MANDIANT

Loss on Demand

Toward Discriminative-Generative Hybrid Models for Malware Classification Confidence

Ethan Rudd, PhD and David Krisiloff, PhD

There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know.

-- Donald Rumsfeld

ML Confidence Analogy

Known Knowns
- Discriminative Models
- Assume all classes are known

Known Unknowns
- “Open Set” classifiers
- Density estimates
- Probability of Sample Inclusion (PSI) estimates

Unknown Knowns
- Potential for transfer learning on new samples

Unknown Unknowns
- Adversarial Samples
- Limitations on model capacity or feature information

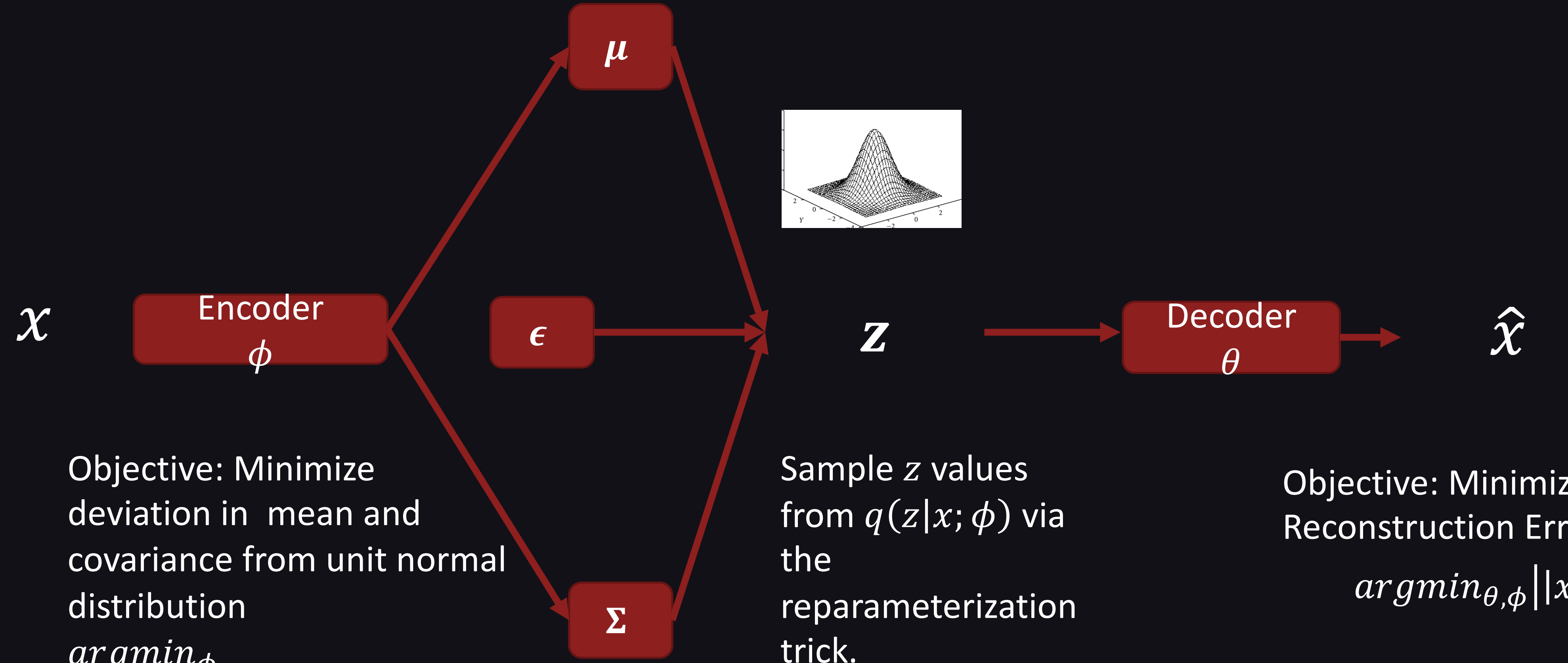
This Talk

- Research Question: Can we bake in better confidence estimation by combining a discriminative model with generative loss functions?

- Design Goals:
  1. Produce sensible confidence scores, incorporating as many confidence types as possible
  2. Don’t add extra baggage to a deployment-ready classifier
  3. Restrict the classifier’s design as little as possible

- Disclaimer: Much research addresses some of these goals but little addresses all three at once!

