Google AI Brain Team

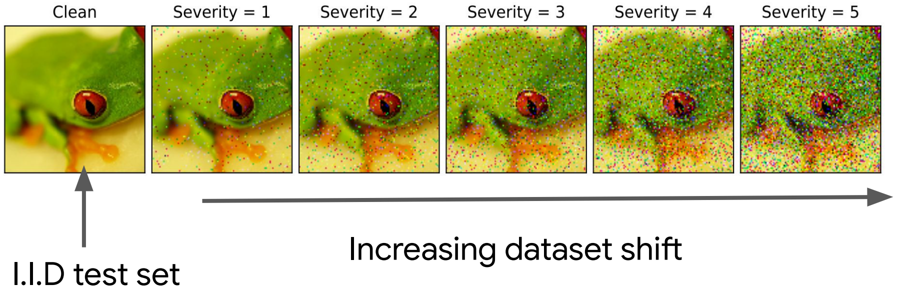
Reliable Deep Anomaly Detection

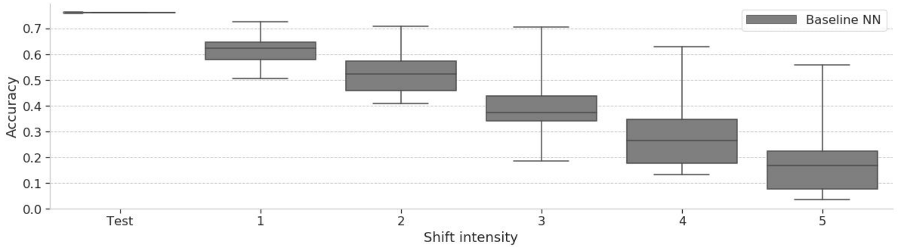
Balaji Lakshminarayanan
balajiln@

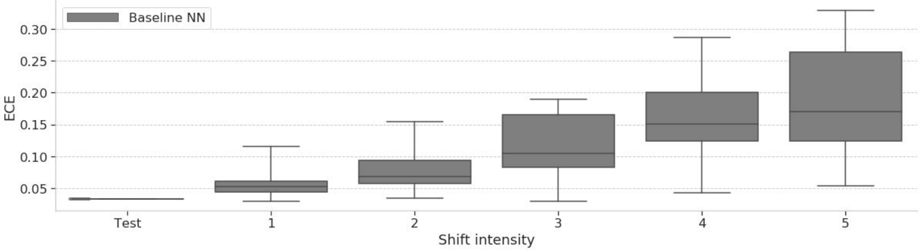
Joint work with awesome collaborators @ Google & DeepMind

Motivation

Why Reliable Deep Learning?







- Independent and identically distributed (IID):

$$p_{TEST}(y,x) = p_{TRAIN}(y,x)$$

- Out-of-distribution(OOD): $p_{TEST}(y,x) \neq p_{TRAIN}(y,x)$

- Accuracy of NNs degrades under dataset shift
  - Imagenet (IID) vs Imagenet-C (OOD)
- Calibration also degrades under dataset shift.

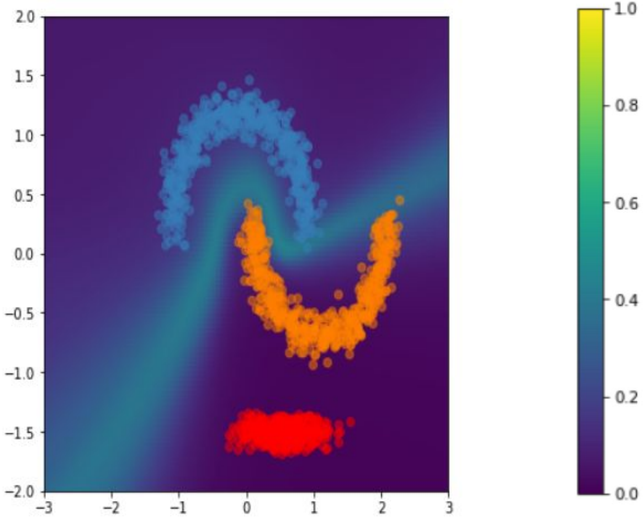
Calibration Error = |Confidence - Accuracy|

predicted probability of correctness

observed frequency of correctness

[See our NeurIPS'2020 tutorial for background]

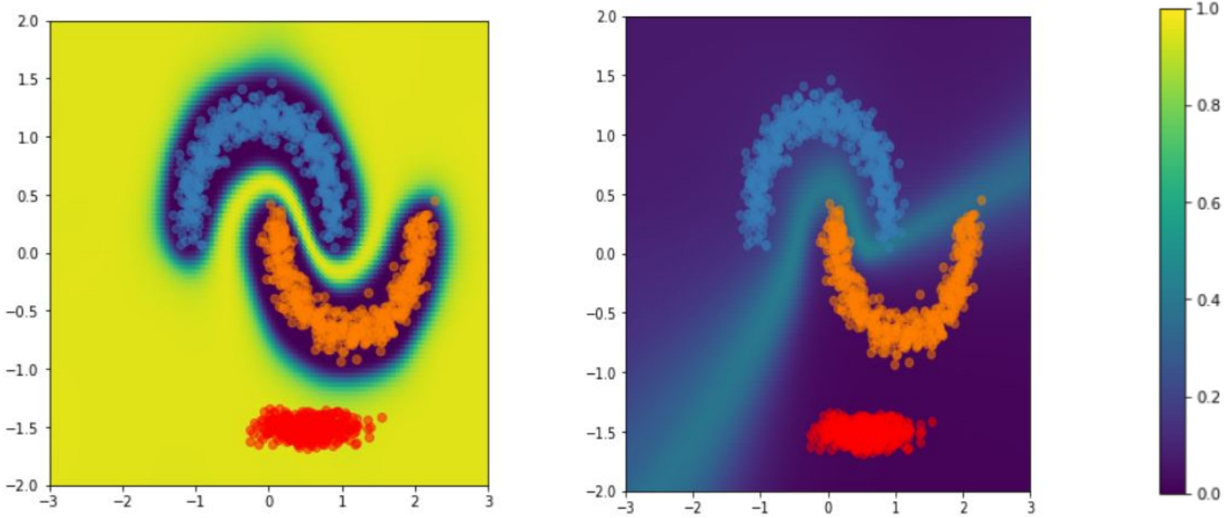
Models assign high confidence predictions to OOD inputs



Deep neural networks

Image source: "Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness" [Liu et al. 2020](Liu et al. 2020)

Models assign high confidence predictions to OOD inputs



High uncertainty
(low confidence)

Low uncertainty
(high confidence)

Ideal behavior

Deep neural networks

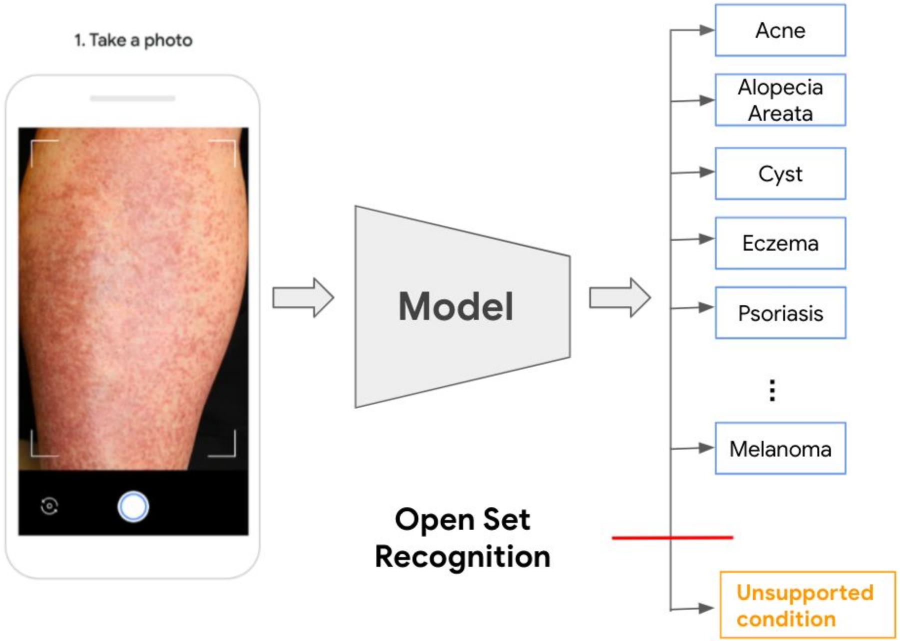
Trust model when x^* is close to $p_{TRAIN}(x,y)$

