南京大学 NANJING UNIVERSITY

RINC Robotic Intelligence & Neural Computing Group

深度学习中的数据优化理论与增强选择方法研究

Research on Data Optimization in Deep Learning: Theory and Methods for Augmentation and Selection

答辩人：杨锁荣

指导老师：申富饶 教授

誠樸雄偉 勵學敦行

RINC
Robotic Intelligence & Neural Computing Group

目录

誠樸雄偉 勵學敦行

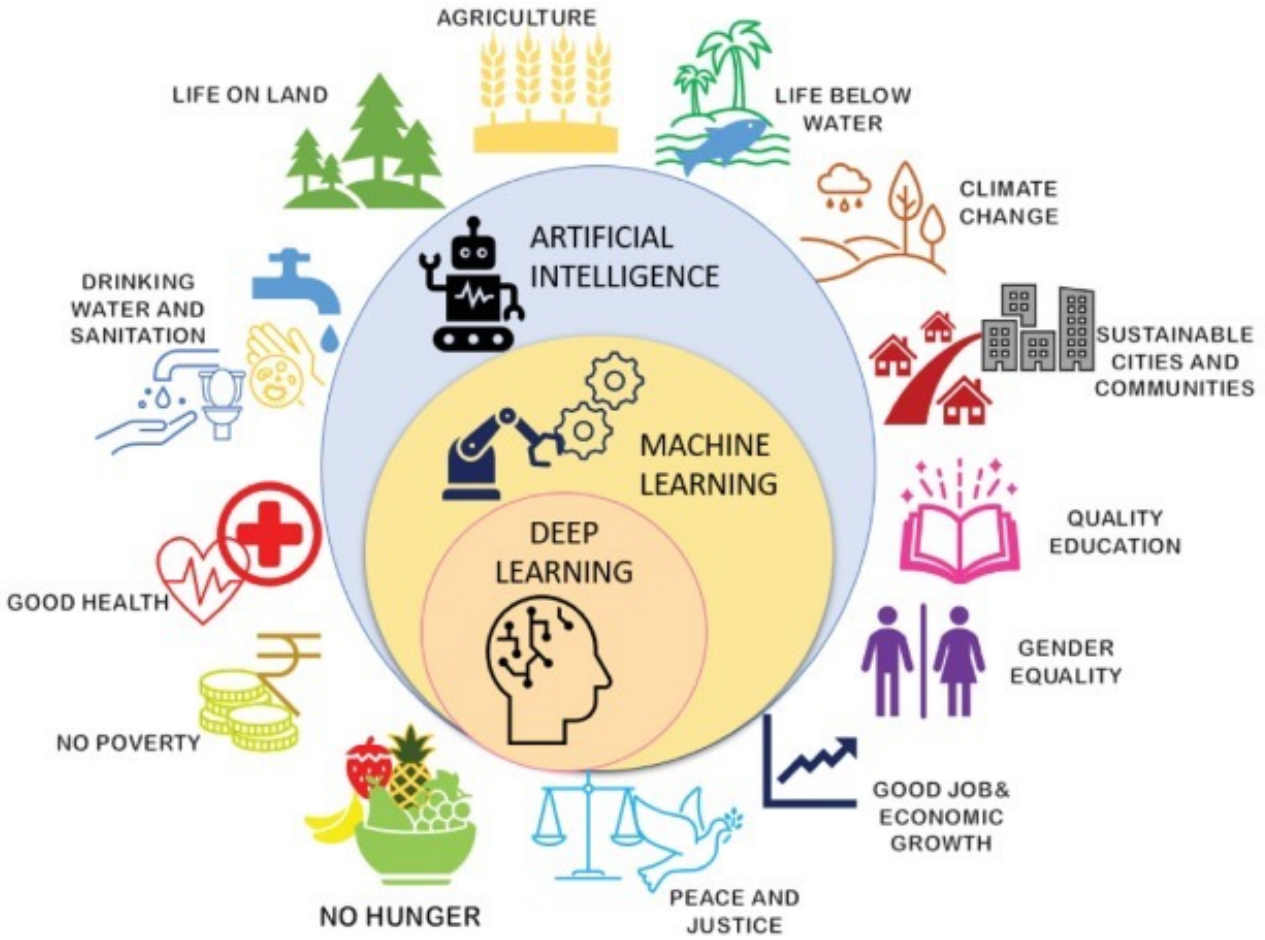
RINC
Robotic Intelligence & Neural Computing Group

以数据为中心的人工智能

Data-centric AI

誠樸雄偉 勵學敦行

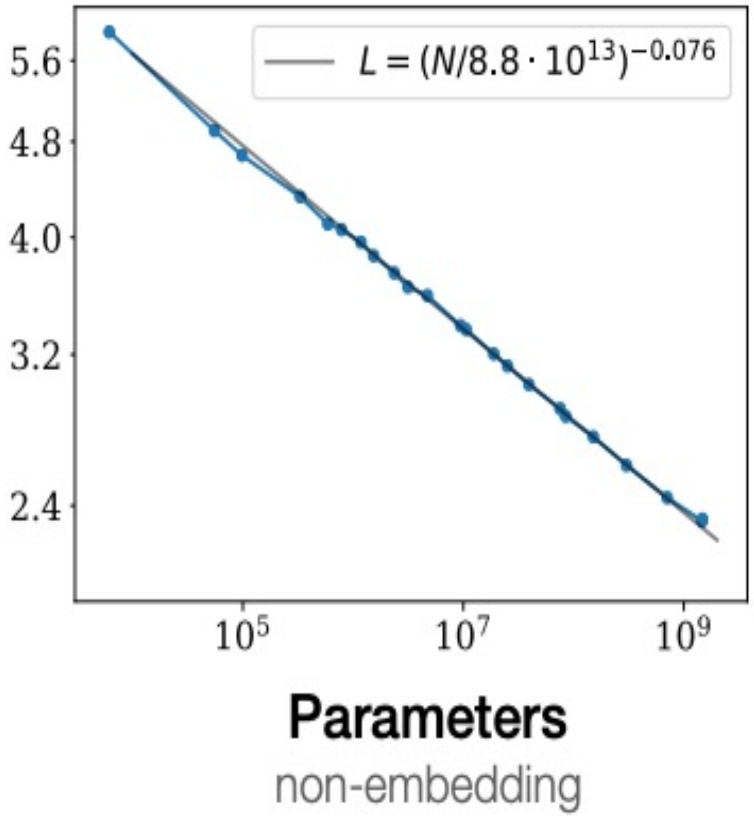
人工智能已经改变了很多领域，从自然科学到气候变化等



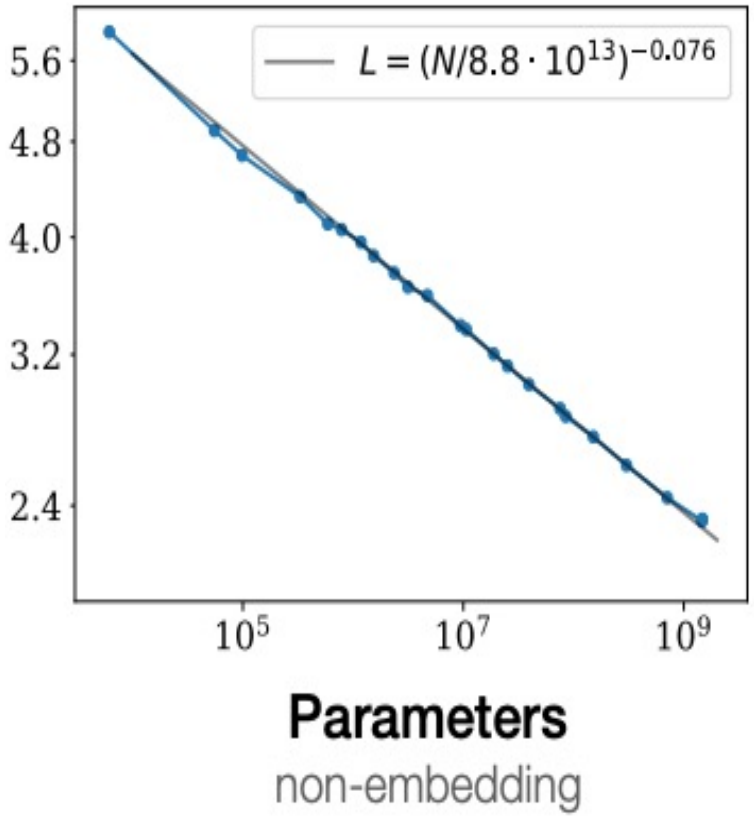
AlphaGo 2016 → Bert 2018 → ChatGPT 2022 → Gemini 2024

这些应用的背后都是海量的数据

Scaling Law： $L \propto N^{-\alpha_N}, D^{-\alpha_D}, C^{-\alpha_C}$



Scaling Law： $L \propto N^{-\alpha_N}, D^{-\alpha_D}, C^{-\alpha_C}$



模型越大，性能越好。 → **模型可以做的更大?**

nature

Explore content ∨ About the journal ∨ Publish with us ∨ Subscribe

nature › news feature › article

NEWS FEATURE | 11 December 2024

The AI revolution is running out of data. What can researchers do?

AI developers are rapidly picking the Internet clean to train large language models such as those behind ChatGPT. Here's how they are trying to get around the problem.

The Data That Powers A.I. Is Disappearing Fast

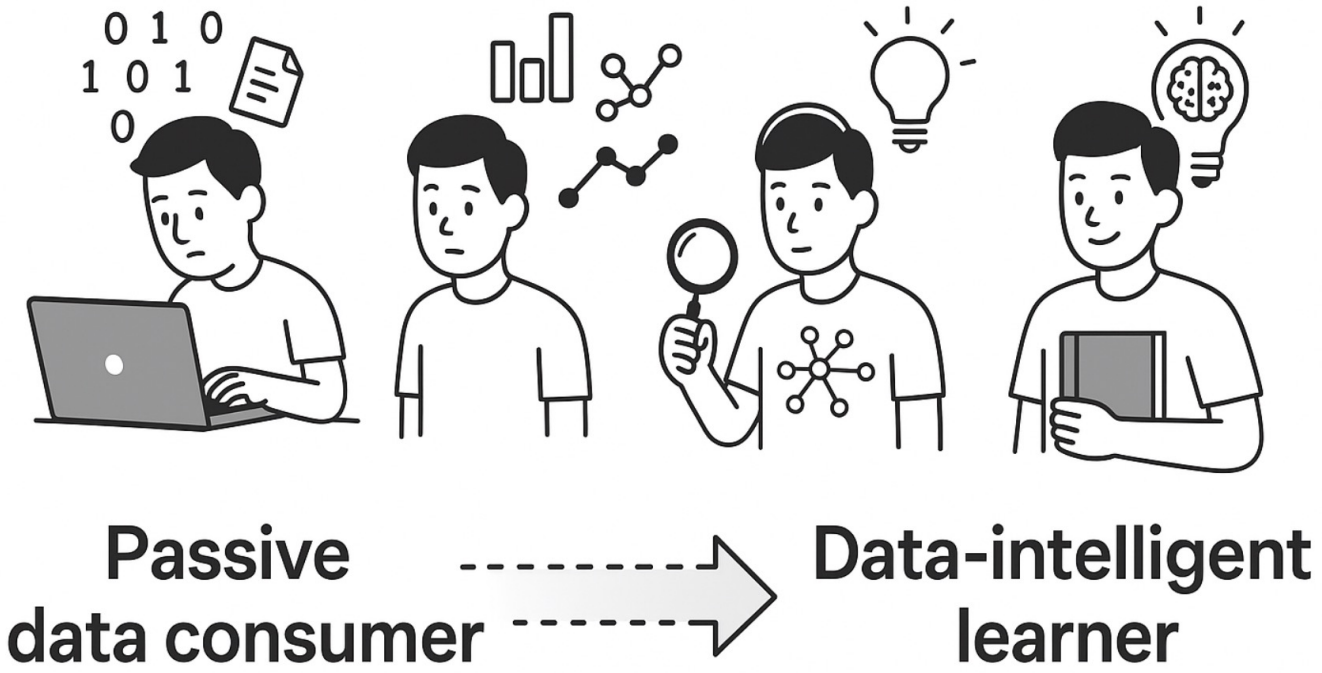
New research from the Data Provenance Initiative has found a dramatic drop in content made available to the collections used to build artificial intelligence.

To drive AI efficiency, CFOs must focus on its 'fuel': OneTrust

As the AI age dawns, CFOs need to be sure they're focusing on what actually makes that technology tick — data, CFO Guido Torrini says.

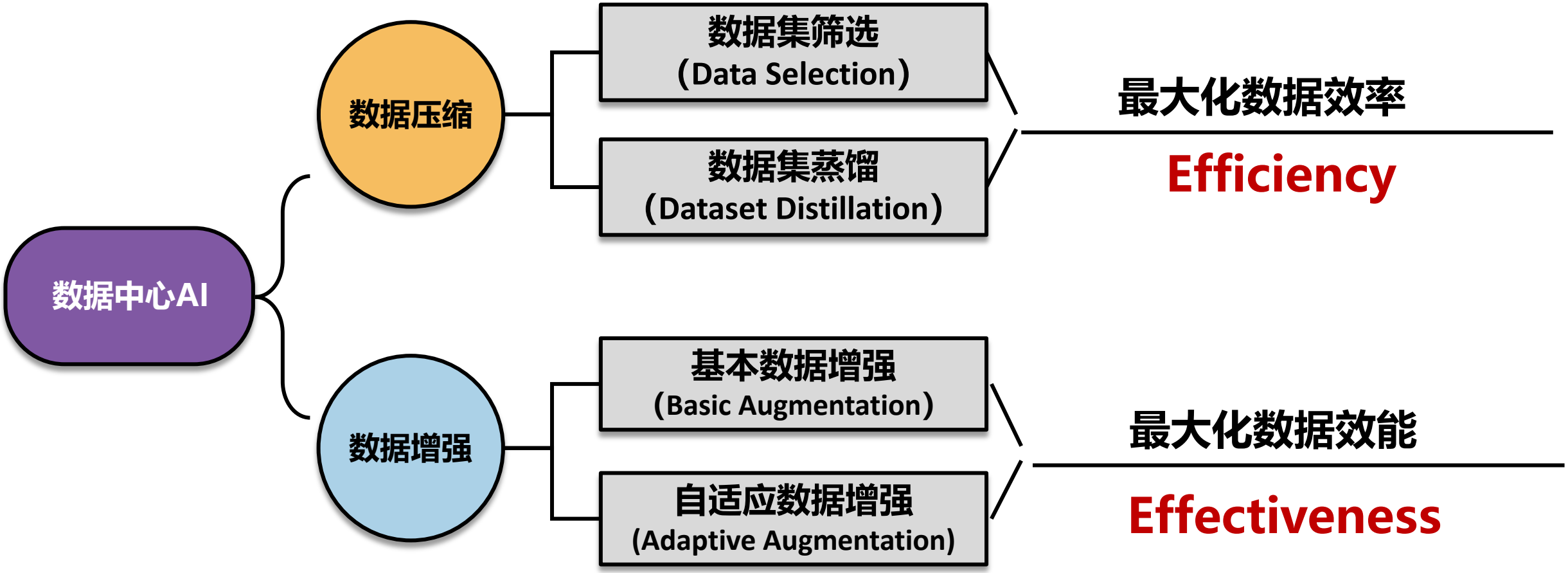
Scaling Compute已无法支撑AI的持续进步，瓶颈转向数据。

经典的训练范式

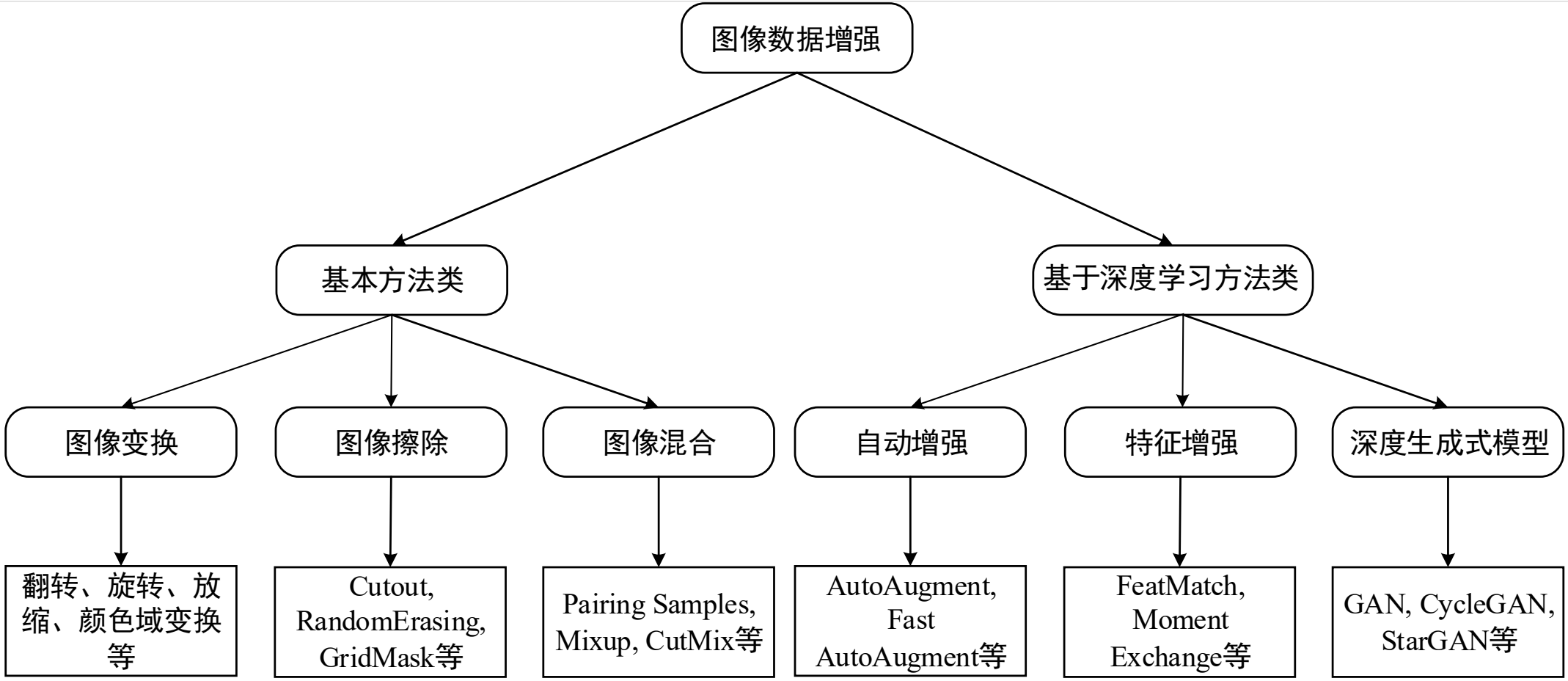


被动的数据消费 → 数据智能的学习

提升数据效率

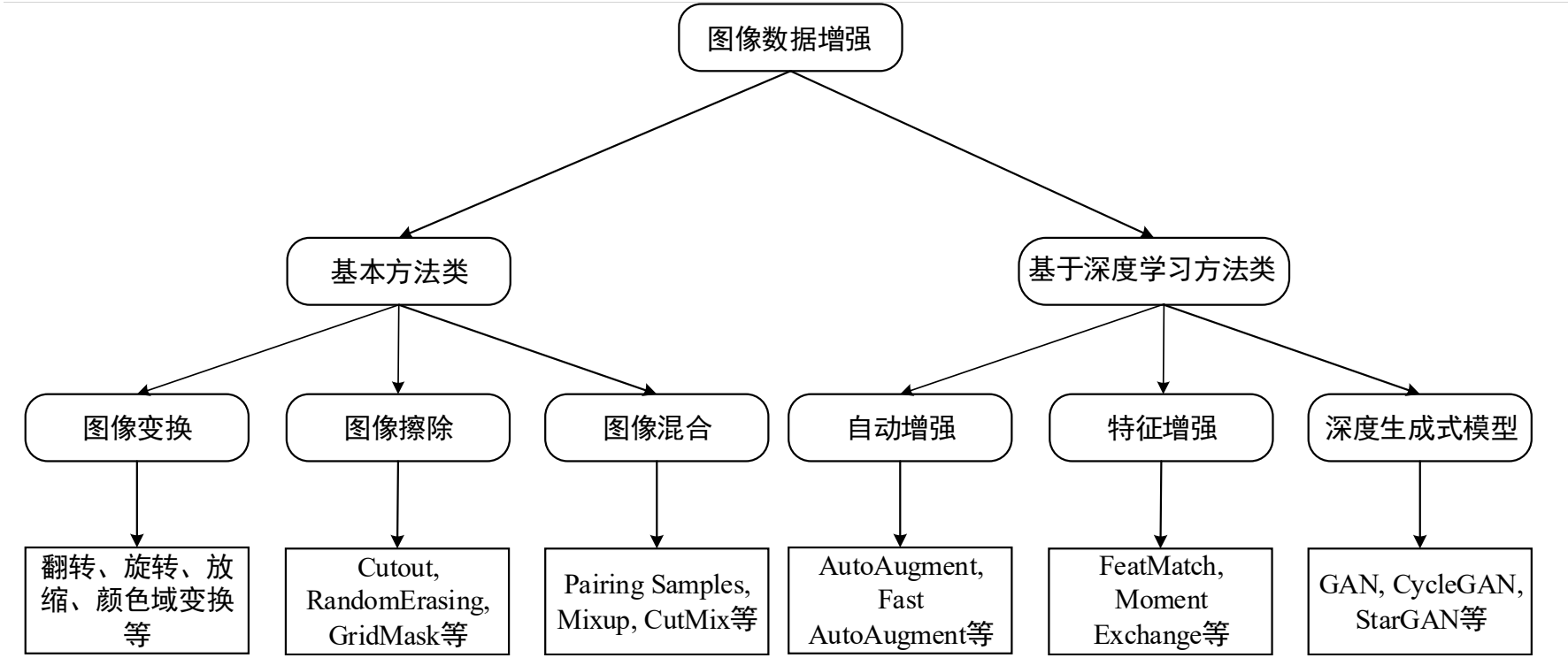

数据中心AI
数据压缩
数据增强
数据集筛选
(Data Selection)
数据集蒸馏
(Dataset Distillation)
基本数据增强
(Basic Augmentation)
自适应数据增强
(Adaptive Augmentation)
最大化数据效率
Efficiency
最大化数据效能
Effectiveness

