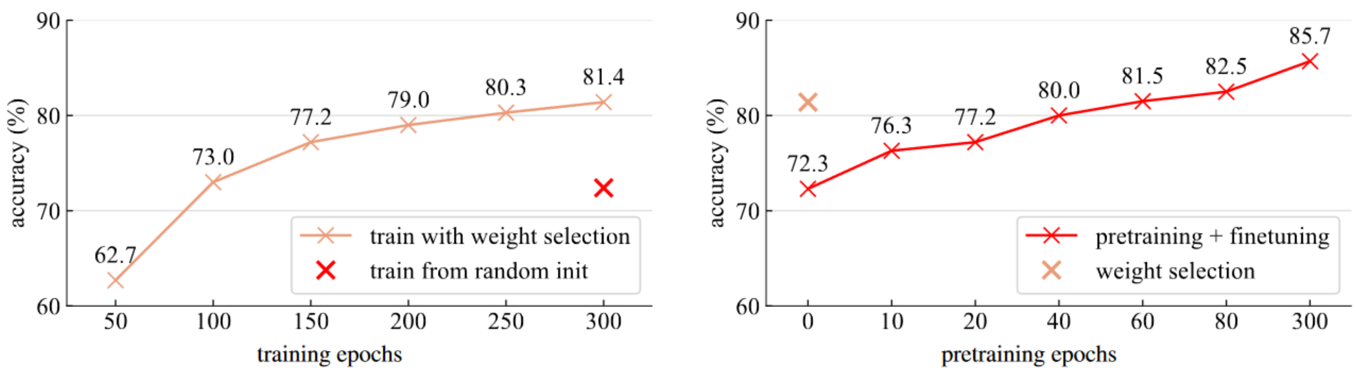
Experiment



(a) ViT-T

(b) ConvNeXt-F

Figure 3: **Training curves on ImageNet-1K.** When initialized using weight selection from ImageNet-21K pretrained models, both ViT-T (from ViT-S) and ConvNeXt-F (from ConvNeXt-T) exhibit superior performance compared to their randomly-initialized counterparts.



(a) Comparison with random initialization

(b) Comparison with pretraining + finetuning

Figure 4: **Faster training.** Compared to random initialization, ViT-T can reach the same performance on CIFAR-100 with only 1/3 epochs compared to training from random initialization. When compared to pretraining (on ImageNet-1K) + finetuning, weight selection is able to match the accuracy at 60 epochs of pretraining, saving 6.12x training time.

消融

init	ViT-T	ConvNeXt-F
timm default (trunc normal)	72.3	81.4
Xavier (Glorot & Bengio, 2010)	72.1	82.8
Kaiming (He et al., 2015)	73	82.5
weight selection (uniform)	81.4	**84.4**
weight selection (consecutive)	81.6	84.0
weight selection (random w/ consistency)	**81.7**	83.9
weight selection (random w/o consistency)	77.4	82.8

Pretrained models	CIFAR-10	CIFAR-100	STL-10
supervised (ImageNet-21K)	95.1	**77.6**	**73.1**
CLIP (Radford et al., 2021)	94.9	77.3	66.0
MAE (He et al., 2022)	**95.9**	77.2	71.0
DINO (Caron et al., 2021)	95.0	75.7	69.4

setting	ViT-T	ConvNeXt-F
random init	72.3	81.4
weight selection	**81.4**	**84.4**
L_1 pruning	79.5	82.8
magnitude pruning	73.8	81.9

setting	ViT-A	ConvNeXt-F
random init	69.6	81.3
first-N layer selection	**77.6**	**84.4**
uniform layer selection	76.7	83.2

teacher	params	test acc
ViT-S	22M	**81.4**
ViT-B	86M	77.6
ViT-L	307M	76.9

消融

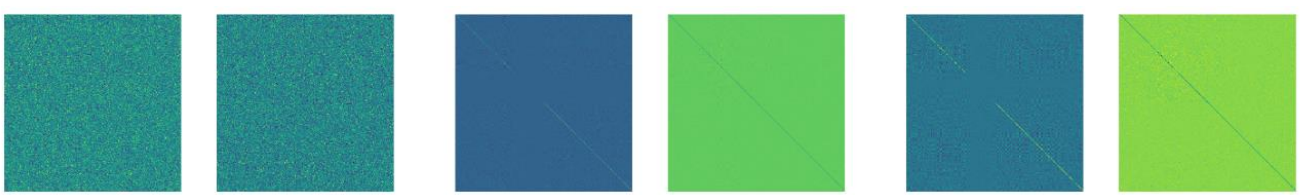
setting	ViT-T	ConvNeXt-F
random init	13.5	7.1
weight selection	**28.2**	**23.6**