Recall the Variational AutoEncoder



Objective: Minimize deviation in mean and covariance from unit normal distribution

$argmin_{\phi} - D_{KL}(q_{\phi}(z|x)||p(z))$

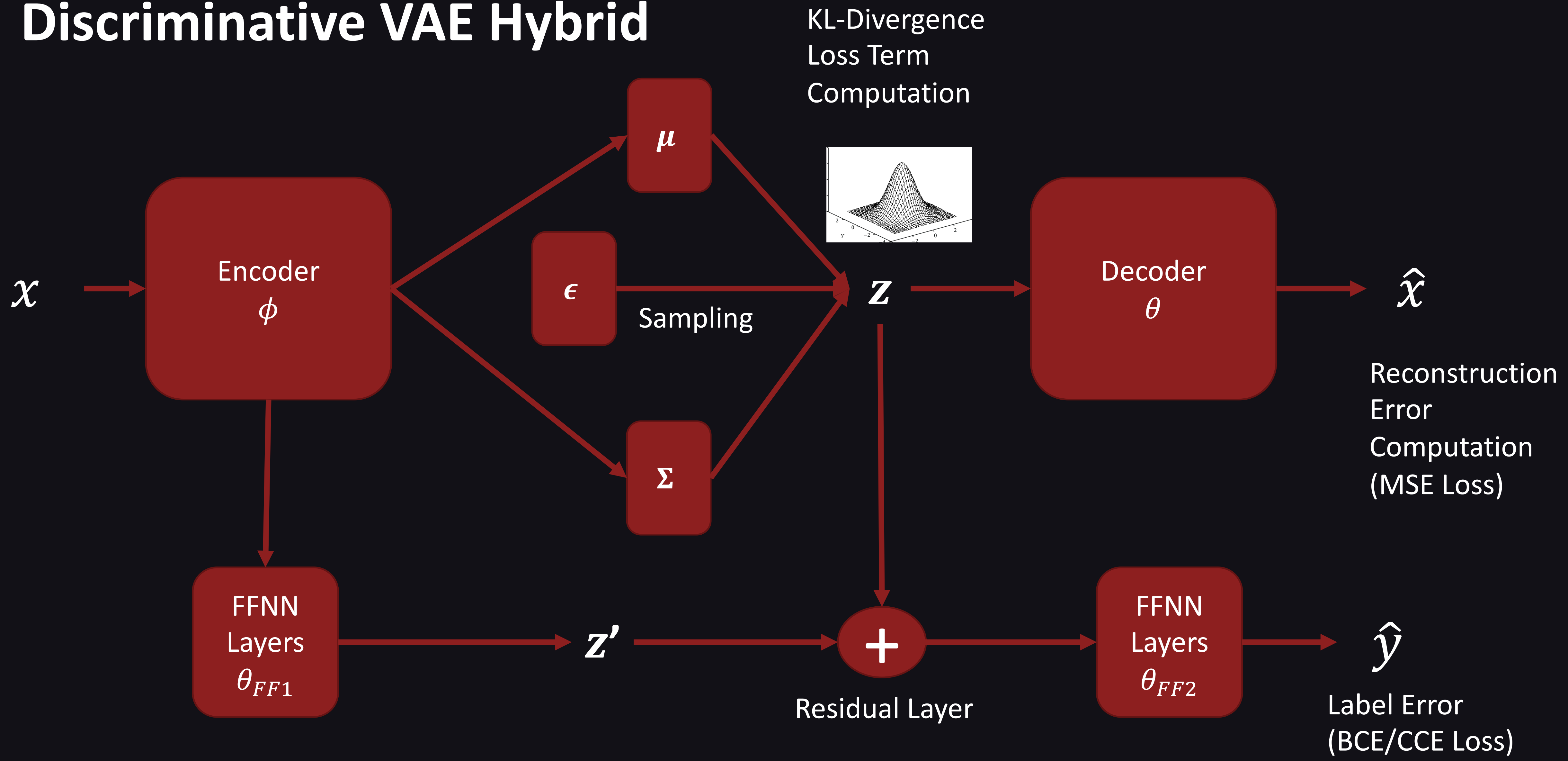
Where: $q(z|x;\phi) \cong p(z|x)$

Sample z values from $q(z|x;\phi)$ via the reparameterization trick.