数据增强

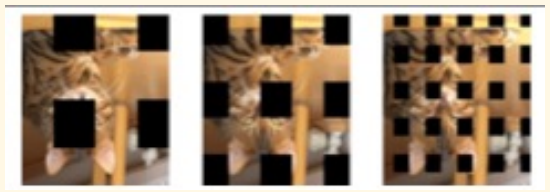


杨锁荣, 杨洪朝, 申富饶, 赵健. 面向深度学习的图像数据增强综述. 软件学报, 2025, 36(3):1390-1412

数据增强

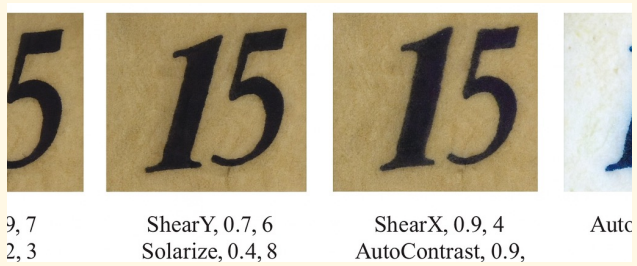


示例



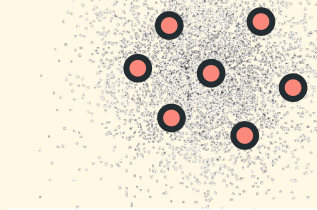
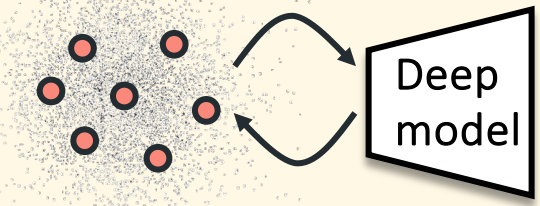
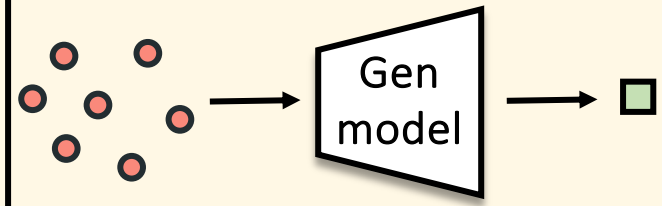
图像擦除

图像混合



自动增强

数据压缩

分类	 $max_{\text{Metric}}\ \mathcal{L}_{\text{model}}\left(\mathcal{D}_{\text{selected}}\right)$ **静态数据选择**	 **动态数据选择**	 **数据集蒸馏**
特点	**静态数据选择**，基于某个重要性指标在训练前筛选		
	动态数据选择，训练过程中动态选择		
	数据集蒸馏，通过生成式模型将数据集信息浓缩到少量样本中		

相似性-多样性理论分析框架

$$\min_{\Theta} \mathbb{E}_{i\sim p_\theta, T\sim q_\phi} \mathcal{L}\left(f_\Theta\left(T(x_i)\right), y_i\right)$$

训练数据的分析算法

实验验证与分析

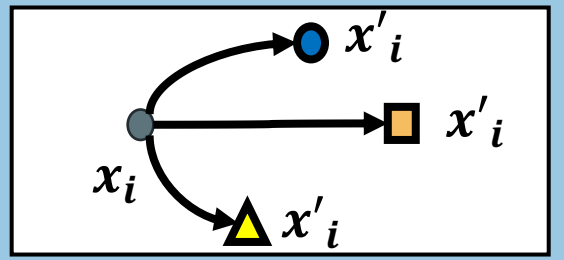
对应第三章内容

多样性提升的数据优化框架

$$\min_{\Theta} \mathbb{E}_{i\sim p_\theta, \mathrm{Ada}T\sim q_\phi} \mathcal{L}\left(f_\Theta\left(T(x_i)\right), y_i\right)$$

+数据增强分布的建模优化

第四章 自适应数据增强算法

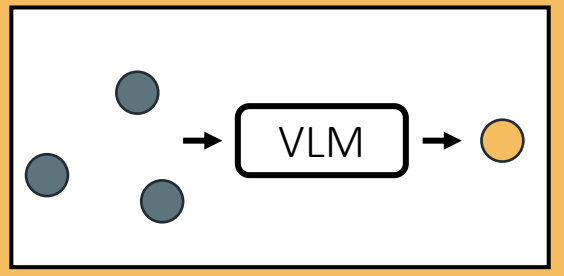


相似性驱动的数据优化框架

$$\min_{\Theta} \mathbb{E}_{i\sim p_\theta, T\sim q_\phi} \mathcal{L}\left(f_\Theta\left(T(x_i)\right), y_i\right)$$
$$\text{s.t. } \mathbb{E}_{i\sim \hat{p}_\theta}[S_A(i)] \geq \tau$$

+跨模态语义一致性度量模型

第五章 多模态数据选择算法

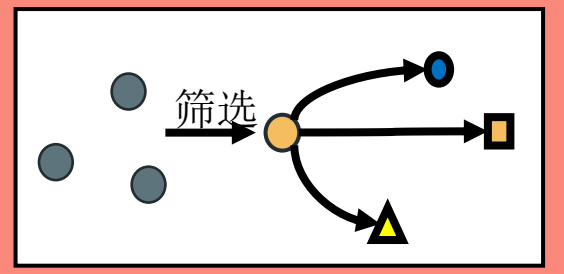


相似性-多样性联合优化框架

$$\min_{\Theta} \mathbb{E}_{i\sim p_\theta, \mathrm{Ada}T\sim q_\phi} \mathcal{L}\left(f_\Theta\left(T(x_i)\right), y_i\right)$$
$$\text{s.t. } \mathbb{E}_{i\sim \hat{p}_\theta}[S_A(i)] \geq \tau,$$

+协同互补的优化框架

第六章 增强与选择协同算法

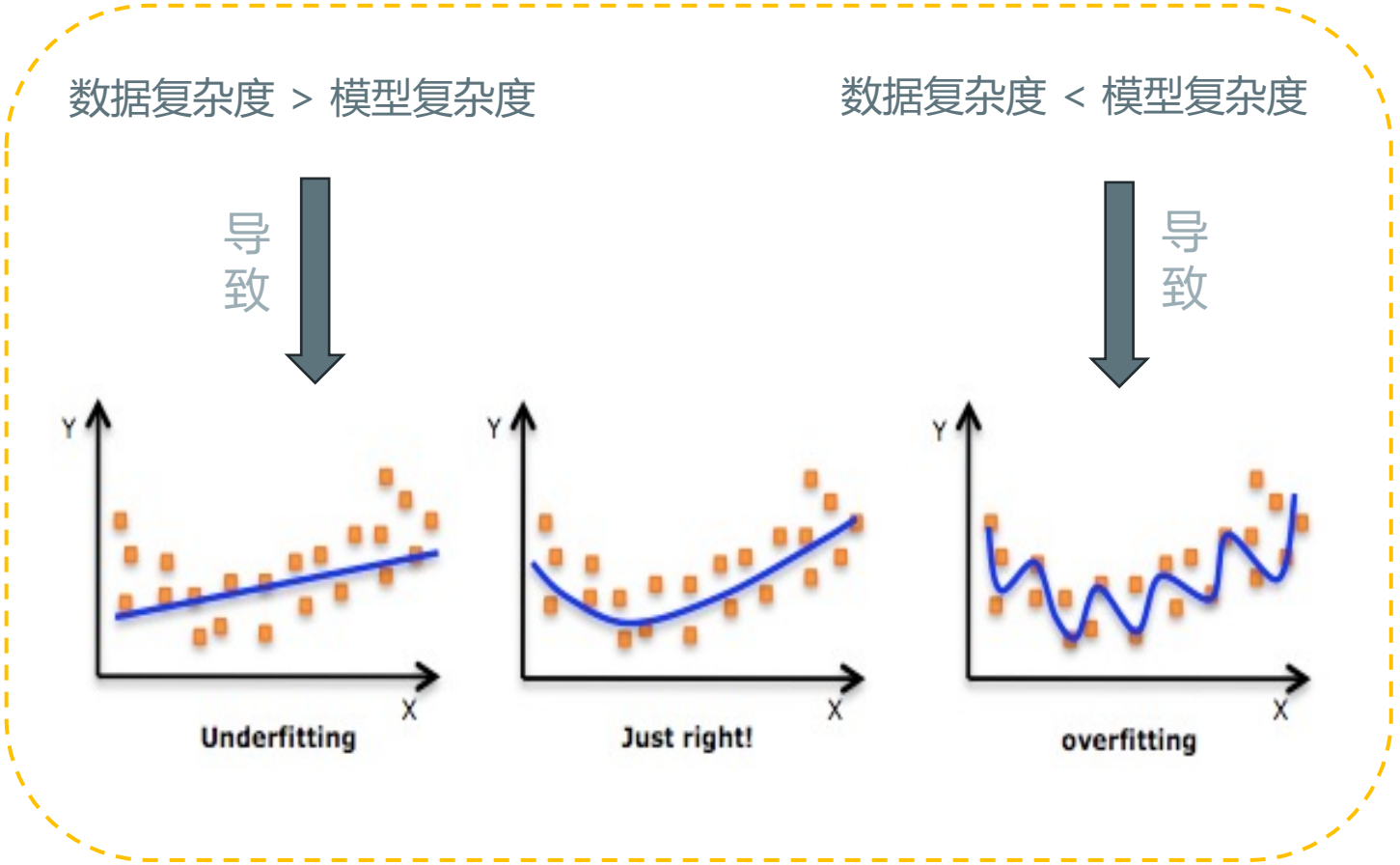


RINC
Robotic Intelligence & Neural Computing Group

基于相似性与多样性的数据分析框架

PART TWO

誠樸雄偉 勵學敦行



多样性
缓解过拟合风险

如何构建数据质量与模型性能之间的关系?

设计目标 增加数据多样性

多样性度量：特征协方差矩阵PCA分解得到的特征向量和特征值

- 生物多样性：物种层面的差异性
- 协方差矩阵构建类内不相关性
- 基于PCA度量有效成分的差异性

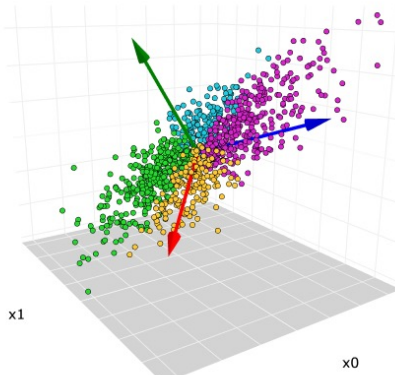
① 构造经验协方差矩阵：刻画类别分布的扩展性和相关性

$$a_{ij}^{*}=\frac{a_{ij}-\overline{a_i}}{\sqrt{s_{ii}}}, i=1,2,\dots,m; j=1,2,\dots,n,$$

$$\mathbf{A}^{*}=\left[a_{ij}^{*}\right]_{m\times n},\quad \mathbf{S}=\frac{1}{n-1}\mathbf{A}^{*}\mathbf{A}^{*T}$$

② 数据信息包含在最大的t个特征值及其特征向量中：

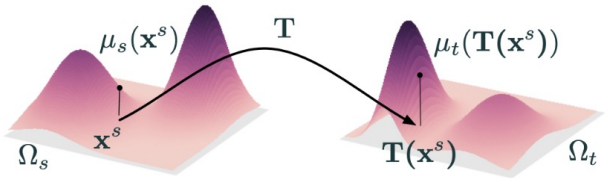
$$\text{diversity}=\frac{1}{k}\sum_{i=1}^{k}\left\|\lambda_i^{*}\cdot\boldsymbol{u}_i^{*}-\lambda_i'^{*}\cdot\boldsymbol{u}_i'^{*}\right\|_2^2,$$



数据分布损害

设计目标

增加数据多样性并保证数据分布的相似性



相似性度量：
最优运输计算特征标签联合分布距离

- **联合样本空间构造**
- **最优运输距离求解**

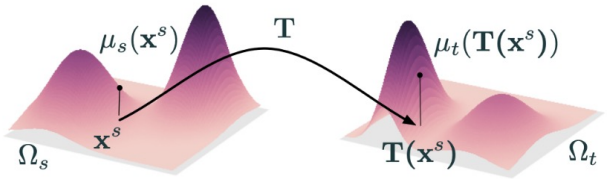
① 对最优运输距离定义：

$$d_{\mathrm{OT}}(\mathcal{D}_t, \mathcal{D}) = \min_{\pi \in \mathcal{V}(\mathcal{C}, \mathcal{C}')} \int_{\mathcal{Z} \times \mathcal{Z}} d_{\mathcal{Z}}(z, z') \pi(z, z')$$

② 相似性度量定义：

$$\mathrm{similarity}(\mathcal{D}_t, \mathcal{D}) = -d_{\mathrm{OT}}(\mathcal{D}_t, \mathcal{D})$$

设计目标 增加**数据多样性**并保证数据分布的**相似性**



相似性度量：最优运输计算特征标签联合分布距离

- **联合样本空间构造**
- **最优运输距离求解**

① **对最优运输距离定义：**

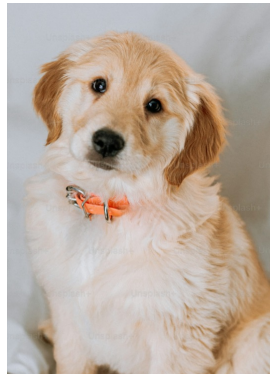
$$d_{\mathrm{OT}}(\mathcal{D}_t, \mathcal{D}) = \min_{\pi \in \mathcal{V}(\mathcal{C}, \mathcal{C}')} \int_{\mathcal{Z} \times \mathcal{Z}} d_{\mathcal{Z}}(z, z') \pi(z, z')$$

② **相似性度量定义：**

$$\mathrm{similarity}(\mathcal{D}_t, \mathcal{D}) = -d_{\mathrm{OT}}(\mathcal{D}_t, \mathcal{D})$$