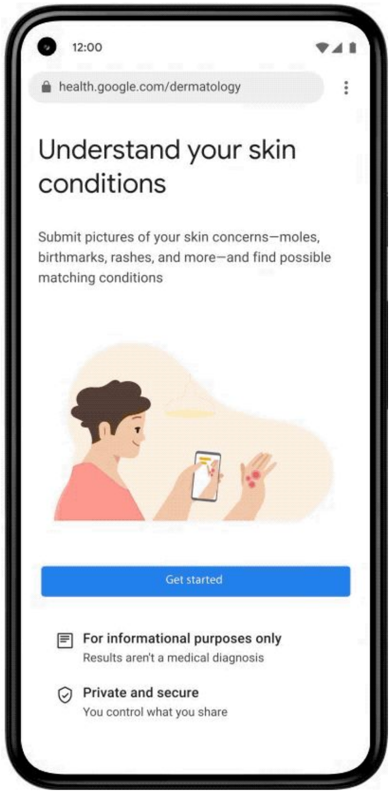
Image source: “Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness” [Liu et al. 2020]

Applications

Open Set Recognition

Google AI
Brain Team





Test input may not belong to one of the K training classes.

Need to be able to say "none-of-the-above".

Image source: https://blog.google/technology/health/ai-dermatology-preview-io-2021/

Open Set Recognition

- Example: Classification of genomic sequences

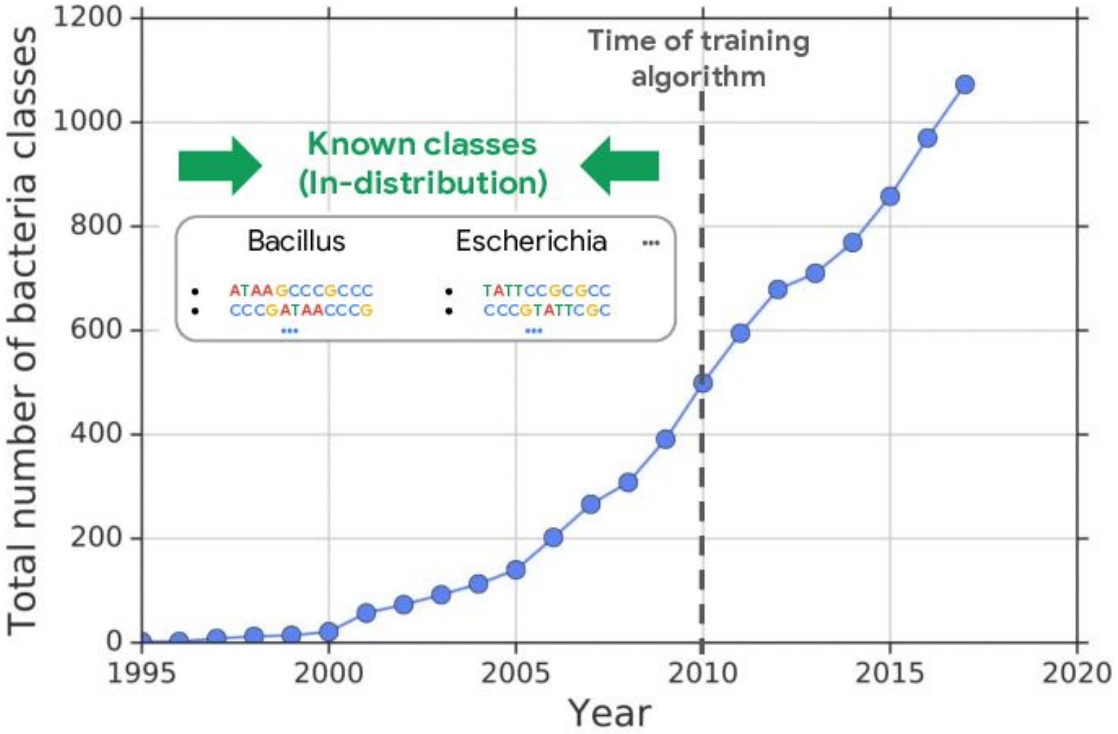


Image source: https://ai.googleblog.com/2019/12/improving-out-of-distribution-detection.html

Open Set Recognition

Google AI Brain Team

- Example: Classification of genomic sequences
- High i.i.d. accuracy on known classes is not sufficient
- Need to be able to detect inputs that do not belong to one of the known classes

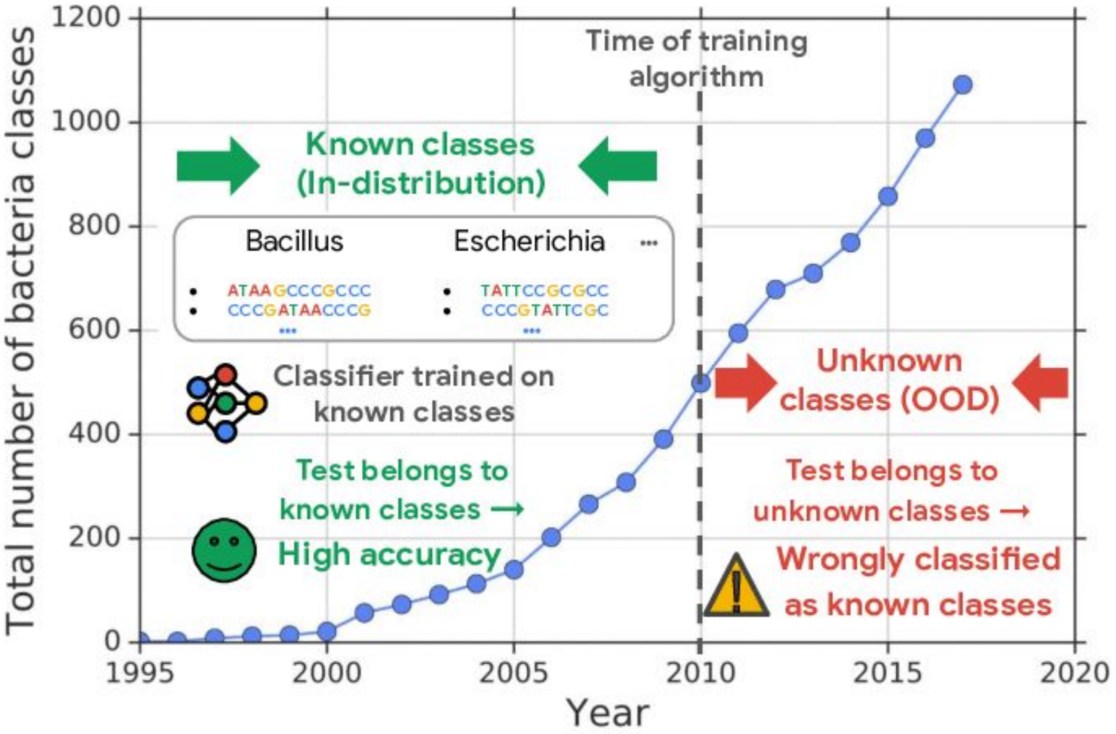
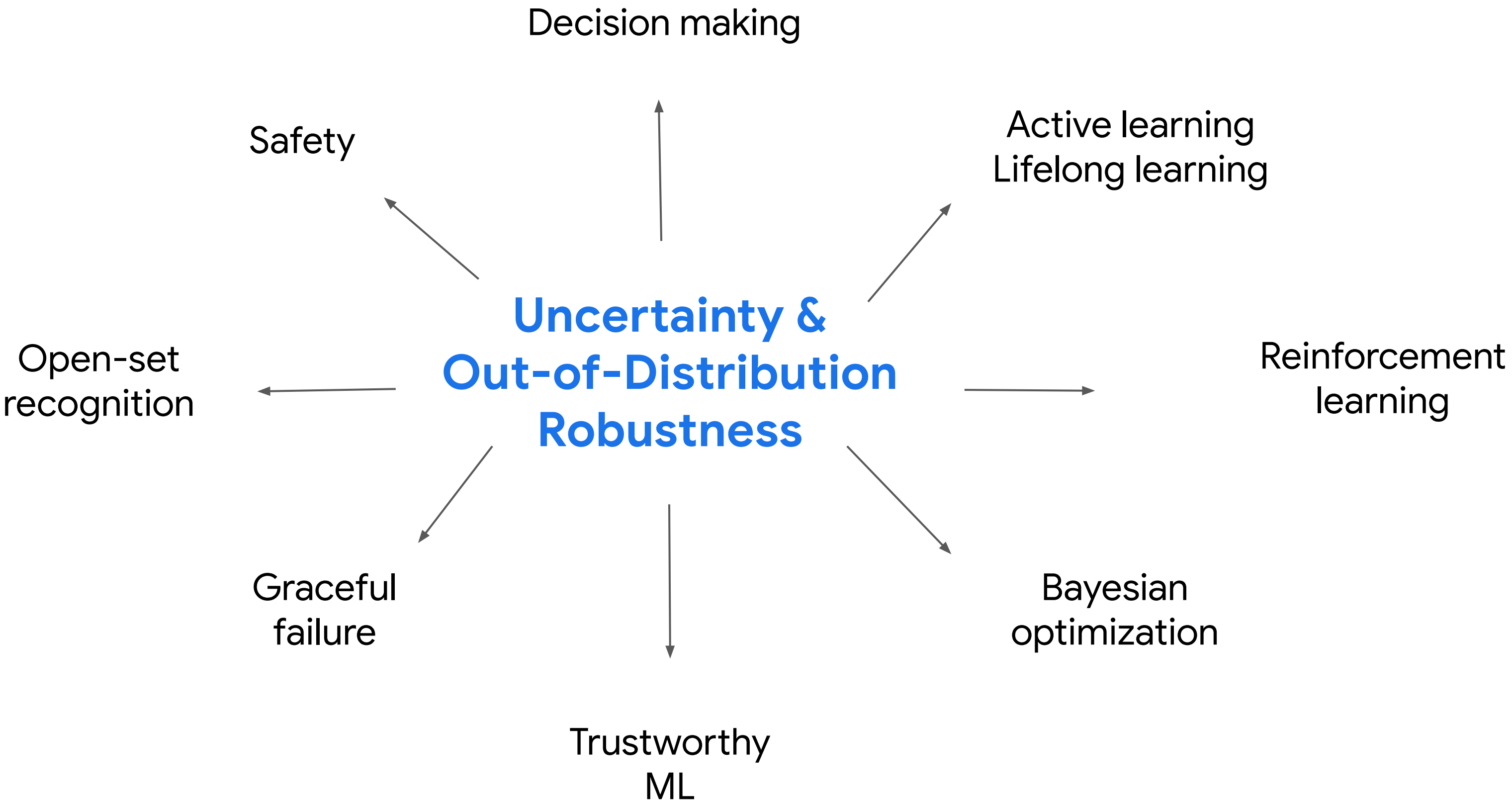