Objective: Minimize Reconstruction Error

$argmin_{\theta,\phi}||x-\hat{x}||^2$

Discriminative VAE Hybrid



Net Loss

$$argmin_{\theta,\phi,\theta_{FF}} \left|\left|x - \hat{x}\right|\right|^2 - D_{KL}(q(z|x;\phi)||p(z)) + CE(y,\hat{y})$$

MSE

KL Divergence

Cross Entropy

ELBO

Can be evaluated during training or deployment

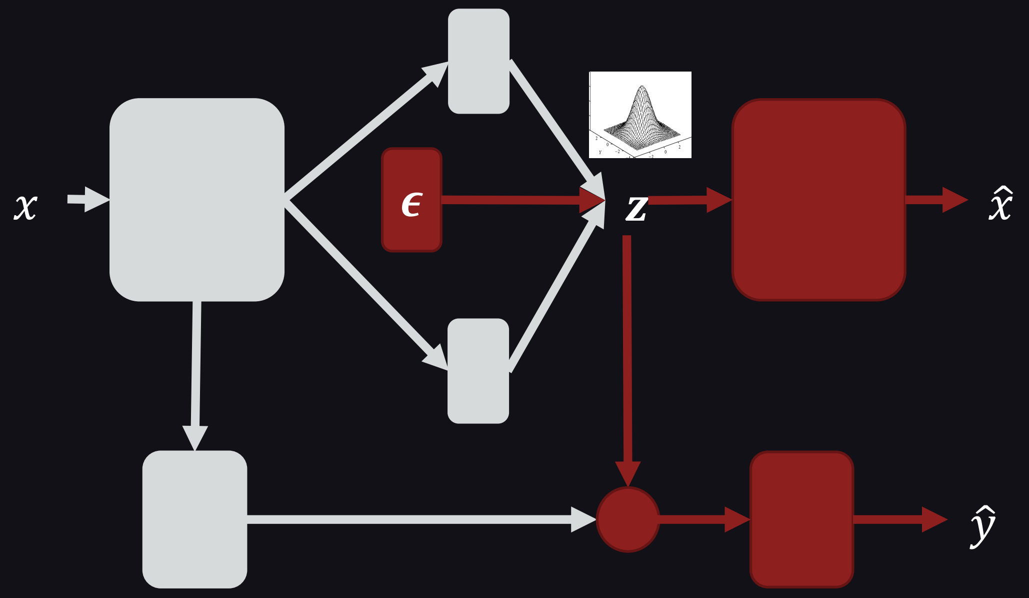
Evaluated only during training

Potential Model Usage

- Sample N z values, multiple forward passes, then …
  - Classification
    - Compute statistics over output scores (e.g., mean and standard deviation)
    - .9897 AUROC on EMBER test vs .9882 baseline model
  - Pointwise density estimates
    - VAE design:

$$\log p(x) = E_z[p(x|z)] \sim -\frac{1}{N}\sum_{i=1}^{N} ||\hat{x}_i - x||^2$$

- KL divergence evaluation
  1. Compute KL-divergence based on parameters returned from the encoder