考虑到不同类距离的尺度不一样

金毛 vs.哈士奇



热狗 vs. 狗

数据增强变换

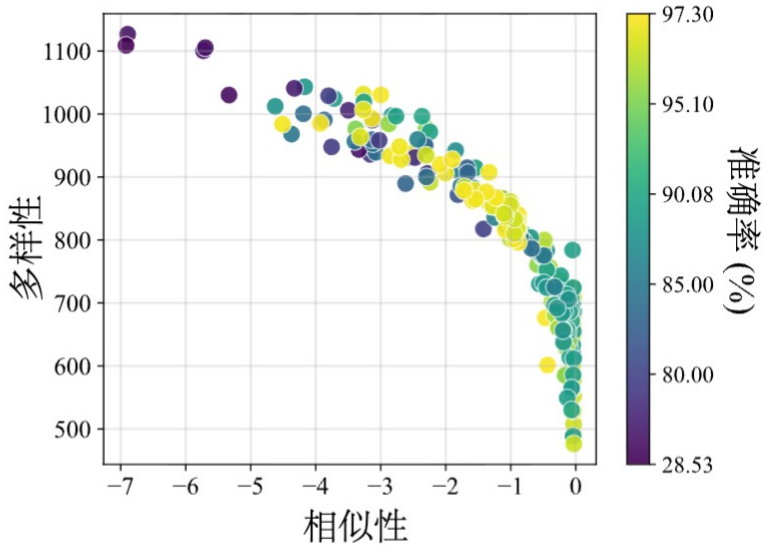
测试了超过220种数据增强方法及其组合

CIFAR10

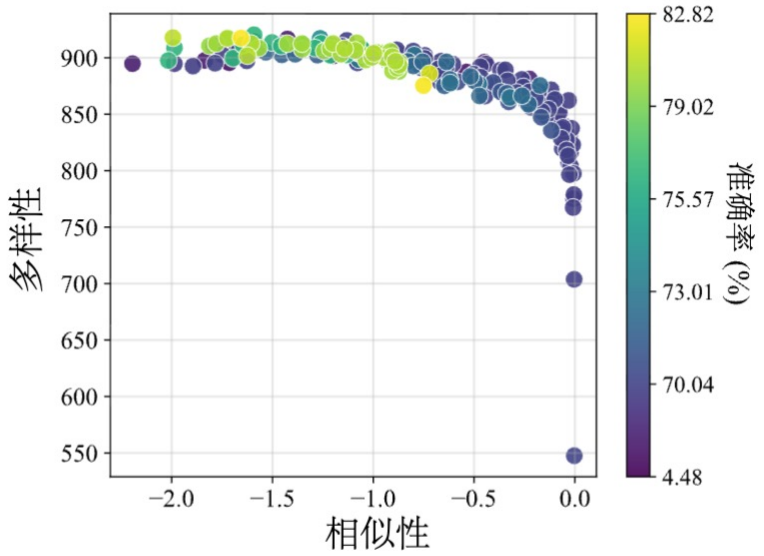
RandAugment(100%)	Rotate(variable, 15deg, 50%)	PatchGaussian(fixed, 8, 0.3, 100%)	Basic()+Color(0.4, 100%)
Basic()	Rotate(variable, 20deg, 100%)	PatchGaussian(fixed, 8, 0.5, 100%)	Basic()+Color(0.5, 100%)
Mixup(100%)	Rotate(variable, 20deg, 50%)	PatchGaussian(fixed, 8, 0.8, 100%)	Basic()+Color(0.6, 100%)
KeepAugment(100%)	Rotate(variable, 20deg, 75%)	PatchGaussian(fixed, 8, 1.0, 100%)	Basic()+Color(0.7, 100%)
Cutmix(100%)	Rotate(variable, 45deg, 50%)	PatchGaussian(fixed, 12, 0.1, 100%)	Basic()+Brightness(0.2, 100%)
Color(0.1, 100%)	Rotate(variable, 5deg, 75%)	PatchGaussian(fixed, 12, 0.2, 100%)	Basic()+Brightness(0.3, 100%)
Color(0.2, 100%)	Rotate(variable, 0deg, 100%)	PatchGaussian(fixed, 12, 0.3, 100%)	Basic()+Brightness(0.5, 100%)
Color(0.3, 100%)	Rotate(variable, 0deg, 50%)	PatchGaussian(fixed, 12, 0.5, 100%)	Basic()+Brightness(0.6, 100%)
Color(0.4, 100%)	Rotate(variable, 20deg, 100%)	PatchGaussian(fixed, 12, 0.8, 100%)	Basic()+Brightness(0.7, 100%)
Color(0.5, 100%)	Rotate(variable, 20deg, 50%)	PatchGaussian(fixed, 12, 1.0, 100%)	Basic()+Invert(100%)
Color(0.6, 100%)	Rotate(variable, 20deg, 75%)	PatchGaussian(fixed, 16, 0.1, 100%)	Basic()+Contrast(0.3, 100%)
Color(0.7, 100%)	Rotate(variable, 45deg, 100%)	PatchGaussian(fixed, 16, 0.2, 100%)	Basic()+Contrast(0.4, 100%)
Brightness(0.1, 100%)	Rotate(variable, 5deg, 100%)	PatchGaussian(fixed, 16, 0.3, 100%)	Basic()+Contrast(0.5, 100%)
Brightness(0.2, 100%)	Rotate(variable, 5deg, 50%)	PatchGaussian(fixed, 16, 0.5, 100%)	Basic()+Contrast(0.6, 100%)
Brightness(0.3, 100%)	Rotate(variable, 60deg, 100%)	PatchGaussian(fixed, 16, 0.8, 100%)	Basic()+Contrast(0.7, 100%)
Brightness(0.4, 100%)	ShearX(fixed, 0.1, 100%)	PatchGaussian(fixed, 16, 1.0, 100%)	Basic()+Posterize(2, 100%)
Brightness(0.5, 100%)	ShearX(fixed, 0.1, 0%)	PatchGaussian(fixed, 20, 0.1, 100%)	Basic()+Posterize(3, 100%)
Brightness(0.6, 100%)	ShearX(fixed, 0.1, 70%)	PatchGaussian(fixed, 20, 0.2, 100%)	Basic()+Posterize(4, 100%)
Brightness(0.7, 100%)	ShearX(fixed, 0.3, 100%)	PatchGaussian(fixed, 20, 0.3, 100%)	Basic()+Posterize(5, 100%)
Contrast(0.1, 100%)	ShearX(fixed, 0.3, 0%)	PatchGaussian(fixed, 20, 0.5, 100%)	Basic()+Posterize(7, 100%)
Contrast(0.2, 100%)	ShearX(fixed, 0.3, 70%)	PatchGaussian(fixed, 20, 0.8, 100%)	Basic()+Sharpness(0.1, 100%)
Contrast(0.4, 100%)	ShearX(variable, 0.1, 100%)	PatchGaussian(fixed, 20, 1.0, 100%)	Basic()+Sharpness(0.2, 100%)
Posterize(3, 100%)	ShearX(variable, 0.1, 0%)	PatchGaussian(fixed, 24, 0.1, 100%)	Basic()+Sharpness(0.3, 100%)
Posterize(4, 100%)	ShearX(variable, 0.1, 70%)	PatchGaussian(fixed, 24, 0.2, 100%)	Basic()+Sharpness(0.5, 100%)
Posterize(5, 100%)	ShearX(variable, 0.3, 100%)	PatchGaussian(fixed, 24, 0.3, 100%)	Basic()+Sharpness(0.6, 100%)
Posterize(6, 100%)	ShearX(variable, 0.3, 0%)	PatchGaussian(fixed, 24, 0.5, 100%)	Basic()+Sharpness(0.7, 100%)
Posterize(7, 100%)	ShearX(variable, 0.3, 70%)	PatchGaussian(fixed, 24, 0.8, 100%)	Basic()+Solarize(0, 100%)
Sharpness(0.1, 100%)	Rotate(variable, 5deg, 10%)	PatchGaussian(fixed, 24, 1.0, 100%)	Basic()+Solarize(100, 100%)
Sharpness(0.2, 100%)	Rotate(variable, 5deg, 20%)	PatchGaussian(fixed, 28, 0.1, 100%)	Basic()+Solarize(150, 100%)
Sharpness(0.3, 100%)	Rotate(variable, 5deg, 30%)	PatchGaussian(fixed, 28, 0.2, 100%)	Basic()+Solarize(200, 100%)
Sharpness(0.4, 100%)	Rotate(variable, 5deg, 40%)	PatchGaussian(fixed, 28, 0.3, 100%)	Basic()+Solarize(250, 100%)
Sharpness(0.5, 100%)	Rotate(variable, 5deg, 50%)	PatchGaussian(fixed, 28, 0.5, 100%)	Basic()+PatchGaussian(fixed, 4, 0.1, 100%)
Sharpness(0.6, 100%)	Rotate(variable, 5deg, 60%)	PatchGaussian(fixed, 28, 0.8, 100%)	Basic()+PatchGaussian(fixed, 4, 0.3, 100%)
Sharpness(0.7, 100%)	Rotate(variable, 5deg, 70%)	PatchGaussian(fixed, 28, 1.0, 100%)	Basic()+PatchGaussian(fixed, 4, 0.8, 100%)
Solarize(0, 100%)	Rotate(variable, 5deg, 80%)	PatchGaussian(fixed, 32, 0.1, 100%)	Basic()+PatchGaussian(fixed, 8, 0.1, 100%)
Solarize(50, 100%)	Rotate(variable, 5deg, 90%)	PatchGaussian(fixed, 32, 0.2, 100%)	Basic()+PatchGaussian(fixed, 8, 0.3, 100%)
Solarize(100, 100%)	Rotate(variable, 5deg, 100%)	PatchGaussian(fixed, 32, 0.3, 100%)	Basic()+PatchGaussian(fixed, 8, 0.8, 100%)
Solarize(150, 100%)	Rotate(variable, 60deg, 10%)	PatchGaussian(fixed, 32, 0.5, 100%)	Basic()+PatchGaussian(fixed, 12, 0.1, 100%)
Solarize(200, 100%)	Rotate(variable, 60deg, 20%)	PatchGaussian(fixed, 32, 0.8, 100%)	Basic()+PatchGaussian(fixed, 12, 0.3, 100%)
Solarize(250, 100%)	Rotate(variable, 60deg, 30%)	PatchGaussian(fixed, 32, 1.0, 100%)	Basic()+PatchGaussian(fixed, 12, 0.8, 100%)
Cutout(fixed, 4, 25%)	Rotate(variable, 60deg, 40%)	Posterize(7, 100%)	Basic()+PatchGaussian(fixed, 16, 0.1, 100%)
Cutout(fixed, 4, 50%)	Rotate(variable, 60deg, 50%)	Basic()+Brightness(0.4, 100%)	Basic()+PatchGaussian(fixed, 16, 0.3, 100%)
Cutout(fixed, 4, 75%)	Rotate(variable, 60deg, 60%)	Basic()+randomerasing(0.8)	Basic()+PatchGaussian(fixed, 16, 0.8, 100%)
Cutout(fixed, 4, 100%)	Rotate(variable, 60deg, 70%)	Basic()+has()	Basic()+PatchGaussian(fixed, 20, 0.1, 100%)
Cutout(fixed, 16, 100%)	Rotate(variable, 60deg, 80%)	AutoAugment(100%)	Basic()+PatchGaussian(fixed, 20, 0.3, 100%)
Cutout(fixed, 16, 25%)	Rotate(variable, 60deg, 90%)	Basic()+Brightness(0.1, 100%)	Basic()+PatchGaussian(fixed, 20, 0.8, 100%)
Cutout(fixed, 16, 50%)	Rotate(variable, 60deg, 100%)	Basic()+Contrast(0.1, 100%)	Basic()+PatchGaussian(fixed, 24, 0.1, 100%)
Cutout(fixed, 16, 75%)	PatchGaussian(fixed, 4, 0.1, 100%)	Basic()+Solarize(50, 100%)	Basic()+PatchGaussian(fixed, 24, 0.3, 100%)
FlipLR(100%)	PatchGaussian(fixed, 4, 0.2, 100%)	Basic()+Contrast(0.2, 100%)	Basic()+PatchGaussian(fixed, 24, 0.8, 100%)
FlipLR(25%)	PatchGaussian(fixed, 4, 0.3, 100%)	Basic()+Posterize(1, 100%)	Basic()+PatchGaussian(fixed, 28, 0.1, 100%)
FlipLR(50%)	PatchGaussian(fixed, 4, 0.5, 100%)	Basic()+Color(0.1, 10%)	Basic()+PatchGaussian(fixed, 28, 0.3, 100%)
FlipUD(100%)	PatchGaussian(fixed, 4, 0.8, 100%)	Basic()+Color(0.2, 10%)	Basic()+PatchGaussian(fixed, 28, 0.8, 100%)
FlipUD(25%)	PatchGaussian(fixed, 4, 1.0, 100%)	Basic()+Color(0.1, 100%)	Basic()+PatchGaussian(fixed, 32, 0.1, 100%)
FlipUD(50%)	PatchGaussian(fixed, 8, 0.1, 100%)	Basic()+Color(0.2, 100%)	Basic()+PatchGaussian(fixed, 32, 0.3, 100%)
Basic()+GridMask(80%)	PatchGaussian(fixed, 8, 0.2, 100%)	Basic()+Color(0.3, 100%)	Basic()+PatchGaussian(fixed, 32, 0.8, 100%)

ImageNet-1k

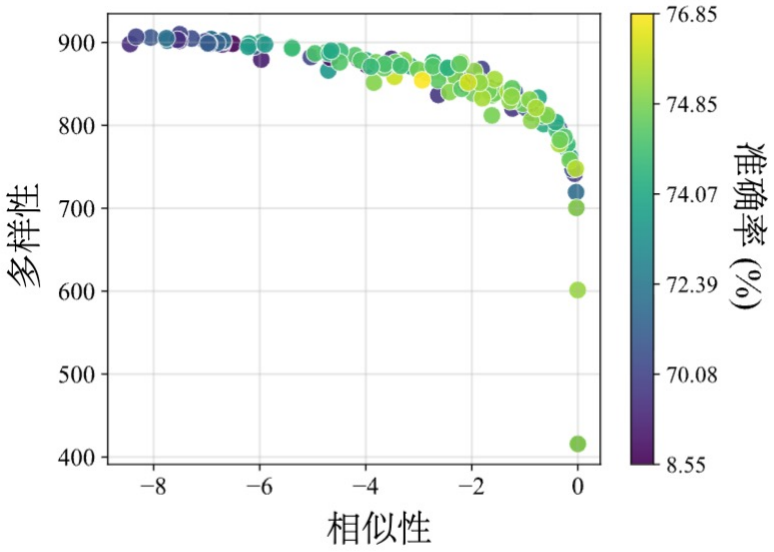
Basic()	PatchGaussian(fixed, 150, 0.5, 100%)	PatchGaussian(fixed, 350, 0.8, 100%)
Color(0.1, 100%)	PatchGaussian(fixed, 150, 0.8, 100%)	PatchGaussian(fixed, 350, 1.0, 100%)
Color(0.2, 100%)	PatchGaussian(fixed, 150, 1.0, 100%)	PatchGaussian(fixed, 350, 2.0, 100%)
Color(0.3, 100%)	PatchGaussian(fixed, 200, 0.2, 100%)	PatchGaussian(fixed, 400, 0.2, 100%)
Color(0.4, 100%)	PatchGaussian(fixed, 200, 0.5, 100%)	PatchGaussian(fixed, 400, 0.5, 100%)
Color(0.5, 100%)	PatchGaussian(fixed, 200, 0.8, 100%)	PatchGaussian(fixed, 400, 0.8, 100%)
Color(0.6, 100%)	PatchGaussian(fixed, 200, 1.0, 100%)	PatchGaussian(fixed, 400, 1.0, 100%)
Color(0.7, 100%)	PatchGaussian(fixed, 250, 0.2, 100%)	PatchGaussian(fixed, 400, 2.0, 100%)
Brightness(0.2, 100%)	PatchGaussian(fixed, 250, 0.5, 100%)	Equalize(100%)
Brightness(0.3, 100%)	PatchGaussian(fixed, 250, 0.8, 100%)	Cutout(fixed, 120, 100%)
Brightness(0.4, 100%)	PatchGaussian(fixed, 250, 1.0, 100%)	Cutout(fixed, 150, 100%)
Brightness(0.5, 100%)	PatchGaussian(fixed, 300, 0.2, 100%)	Cutout(fixed, 180, 100%)
Brightness(0.6, 100%)	PatchGaussian(fixed, 300, 0.5, 100%)	Cutout(fixed, 30, 100%)
Brightness(0.7, 100%)	PatchGaussian(fixed, 300, 0.8, 100%)	Cutout(fixed, 60, 100%)
Contrast(0.1, 100%)	PatchGaussian(fixed, 300, 1.0, 100%)	Cutout(fixed, 90, 100%)
Contrast(0.2, 100%)	PatchGaussian(fixed, 350, 0.5, 100%)	Rotate(variable, 0deg, 100%)
Contrast(0.3, 100%)	PatchGaussian(fixed, 350, 0.8, 100%)	FullGaussian(0.1, 100%)
Contrast(0.4, 100%)	PatchGaussian(fixed, 350, 1.0, 100%)	FullGaussian(0.2, 100%)
Contrast(0.5, 100%)	PatchGaussian(fixed, 400, 0.2, 100%)	FullGaussian(0.3, 100%)
Contrast(0.6, 100%)	PatchGaussian(fixed, 400, 0.5, 100%)	FullGaussian(0.5, 100%)
Contrast(0.7, 100%)	PatchGaussian(fixed, 400, 0.8, 100%)	FullGaussian(0.8, 100%)
FlipUD(100%)	PatchGaussian(fixed, 400, 1.0, 100%)	FullGaussian(1.0, 100%)
Invert(100%)	PatchGaussian(fixed, 100, 0.2, 100%)	TranslateX(70, 100%)
Sharpness(0.1, 100%)	PatchGaussian(fixed, 100, 0.5, 100%)	TranslateX(40, 100%)
Sharpness(0.2, 100%)	PatchGaussian(fixed, 100, 0.8, 100%)	TranslateX(90, 100%)
Sharpness(0.3, 100%)	PatchGaussian(fixed, 100, 1.0, 100%)	TranslateX(60, 100%)
Sharpness(0.4, 100%)	PatchGaussian(fixed, 100, 2.0, 100%)	Rotate(variable, 25deg, 100%)
Sharpness(0.5, 100%)	PatchGaussian(fixed, 150, 0.2, 100%)	TranslateX(80, 100%)
Sharpness(0.6, 100%)	PatchGaussian(fixed, 150, 0.5, 100%)	Rotate(variable, 15deg, 100%)
Sharpness(0.7, 100%)	PatchGaussian(fixed, 150, 0.8, 100%)	RandomErasing(fixed, 180, 100%)
Solarize(0, 100%)	PatchGaussian(fixed, 150, 1.0, 100%)	RandomErasing(fixed, 90, 100%)
Solarize(50, 100%)	PatchGaussian(fixed, 150, 2.0, 100%)	AutoContrast(100%)
Solarize(100, 100%)	PatchGaussian(fixed, 200, 0.2, 100%)	TranslateX(10, 100%)
Solarize(150, 100%)	PatchGaussian(fixed, 200, 0.5, 100%)	Rotate(variable, 0deg, 100%)
Solarize(200, 100%)	PatchGaussian(fixed, 200, 0.8, 100%)	TranslateX(50, 100%)
Solarize(250, 100%)	PatchGaussian(fixed, 200, 1.0, 100%)	Rotate(variable, 5deg, 100%)
Posterize(1, 100%)	PatchGaussian(fixed, 200, 2.0, 100%)	TranslateX(20, 100%)
Posterize(2, 100%)	PatchGaussian(fixed, 250, 0.2, 100%)	Rotate(variable, 10deg, 100%)
Posterize(3, 100%)	PatchGaussian(fixed, 250, 0.5, 100%)	RandomErasing(fixed, 30, 100%)
Posterize(4, 100%)	PatchGaussian(fixed, 250, 0.8, 100%)	Rotate(variable, 20deg, 100%)
Posterize(5, 100%)	PatchGaussian(fixed, 250, 1.0, 100%)	TranslateX(30, 100%)
Posterize(6, 100%)	PatchGaussian(fixed, 250, 2.0, 100%)	RandomErasing(fixed, 60, 100%)
Posterize(7, 100%)	PatchGaussian(fixed, 300, 0.2, 100%)	Rotate(variable, 30deg, 100%)
PatchGaussian(fixed, 100, 0.2, 100%)	PatchGaussian(fixed, 300, 0.5, 100%)	RandomErasing(fixed, 150, 100%)
PatchGaussian(fixed, 100, 0.5, 100%)	PatchGaussian(fixed, 300, 0.8, 100%)	RandomErasing(fixed, 120, 100%)
PatchGaussian(fixed, 100, 0.8, 100%)	PatchGaussian(fixed, 300, 1.0, 100%)	Mixup()
PatchGaussian(fixed, 100, 1.0, 100%)	PatchGaussian(fixed, 300, 2.0, 100%)	Cutmix()
PatchGaussian(fixed, 150, 0.2, 100%)	PatchGaussian(fixed, 350, 0.2, 100%)	



(a) CIFAR-10 数据集



(b) CIFAR-100 数据集



(c) ImageNet 数据集

实验结论

① 不同数据集对于训练数据的相似性-多样性有不同偏好

② 并不是多样性或者相似性越高越好，数据的有效性关键在于二者之间的平衡

③ 核心是在保证整体相似性的前提下，提高数据集内的多样性

(c) CIFAR-100: 相似性　　(d) CIFAR-100: 多样性

第三章

构建了一个以相似性与多样性为基础的数据分析框架

- ✓ **验证了训练数据的有效性来源**
- ✓ **分析了相似性、多样性作为单一学习目标的不足**
- ✓ **该工作对应论文成果：**

Suorong Yang, Suhan Guo, Furao Shen, & Jian Zhao, Investigating the Effectiveness of Data Augmentation from Similarity and Diversity: An Empirical Study. Pattern Recognition 2024 , 148, 110204.

使用上述框架在不同数据集不同数据变换下进行了测试

RINC
Robotic Intelligence & Neural Computing Group

基于多样性提升的自适应数据增强研究

PART THREE

Suorong Yang, Peijia Li, Furao Shen, Jian Zhao, AdaAugment: A Tuning-Free and Adaptive Approach to Enhance Data Augmentation. IEEE Transactions on Image Processing （TIP）2025.