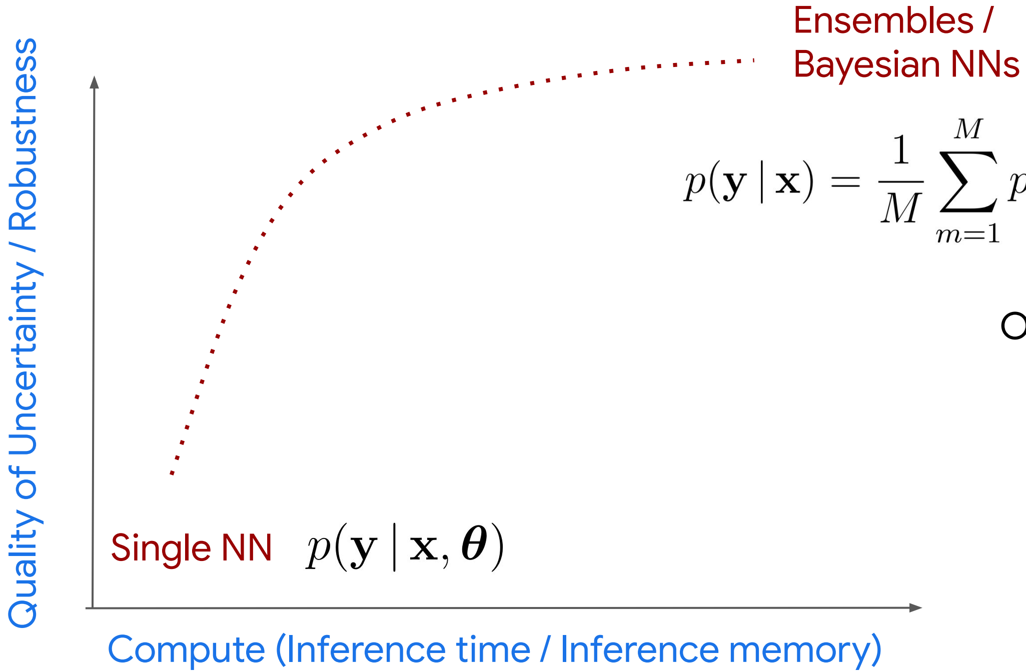
Image source: https://ai.googleblog.com/2019/12/improving-out-of-distribution-detection.html

All models are wrong, but ~~some~~ models that know when they are wrong, are useful.



Methods

Cartoon: Uncertainty/Robustness vs Compute frontier



$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_m)$$

Orthogonal ways of improving performance

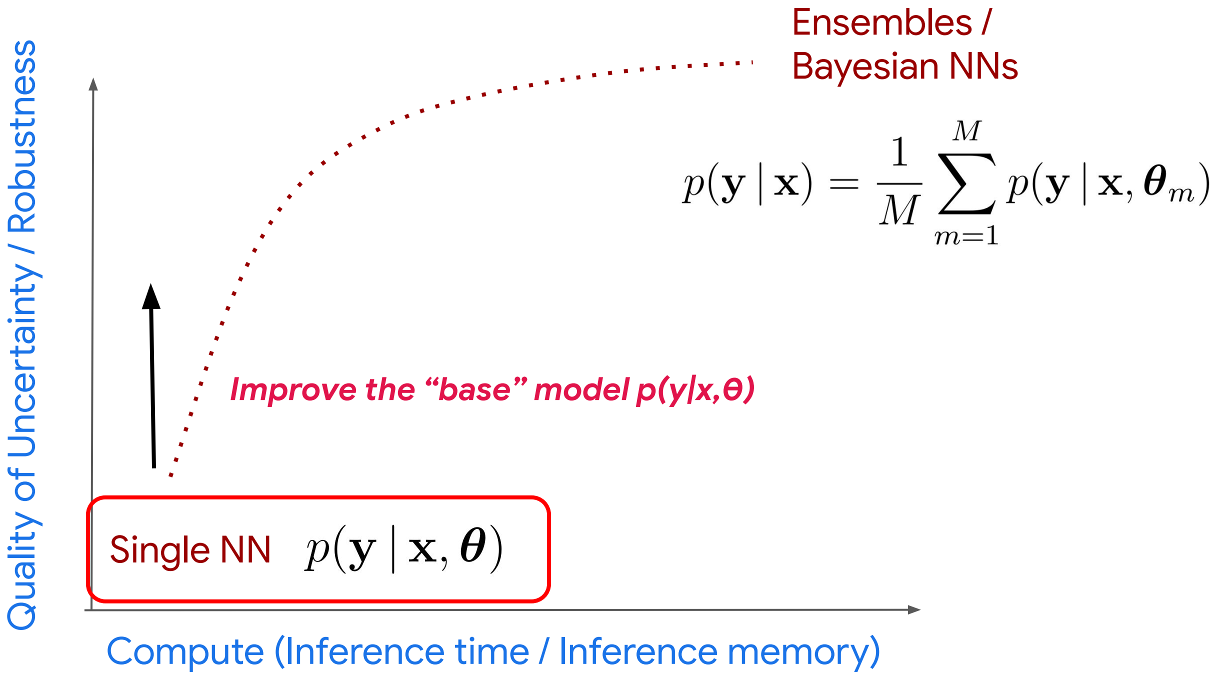
- Improve the single model p(y|x,θ)
- Better representation learning (e.g. pre-training)
- Average predictions over diverse set of functions $\theta_1, \theta_2 \ldots \theta_M$
- Outlier exposure
- Density modeling in latent space

