SAND2020-3014C

Defense Nuclear Nonproliferation Research & Development

Nuclear Explosion Monitoring Program Review
NEM2020

Confidence Baselines for Seismic Event Discrimination

Lisa Linville
Sandia National Laboratory

18 March 2020

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Background and Goals

- **Supervised-learning with deep neural networks (DNNs) works well for classification tasks using seismic data** i.e denoising, detection, phase picking, pick refinement, pick association and source type classification

- **These and other new methods have the ability to greatly increase the number of identified seismic events**

- **While clear demonstrations that DNNs outperform current practice can justify their use in parts of the pipeline** (phase refinement), **terminal and authoritative decisions** (source type) **typically require humans**

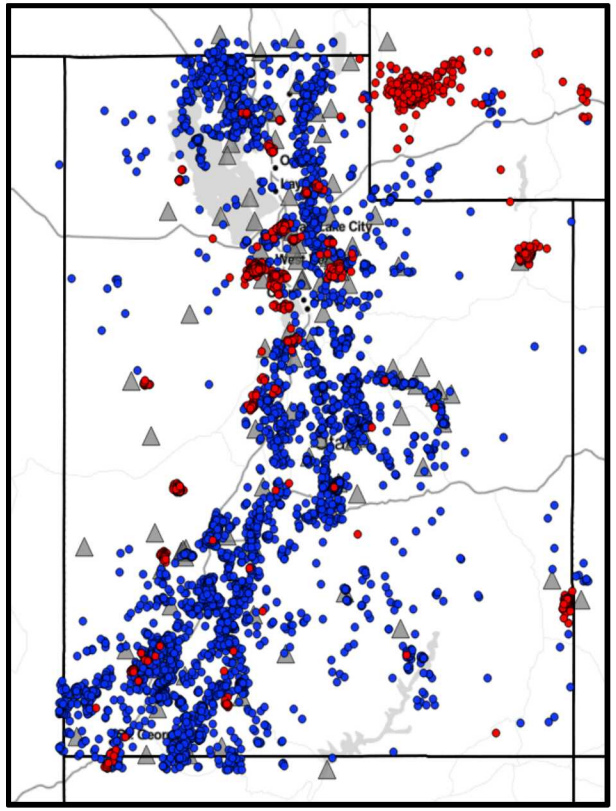
- **This study explores how regularization techniques for DNNs affect decision confidence.**

Can classification confidence produced by a DNN model for each decision be used to help prioritize human analyst resources?

Datasets

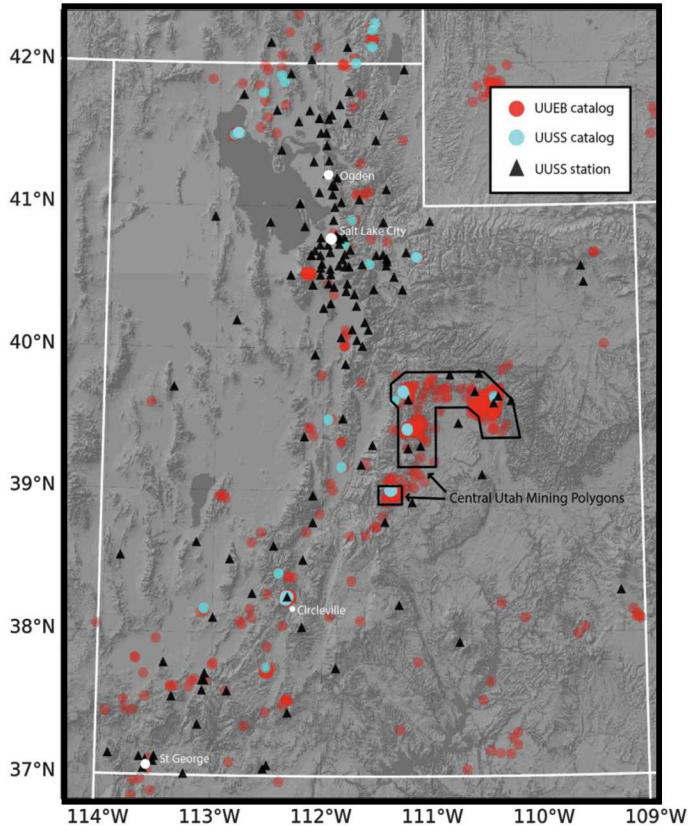
UUSS

Linville, L., Pankow, K., & Draelos, T. (2019b). Deep Learning Models Augment Analyst Decisions for Event Discrimination. *Geophysical Research Letters*, *46*(7), 3643-3651.