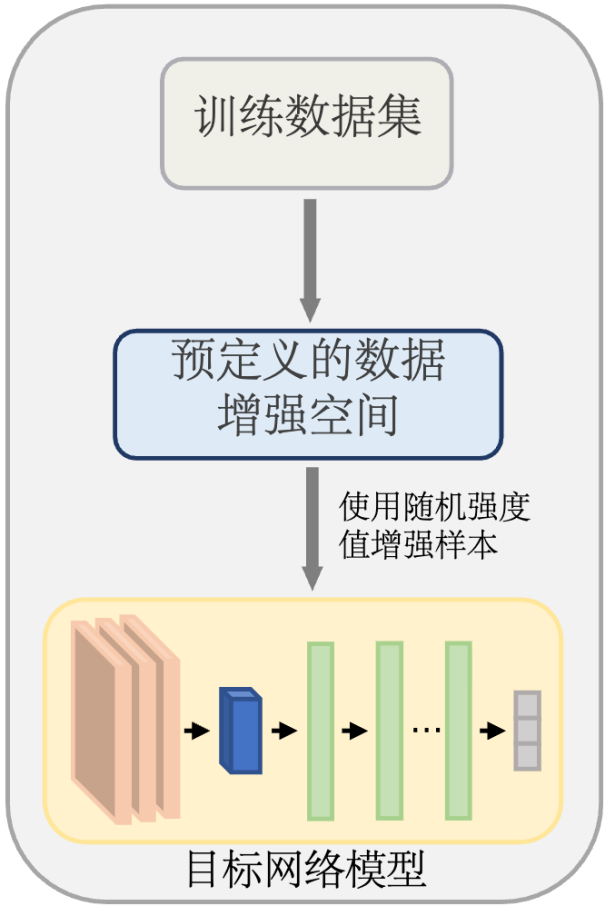
誠樸雄偉 勵學敦行

问题动机 **现有数据增强方法采用固定或随机增强，导致训练数据的多样性幅度不可控**

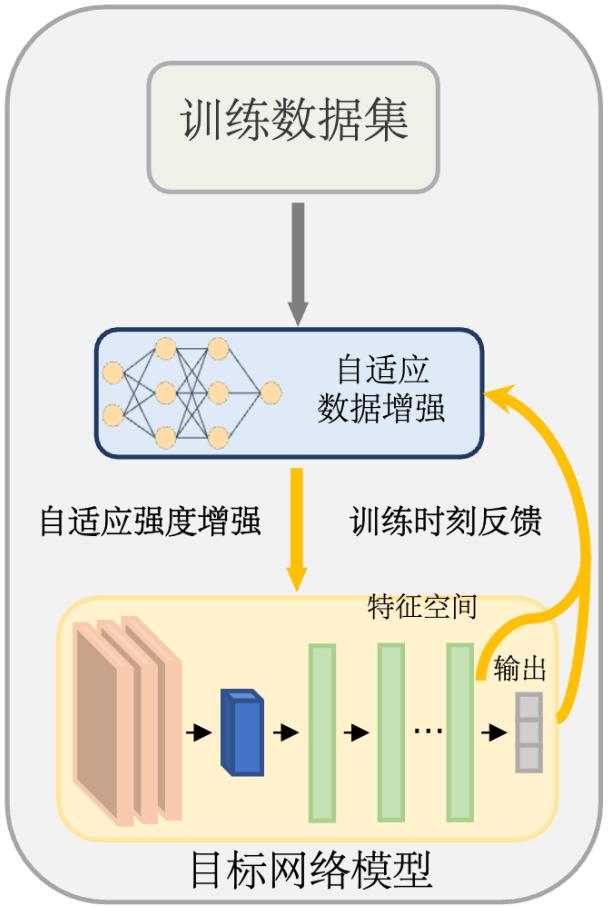
	Operation 1	Operation 2
Sub-policy 0	(Posterize,0.4,8)	(Rotate,0.6,9)
Sub-policy 1	(Solarize,0.6,5)	(AutoContrast,0.6,5)
Sub-policy 2	(Equalize,0.8,8)	(Equalize,0.6,3)
Sub-policy 3	(Posterize,0.6,7)	(Posterize,0.6,6)
Sub-policy 4	(Equalize,0.4,7)	(Solarize,0.2,4)
Sub-policy 5	(Equalize,0.4,4)	(Rotate,0.8,8)
Sub-policy 6	(Solarize,0.6,3)	(Equalize,0.6,7)
Sub-policy 7	(Posterize,0.8,5)	(Equalize,1.0,2)
Sub-policy 8	(Rotate,0.2,3)	(Solarize,0.6,8)
Sub-policy 9	(Equalize,0.6,8)	(Posterize,0.4,6)
Sub-policy 10	(Rotate,0.8,8)	(Color,0.4,0)
Sub-policy 11	(Rotate,0.4,9)	(Equalize,0.6,2)
Sub-policy 12	(Equalize,0.0,7)	(Equalize,0.8,8)
Sub-policy 13	(Invert,0.6,4)	(Equalize,1.0,8)
Sub-policy 14	(Color,0.6,4)	(Contrast,1.0,8)
Sub-policy 15	(Rotate,0.8,8)	(Color,1.0,2)
Sub-policy 16	(Color,0.8,8)	(Solarize,0.8,7)
Sub-policy 17	(Sharpness,0.4,7)	(Invert,0.6,8)
Sub-policy 18	(ShearX,0.6,5)	(Equalize,1.0,9)
Sub-policy 19	(Color,0.4,0)	(Equalize,0.6,3)
Sub-policy 20	(Equalize,0.4,7)	(Solarize,0.2,4)
Sub-policy 21	(Solarize,0.6,5)	(AutoContrast,0.6,5)
Sub-policy 22	(Invert,0.6,4)	(Equalize,1.0,8)
Sub-policy 23	(Color,0.6,4)	(Contrast,1.0,8)

Table 10: AutoAugment policy found on reduced ImageNet.

AutoAugment**所使用的增强空间**



(a) 传统数据增强

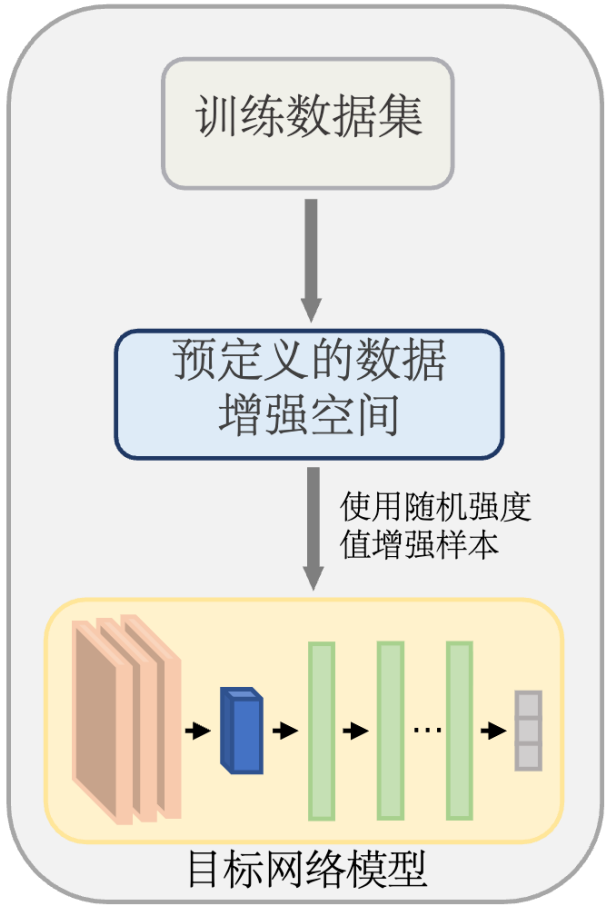


(b) 自适应数据增强

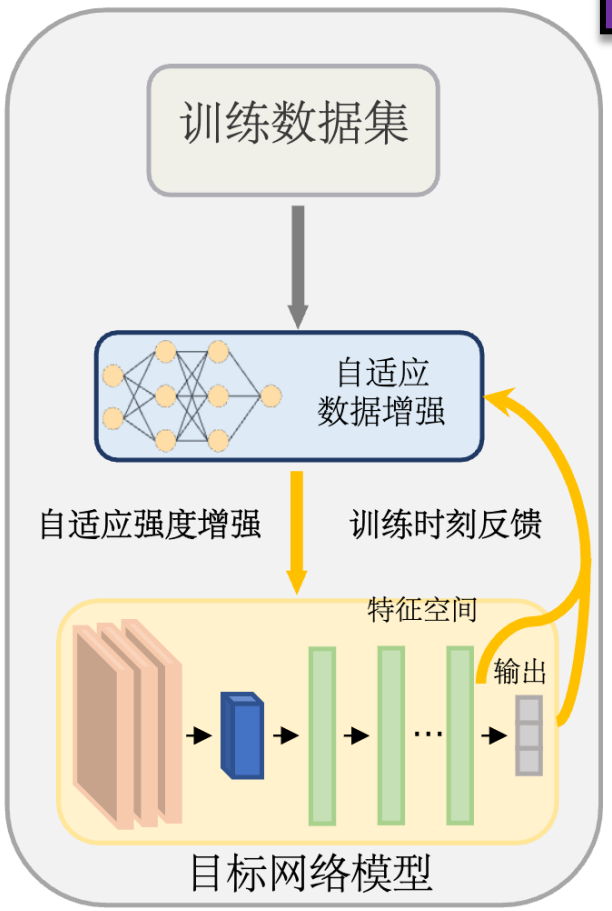
关键点：

在训练过程中根据模型状态自适应调整数据的多样性偏好。

问题动机 现有数据增强方法采用固定或随机增强，导致训练数据的多样性幅度不可控



(a) 传统数据增强



(b) 自适应数据增强

问题定义

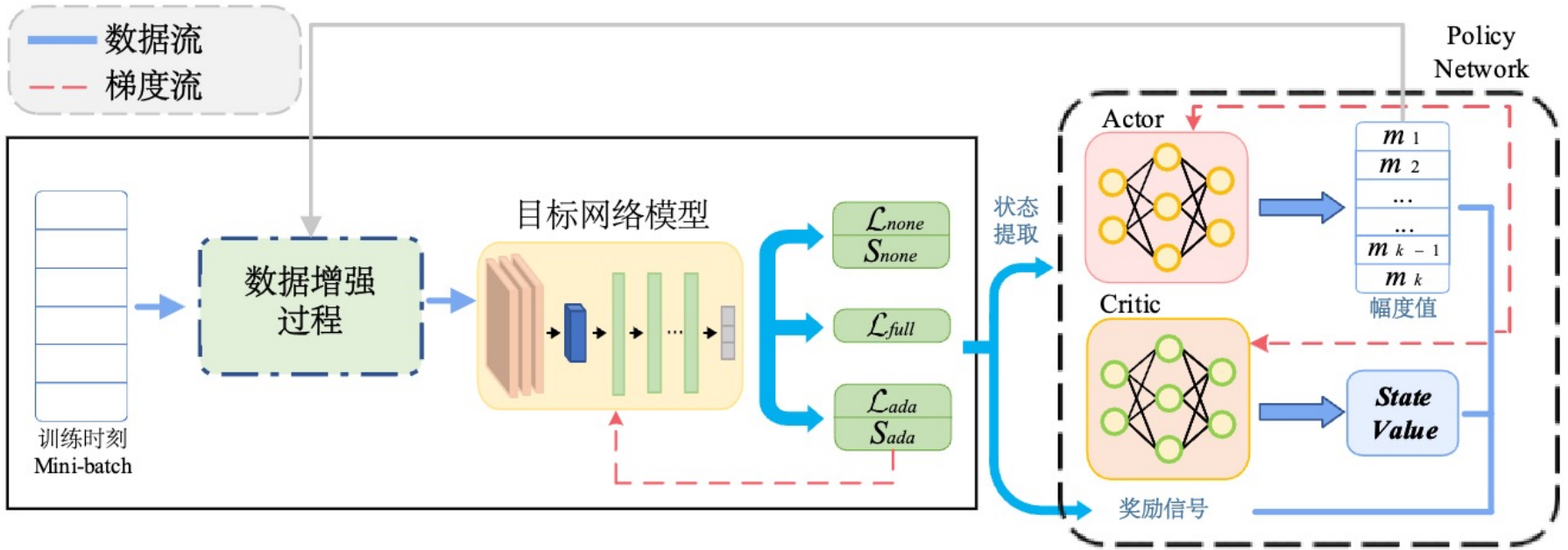
通过自适应增强调整训练数据的分布

样本难度　　模型训练状态

决定数据多样性 → 本质上是决策问题

强化学习！

框架设计



增强调控

操作类型	S_{Max}	对称性
identity	-	-
auto contrast	-	-
equalize	-	-
color	1.9	-
contrast	1.9	-
brightness	1.9	-
sharpness	1.9	-
rotation	30°	±
translate_x	10	±
translate_y	10	±
shear_x	0.3	±
shear_y	0.3	±
solarize	256	-
posterize	4	-

核心：双模型架构，通过RL模块调整训练数据分布。

状态设计 → 通过获取**目标网络在样本上的特征表示**

s_{none} 未增强样本的特征表示

反映样本的固有难度

s_{ada} 自适应增强样本的特征表示

反映增强和模型的交互

状态设计 → 通过获取**目标网络在样本上的特征表示**

s_{none} 未增强样本的特征表示

反映样本的固有难度

s_{ada} 自适应增强样本的特征表示

反映增强和模型的交互

动作设计 → 策略网络确定**数据的增强强度m**

自适应增强操作：$\tilde{x} = e(m, x)$,

- **降低决策复杂度**
- **更直接的数据多样性控制**

状态设计 通过获取**目标网络在样本上的特征表示**

s_{none} 未增强样本的特征表示

反映样本的固有难度

s_{ada} 自适应增强样本的特征表示

反映增强和模型的交互

动作设计 策略网络确定**数据的增强强度m**

自适应增强操作：$\tilde{x} = e(m, x)$,

- **降低决策复杂度**
- **更直接的数据多样性控制**

奖励函数设计 基于**目标模型反馈**调整**训练数据多样性**，训练早期加速收敛，后期缓解过拟合

$\mathcal{L}_{full}(f_\theta(x^+), y)$ 最大增强强度样本的损失

$\mathcal{L}_{ada}(f_\theta(\tilde{x}, y))$ 自适应增强样本的损失

$\mathcal{L}_{none}(f_\theta(x^-), y)$ 未增强样本的损失

奖励函数：

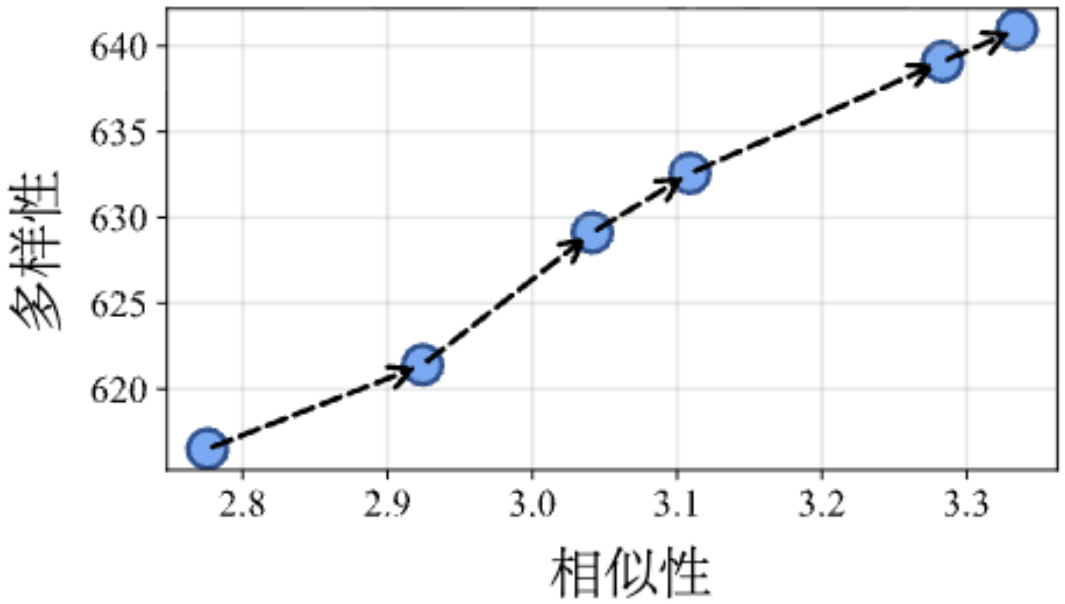
$$r = \lambda(\mathcal{L}_{full} - \mathcal{L}_{ada}) + (1 - \lambda)(\mathcal{L}_{ada} - \mathcal{L}_{none}),$$

λ是调节因子，从1衰减到0.

训练初期施加较弱的增强 加速模型学习

训练后期施加较强的增强 提高样本多样性

训练数据的相似性-多样性演进



- 强化学习模型能够自发的调整训练数据的多样性-相似性偏好
- 训练数据的多样性不断提升，与不断演化的学习状态保持协调
- 无需人工干预，端到端多样性提升数据增强方法

数据集

①粗粒度分类任务：CIFAR-10/100、Tiny-ImageNet、ImageNet-1k

②长尾分类任务：ImageNet-LT、Places-LT

③细粒度分类任务： Oxford Flowers、Oxford-IIIT Pets、FGVC-Aircrafts、Stanford Cars

对比方法

HaS、Fast-AutoAugment (FAA)、 DADA、 Cutout、 CutMix、 MADAug、AdvMask、 GridMask、 AutoAugment、 RandAugment、 TeachAugment、TrivialAugment 和 RandomErasing.

评价指标

分类准确率、迁移学习准确率、复杂度分析等。

在不同架构下的分类准确率

方法	CIFAR-10				CIFAR-100			
	ResNet-18	ResNet-50	WRN-28-10	ShakeShake	ResNet-18	ResNet-50	WRN-28-10	ShakeShake
baseline	95.28±0.14*	95.66±0.08*	95.52±0.11*	94.90±0.07*	77.54±0.19*	77.41±0.27*	78.96±0.25*	76.65±0.14*
HaS[29]	96.10±0.14*	95.60±0.15	96.94±0.08	96.89±0.10*	78.19±0.23	78.76±0.24	80.22±0.16	76.89±0.33
FAA[41]	95.99±0.13	96.69±0.16	97.30±0.24	96.42±0.12	79.11±0.09	79.08±0.12	79.95±0.12	81.39±0.16
DADA[19]	95.58±0.06	95.61±0.14	97.30±0.13*	97.30±0.14*	78.28±0.22	80.25±0.28	82.50±0.26*	80.98±0.15
Cutout[13]	96.01±0.18*	95.81±0.17	96.92±0.09	96.96±0.09*	78.04±0.10*	78.62±0.25	79.84±0.14	77.37±0.28
CutMix[15]	96.64±0.62*	96.81±0.10*	96.93±0.10*	96.47±0.07	79.45±0.17	81.24±0.14	82.67±0.22	79.57±0.10
MADAug[117]	96.49±0.10	97.12±0.11	97.48±0.12	97.37±0.11	79.39±0.19	81.40±0.10	83.01±0.11	81.67±0.18
AdvMask[30]	96.44±0.15*	96.69±0.10*	97.02±0.05*	97.03±0.12*	78.43±0.18*	78.99±0.31*	80.70±0.25*	79.96±0.27*
GridMask[27]	96.38±0.17	96.15±0.19	97.23±0.09	96.91±0.12	75.23±0.21	78.38±0.22	80.40±0.20	77.28±0.38
AutoAugment[16]	96.51±0.10*	96.59±0.04*	96.99±0.06	97.30±0.11	79.38±0.20	81.34±0.29	82.21±0.17	82.19±0.19
RandAugment[17]	96.47±0.32	96.25±0.06	96.94±0.13*	97.05±0.15	78.30±0.15	80.95±0.22	82.90±0.29*	80.00±0.29
TeachAugment[18]	96.47±0.09	96.40±0.10	97.50±0.12	97.29±0.14	79.27±0.17	80.54±0.23	82.81±0.20	81.30±0.20
TrivialAugment[43]	96.28±0.10	97.07±0.08	97.18±0.11	97.30±0.10	78.67±0.19	81.34±0.18	82.75±0.26	82.14±0.16
RandomErasing[28]	95.69±0.10						80.57±0.15	77.30±0.18
AdaAugment	**96.75**±0.06						**83.23**±0.23	**82.82**±0.25

- 在多个数据集下能够取得优越的泛化性
- 可以泛化到多样化的深度模型架构上

ImageNet-1k

baseline	HaS	FAA	DADA	Cutout	CutMix	MADAug	GridMask	AA	RA	TeachAug	TA	RE	AdaAugment
77.1	77.2	77.6	77.5	77.1	77.2	**78.3**	77.9	77.6	77.8	78.0	77.9	77.3	**78.3**

迁移学习实验结果

baseline	HaS[29]	FAA[41]	DADA[19]	Cutout[13]	CutMix[15]	MADAug[117]	GridMask[27]	AA[16]	RA[17]	TeachAug[18]	TA[43]	RE[28]	AdaAugment
$91.53_{\pm.03}$	$92.51_{\pm.24}$	$92.28_{\pm.13}$	$92.58_{\pm.09}$	$92.42_{\pm.20}$	$92.81_{\pm.47}$	$92.84_{\pm.10}$	$91.49_{\pm.10}$	$92.82_{\pm.04}$	$92.78_{\pm.23}$	$92.83_{\pm.18}$	$92.80_{\pm.16}$	$92.55_{\pm.05}$	$\mathbf{93.06}_{\pm.25}$
$64.02_{\pm.05}$	$66.84_{\pm.06}$	$70.32_{\pm.63}$	$69.04_{\pm.43}$	$65.54_{\pm.75}$	$69.29_{\pm.09}$	$72.82_{\pm.32}$	$64.88_{\pm.43}$	$69.53_{\pm.53}$	$64.68_{\pm.97}$	$69.98_{\pm.17}$	$71.53_{\pm.35}$	$64.56_{\pm.27}$	$\mathbf{76.86}_{\pm.12}$