Not covered in this talk: Unsupervised anomaly detection

- Anomaly Detection using Deep Generative Models: Pitfalls and Promises (overview talk that covers the following papers)

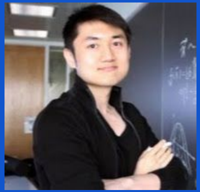
- **Do deep generative models know what they don't know?** E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, *B. Lakshminarayanan*. *ICLR 2019*.
- **Likelihood ratios for out-of-distribution detection**. J. Ren, P. Liu, E. Fertig, J. Snoek, R. Poplin, M. DePristo, J. Dillon, *B. Lakshminarayanan*. *NeurIPS 2019*.
  - *See also **A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection** where we extended this to density models of representations*
- **Detecting out-of-distribution inputs to deep generative models using a test for typicality.** E. Nalisnick, A. Matsukawa, Y. W. Teh, *B. Lakshminarayanan*. *arXiv 2019*.
- **Density of States Estimation for Out-of-Distribution Detection** W. R. Morningstar, C. Ham, A. G. Gallagher, B. Lakshminarayanan, A. A. Alemi, J. V. Dillon AISTATS 2021

Improving single model performance

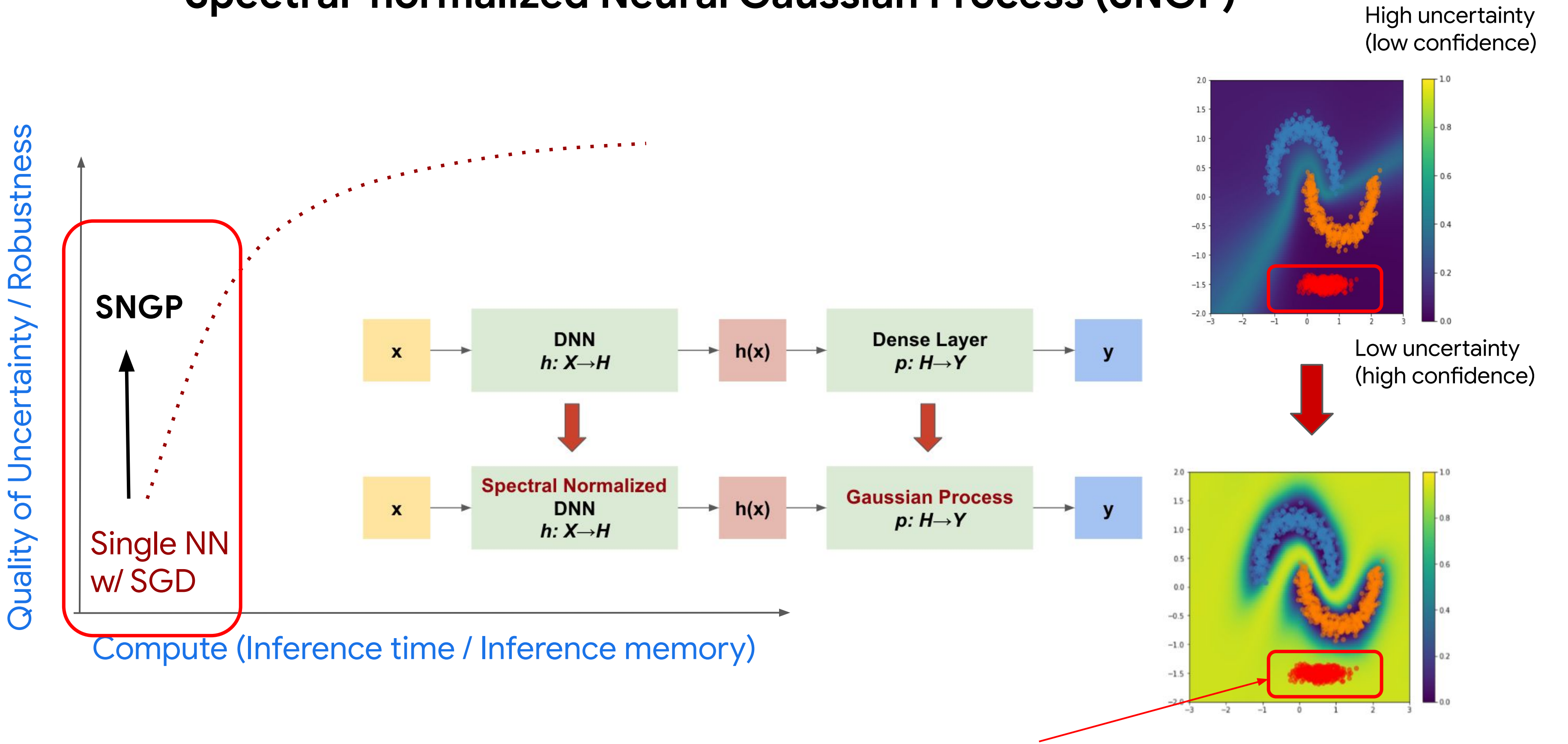


Improving Single Model Uncertainty via Distance Awareness

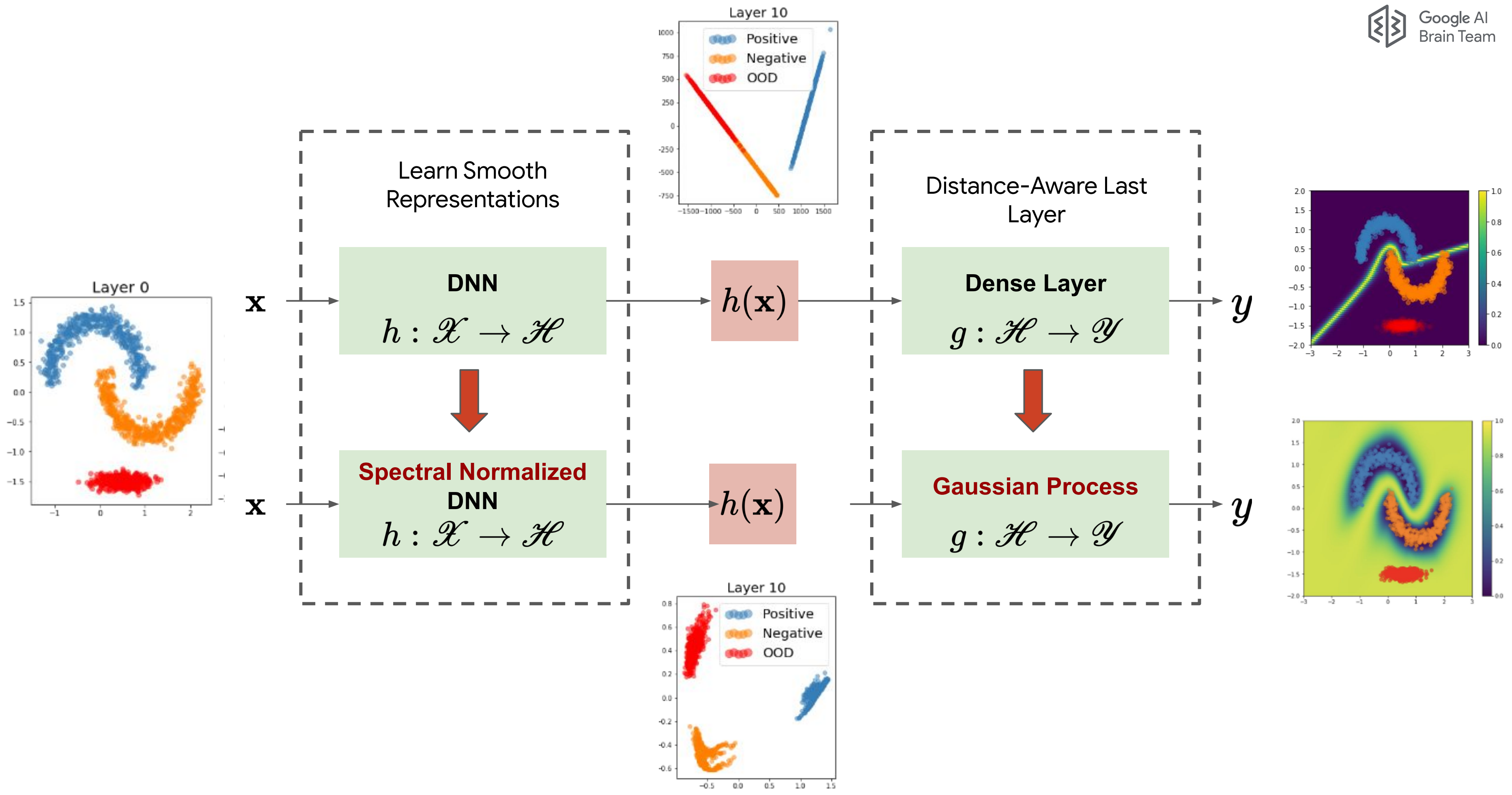
Jeremiah Liu (jereliu@) et al.



