CERIAS

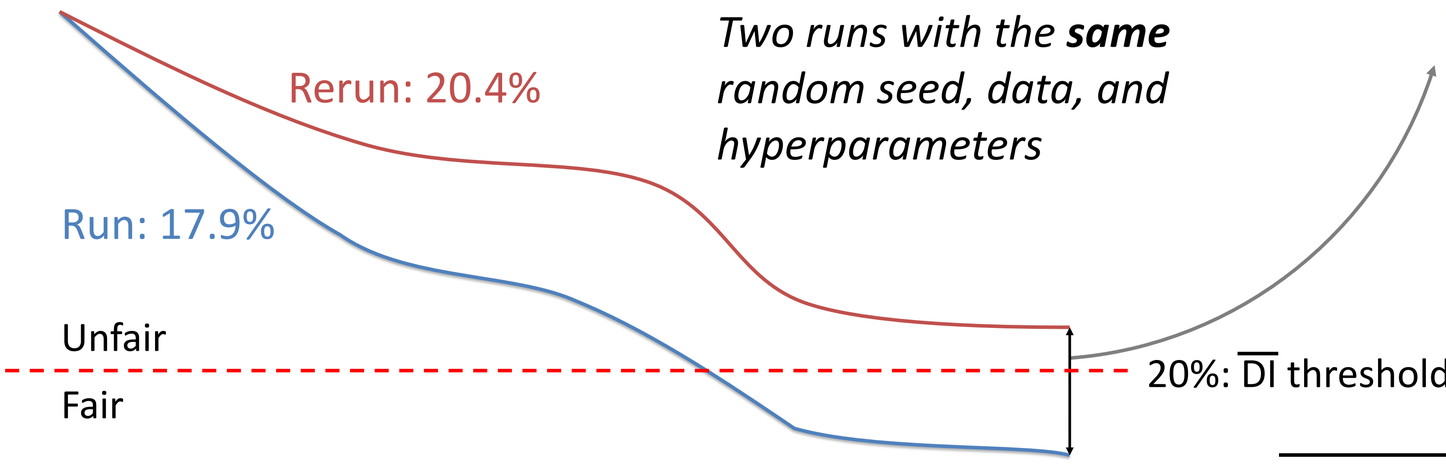
The Center for Education and Research in Information Assurance and Security

Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training

Shangshu Qian[1], Hung Viet Pham[2], Thibaud Lutellier[2], Zeou Hu[2], Jungwon Kim[1], Lin Tan[1], Yaoliang Yu[2], Jiahao Chen[3], Sameena Shah[3]

[1]Purdue University [2]University of Waterloo [3]J.P. Morgan AI Research

Motivation: Deep learning (DL) training is non-deterministic even with a **fixed** random seed.



*Two runs with the **same** random seed, data, and hyperparameters*

2.5% bias difference is caused by DL **software implementations** (e.g., TensorFlow).

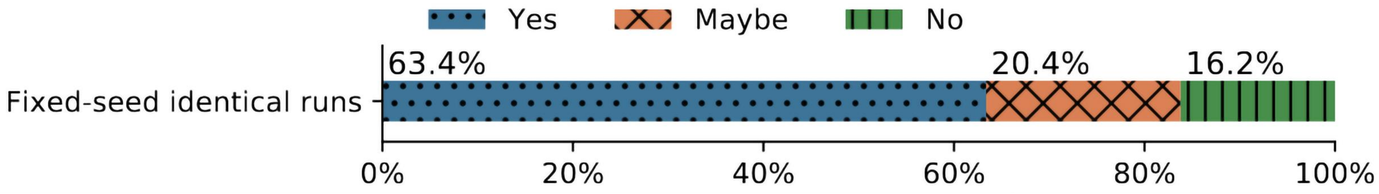
- Floating-point calculations are not associative.
- DL software selects primitive operations at runtime.

Variance **affects** the **evaluation** of debiasing algorithms: debiasing algorithms **increase** bias amplification.