- **提升特征可迁移性：**通过动态调控增强幅度，AdaAugment 能够使预训练模型学到的表示更具普适性，而不是过度依赖特定数据集的分布特征
- **增强泛化性能：**在微调阶段，迁移后的模型在新任务上表现出更好的泛化能力。这说明 AdaAugment 有效缓解了源任务中的过拟合风险，使得模型在面对新的数据分布时具备更强的适应性。

迁移学习实验结果

baseline	HaS[29]	FAA[41]	DADA[19]	Cutout[13]	CutMix[15]	MADAug[117]	GridMask[27]	AA[16]	RA[17]	TeachAug[18]	TA[43]	RE[28]	AdaAugment
91.53±.03	92.51±.24	92.28±.13	92.58±.09	92.42±.20	92.81±.47	92.84±.10	91.49±.10	92.82±.04	92.78±.23	92.83±.18	92.80±.16	92.55±.05	**93.06**±.25
64.02±.05	66.84±.06	70.32±.63	69.04±.43	65.54±.75	69.29±.09	72.82±.32	64.88±.43	69.53±.53	64.68±.97	69.98±.17	71.53±.35	64.56±.27	**76.86**±.12

长尾分类任务

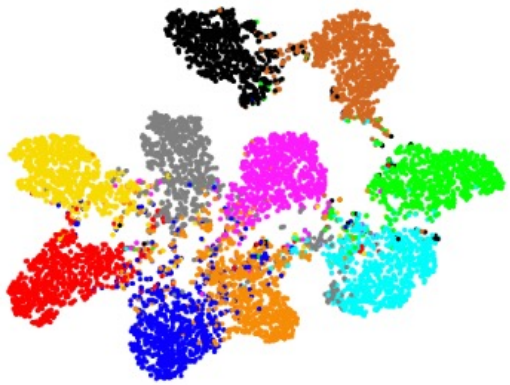
数据集	方法	闭集设置				开集设置			
		Many-shot	**Medium-shot**	**Few-shot**	**Overall**	**Many-shot**	**Medium-shot**	**Few-shot**	**F-measure**
ImageNet-LT	OLTR	43.2±0.1*	35.1±0.2*	18.5±0.1*	35.6±0.1*	41.9±0.1*	33.9±0.1*	17.4±0.2*	44.6±0.2*
	OLTR+**AdaAugment**	**45.9**±0.1	**38.3**±0.1	**22.0**±0.2	**39.0**±0.1	**44.1**±0.1	**36.8**±0.1	**20.8**±0.2	**45.8**±0.1
Places-LT	OLTR	**44.7**±0.1*	37.0±0.2*	25.3±0.1*	35.9±0.1*	**44.6**±0.1*	36.8±0.1*	25.2±0.2*	46.4±0.1*
	OLTR+**AdaAugment**	43.7±0.1	**41.1**±0.1	**29.5**±0.2	**39.6**±0.1	43.9±0.1	**40.8**±0.1	**28.9**±0.1	**50.4**±0.1

- **在闭集设置下，可以带来超过3%的性能提升，表明可以缓解长尾分布带来的不平衡问题**
- **动态调节数据相似性多样性偏好，改善头部类和尾部类的表示学习能力**

细粒度分类任务

数据集	基线	AdaAugment
Oxford Flowers[120]	89.47±0.08	**97.17**±0.14
Oxford-IIIT Pets[121]	89.73±0.18	**91.95**±0.24
FGVC-Aircraft[122]	77.25±0.09	**90.92**±0.05
Stanford Cars[123]	82.13±0.03	**84.76**±0.20

可视化实验



(a) 基线

(b) AdaAugment

- AdaAugment得到的特征分布呈现更优的类内聚合度和类间散度。

复杂度分析

模型	FLOPs	参数量	GPU 耗时	准确率提升
ResNet-18	1.82G	+0.15M	$+0.41h_{\pm 0.03}$	$+1.47\%_{\pm 0.06}$
ResNet-50	4.14G	+0.60M	$+0.49h_{\pm 0.03}$	$+1.68\%_{\pm 0.13}$
WRN-28-10	5.25G	+0.19M	$+0.43h_{\pm 0.03}$	$+2.04\%_{\pm 0.07}$

A2C网络结构

	层序	层类型	维度
Actor	1	线性层	(512, 512)
	2	线性层	(512, 256)
	3	线性层	(256, 1)
Critic	1	线性层	(512, 512)
	2	线性层	(512, 256)
	3	线性层	(256, 1)

- 相比于常用网络模型，参数量增加极为有限，不到2.5%；
- 在有限增加训练开销的前提下，显著改进准确率，在性能与效率之间达到了卓越平衡；

第四章

提出了自适应数据增强方法AdaAugment来提升训练数据多样性

- ✓ **不依赖人工启发式设计，端到端完成**
- ✓ **该工作对应论文成果：**

Suorong Yang, Peijia Li, Furao Shen, Jian Zhao, AdaAugment: A Tuning-Free and Adaptive Approach to Enhance Data Augmentation. IEEE Transactions on Image Processing （TIP）2025.

基于第三章的相似性-多样性分析框架，实现了“合理的训练数据应在相似性与多样性之间保持平衡”，在不同阶段自适应提升数据多样性来改善泛化

RINC
Robotic Intelligence & Neural Computing Group

基于相似性驱动的多模态数据选择研究

PART FOUR

誠樸雄偉 勵學敦行

Suorong Yang, Peng Ye, Wanli Ouyang, Dongzhan Zhou, Furao Shen, A CLIP-Powered Framework for Robust and Generalizable Data Selection. **ICLR 2025 Spotlight**

