山东大学
人工智能研究中心

TIME 机器学习与数据挖掘实验室

NEURAL INFORMATION PROCESSING SYSTEMS

From Pretraining to Pathology: How Noise Leads to Catastrophic Inheritance in Medical Models

NeurIPS 2025

Hao Sun, Zhongyi Han*, Hao Chen, Jindong Wang, Xin Gao, Yilong Yin*

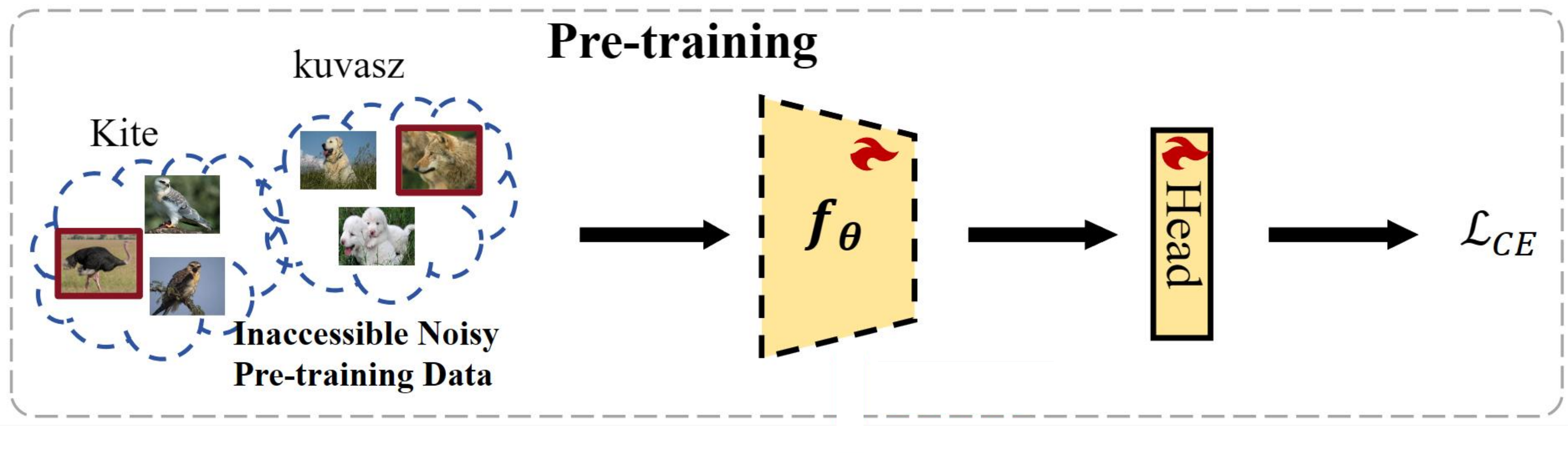
*corresponding author

学无止境 气有浩然

Motivation

山东大学 人工智能研究中心

Catastrophic Inheritance in Medical Models.



- **Does pretraining noise affect downstream medical performance?**
- **Why does such degradation emerge?**
- **How can we mitigate it efficiently without retraining the foundation model?**

Influence

The label noise in pre-training induces structural degradation in down-stream medical tasks.

- Controlled pre-training with varying noise ratios (0–30%) on ImageNet-1K.
- Evaluated via linear probing on Camelyon17, HAM10000, and NIH ChestXray.