Setting	CIFAR-10	CIFAR-100	STL-10
random init	92.4	72.3	61.5
weight selection	**97.0**	**81.4**	**83.4**
w/o patch embed	96.8	79.5	77.1
w/o pos embed	95.6	78.4	80.2
w/o attention	96.2	77.3	80.5
w/o normalization	96.2	79.0	79.8
w/o mlp	95.6	78.8	74.2

Table 10: **ViT component ablation.** Using all components from pretrained models is the best.

setting	ViT-T		ConvNeXt-F	
	test acc	change	test acc	change
random init	73.9	-	76.1	-
weight selection	**75.5**	↑1.6	**76.4**	↑0.3
random init (longer training)	76.3	-	77.5	-
weight selection (longer training)	**77.4**	↑1.1	**77.7**	↑0.2

setting	CIFAR-10	CIFAR-100	STL-10
random init	92.4	72.3	61.5
mimetic init	93.3	74.7	67.5
weight selection	**97.0**	**81.4**	**83.4**



(a) random init (b) pretrained ViT-S (c) weight selection

Figure 5: **Visualization of self-attention layers.** Visualization of $W_q W_k^T$ (left) and VW_{proj} (right) for ViT-T with random initialization, pretrained ViT-S, and ViT-T with weight selection. Weight selection can inherit the diagonal property of self-attention layers that only exists in pretrained ViTs.

消融

setting	CIFAR-100 test acc
first-N layer selection	**81.6**
mid-N layer selection	68.3
last-N layer selection	62.0
uniform layer selection	76.3

Table 16: **Layer selection.** First-N layer selection performs significantly better than uniform layer selection when ruling out the effect of element selection.

setting	CIFAR-100 test acc
first-N layer selection	76.9
mid-N layer selection	75.9
last-N layer selection	77.1
uniform layer selection	**77.5**

Table 17: **Layer selection (ViT-L as teacher).** Uniform layer selection yields slightly better results than first-N layer selection when student's ratio to teacher is small.

More

Model Configuration

Architecture	student	# of parameters	teacher	# of parameters
ResNet	ResNet-18	11.7M	ResNet-34	21.8M
Mlp-Mixer	Mixer-T/32	5.4M	Mixer-S/32	19.1M
Swin-Transformer	Swin-F	7.53M	Swin-T	28.5M
Pyramid Vision transformer	PVT-v2-b0	3.7M	PVT-v2-b1	14.0M

CIFAR-10

Setting / Model	ResNet	Mlp-Mixer	Swin Transformer	PVT
random init	96.4	90.8	94.9	96.3
weight selection	97.1	95.1	96.5	97.4

CIFAR-100

Setting / Model	ResNet	Mlp-Mixer	Swin Transformer	PVT
random init	80.3	72.3	79.0	81.5
weight selection	82.3	77.9	81.7	83.4

Setting	CIFAR-100 test acc
random init	72.3
ViT-B -> ViT-T	77.6
ViT-B -> ViT-S -> ViT-T	80.4

与知识蒸馏的兼容性

setting	ImageNet-1K (logit-based distillation)		CIFAR-100 (feature-based distillation)	
	test acc	change	test acc	change
baseline	73.9	-	72.3	-
distill	74.8	↑0.9	78.4	↑6.4
weight selection	**75.5**	↑1.6	**81.4**	↑9.1
distill + weight selection	**76.0**	↑2.1	**83.9**	↑11.6

Table 4: **Compatibility with knowledge distillation.** Weight selection is useful as an independent technique, and can be combined with knowledge distillation to achieve the best performance.

$$\mathcal{L} = \mathcal{L}_{class} + \alpha \cdot KL(p_t || p_s) \qquad \mathcal{L} = \mathcal{L}_{class} + \alpha \cdot L_1(O_t, MLP(O_s))$$

- 作者介绍
- 研究背景
- 方法
- 实验效果
- 总结

总结反思

- 一种无需预训练模型参与训练过程中的、权重初始化方式

仍有很多改进的地方：

- 选择的方式会导致信息的丢失--无损的参数选择与压缩
- 依靠经验性规则选取参数--带有语义的参数选择
- 受限于同种模型的初始化--任意模型到任意模型的初始化
- 太大的模型无法蒸馏到小模型--提升初始化的scaling性能

谢谢！