问题动机

- 数据集存在冗余信息

影响训练效率，显著增加训练成本

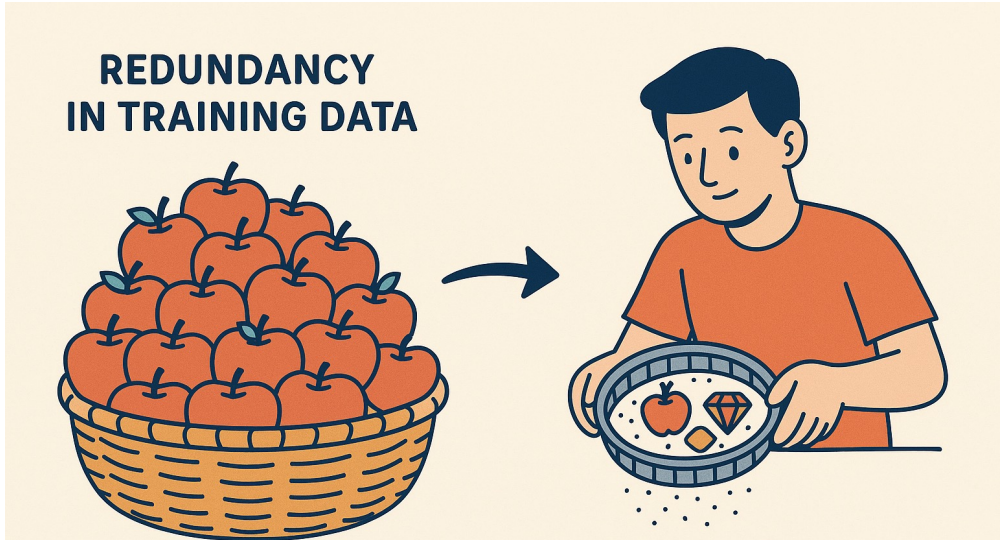
冗余样本使模型过拟合，影响泛化

问题动机

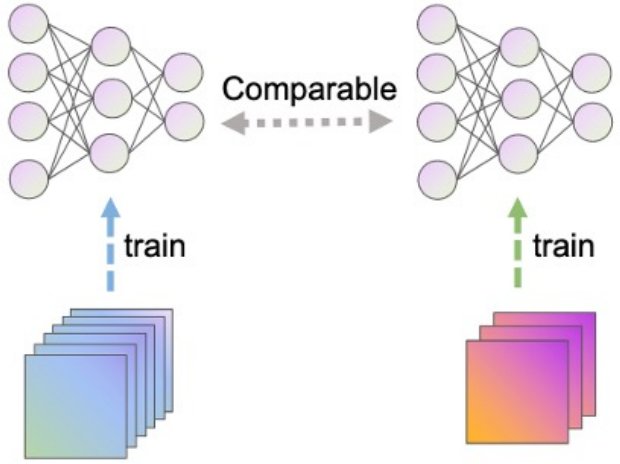
- 数据集存在冗余信息

影响训练效率，显著增加训练成本

冗余样本使模型过拟合，影响泛化



- 数据筛选技术选出具有一个代表性的高质量核心集



问题动机

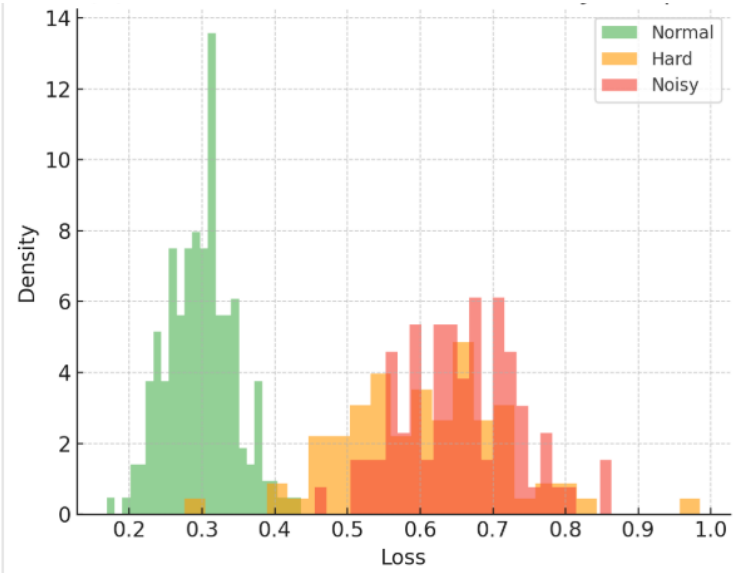
- 真实世界收集的数据存在噪声

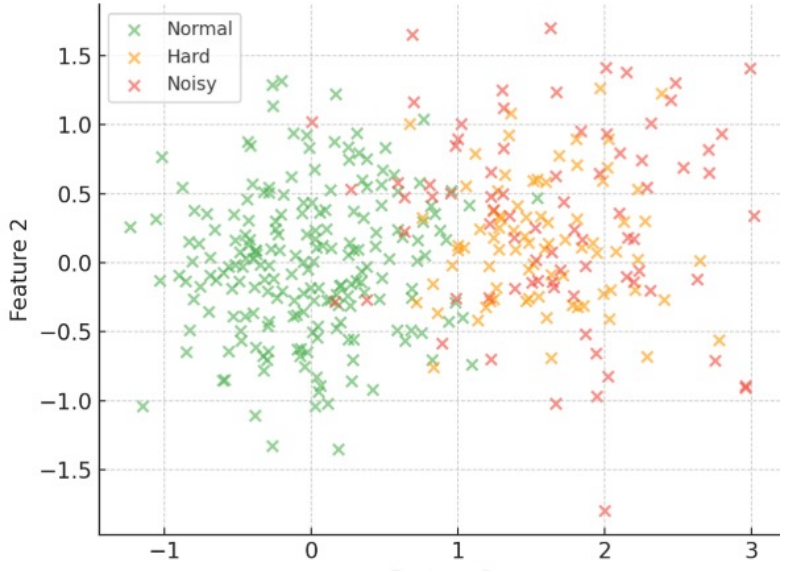
标签错误

图片信息受损

数据筛选

偏好难样本，含有更多的信息让模型学习

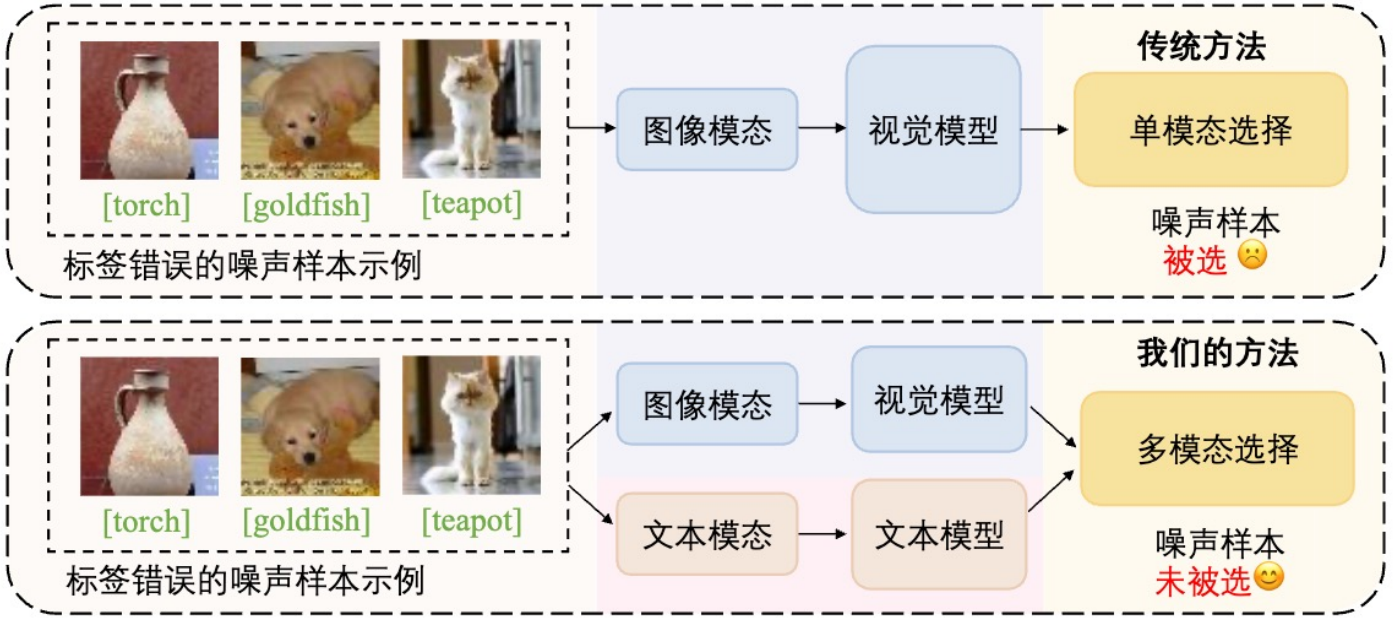




问题定义

噪声样本作为**难样本被选入训练集用于模型训练！**

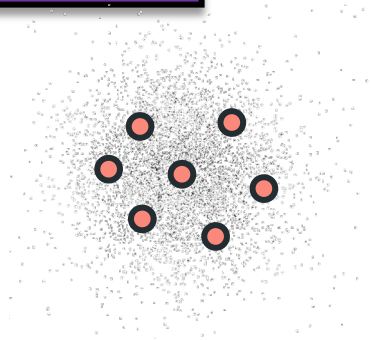
区分难样本和噪声样本



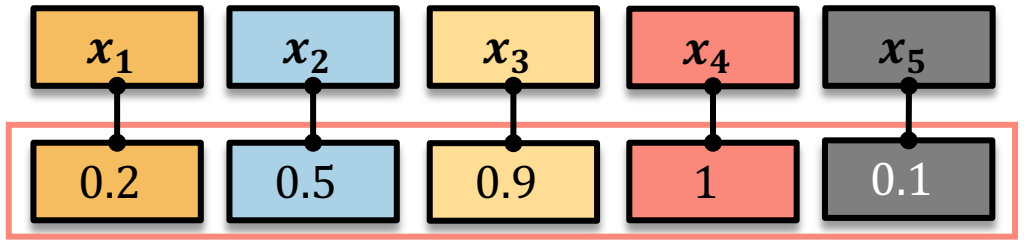
多模态相似性驱动的优势

- 文本语义和图像语义的跨模态一致性评估
- 有效过滤噪声和语义失配的噪声样本

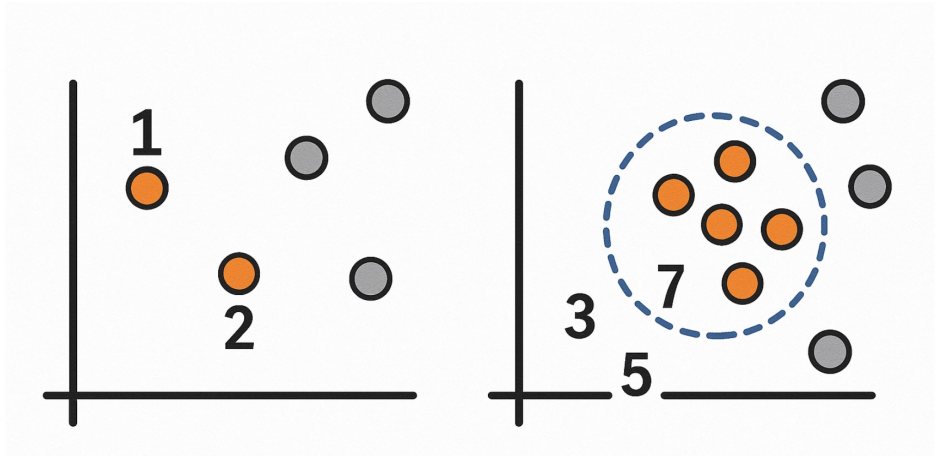
问题动机



重要性指标

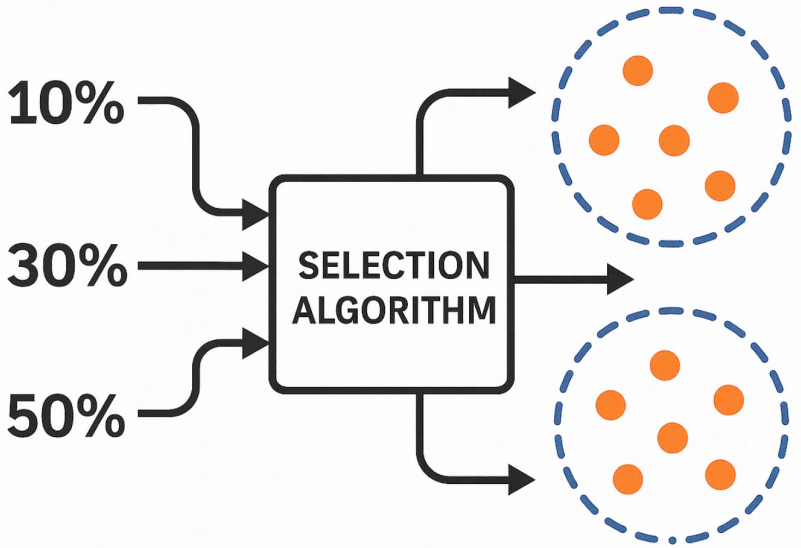


- 数据集的群体效应



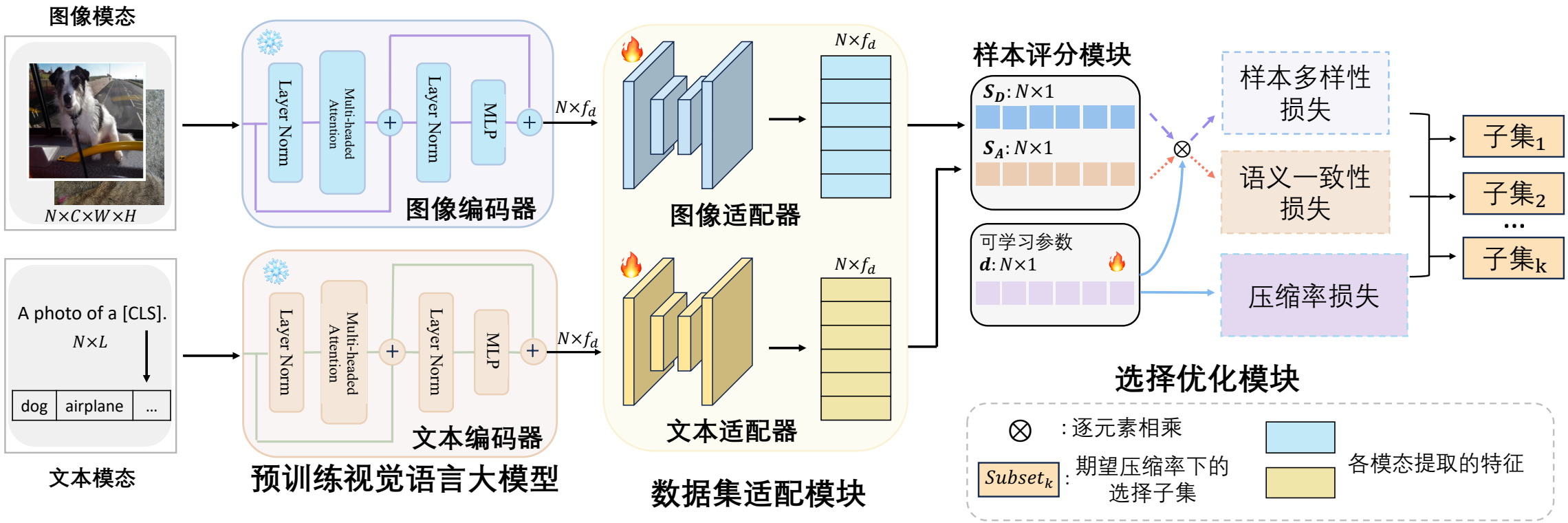
问题定义

独立地评估每个样本的重要性往往会忽略样本间的关系！

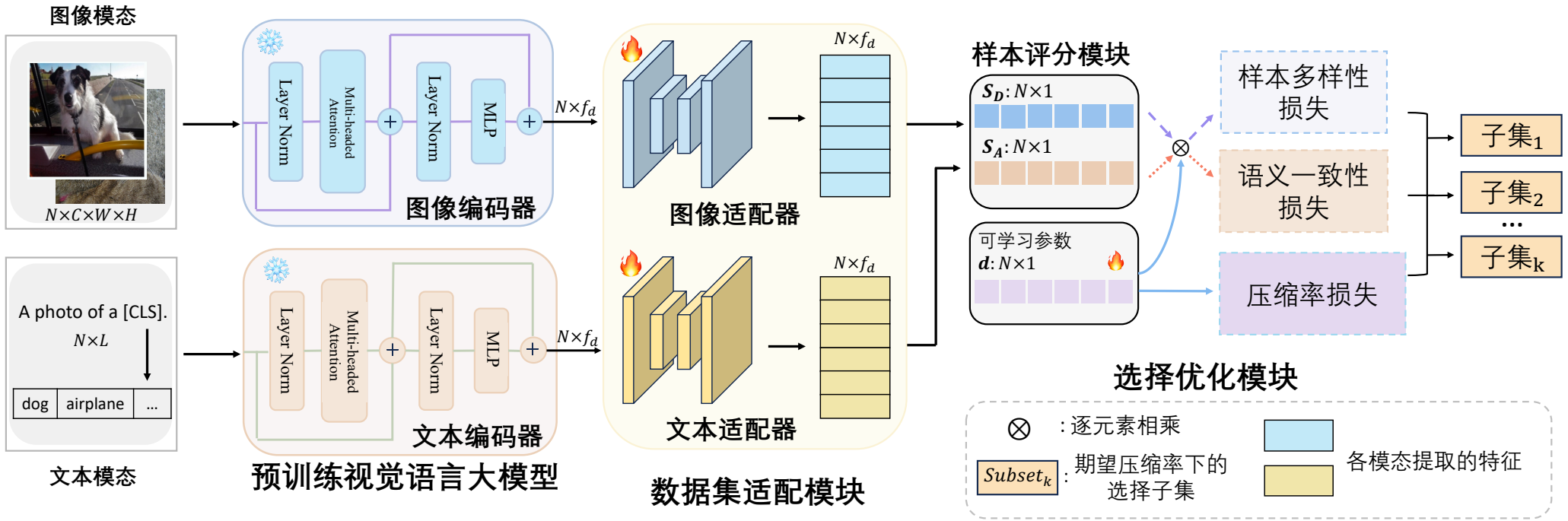


找到给定数据集大小下的最优样本组合

框架设计


图像模态
$N\times C\times W\times H$
A photo of a [CLS].
$N\times L$
dog
airplane
...
文本模态
Layer Norm
Multi-headed Attention
Layer Norm
MLP
图像编码器
文本编码器
预训练视觉语言大模型
$N\times f_d$
图像适配器
文本适配器
数据集适配模块
样本评分模块
$S_D: N\times 1$
$S_A: N\times 1$
可学习参数
$d: N\times 1$
样本多样性损失
语义一致性损失
压缩率损失
子集$_1$
子集$_2$
...
子集$_k$
选择优化模块
: 逐元素相乘
$Subset_k$: 期望压缩率下的选择子集
各模态提取的特征

框架设计

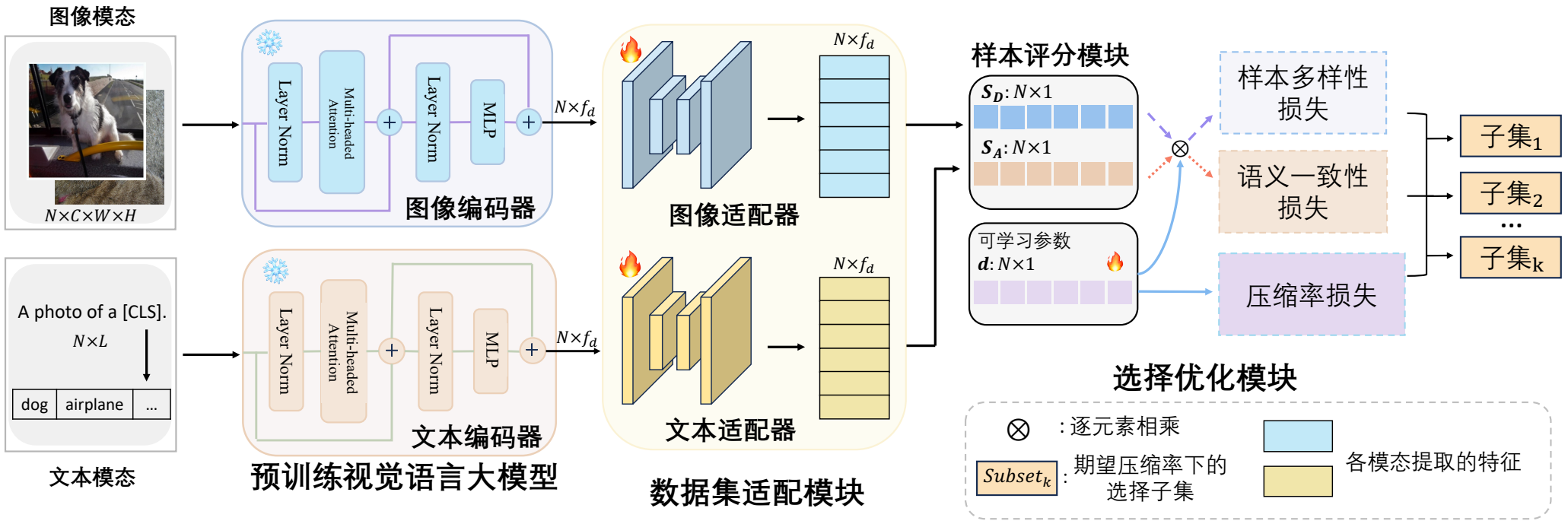


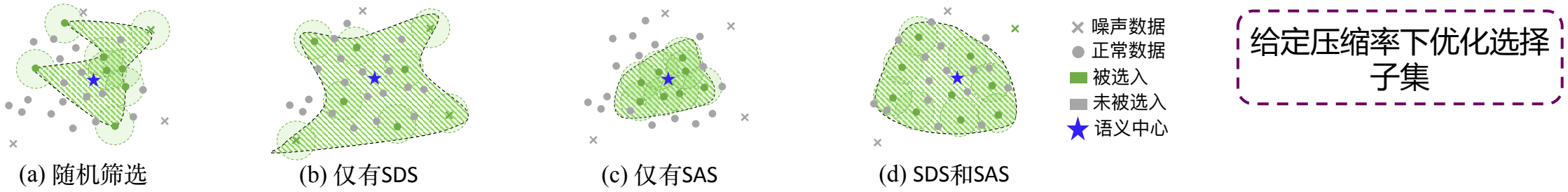
高效构建文本模态

将预训练知识迁移到目标数据集

给定压缩率下优化选择子集

框架设计





(a) 随机筛选 (b) 仅有SDS (c) 仅有SAS (d) SDS和SAS

多样性损失

$$\mathcal{L}_{sd} = -\frac{1}{N}\sum \text{sigmoid}(\boldsymbol{d}) * \boldsymbol{S}_{Di}.$$

优先高SD样本

相似性损失

$$\mathcal{L}_{sa} = -\frac{1}{N}\sum_{i}^{N} \text{sigmoid}(\boldsymbol{d}) * \boldsymbol{S}_{Ai},$$

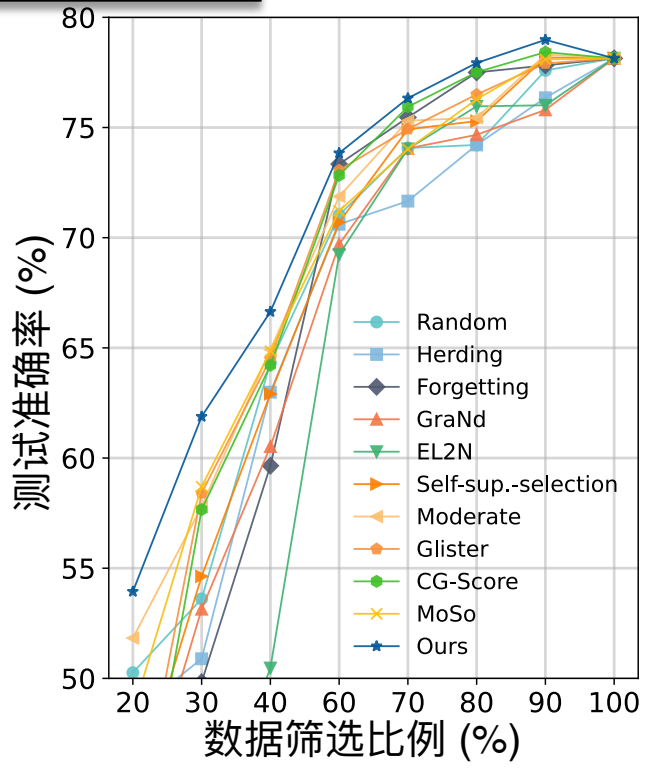
优先高SA样本

剪枝损失

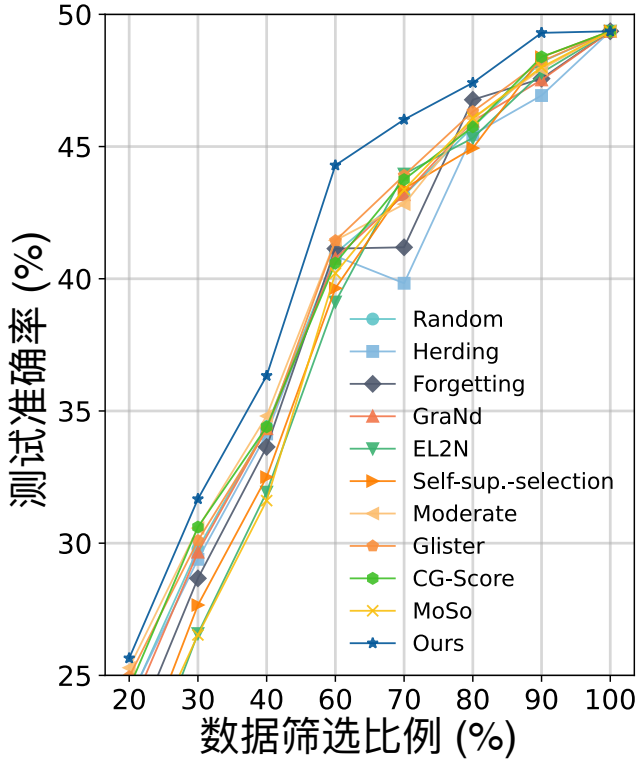
$$\mathcal{L}_{s} = \sqrt{\left[\frac{1}{N}\sum_{i} \text{STE}\left[\mathbb{I}\left(\text{sigmoid}(\boldsymbol{d}_i)_{\text{i}} > 0.5\right)\right] - s_r\right]^2},$$

保证压缩率约束

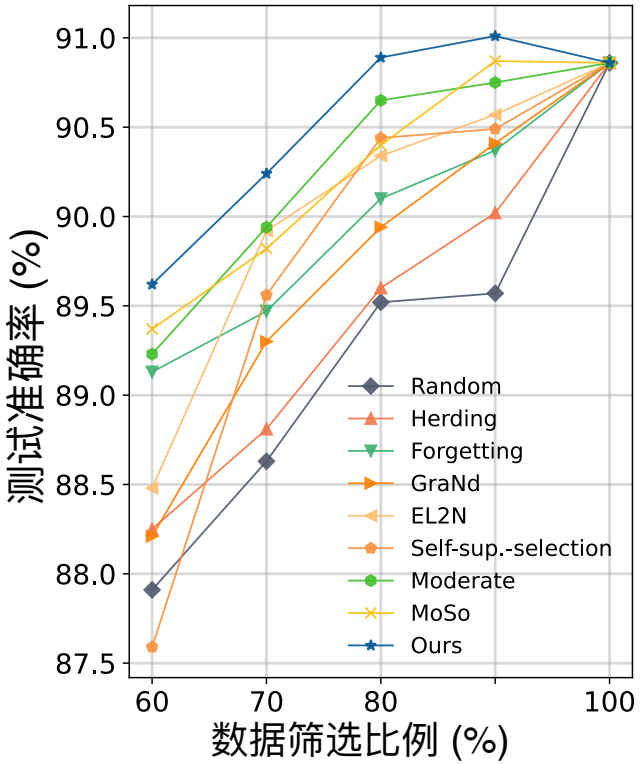
基准数据集



CIFAR100



Tiny-ImageNet



ImageNet-1k

①在所有压缩率上**优于其他算法**

② 数据量的减少会带来性能损失，我们的方法带来的**损失是最小**

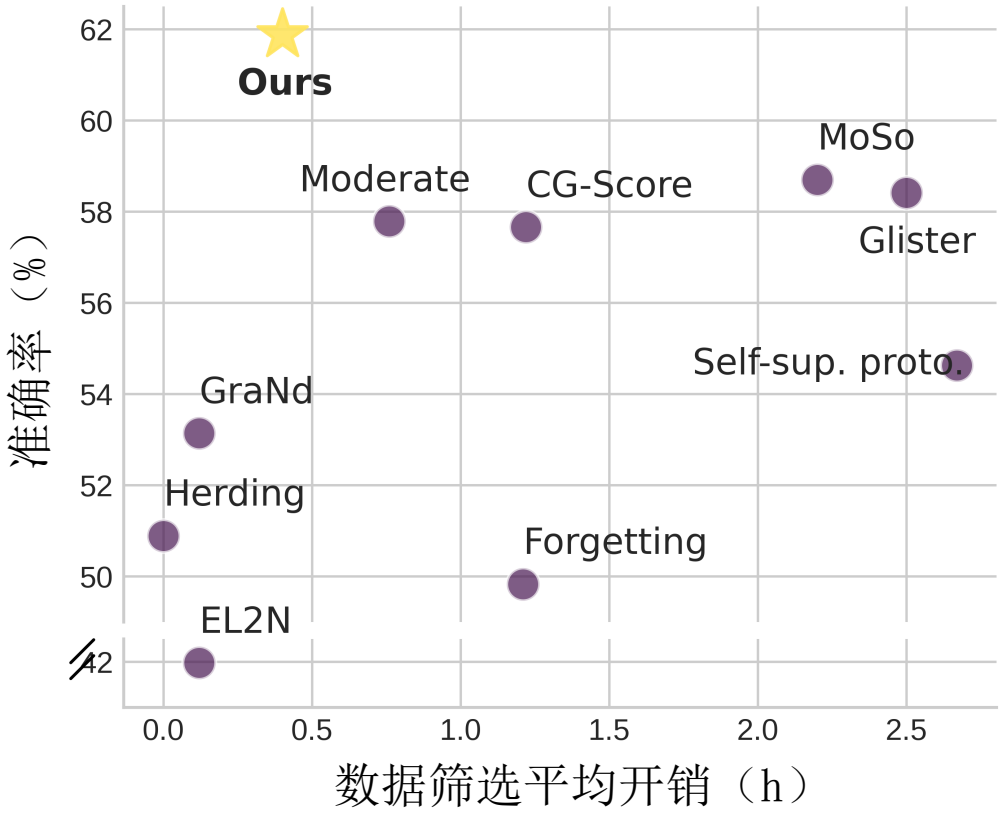
跨架构泛化性

方法 / 选择比例	VGG-16				Densenet-121			
	70%	80%	90%	100%	70%	80%	90%	100%
Random	47.39±2.72	49.38±0.23	51.15±0.64	57.23±1.08	59.55±0.20	60.78±0.18	61.03±0.22	62.22±0.23
EL2N	48.30±2.95	48.75±1.65	49.01±1.31	57.23±1.08	59.61±0.00	60.38±0.04	61.16±0.47	62.22±0.23
GraNd	50.79±1.26	46.84±1.38	54.73±0.49	57.23±1.08	59.62±0.02	60.84±0.09	61.10±0.05	62.22±0.23
MoSo	50.47±1.01	50.12±0.83	50.07±0.43	57.23±1.08	59.27±0.33	59.86±0.07	60.00±0.37	62.22±0.23
Herding	48.59±0.07	45.77±0.12	50.77±1.24	57.23±1.08	59.00±0.28	60.03±0.35	61.15±0.12	62.22±0.23
Glister	48.74±2.29	50.05±0.02	49.42±1.81	57.23±1.08	59.98±0.01	60.62±0.34	61.28±0.18	62.22±0.23
CG-Score	48.73±2.70	48.49±1.88	49.62±1.08	57.23±1.08	59.74±0.15	60.55±0.20	61.14±0.11	62.22±0.23
Self-sup. prototypes	48.38±1.38	49.98±1.49	54.71±0.84	57.23±1.08	59.56±0.03	60.22±0.12	60.91±0.29	62.22±0.23
Forgetting	47.50±2.43	48.59±1.77	49.82±0.62	57.23±1.08	58.54±0.15	60.39±0.46	61.12±0.10	62.22±0.23
Moderate-DS	50.78±0.93	49.31±0.41	49.25±0.77	57.23±1.08	59.41±0.18	60.42±0.14	61.44±0.11	62.22±0.23
Ours	**53.40**±3.20	**52.25**±0.58	**56.34**±2.93	57.23±1.08	**60.12**±0.06	**60.93**±0.03	**61.59**±0.03	62.22±0.23

①在结构差异较大的VGG-16和Densenet-121上均展现良好的泛化性

②本章工作所选子集具有较强的泛化性

训练效率对比



效率来源分析：

- **极简的Adapter架构设计；**
- **高效的数值优化过程；**

①相比于基于复杂优化过程的方法，具有明显的效率优势

②相比于one-shot方法具有相似水平的开销

③同时具有最好的性能表现

噪声场景的鲁棒性

方法 / 选择比例	CIFAR-100 (标签噪声)		Tiny-ImageNet (标签噪声)		噪声比例	
	20%	30%	20%	30%	20%	30%
Random	$34.47_{\pm 0.64}$	$43.26_{\pm 1.21}$	$17.78_{\pm 0.44}$	$23.88_{\pm 0.42}$	20.80	19.83
MoSo	$31.01_{\pm 0.67}$	$43.73_{\pm 0.14}$	$21.55_{\pm 0.37}$	$27.80_{\pm 0.16}$	7.78	8.82
Moderate-DS	$40.25_{\pm 0.12}$	$48.53_{\pm 1.60}$	$19.64_{\pm 0.40}$	$24.96_{\pm 0.30}$	0.30	0.31
Glister	$28.51_{\pm 1.46}$	$43.16_{\pm 1.31}$	$21.61_{\pm 0.19}$	$25.45_{\pm 0.23}$	21.21	21.95
Herding	$42.29_{\pm 1.75}$	$50.52_{\pm 3.38}$	$18.98_{\pm 0.44}$	$24.23_{\pm 0.29}$	35.00	30.56
Forgetting	$36.53_{\pm 1.11}$	$45.78_{\pm 1.04}$	$13.20_{\pm 0.38}$	$21.79_{\pm 0.43}$	23.00	21.76
GraNd	$31.72_{\pm 0.67}$	$42.80_{\pm 0.30}$	$18.28_{\pm 0.32}$	$23.72_{\pm 0.18}$	5.00	5.14
EL2N	$29.82_{\pm 1.19}$	$33.62_{\pm 2.35}$	$13.93_{\pm 0.69}$	$18.57_{\pm 0.31}$	22.00	21.80
Self-sup. prototypes	$31.08_{\pm 0.78}$	$41.87_{\pm 0.63}$	$15.10_{\pm 0.73}$	$21.01_{\pm 0.36}$	21.70	20.21
CG-Score	$6.82_{\pm 1.60}$	$20.07_{\pm 0.45}$	$8.35_{\pm 0.65}$	$15.31_{\pm 0.90}$	45.09	39.69
Ours	**$46.05_{\pm 0.21}$**	**$58.34_{\pm 0.36}$**	**$26.09_{\pm 0.12}$**	**$33.13_{\pm 0.25}$**	**0.25**	**0.32**