	Baseline	Debiased
One run	7.8%	6.1% More fair
Avg. of 16 runs	7.4%	8.7% Less fair

Motivation: Over **80%**[1] of DL researchers and practitioners are **unaware or unsure** about variance in DL model training!

Q26[1] - Do you expect fixed-seed identical DL training runs to be deterministic?



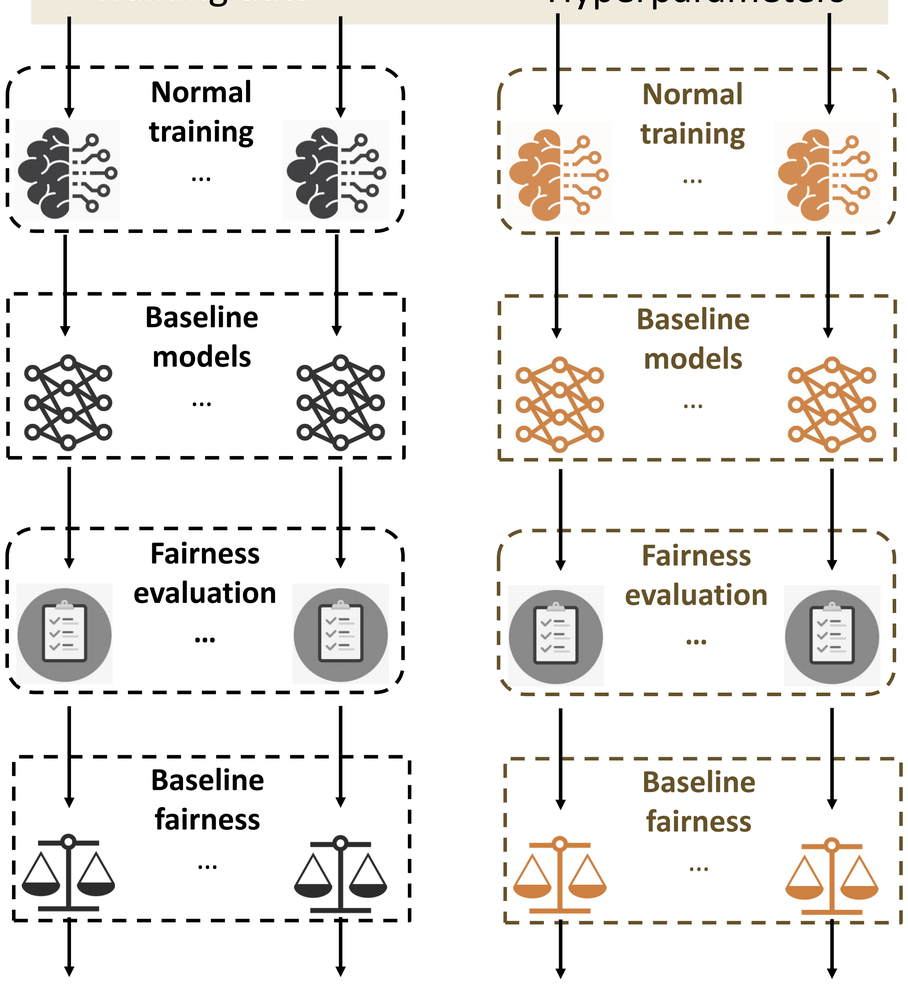
A variance analysis on DL models' fairness is needed!

Approach: Fairness Variance Analysis

Fixed-Seed **I**dentical **T**raining (FIT) Runs

Using the **same**

- Random seed
- Training data
- Train test split
- Hyperparameters



Statistical analysis (baseline VS debiased)

