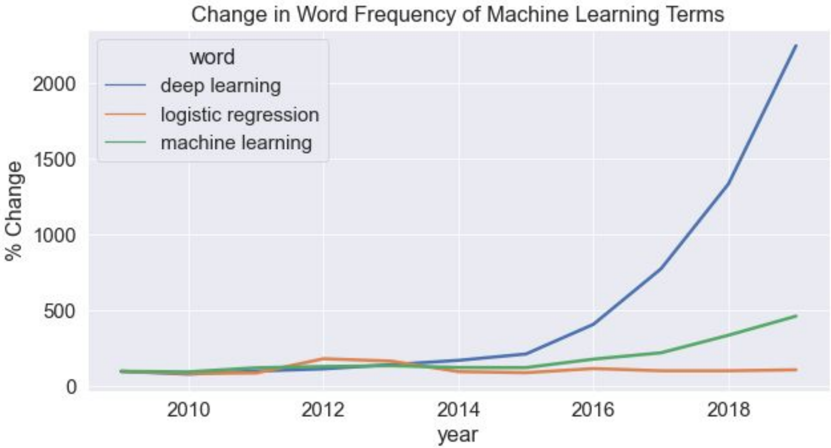
Measure Twice, Cut Once: Quantifying Bias and Fairness in Deep Networks

AFCR 2021

Presented by: Cody Blakeney and Gentry Atkinson

Impact of bias and past work.

- Machine Learning is becoming larger and less interpretable.
- Biased models can perform well in testing but fail in the real world.
- Unfair models degrade public trust of ML.
- Many metrics exist to measure fairness in ML, but the focus has been overwhelmingly on binary classification.



What are Bias and Fairness?

- Bias when used in this presentation -> performance by a classifier that varies greatly in one or several classes.
- Fairness:
    - *Many* possible definitions. No one system, metric, or platform is going to address every possible fairness.
    - One common taxonomy: individual fairness, group fairness(possible intersectional), and sub-group fairness.

CEV and SDE

Combined Error Variance measures how much class-wise error rates change relative to the mean when comparing to model.

$$\delta X_{ie} = \frac{X_{ie} - \hat{X_{ie}}}{\hat{X_{ie}}} \qquad (1) \qquad \delta X_{\mu e} = \frac{1}{n}\sum_{i=0}^{n}(\delta X_{ie})$$

$$cev = \frac{1}{n}\sum_{i=1}^{n}(dist((\delta X_{\mu pos}, \delta X_{\mu neg}), (\delta X_{ipos}, \delta X_{ineg})))^2$$

Symmetric Distance Error measures how much error rates shift towards false positives or negatives when comparing two models.

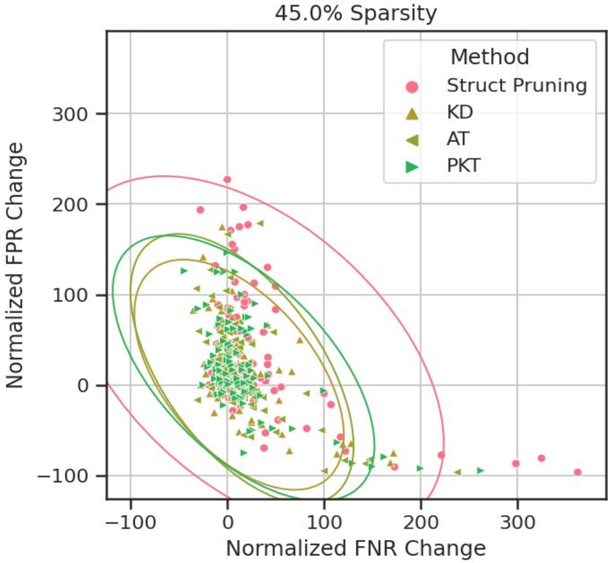
$$sde = \frac{1}{n}\sum_{i=0}^{n}|\delta FNR_i - \delta FPR_i|$$