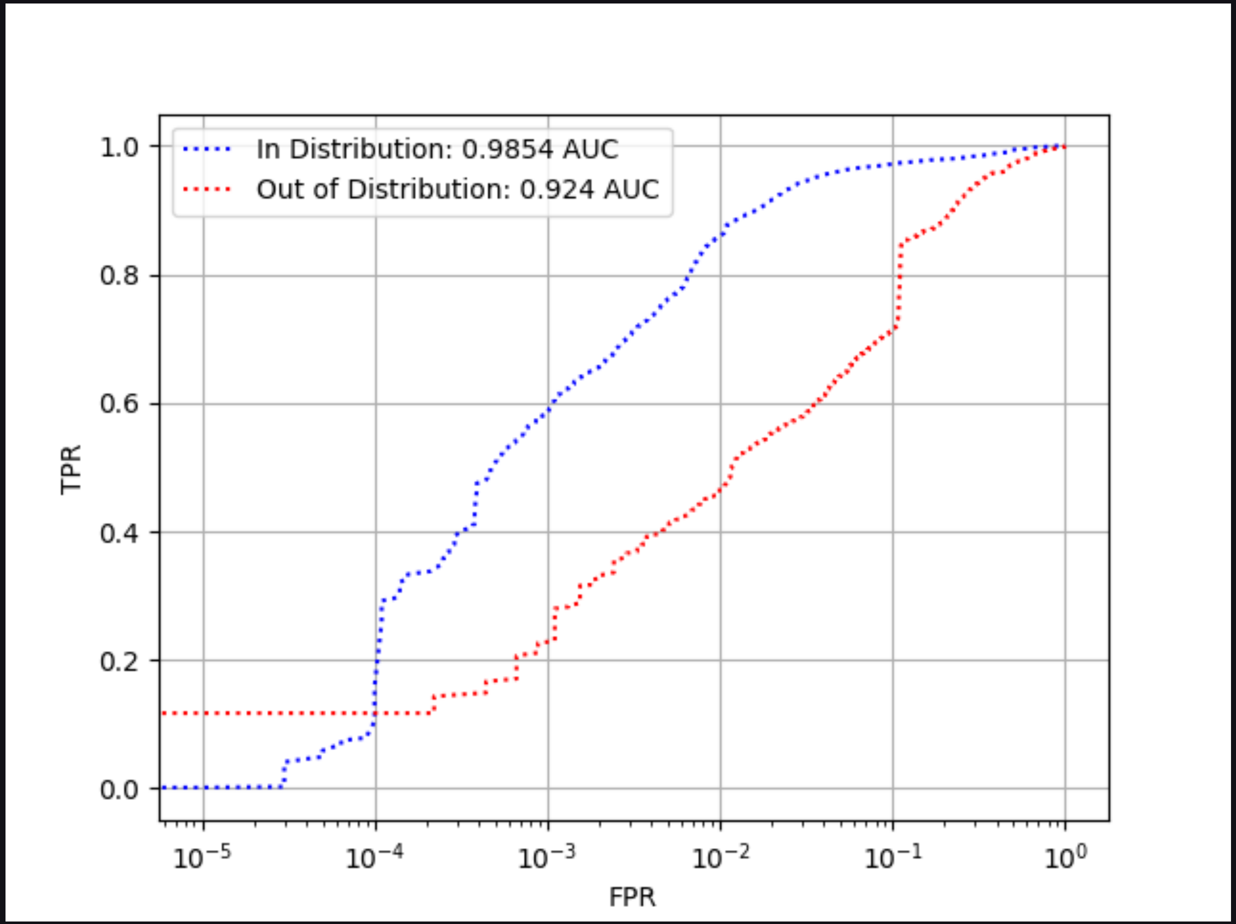


Segments in red have stochastic outputs

“Opening up” the EMBER 2018 Dataset

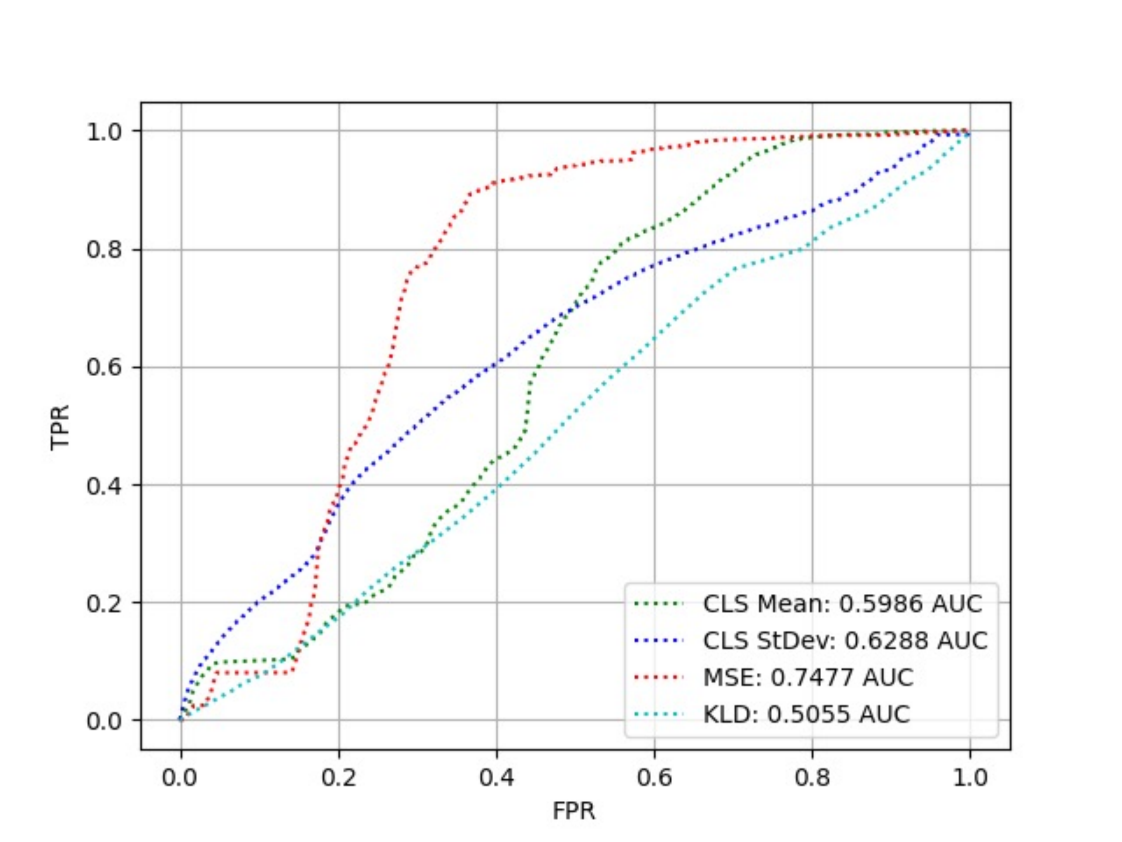
- Highly performant models suggest strong similarity between train and test distributions.
- How do we test turn EMBER into an “Open Set” dataset?
- Solution: CAPA (https://github.com/mandiant/capa)
  - Open source tool from Mandiant’s FLARE team for PE, ELF, and shellcode capabilities analysis
  - Outputs capabilities it “thinks” a file has, based on disassembly, heuristics, and a rule-based engine.
- We remove all samples w/ packing/unpacking capabilities during train and flag test samples with these capabilities as outside the training distribution.
  - 41,276 samples in train; 12,062 samples in test