①显著抑制噪声数据的引入；

②语义对齐有效过滤图像内容与标签之间的不匹配的情况；

③噪声场景下依然可以选择出最好的数据子集用于训练；

对图像受损的鲁棒性

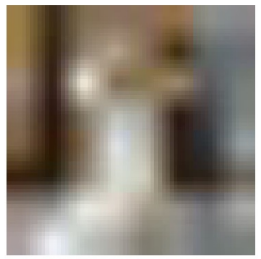

(a) 原图


(b) 雾化

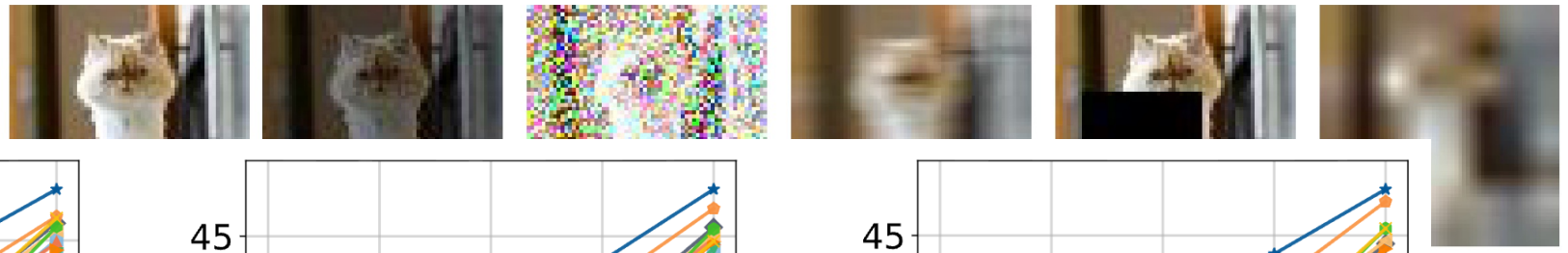

(c) 高斯噪声

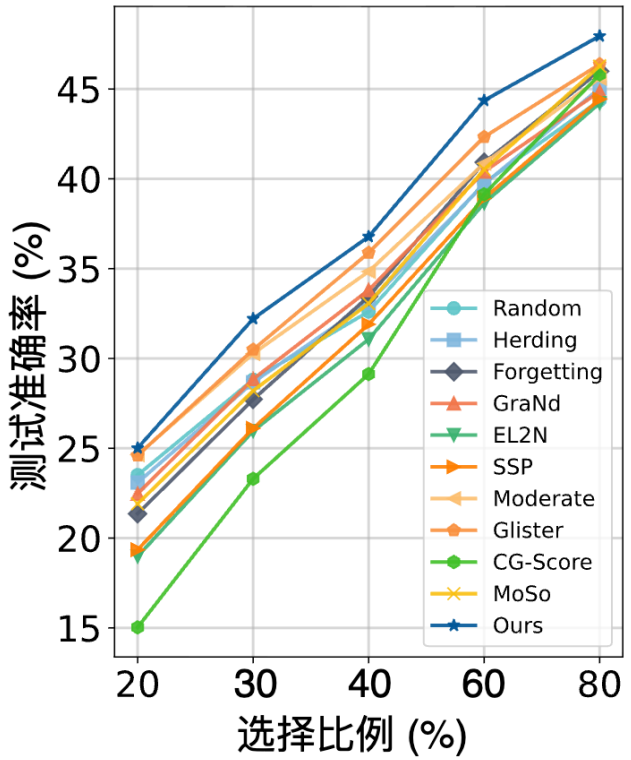

(d) 运动模糊


(e) 随机遮挡

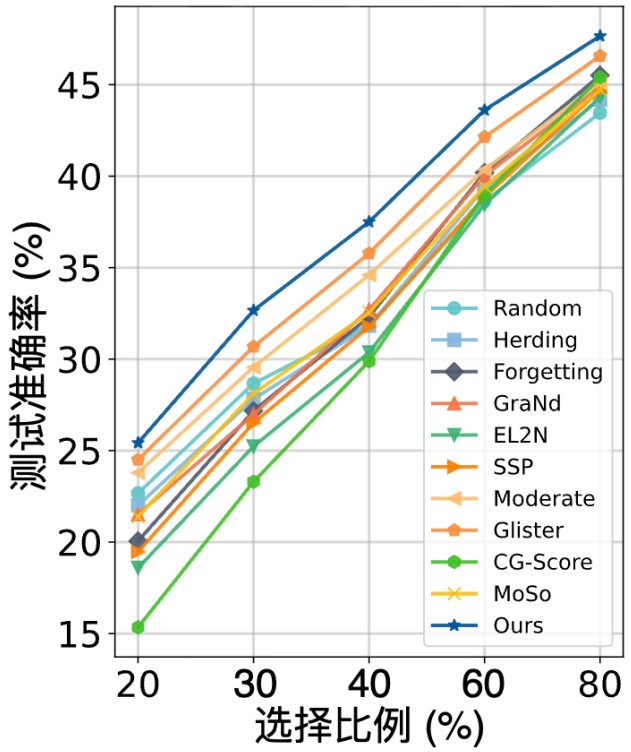

(f) 分辨率降低

对图像受损的鲁棒性

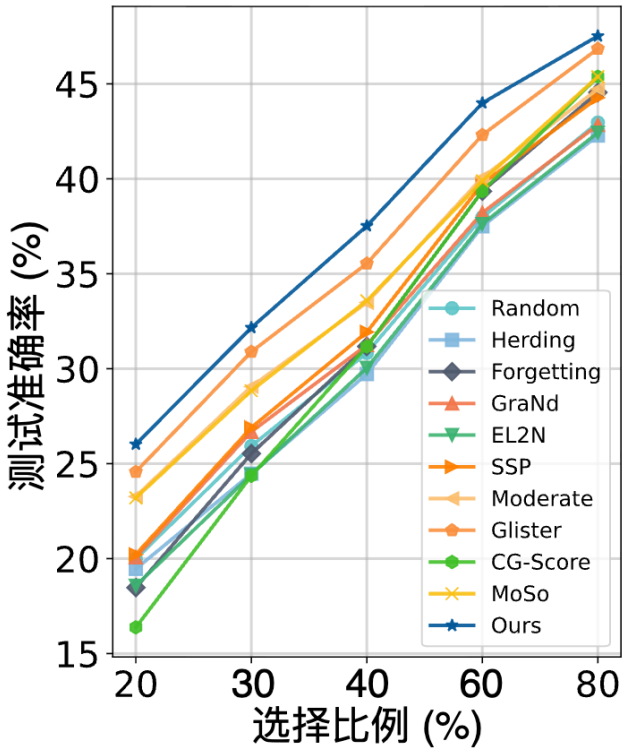




(a) 5% 受损率



(b) 10% 受损率



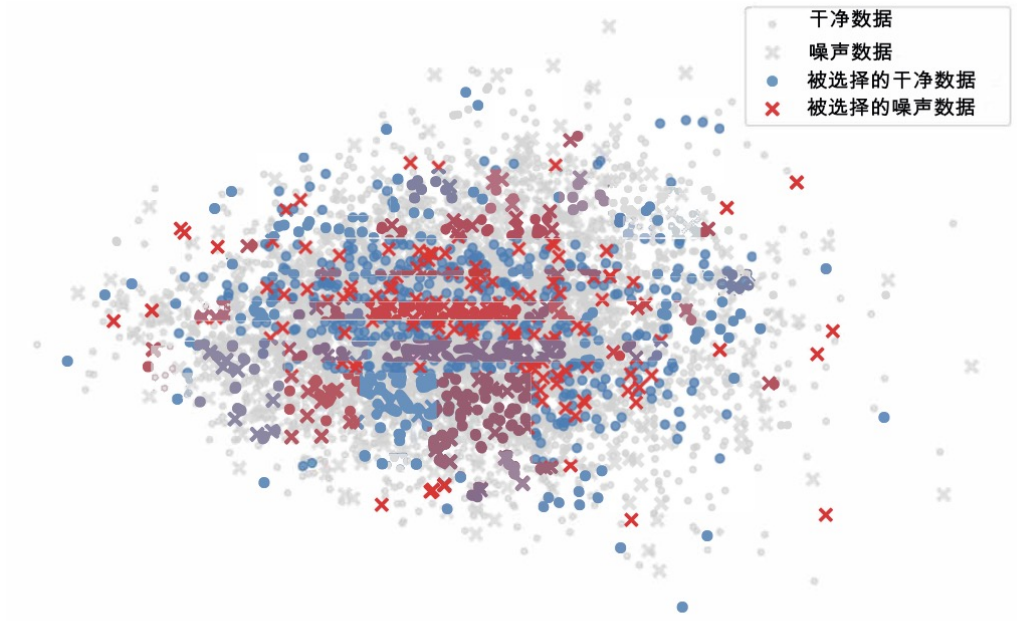
(c) 20% 受损率

率降低

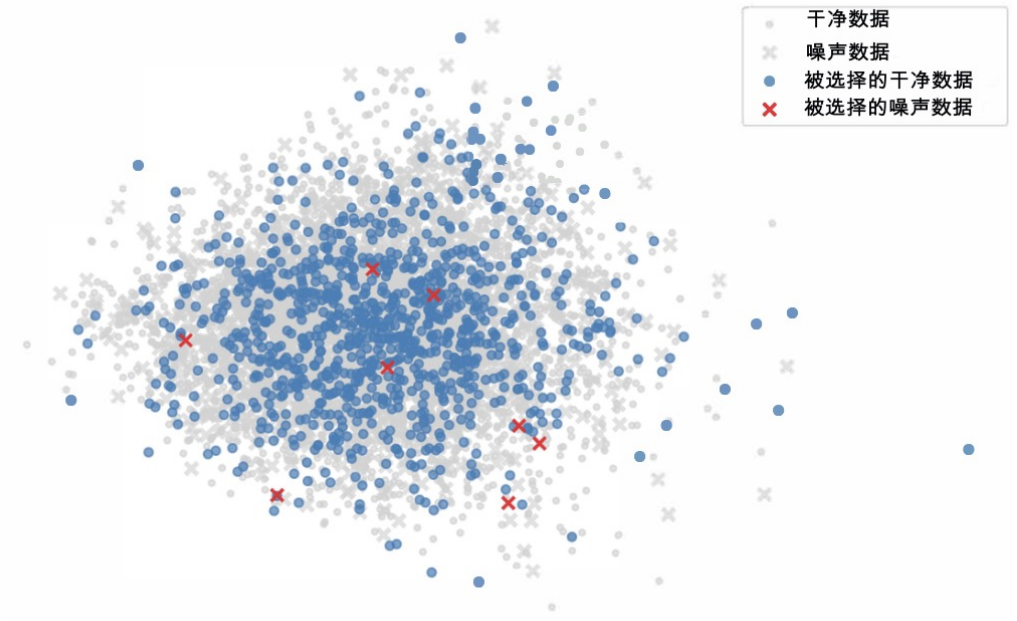
①可以应对**多种形式下的图像内容受损**；

②在**受损率高达20%**的情况下，仍能保持泛化；

噪声场景下的可视化分析



(a) 随机选择



(b) 本章的方法

①随机选择会引入**大量噪声数据**；

②被选择的噪声数量极少，所选数据能**覆盖大部分特征空间**；

第五章

提出了基于相似性驱动的多模态数据选择算法

- ✓ 围绕基于相似性与多样性分析的数据研究框架，提出一种全新的致力于相似度提升的数据选择方法，实现了更加鲁棒且泛化的数据筛选机制

- ✓ 该工作对应论文成果：

Suorong Yang, Peng Ye, Wanli Ouyang, Dongzhan Zhou, & Furao Shen, A CLIP-Powered Framework for Robust and Generalizable Data Selection. ICLR 2025.

基于第三章的相似性-多样性分析框架，通过相似性驱动来提升训练数据集的质量

RINC
Robotic Intelligence & Neural Computing Group

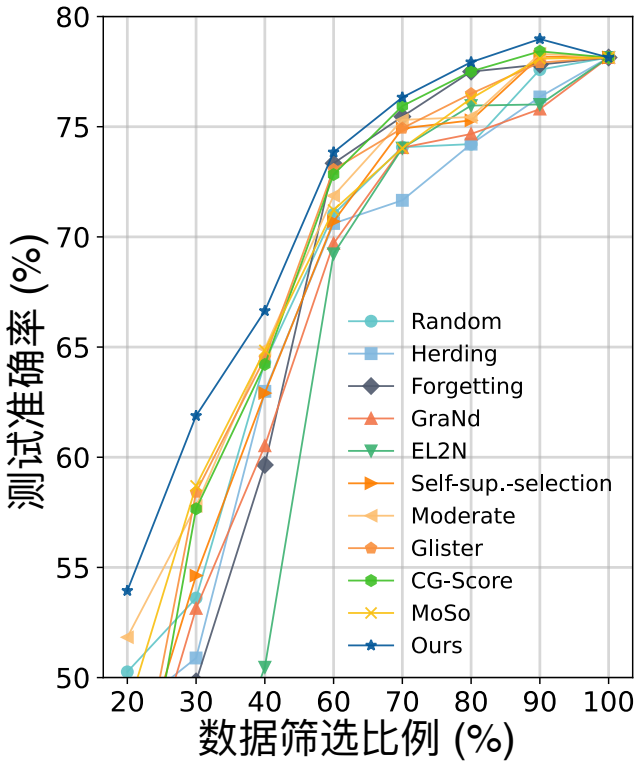
基于相似性-多样性联合优化的增强与选择协同研究

PART FIVE

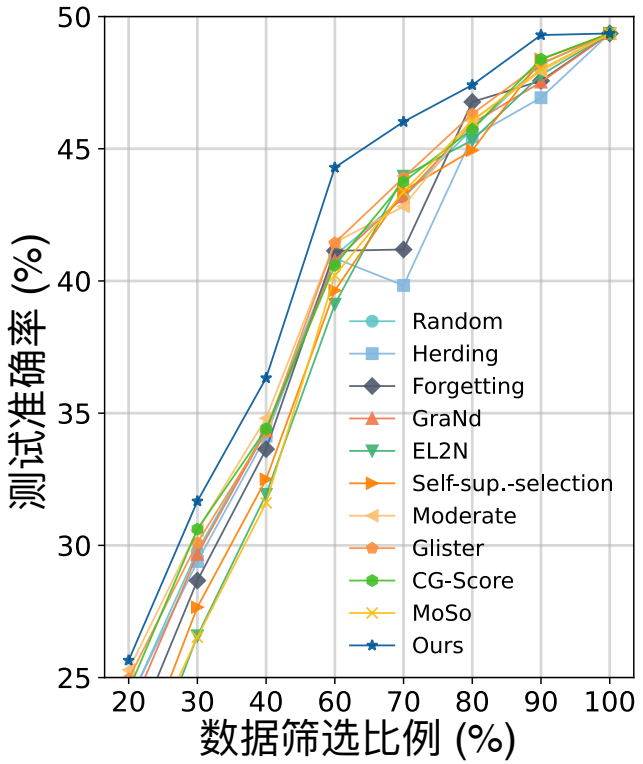
誠樸雄偉 勵學敦行

Suorong Yang, Peng Ye, Furao Shen, Dongzhan Zhou, When Dynamic Data Selection Meets Data Augmentation: Achieving Enhanced Training Acceleration. ICML 2025.

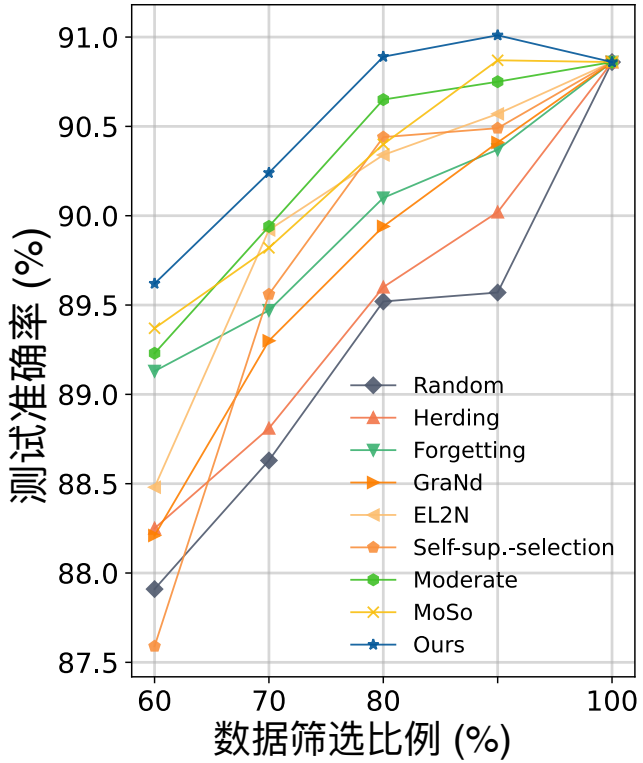
问题动机



CIFAR100

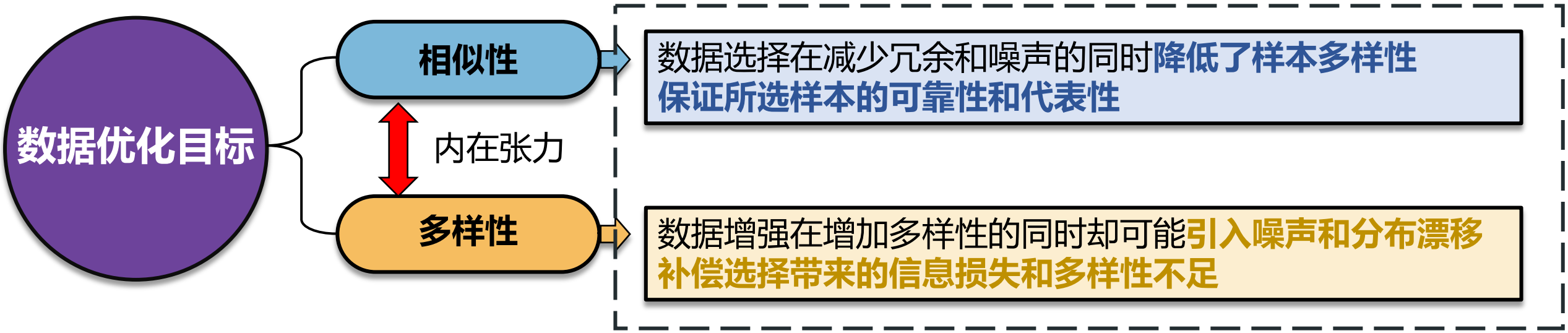


Tiny-ImageNet



ImageNet-1k

- **虽然数据筛选可以去冗余，但是当数据压缩率进一步下降时，性能下降显著**


数据优化目标
相似性
内在张力
多样性
数据选择在减少冗余和噪声的同时降低了样本多样性
保证所选样本的可靠性和代表性
数据增强在增加多样性的同时却可能引入噪声和分布漂移
补偿选择带来的信息损失和多样性不足

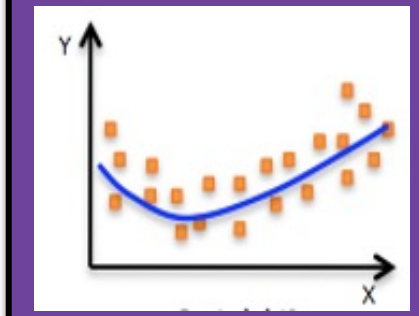
数据优化目标

相似性 ⇔ 内在张力 ⇔ **多样性**

数据选择在减少冗余和噪声的同时**降低了样本多样性**
保证所选样本的可靠性和代表性

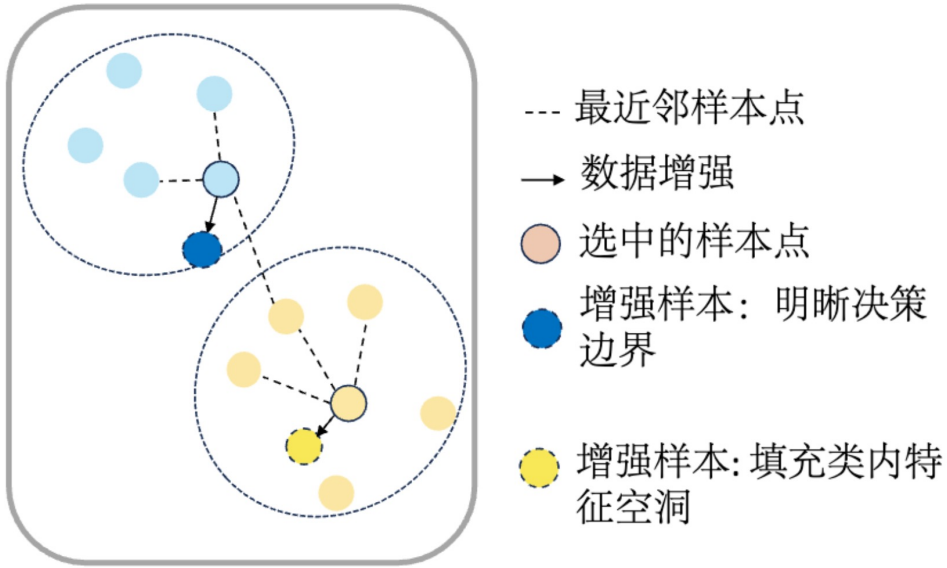
数据增强在增加多样性的同时却可能**引入噪声和分布漂移**
补偿选择带来的信息损失和多样性不足

选择什么样的样本用于增强?