UUEB

Linville, L., Brogan, R. C., Young, C., & Aur, K. A. (2019). Global-and Local-Scale High-Resolution Event Catalogs for Algorithm Testing. *Seismological Research Letters*, *90*(5), 1987-1993.



Using two seismic datasets our DNN classifies earthquakes and non-earthquakes

Model input: 1-3 channel spectrogram from any station in the network

Model output: [1-x, x]; sum=1

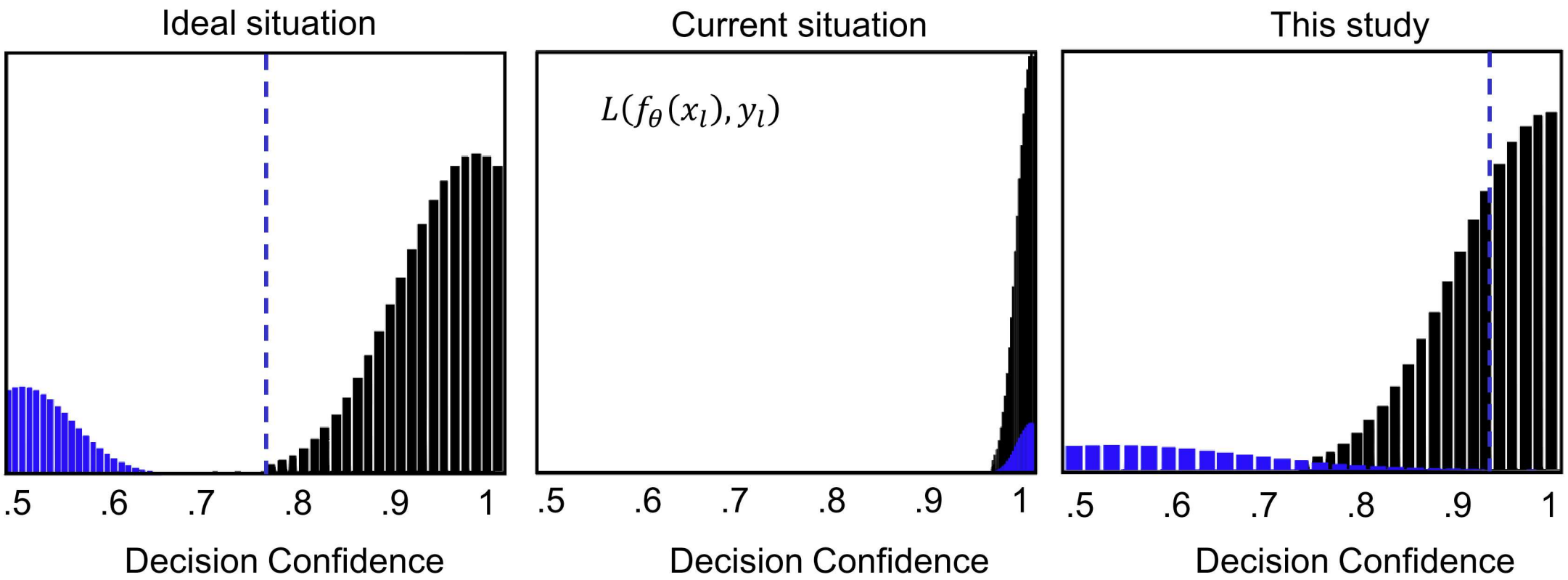
example of earthquake output:

[.95,.05]

The higher of the two values decides the classification label.