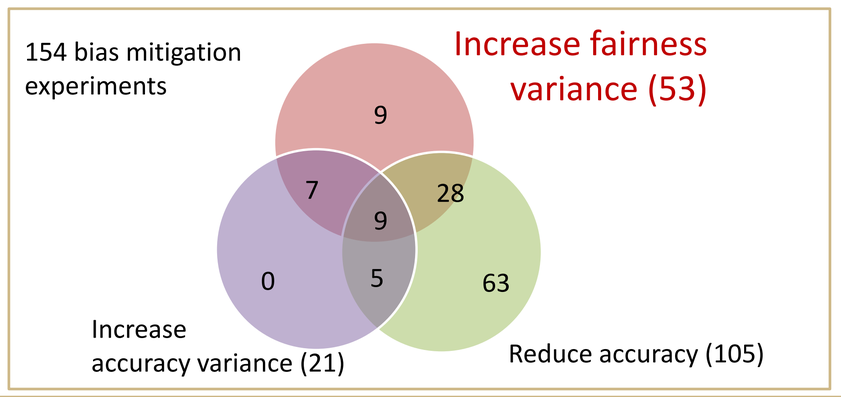
- Mann-Whitney U-test for mean value
- Levene's test for variance

Finding: Software alone causes large fairness gap (**up to 12.6%**)!

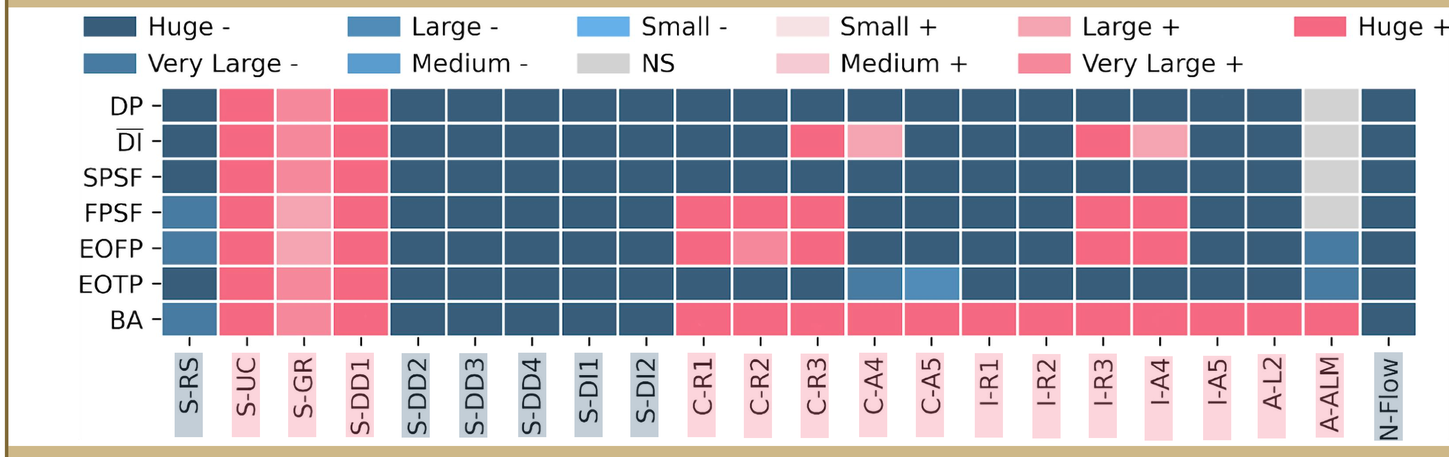
Technique	Metric	MaxDiff (%)	Max (%)	Min (%)	Avg (%)
A-L2	DP	**12.6**	39.9	27.3	35.2
S-GR	$\overline{DI}$	**11.8**	31.1	21.3	28.5

Finding: Hidden cost of debiasing includes higher fairness variance.

About **one third** (53/154) of the bias mitigation experiments **increases** fairness variance compared with the corresponding baseline.



Finding: Most (**15 out of 22**) debiasing techniques increase at least one bias metric.



We call for awareness of **implementation-level non-determinism**: **using proper statistical tests** to ensure the validity of deep learning experiments and more!

Reference

[1] Pham, Qian, Wang, Lutellier, Rosenthal, Tan, Yu, & Nagappan. Problems and opportunities in training deep learning software systems: an analysis of variance. ASE 2020.

Check out our paper here!