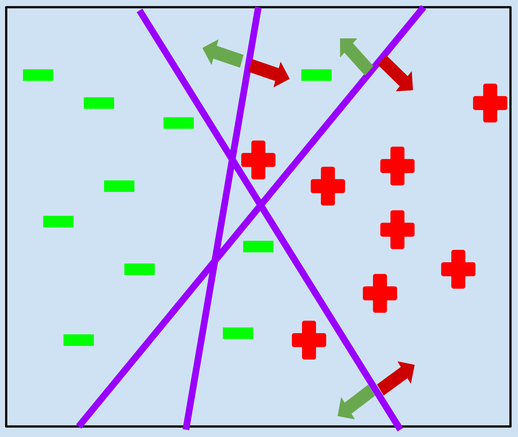
An Empirical Investigation into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification

Prakhar Ganesh, McGill University & Mila

What is Model Multiplicity?

Model multiplicity is the existence of multiple high-performing models with diverse predictive behaviours due to varying learned functions.



Many different decision boundaries can still have 'similar accuracy' on some benchmark, and thus making the right choice is important to avoid failure at deployment.

Model Multiplicity and Trustworthy ML

Model multiplicity presents an opportunity to deploy trustworthy models without compromising their utility. Thus, addressing the challenges of multiplicity can mark a crucial step towards creating trustworthy systems.

Accuracy Under Intervention

Security Metric: Adversarial Distance → Intervention: Accuracy Under Adversarial Attack

Not comparable

Comparable!

Robustness Metric: OOD Performance → Intervention: Accuracy on OOD Dataset

Multiplicity Sheets

Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	92.85	92.85	92.37	0.49	92.85	92.81	92.18	0.68	92.85	92.51	0.34	92.85	92.60	0.25	92.85	92.22	92.13	0.72
	92.89	92.64	92.60	0.30	92.89	92.98	92.62	0.36	92.89	92.89	0.00	92.89	92.55	0.34	92.89	92.49	92.05	0.84
	92.47	92.51	92.43	0.08	92.47	92.66	92.60	0.19	92.47	92.87	0.40	92.47	92.68	0.21	92.47	92.24	92.45	0.23
	93.17	92.91	92.64	0.53	93.17	93.08	92.68	0.49	93.17	92.79	0.38	93.17	92.87	0.30	93.17	92.05	92.20	1.12
	92.60	92.98	92.66	0.38	92.60	92.32	92.07	0.53	92.60	92.87	0.27	92.60	92.45	0.15	92.60	92.30	92.18	0.42
Δ_{max}	0.70	0.46	0.30		0.70	0.76	0.61		0.70	0.38		0.70	0.42		0.70	0.44	0.40	

92 94 0 2

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **Accuracy**

Dataset: **UTKFace**

Δ^{all}_{max}: **1.12**

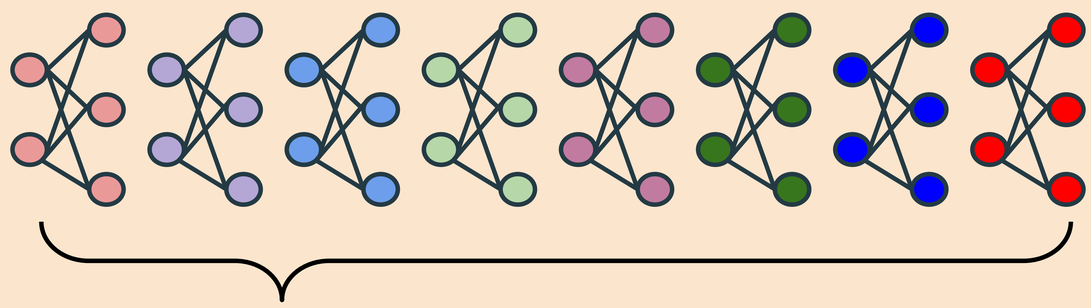
First Level of Readability: All raw scores are present in the multiplicity sheet for reference.

Second Level of Readability: We also record Δ_{max} values across each axis to capture trends of every single choice.

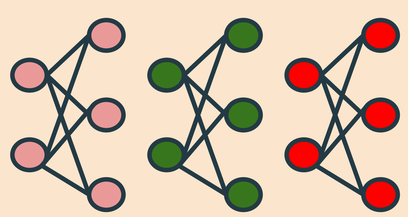
Third Level of Readability: Finally, we record Δ^{all}_{max} to record the overall variance in the given metric, in an attempt to capture multiplicity in a single score.

We create multiplicity sheets for fairness, robustness, privacy, and security; across different learning rates, batch sizes, augmentations, optimizers, and architectures.

Model Selection and Unseen Failures



Model Selection Based on Some Trustworthy Metrics



How will this Selection Impact Other Unseen Metrics?

Despite combining several trustworthy metric for model selection, we see persistent multiplicity on unseen trustworthy metrics, underscoring the need for better safeguards against multiplicity when deploying models.