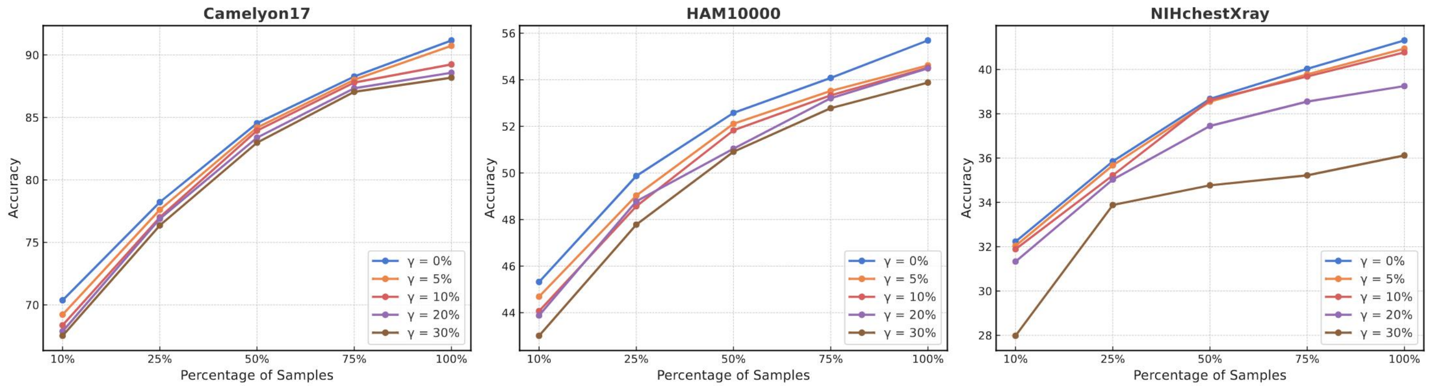


Figure 2: Average evaluation results of ImageNet-1K (IN-1K) fully supervised pre-training on downstream tasks with various percentages of data using ResNet-50. The robustness performance constantly decreases once noise is introduced in pre-training.

Analysis

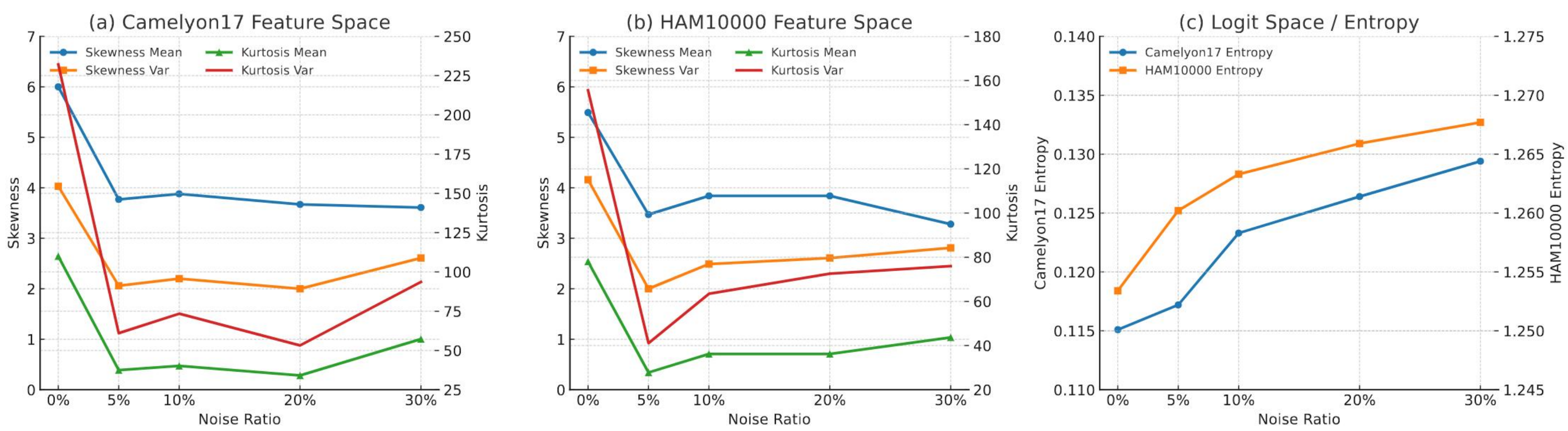
The pre-training noise flattens the representational space by reducing skew-ness and kurtosis.