Adding *distance-awareness* using
Spectral-normalized Neural Gaussian Process (SNGP)



SNGP assigns lower confidence predictions to inputs far away from the training data



Google AI
Brain Team
Layer 0
$\mathbf{x}$
Learn Smooth Representations
DNN
$h : \mathscr{X} \rightarrow \mathscr{H}$
Spectral Normalized DNN
$h : \mathscr{X} \rightarrow \mathscr{H}$
Layer 10
Positive
Negative
OOD
$h(\mathbf{x})$
Distance-Aware Last Layer
Dense Layer
$g : \mathscr{H} \rightarrow \mathscr{Y}$
Gaussian Process
$g : \mathscr{H} \rightarrow \mathscr{Y}$
y

Spectral-normalized Neural Gaussian process (SNGP)

Google AI
Brain Team

“Models should be distance aware: uncertainty should increase as we move farther from training data.”

Key idea:

1. Replace linear dense layer with “GP layer”.
2. Apply spectral normalization to encourage smooth representations (bi-Lipschitz regularization) and avoid “feature collapse”.

See also [van Amersfoort+ 2020].

Method	Acc (↑)	ECE (↓)	Corrupted Acc/ECE	OOD AUPR (↑) SVHN	OOD AUPR (↑) CIFAR-100	Latency (ms)
Deterministic	96.0	0.023	72.9 / 0.153	0.7810	0.8352	**3.91**
MCD-GP	95.5	0.024	70.0 / 0.100	0.9599	0.8631	29.53
DUQ	94.7	0.034	71.6 / 0.183	0.9733	0.8537	8.68
MC Dropout	96.0	0.024	70.0 / 0.116	0.9714	0.8320	27.10
Deep Ensembles	**96.6**	**0.010**	**77.9 / 0.087**	0.9640	0.8875	38.10
SNGP (Ours)	95.9	0.018	74.6 / 0.090	**0.9901**	**0.9050**	6.25

Results on CIFAR-10 using Wide ResNet

BERT on an intent detection benchmark

Method	Accuracy (↑)	ECE (↓)	OOD AUROC (↑)	OOD AUPR (↑)	Latency (ms / example)
Deterministic	96.5	0.0236	0.8970	0.7573	**10.42**
MCD-GP	95.9	0.0146	0.9055	0.8030	88.38
DUQ	96.0	0.0585	0.9173	0.8058	15.60
MC Dropout	96.5	0.0210	0.9382	0.7997	85.62
Deep Ensemble	**97.5**	0.0128	0.9635	0.8616	84.46
SNGP	96.6	**0.0115**	**0.9688**	**0.8802**	17.36

[Liu+ 2020]

Exploring the limits of OOD detection

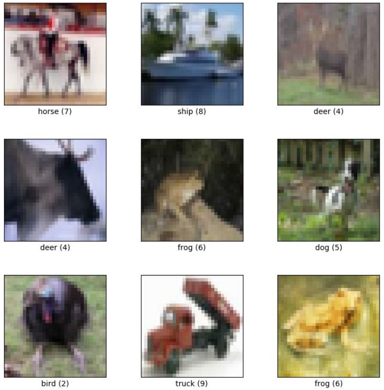
Stanislav Fort*, Jie Ren* et al.

Google AI
Brain Team

Goal: Improve SOTA on hard OOD detection tasks

Far OOD, AUROC = 99%

Near OOD, AUROC = 80%



CIFAR-10 (ID)

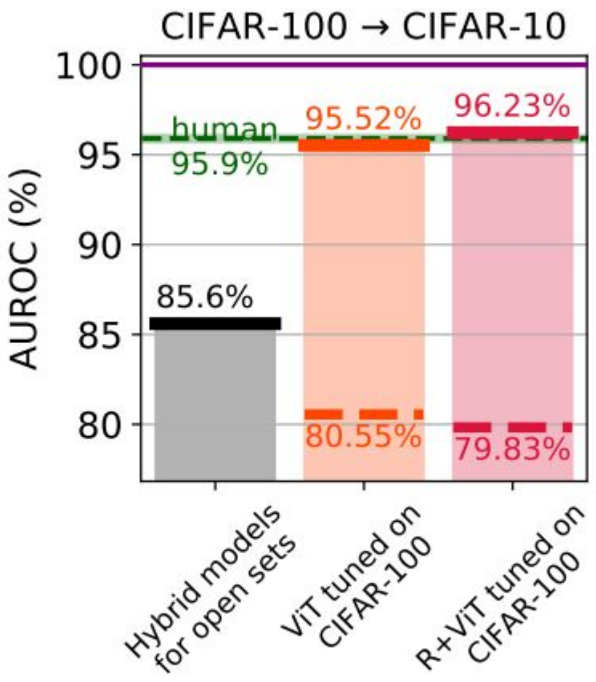
SVHN (OOD)

CIFAR-100 (ID)

CIFAR-10 (OOD)

- OOD: test input $\{X', y'\}, y' \notin Y_{\text{ID}}$, new class, shift in y
- In-distribution shift: $\{X', y'\}, y' \in Y_{\text{ID}}$, same class, shift in x

Pre-trained ViT improves near-OOD detection



In-dist.	Test Accuracy	Mahalanobis AUROC	MSP AUROC
WRN training from scratch	79.80%	74.91%	75.40%
Pretrain+finetune ViT	**91.67%**	**96.23%**	**92.08%**

We improve SOTA AUROC on CIFAR-100 vs CIFAR-10 from ~85% to ~96% using fine-tuned ViT