How far the higher value is from .5 describes how confident the model is in the classification label (decision confidence)

Overview



Not real distributions- for illustration only

Approach

Cross-entropy Loss

$$\mathcal{L}_v = L(f_\theta(x_l), y_l)$$

Adversarial Loss

$$\mathcal{L}_{adv} = L(f_\theta(x_l'), y_l)$$

$$x_l' = x_l + \epsilon \text{sign}(\nabla_{x_l} L(f_\theta(x_l), y_l))$$

Station Consistency

$$\mathcal{L}_{sc} = \frac{1}{\mathrm{E}} \sum_{j}^{\mathrm{E}} \sqrt{\frac{\sum_i^{\mathrm{N}} \left(f_\theta(x_{ji}) - f_\theta(\bar{x}_j)\right)^2}{\mathrm{N}-1}}$$

Training-based regularization

$$\text{vanilla} = \mathcal{L}_v$$

$$adv = \mu\mathcal{L}_v + (1-\mu)\mathcal{L}_{adv}$$

$$sc = \mu\mathcal{L}_v + (1-\mu)\mathcal{L}_{sc}$$

$$advsc = \mathcal{L}_v + \mathcal{L}_{sc} + \mathcal{L}_{adv}$$

Model-based regularization

Dropout

fraction of the nodes in each layer of the network are ignored during training

$adv\text{drop}$

$advscdrop$

Ensembles

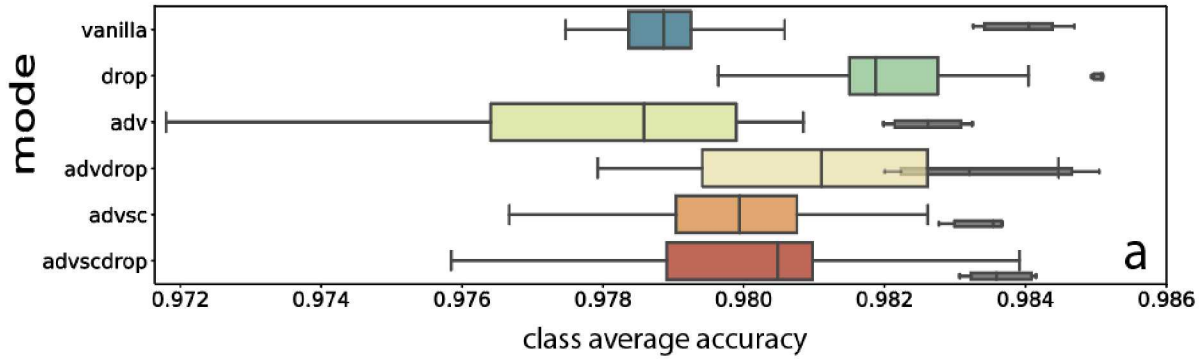
Results

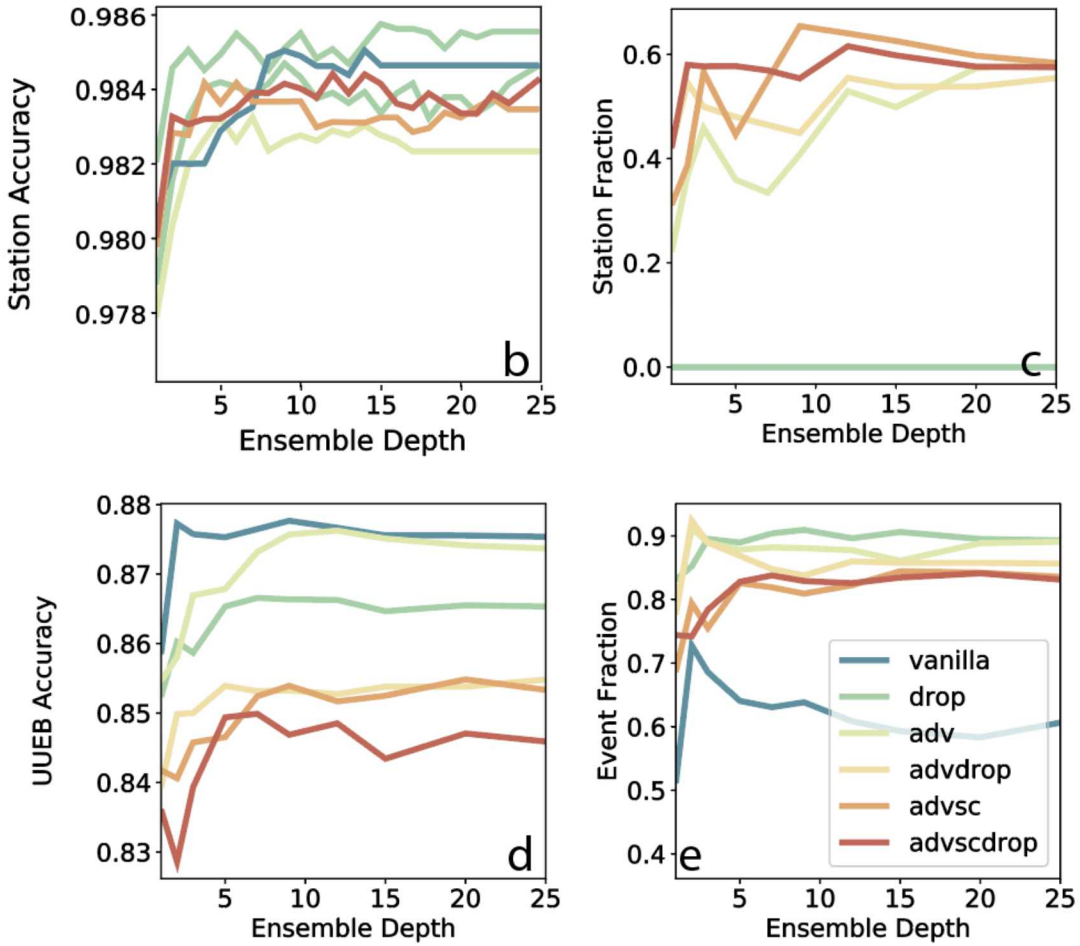
- **Dropout alone can not reduce decision confidence on incorrect examples**
- **Tradeoffs exist for most metrics**
- **High decision performance and better decision confidence can be achieved through the use of regularization**

vanilla	drop	adv	advdrop	advsc	advscdrop
Unregularized VGG11 model	VGG11 with fixed dropout rate at each layer	VGG11 with no dropout using adversarial training (eps=.5)	VGG11 with dropout using adversarial training (eps=.5)	VGG11 no dropout adversarial training (eps =.5)	VGG11 dropout adversarial training (eps=.5)