ROC Comparison – Malware Detection on Open Set EMBER



- In-distribution malware detection performance remains relatively consistent.
- Significant performance decline for OOD (packed) malware.

ROC Comparison: Out of Distribution Detection



- As expected, using the density estimation head allows best detection of "known unknowns"
- Efficacy of thresholding on standard deviation over the classifier prediction suggests some level of score-level variability at the margin.

Conclusions

- We introduced a performant hybrid architecture with unique measures of “known known” and “known unknown” confidence

- Classifier estimate similar to “Dropout as a Bayesian Estimator” approach, but uses sampling from the latent distribution for stochasticity
  - Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. PMLR, 2016.

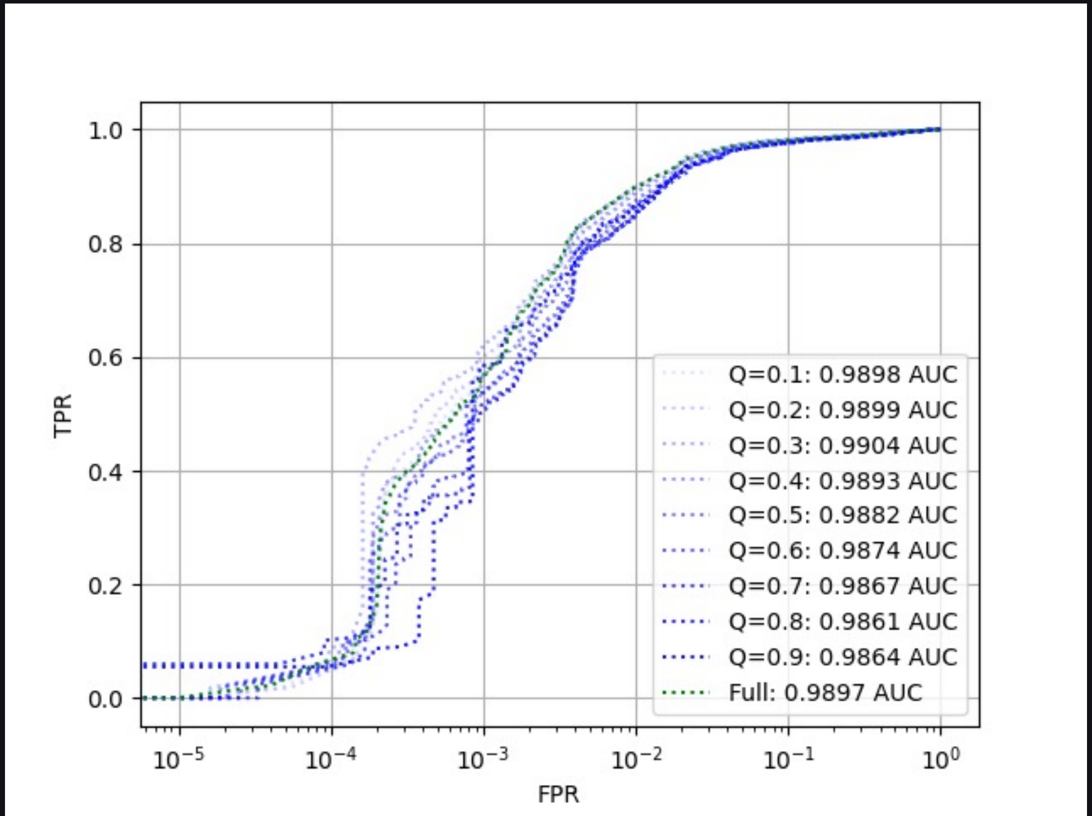
- Introduced an approach to turn EMBER (or other executable malware datasets) into Open Set benchmarks