*OOD score: Mahalanobis distance based on last layer embeddings

Pre-trained ViT improves near-OOD detection

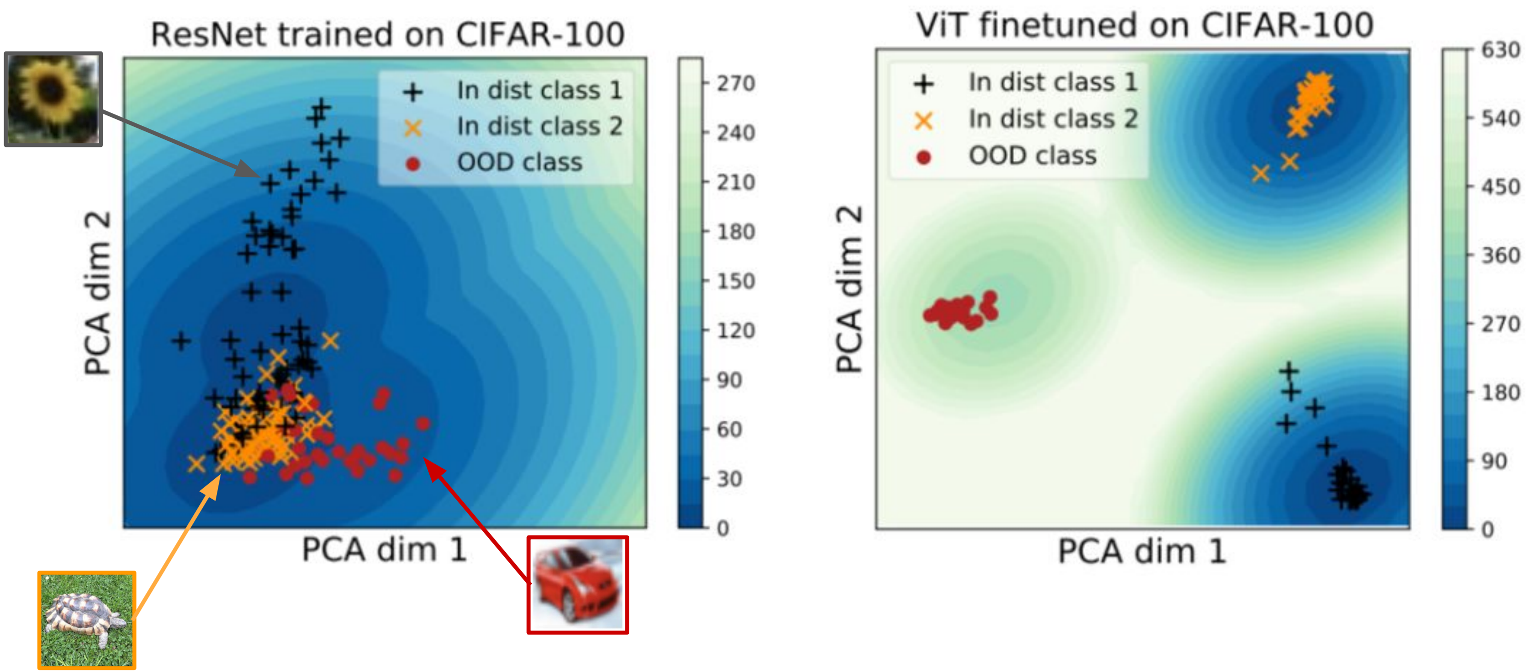
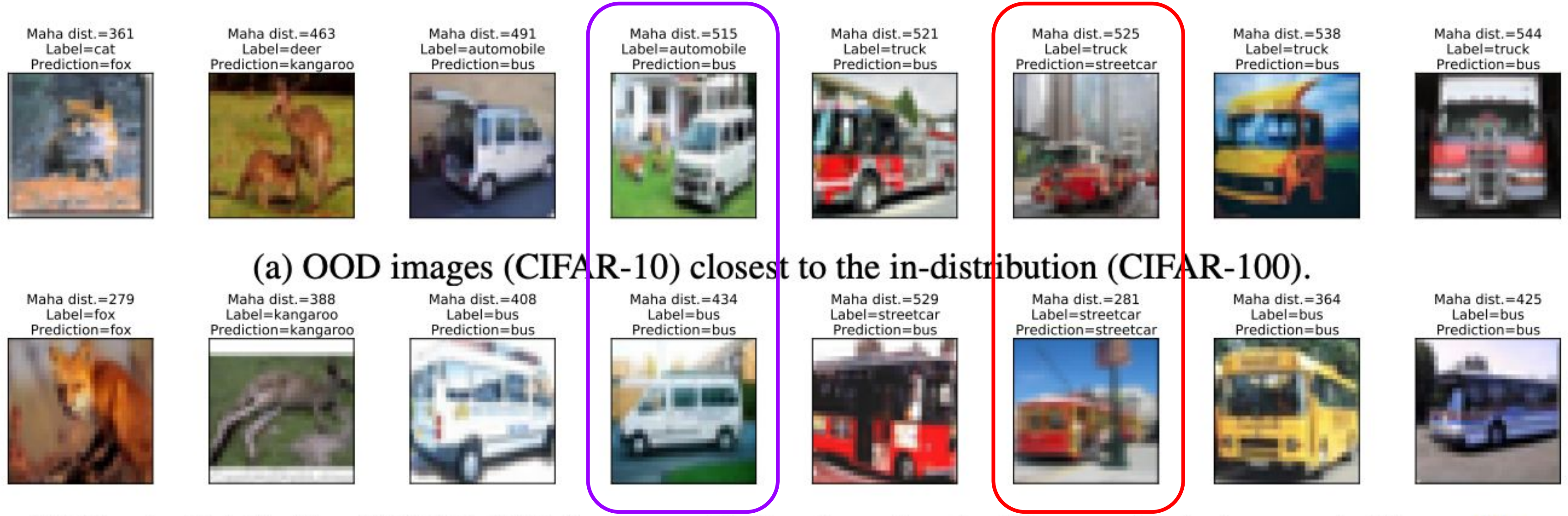


Figure: 2D PCA project of the space of embedding.
Color coding shows Mahalanobis outlier score.

Google AI
Brain Team

Qualitative failure cases of ViT OOD detection

- Most false positives are due to mislabeling or ambiguity

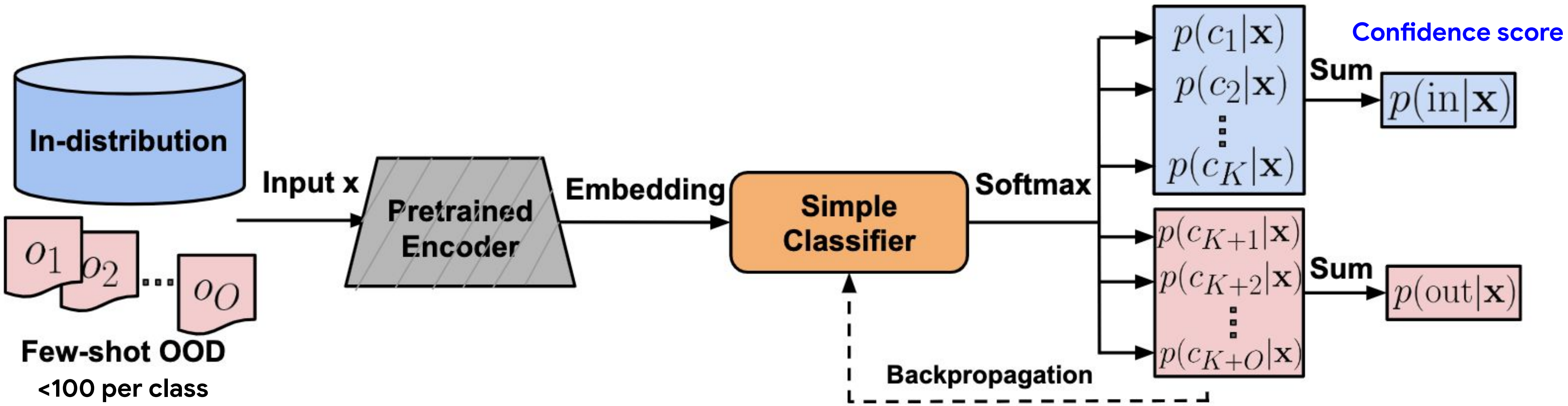


(a) OOD images (CIFAR-10) closest to the in-distribution (CIFAR-100).

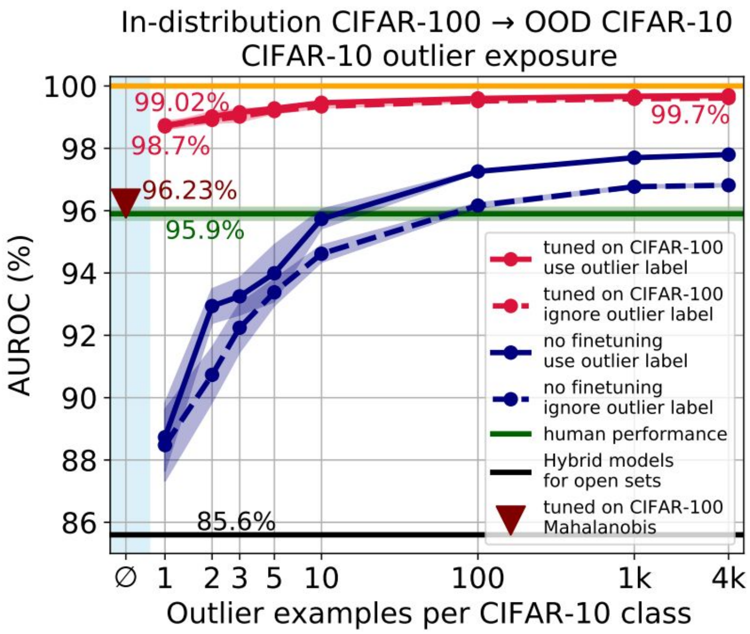
(b) The in-distribution (CIFAR-100) images with the closest embedding vector to images in Figure 11a.