Definition 1 (Feature-wise Skewness). The skewness of feature dimension j is defined as:

$$Skew(F_{:,j}) = \frac{\mathrm{M}}{(\mathrm{M}-1)(\mathrm{M}-2)} \sum_{i=1}^{M} \left(\frac{F_{i,j}-\mu_j}{\sigma_j}\right)^3$$

Definition 2 (Feature-wise Kurtosis). The kurtosis of feature dimension j is defined as:

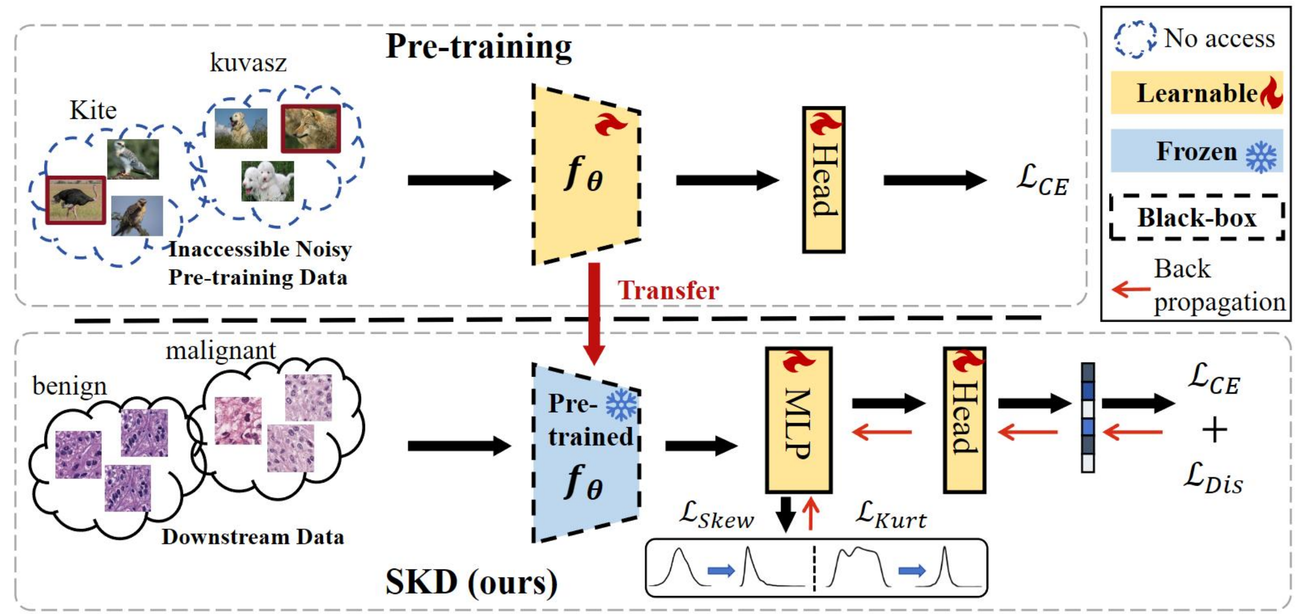
$$Kurt(F_{:,j}) = \frac{\mathrm{M}(\mathrm{M}+1)}{(\mathrm{M}-1)(\mathrm{M}-2)(\mathrm{M}-3)} \sum_{i=1}^{M} \left(\frac{F_{i,j}-\mu_j}{\sigma_j}\right)^4 - \frac{3(M-1)^2}{(M-2)(M-3)}$$



Mitigation

山东大学 人工智能研究中心

Mitigating the Noise with Regularization on Distributional Shape



$$\mathcal{L}_{\text{skew}} = \frac{1}{D}\sum_{j=1}^{D} |\text{Skew}(F_{:,j}) - \tau_s|$$

$$\mathcal{L}_{\text{kurt}} = \frac{1}{D}\sum_{j=1}^{D} |\text{Kurt}(F_{:,j}) - \tau_k|$$

$$\mathcal{L}_{\text{dis}}(x, y) = \frac{1}{\log 2} \log\left(1 + \exp\left(h(x)_y - \frac{1}{|\mathcal{Y}| - 1}\sum_{\hat{y} \neq y} h(x)_{\hat{y}}\right)\right)$$

Overall Objective

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_s \mathcal{L}_{\text{skew}} + \lambda_k \mathcal{L}_{\text{kurt}} + \lambda_d \mathcal{L}_{\text{dis}}$$