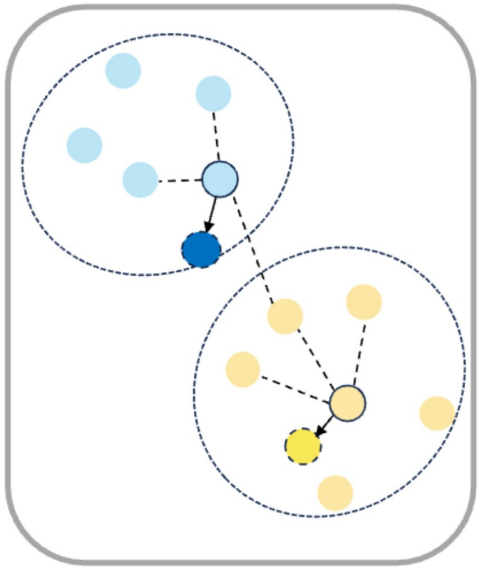


模型分布学习数据分布

关注欠表示、欠学习区域 → 低密度区域



基于密度分布的样本选择



用近邻样本来估计局部密度

$$\rho_{x_i} = \frac{1}{k} \sum_{j \in NN(x_i)} ||x_i - x_j||,$$

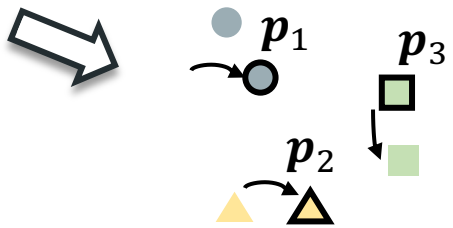
样本的局部密度反比于到其近邻的距离

噪声样本、离群样本→低密度

$$con(x_i) = \ell_{cos}\left(E_I(x_i), E_T(y_i)\right),$$

跨模态的相似性驱动

数据增强器



$$p_{sel}(x_i) = p_\rho(x_i) * p_{con}(x_i).$$

联合分布：结构稀疏性+语义相似性

性能测评

Dataset	CIFAR-10			CIFAR-100			Tiny-ImageNet		
Whole Dataset	95.6			78.2			45.0		
Selection Ratio (%)	30	50	70	30	50	70	30	50	70
Random	90.2	92.3	93.9	69.7	72.1	73.8	29.8	37.2	42.2
Herding[67]	80.1	88.0	92.2	69.6	71.8	73.1	29.4	31.6	39.8
EL2N[53]	91.6	95.0	95.2	69.5	72.1	77.2	26.6	37.1	44.0
GraNd[53]	91.2	94.6	95.3	68.8	71.4	74.6	29.7	36.3	43.2
Glister[62]	90.9	94.0	95.2	70.4	73.2	76.6	30.1	39.5	43.9
Forgetting[20]	91.7	94.1	94.7	69.9	73.1	75.3	28.7	33.0	41.2
Moderate-DS[52]	91.5	94.1	95.2	70.2	73.4	77.3	30.6	38.2	42.8
Self-sup. prototypes[71]	91.0	94.0	95.2	70.0	71.7	76.8	27.7	37.9	43.4
MoSo[21]	91.1	94.2	95.3	70.9	73.6	77.5	31.2	38.5	43.4
DP[51]	90.8	93.8	94.9	-	73.1	77.2	-	-	-
Random*	93.0	94.5	94.8	74.4	75.3	77.3	41.5	42.8	43.1
UCB[54]	93.9	94.7	95.3	-	75.3	77.3	-	-	-
ϵ-Greedy[54]	94.1	94.9	95.2	-	74.8	76.4	-	-	-
InfoBatch[23]	94.7	95.1	95.6	76.5	78.1	78.2	42.2	43.2	43.8
Ours	**94.9**	**95.5**	**96.0**	**77.6**	**78.9**	**79.5**	**44.9**	**47.0**	**49.4**

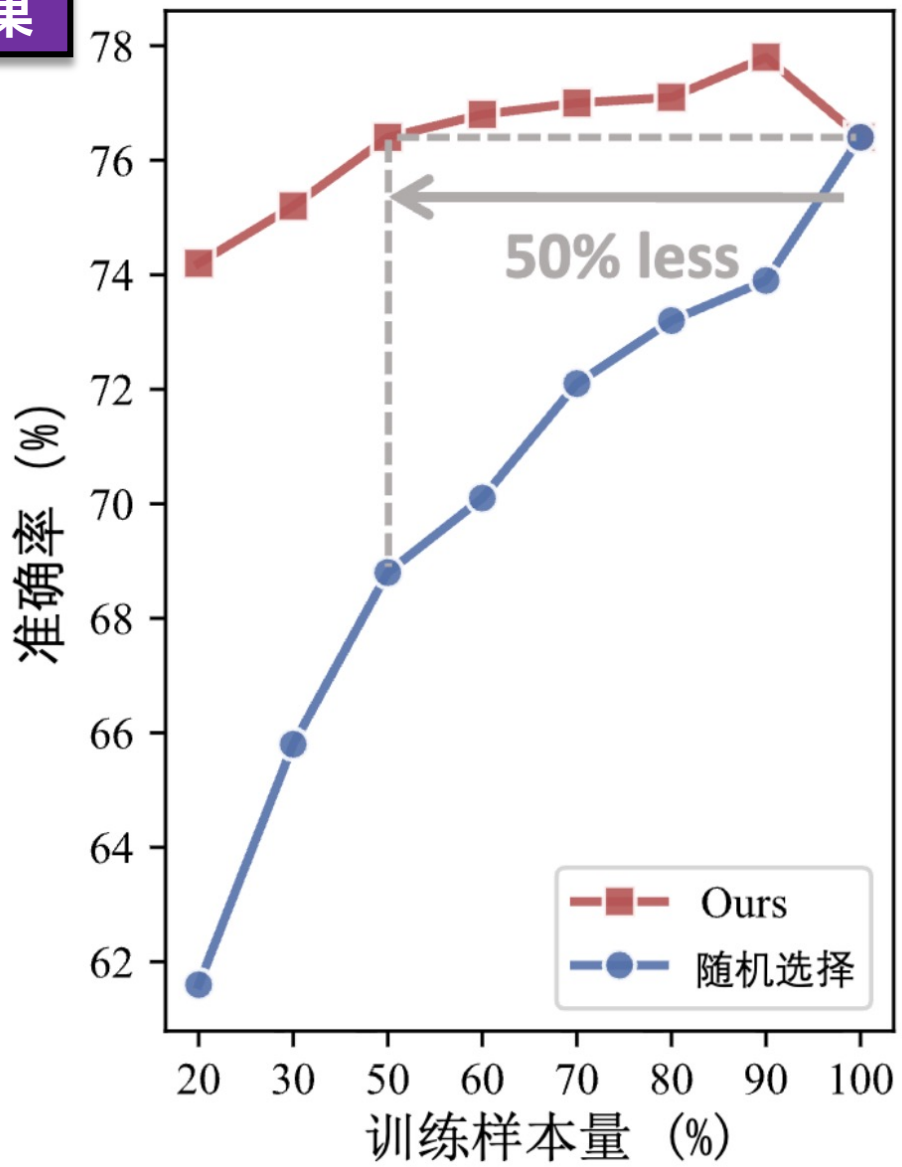
① 本章工作与其他baseline对比

在所有数据集和所有压缩率下的分类准确性上优于其他算法

② 本章工作的性能

可以显著降低无损压缩率，例如在Tiny—ImageNet上只用30%的训练成本即可以实现全集训练的效果

训练加速效果



① 可以显著降低由于数据量压缩带来的信息损失

② 首个在ImageNet-1k上实现50%无损压缩率的方法

③ 对应的训练时间减少约70个GPU小时的同时性能不会降低

噪声场景下的鲁棒性实验

方法 / 选择比例 (%)	Noisy		Corrupted	
	20	30	20	30
Random	17.8	23.9	20.0	25.9
Herding	19.0	24.2	35.0	30.6
Moderate-DS	19.6	25.0	23.3	29.1
EL2N	13.9	18.6	18.6	24.4
GraNd	18.3	23.7	20.0	26.7
Forgetting	13.2	21.8	18.5	25.5
Self-sup. prototypes	15.1	21.0	20.2	26.9
CG-Score	8.4	15.3	16.4	24.4
Glister	21.6	25.5	21.2	22.0
MoSo	7.4	11.3	23.1	28.8
Random*	33.8	36.5	35.1	36.9
InfoBatch	34.9	37.1	35.1	38.1
Ours	**35.9**	**39.6**	**39.1**	**42.0**

ViT架构下的泛化性实验

S_r(%)	50	60	70	80	90	全集
ViT-B	82.6	82.9	83.2	83.2	**83.3**	82.5
ViT-L	85.2	85.3	85.6	**85.7**	**85.7**	84.6
Swin-T	84.1	84.1	84.2	84.2	**84.3**	84.2

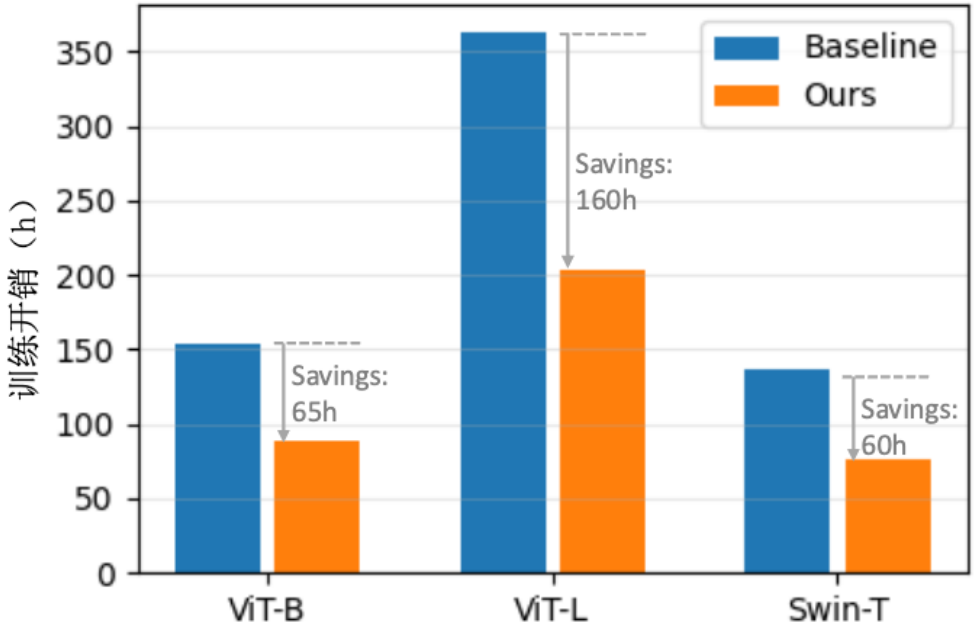
① 本章工作与其他baseline对比

在噪声场景下的鲁棒性显著优于其他算法

② 本章工作在ViT更大的模型架构下，稳定取得有效的训练加速

在ViT-B上节省65h、ViT-L上节省160h、Swin-T上节省60h的GPU开销

噪声场景下的鲁棒性实验



MoSo	7.4	11.3	23.1	28.8
Random*	33.8	36.5	35.1	36.9
InfoBatch	34.9	37.1	35.1	38.1
Ours	**35.9**	**39.6**	**39.1**	**42.0**

ViT架构下的泛化性实验

$S_r(\%)$	50	60	70	80	90	全集
ViT-B	82.6	82.9	83.2	83.2	**83.3**	82.5
ViT-L	85.2	85.3	85.6	**85.7**	**85.7**	84.6
Swin-T	84.1	84.1	84.2	84.2	**84.3**	84.2

① 本章工作与其他baseline对比

在噪声场景下的鲁棒性显著优于其他算法

② 本章工作在ViT更大的模型架构下，取得有效的训练加速

在ViT-B上节省65h、ViT-L上节省160h、Swin-T上节省60h的GPU开销

第六章

提出了基于相似性-多样性联合优化的增强与选择协同算法

- ✓ **主动识别最需要增强的样本，结合语义相似性的约束，在显著提升样本多样性的同时降低训练开销**
- ✓ **该工作对应论文成果：**

Suorong Yang, Peng Ye, Furao Shen, & Dongzhan Zhou, When Dynamic Data Selection Meets Data Augmentation: Achieving Enhanced Training Acceleration. **ICML 2025.**

通过公开数据集指标评估验证了算法的有效性

RINC
Robotic Intelligence & Neural Computing Group

总结与展望

PART SIX

誠樸雄偉 勵學敦行

- 形式化了**一般的数据分析框架**
- 提供方法设计的理论动机与可解释性支撑

- 提出了**多样性提升的数据优化框架**
- 结合强化学习提出了自适应多样性提升的**数据增强算法**

- 提出了**相似性驱动的数据优化框架**
- 结合多模态大模型提出了**数据质量筛选算法**

- 提出了**相似性-多样性联合优化框架**
- 整合增强与选择提出了**协同数据优化算法**

展望1

扩展到更多模态的通用框架值得期待

本文验证了该数据优化框架在图像任务上的性能，未来将其应用到更多模态，如文本、语音等。

展望2

多模态表征学习的优化

如何在多模态异构数据，如视频-文本场景中，如何高效构建样本间的交互仍是一个开放性问题。

展望3

自适应机制的扩展版本

AdaAugment的研究范围集中于图像分类任务，对将其应用于更大规模的训练场景可进行进一步研究。

展望4

数据智能系统与应用的进一步探索

本文的研究主要集中在方法论与实验验证层面，未来可在实际应用场景中推进数据智能系统的落地……

博士期间部分已发表论文：

在机器学习、计算机视觉顶级会议（ICLR，ICML，NeruIPS，ICCV，ECCV，CVPR）和期刊（TPAMI，TIP，PR）发表论文十余篇，累计影响因子76+。

1. **Suorong Yang,** Hongchao Yang, Suhan Guo, Furao Shen, & Jian Zhao (2025). IPF-RDA: An Information-Preserving Framework for Robust Data Augmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI，CCF-A期刊**，SCI Q1 Top).
2. **Suorong Yang**, Peijia Li, Furao Shen, & Jian Zhao, AdaAugment: A Tuning-Free and Adaptive Approach to Enhance Data Augmentation. IEEE Transactions on Image Processing (**TIP，CCF-A 期刊**，SCI Q1 Top).
3. **Suorong Yang**, Peng Ye, Wanli Ouyang, Dongzhan Zhou, & Furao Shen, A CLIP-Powered Framework for Robust and Generalizable Data Selection. （**ICLR 2025**，Spotlight top 5%）
4. **Suorong Yang**, Peng Ye, Furao Shen, & Dongzhan Zhou, When Dynamic Data Selection Meets Data Augmentation: Achieving Enhanced Training Acceleration. (**ICML 2025, CCF-A会议**)
5. **Suorong Yang**, Peijia Li, Furao Shen, & Jian Zhao, RL-Selector: Reinforcement Learning-Guided Data Selection via Redundancy Assessment. （**ICCV 2025，CCF-A会议**）

6. **Suorong Yang**, Furao Shen, & Jian Zhao, EntAugment: Entropy-Driven Adaptive Data Augmentation Framework for Enhancing Generalization in Image Classification. （**ECCV 2024，CCF-B会议**）
7. **Suorong Yang**, Jinqiao Li, Tianyue Zhang, Jian Zhao, & Furao Shen (2023). AdvMask: A sparse adversarial attack-based data augmentation method for image classification. Pattern Recognition（**CCF-B期刊，** SCI Q1 Top ）
8. **Suorong Yang**, Suhan Guo, Furao Shen, & Jian Zhao, Investigating the Effectiveness of Data Augmentation from Similarity and Diversity: An Empirical Study. Pattern Recognition（**CCF-B期刊，** SCI Q1 Top ）
10. **Suorong Yang**, Tianyue Zhang, Zhiming Xu, Leijia Li, Baile Xu, Furao Shen, & Jian Zhao, Supervised Contrastive Learning with Prototype Distillation for Data Incremental Learning. Neural Networks (**CCF-B期刊，** SCI Q1)
11. **Suorong Yang**, Hongchao Yang, Jian Zhao, & Furao Shen, Image data augmentation for deep learning: A survey. Journal of Software. （**CCF-A期刊**）

部分已授权专利：

1. 申富饶，**杨锁荣**，李俊，赵健，“一种基于实时定位轨迹数据进行动态滤波优化的方法”. CN201911005643.4
2. 申富饶，**杨锁荣**，李春华，秦辞海，杨洪朝，张天玥，陈昊，“一种基于图像配准的视频流变化检测方法和系统”. CN202211342122.X
3. 申富饶，**杨锁荣**，李俊，赵健，“实时定位轨迹数据进行局部线性插值和预测的方法”. CN202010316990.5
4. 申富饶，陈昊，**杨锁荣**，杨洪朝，卢俣金，张凌茗，刘佩涵，李若彤，赵健，“一种基于对比检测的机器人巡检系统”. CN202310111632.4

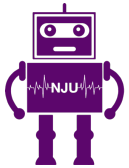
参与项目：

- 2023.1-至今：**国家自然基金面上项目**，面向增量式无监督学习的新型神经网络研究
  项目职责：学术研究，项目报告撰写
- 2021.1-至今：**科技部重大项目-科技创新 2030 项目**，基于神经可塑性的脉冲神经网络高效学习机制与类脑智能系统
  项目职责：算法设计，项目报告撰写
- 2019.1-2022.12: **国家自然基金面上项目**，基于深度感知增量式联想记忆神经网络的信息融合系统研究
  项目职责：算法设计，项目报告撰写
- **上海市科技重大专项**

部分获奖情况：

南京大学2023、2024年度优秀研究生、博士英才奖学金等。

南京大学
NANJING UNIVERSITY



RINC
Robotic Intelligence & Neural Computing Group

感谢各位老师
敬请批评指正

誠樸雄偉　勵學敦行