Further improvement: Few-shot Outlier Exposure

- When only a handful of known outlier examples is available
  - Either collected intentionally or collected from failure cases

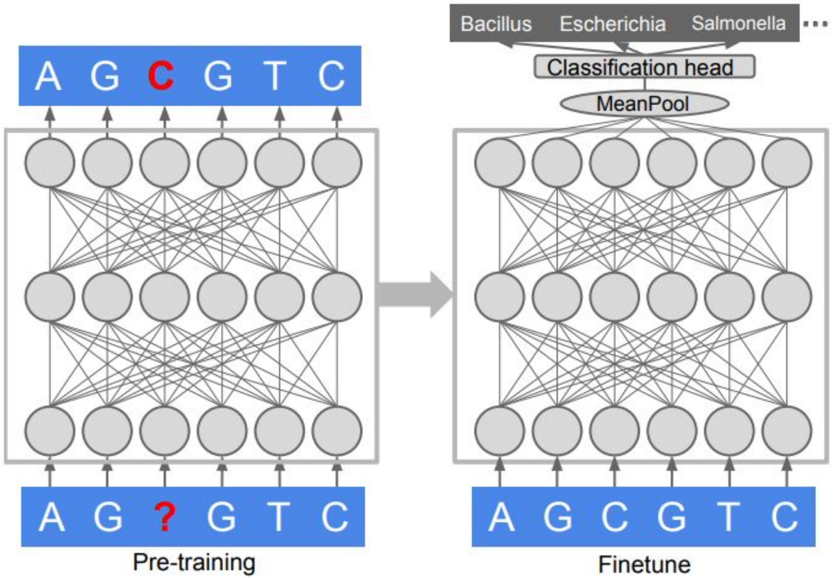


Few-shot Outlier Exposure



Few shot outlier exposure further improves AUROC on CIFAR-100 vs CIFAR-10 to ~99% with just 1 labeled example per outlier class

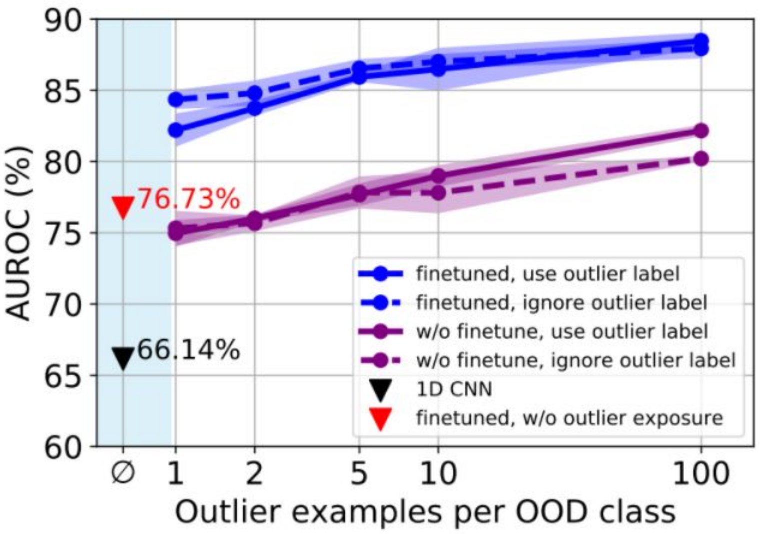
Challenging genomics near-OOD benchmark



Model	Test accuracy	Mahalanobis AUROC	MSP AUROC
1D CNN [Ren et al., 2019]	85.93%	64.75%	65.84%
BERT pretrain and finetune	**89.84%**	**77.49%**	**73.53%**

Recipe translates to other data modalities

Google AI
Brain Team



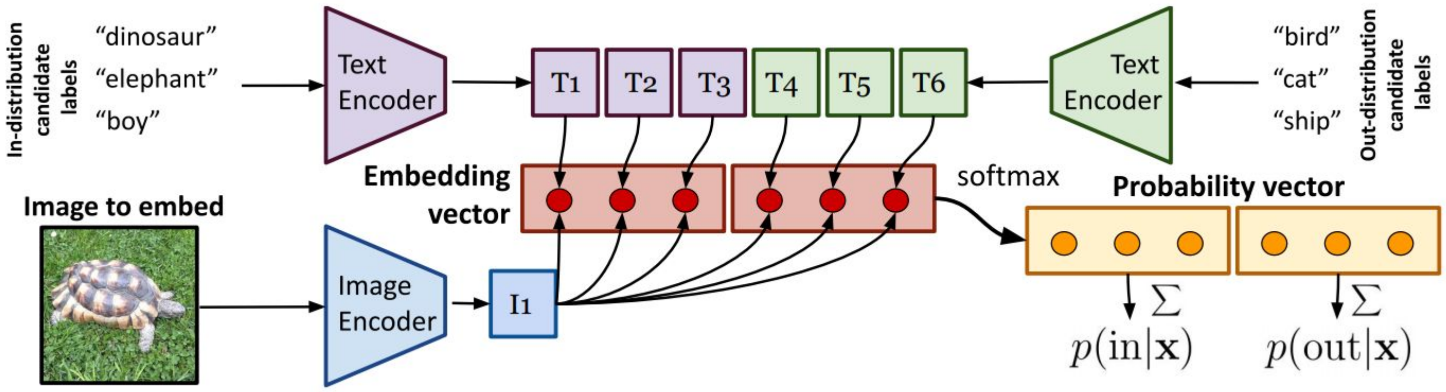
Challenging genomics near-OOD benchmark:

66% (current SOTA)

-> 77% (pre-trained transformer)

-> 88% (few-shot outlier exposure)

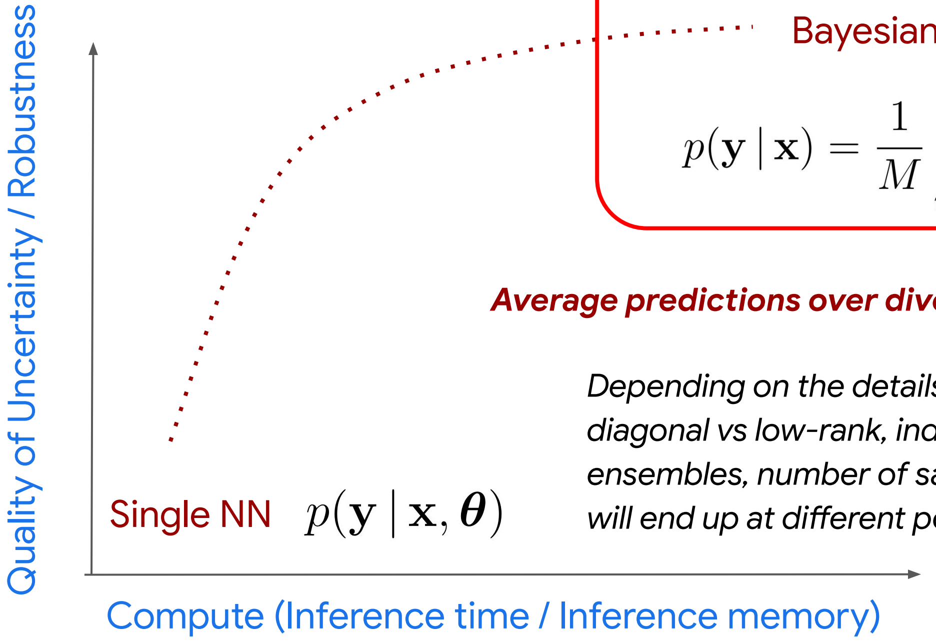
Zero-shot Outlier Exposure using CLIP



Distribution 1	Distribution 2	Labels 1	Labels 2	AUROC
CIFAR-100	CIFAR-10	CIFAR-100 names	—	69.49%
CIFAR-100	CIFAR-10	CIFAR-100 names	CIFAR-10 names	94.68%
CIFAR-10	CIFAR-100	CIFAR-10 names	—	89.17%
CIFAR-10	CIFAR-100	CIFAR-10 names	CIFAR-100 names	94.68%

*We do not finetune CLIP
Just use the names of known outliers without any accompanying images

Improving the quality of model uncertainty



Ensembles /
Bayesian NNs

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_m)$$

Average predictions over diverse set of functions $\theta_1, \theta_2 \dots \theta_M$

Depending on the details of the approximation (e.g. diagonal vs low-rank, independent ensembles vs efficient ensembles, number of samples/ensemble members), we will end up at different points along this curve.

[*Does Your Dermatology Classifier Know What It Doesn't Know? Detecting the Long-Tail of Unseen Conditions*]

Abhijit Guha Roy*, Jie Ren*, et al.