Experiments

山东大学
人工智能研究中心

Main Results

Pretrained	Dataset	Method	0%	5%	10%	20%	30%	Avg Gain
CLIP	Camelyon17	LP	83.94	81.88	81.71	81.06	80.18	-
		GCE	83.12	82.74	82.21	81.57	80.68	0.31
		NML	80.61	84.21	83.32	84.68	85.61	1.93
		SKD	**89.48**	**89.59**	**89.50**	**86.08**	**86.65**	**6.51**
	HAM10000	LP	49.36	47.78	45.88	45.07	45.26	-
		GCE	50.54	48.44	47.20	45.89	45.98	0.94
		NML	50.76	51.35	49.03	52.78	**50.04**	4.12
		SKD	**55.19**	**55.88**	**55.75**	**53.32**	49.84	**7.33**
	ChestX-ray	LP	44.75	42.00	42.78	41.58	41.75	-
		GCE	45.42	42.71	43.11	42.06	42.13	0.51
		NML	36.02	35.71	35.92	36.58	37.19	-6.29
		SKD	**45.92**	**45.81**	**45.48**	**45.45**	**45.86**	**3.13**
ResNet50	Camelyon17	LP	91.16	90.73	89.24	88.57	88.18	-
		GCE	91.11	91.43	89.48	89.12	88.64	0.38
		NML	89.27	92.44	88.09	**90.51**	**91.29**	0.74
		SKD	**91.76**	**92.50**	**91.39**	89.12	89.02	**1.18**
	HAM10000	LP	55.69	54.62	54.52	54.49	53.88	-
		GCE	55.73	54.79	54.52	54.51	54.46	0.16
		NML	54.71	55.16	54.21	54.77	50.67	-0.74
		SKD	**58.25**	**57.54**	**56.94**	**58.37**	**57.54**	**3.09**
	ChestX-ray	LP	41.31	35.94	40.77	39.25	36.12	-
		GCE	41.37	36.28	41.23	40.02	36.67	0.44
		NML	38.94	36.40	36.55	37.38	38.79	-1.07
		SKD	**44.81**	**43.37**	**44.22**	**44.25**	**45.39**	**5.73**

Experiments

山东大学
人工智能研究中心

Real-world validation results

Table 2: Real-world evaluation on PLIP using its original medical datasets. SKD consistently outperforms baselines across F1 and accuracy.

Model	Dataset	Method	F1	Accuracy
PLIP	Kather colon	Zero-Shot	0.565	-
		LP(origin)	0.877	-
		LP	0.899	0.895
		NML	0.931	0.929
		SKD	**0.959**	**0.959**
	PanNuke	Zero-Shot	0.656	-
		LP(origin)	0.902	-
		LP	0.930	0.930
		NML	0.948	0.948
		SKD	**0.956**	**0.956**
	DigestPath	Zero-Shot	0.832	-
		LP(origin)	0.856	-
		LP	0.968	0.968
		NML	**0.979**	**0.979**
		SKD	0.976	0.969
	WSSS4LUAD	Zero-Shot	0.734	-
		LP(origin)	0.927	-
		LP	0.952	0.952
		NML	0.956	0.956
		SKD	**0.958**	**0.958**

Table 3: Real-world evaluation on biomedical NER tasks using PubMedBERT across five datasets. SKD consistently improves both F1 and accuracy over LP and NML, demonstrating its effectiveness beyond medical imaging.

Dataset	Method	F1	Accuracy
BC2GM	LP	0.9053	0.9222
	NML	0.9187	0.9271
	SKD	**0.9459**	**0.9501**
NCBI-disease-IOB	LP	0.9330	0.9355
	NML	0.9378	0.9402
	SKD	**0.9459**	**0.9471**
JNLPBA	LP	0.8895	0.8825
	NML	0.9032	0.8957
	SKD	**0.9211**	**0.9128**
BC4CHEMD	LP	0.9373	0.9485
	NML	0.9506	0.9567
	SKD	**0.9706**	**0.9725**
BioNLP11EPI-IOB	LP	0.9286	0.9401
	NML	0.9362	0.9426
	SKD	**0.9481**	**0.9492**

Q & A

Thanks!

学无止境 气有浩然