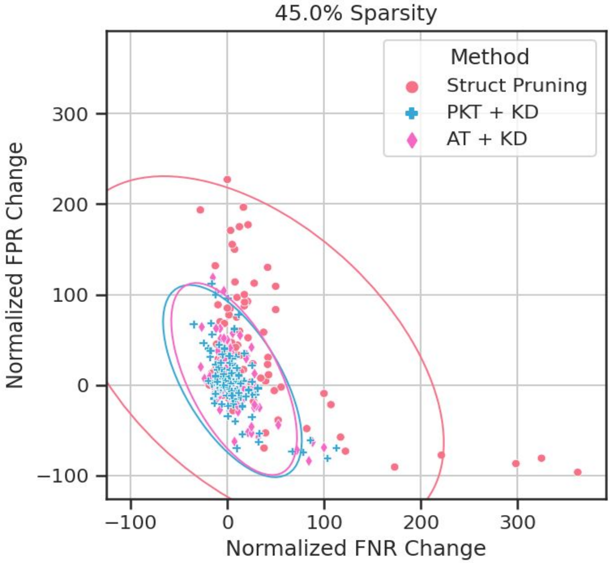
Compression

- We combined knowledge distillation and pruning on a biased CIFAR100 set while applying compression
- Accuracy alone is a poor selection criteria when considering compression methods

Method	# of CIEs	CEV	SDE	Accuracy
AT + KD	742	0.00187	0.13173	77.100
PKT + KD	748	0.00199	0.13098	77.335
SP + KD	768	0.00331	0.16162	76.927
FSP + KD	742	0.00333	0.16002	75.285
KD (Hinton et al., 2015)	770	0.00338	0.16065	78.142
AT (Zagoruyko & Komodakis, 2017)	909	0.00430	0.19306	78.097
PKT (Passalis & Tefas, 2018)	881	0.00481	0.19891	78.963
SP (Tung & Mori, 2019)	838	0.00583	0.21591	78.520
FSP (Yim et al., 2017)	877	0.00638	0.22525	78.413
Struct Pruning	887	0.00931	0.26687	77.242

Compression





Binary Fairness: Titanic Dataset

- A binary classification dataset was chosen to compare our metrics to existing metrics of bias.
- 3 ML models were trained to predict passengers as having "Survived" or "Not Survived": a shallow NN, an SVM, and a Gradient Tree Boosting classifier.
- Fairness was measured using the protected feature Passenger Sex, reported as a binary value recorded in historical records, which was excluded from training.

Model	Our Metrics				Existing Metrics			
-	CEV		SDE		ERED		DEV	
-	All→Men	All→Women	All→Men	All→Women	FPED	FNED	DIMS	DIAMR
NN	0.013557	0.012737	0.115002	0.093218	0.548443	0.458016	-0.269742	0.288790
SVM	0.012089	0.000736	0.109744	0.027081	0.412500	0.593508	-0.067460	0.491071
GTB	0.000107	0.000941	0.010341	0.030619	0.458462	0.513932	-0.193700	0.364831

Multi-Class Fairness: CelebA Groups

- Finding bias doesn't mean finding bias *against*
- A Deep NN was trained to identify the hair color (Brown, Blond, Gray, or Black) of Celebrity Headshots.
- CEV and SDE indicate a substantial bias for instances from the dataset that have the Male feature.

Protected Attribute	Top-1	CEV	SDE	Change in FPR	Change in FNR
Full Test Set	0.9212				
Attractive	0.9222	0.0015	0.0331	-31.0809	80.4380
Male	0.9225	0.1413	0.2205	12.8440	77.8003
Pale Skin	0.9224	0.0035	0.0465	-43.8572	-33.9335
Young	0.9215	0.0002	0.0082	-27.6765	150.5065
Not Attractive	0.9208	0.0034	0.0493	45.6423	6.8297
Not Male	0.9207	0.0053	0.0562	1.2762	47.2981
Not Pale_Skin	0.9207	0.0000	0.0021	1.9565	1.4648
Not Young	0.9213	0.0035	0.0313	146.3057	0.2381

Multi-Class Fairness: CelebA Groups

Discussion + Conclusion

- These metrics were developed to measure bias specifically, but we have demonstrated their usability as metrics of fairness.
- Our metrics only indicate the *presence* of bias, not whether a particular group or class is advantaged or disadvantaged.
- Our metrics only indicate the presence of bias relatively between two models, not absolutely.
- All issues of fairness still require a healthy presence of human judgement.
- Future Work:
    - Can a hard threshold be determined?
    - Improve meaning and interpretability of metrics.
    - Does mitigating bias necessarily improve fairness?

Code available at:

Check https://arxiv.org/abs/2110.04397 for an updated paper.