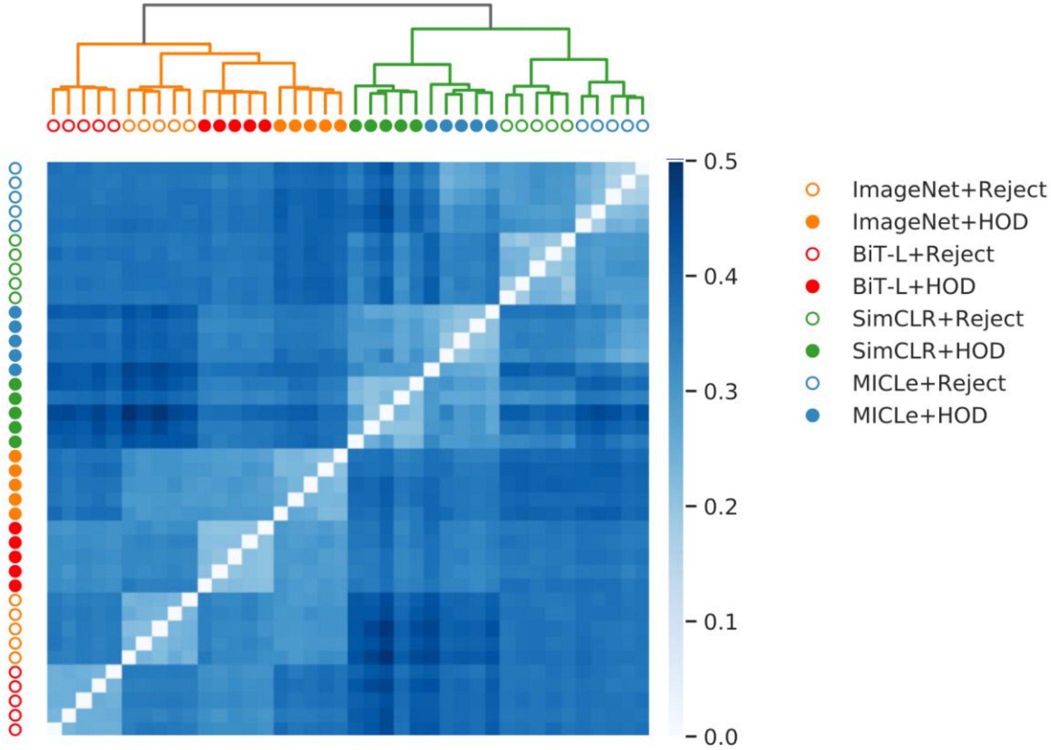
Diverse Ensembles Improve OOD detection

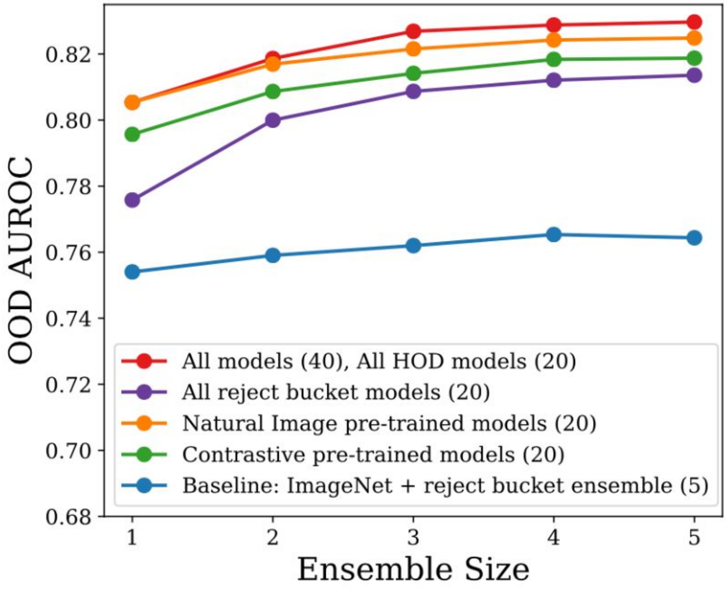
Google AI
Brain Team

Method	OOD detection metrics			
	AUROC (↑)	FPR @ 0.95 TPR (↓)	AUPR-in (↑)	Inlier accuracy (↑)
ImageNet + Ensemble	76.4	75.3	79.9	72.9
ImageNet + HOD + Ensemble	79.2	70.6	81.8	70.9
BiT-JFT + Ensemble	77.8	71.0	80.6	73.8
BiT-JFT + HOD + Ensemble	81.6	62.6	83.9	75.6
SimCLR + Ensemble.	77.0	76.4	79.8	75.1
SimCLR + HOD + Ensemble	78.7	70.4	81.5	71.3
MICLe + Ensemble	79.0	71.5	82.1	75.8
MICLe + HOD + Ensemble	80.5	67.5	83.2	72.4
Diverse ensemble	**83.0**	**61.4**	**85.8**	**76.3**

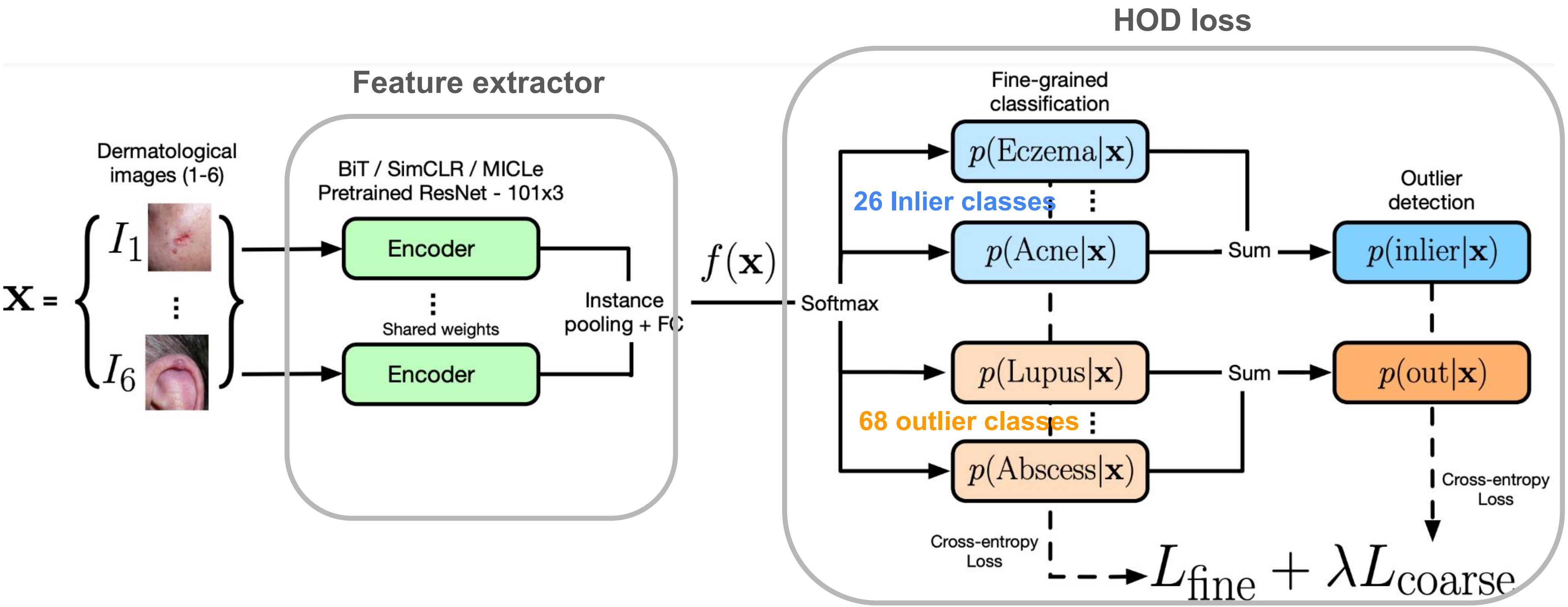
Diverse Ensembles

Google AI
Brain Team





Hierarchical Outlier Detection Loss (HODL)



Takeaways

- Uncertainty & robustness are critical problems in AI and machine learning.
- Orthogonal directions to improve performance:
    - Improving single model uncertainty via distance awareness
    - Pre-training and few-shot outlier exposure
    - Diverse ensembles
    - Understanding failure modes of anomaly detection in deep generative models (link)
- Links to papers available in my webpage: http://www.gatsby.ucl.ac.uk/~balaji/
    - Uncertainty baselines code: github.com/google/uncertainty-baselines
    - Robustness metrics code: github.com/google-research/robustness_metrics