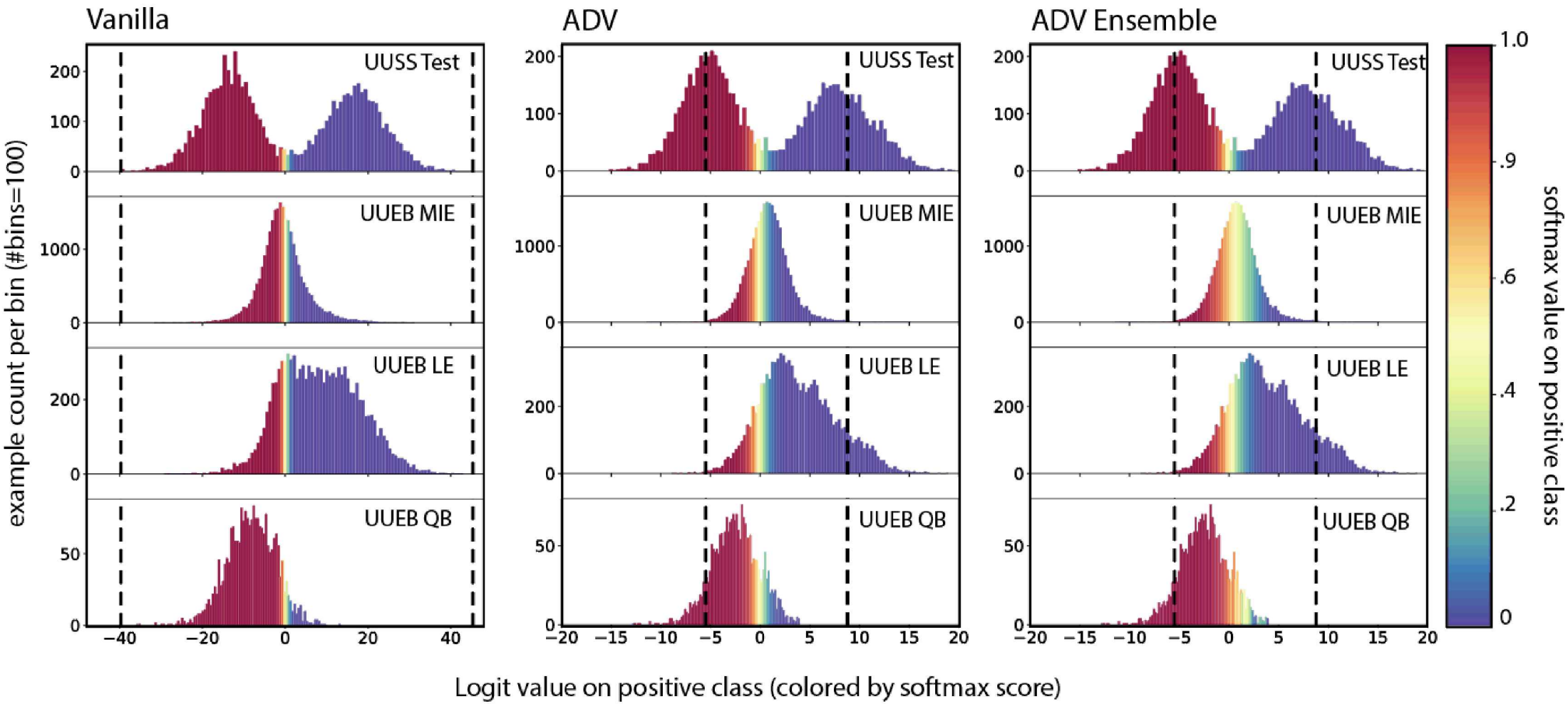
Mode	Accuracy	Fraction of Examples Assigned	Fraction of Events Assigned	UUEB Event Accuracy	UUEB Station Accuracy
vanilla	97.75%	0.00%	72.04%	87.92%	94.83%
drop	98.18%	0.00%	92.89%	85.74%	94.83%
adv	97.18%	61.84%	96.33%	84.82%	95.86%
advsc	98.04%	67.39%	93.96%	85.06%	92.88%
advdrop	98.35%	59.26%	97.63%	85.58%	95.34%
advscdrop	97.79%	63.75%	98.82%	84.10%	95.08%
***max assigned examples for individual models**					
vanilla	98.47%	0.00%	89.34%	87.77%	94.18%
drop	98.58%	0.00%	96.56%	86.46%	94.44%
adv	98.32%	33.44%	97.63%	87.32%	93.92%
advsc	98.42%	51.19%	98.34%	84.75%	92.11%
advdrop	98.51%	55.46%	97.51%	85.48%	94.31%
advscdrop	98.44%	61.56%	98.82%	84.85%	93.53%
***max average accuracy for all models and ensembles**					

Discussion


vanilla
drop
adv
advdrop
advsc
advscdrop
mode
0.972
0.974
0.976
0.978
0.980
0.982
0.984
0.986
class average accuracy
a


Station Accuracy
0.986
0.984
0.982
0.980
0.978
b
5
10
15
20
25
Ensemble Depth
Station Fraction
0.6
0.4
0.2
0.0
c
Ensemble Depth
UUEB Accuracy
0.88
0.87
0.86
0.85
0.84
0.83
d
Ensemble Depth
Event Fraction
0.9
0.8
0.7
0.6
0.5
0.4
e
vanilla
drop
adv
advdrop
advsc
advscdrop
Ensemble Depth

More Discussion



Conclusion

- **Regularization approaches make decision confidence more meaningful**
- **We have some evidence to suggest that confidence based thresholds can be used to identify when a model is extrapolating beyond training data**
- **Conservative thresholds using basic implementations and minimal optimization of all these concepts gives us single station review fractions of ~65% of test catalogs.** Not bad for sparse networks trying to recover small events.
- **Network decisions increase the review fractions considerably~ 98%.** Not bad for dense networks trying to maximize event recovery.

Regularization and ensembles are cheap and scalable approaches to actionable classification confidence.

- **Future (but not for me): can bayesian methods for DNNs outperform confidence for uncertainty estimation?**