- We also ran an experiment to reject OOD based on the KL divergence
  - Potentially higher confidence w/in a specific KL Divergence range
  - The effect is very slight; needs further investigation, potentially on another dataset to determine if it addresses “unknown unknowns”

MANDIANT

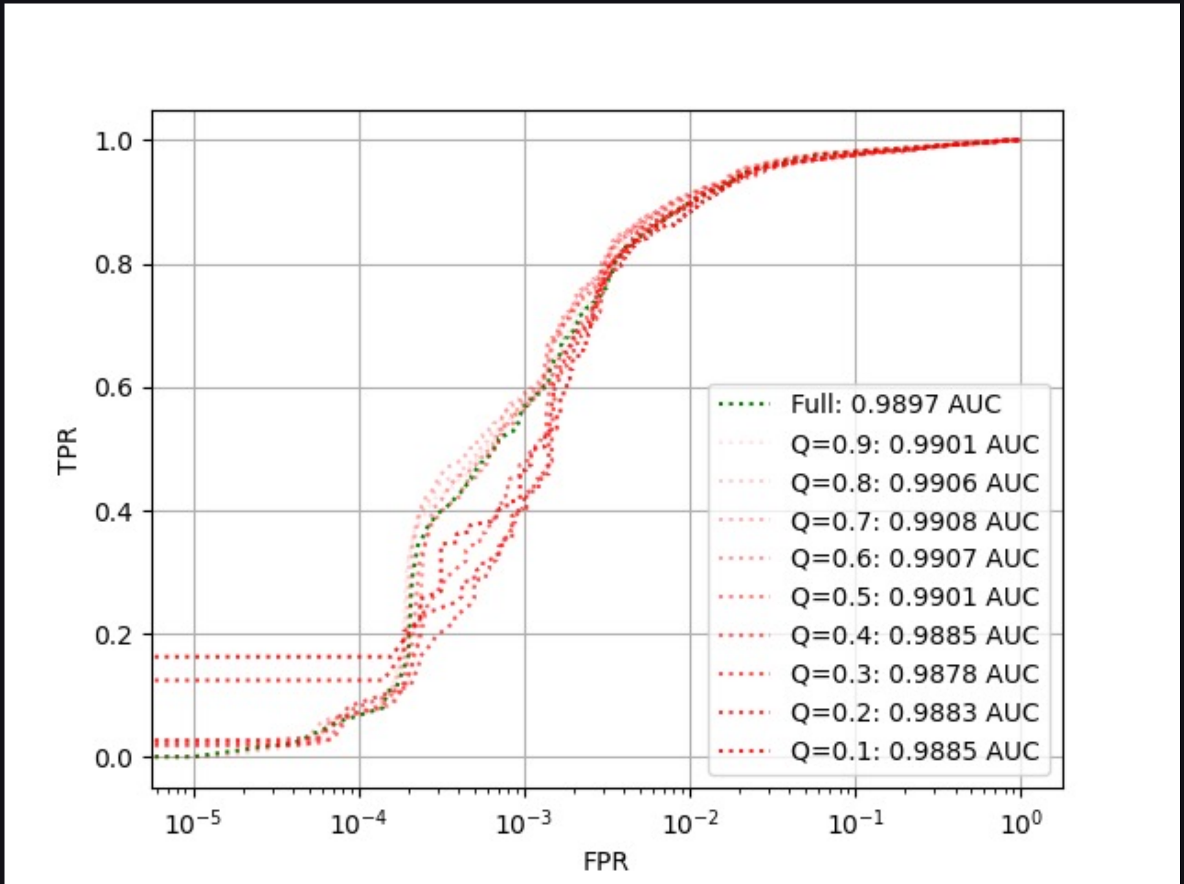
YOUR CYBERSECURITY ADVANTAGE

“Unknown Unknown” Detection by thresholding KL Divergence on EMBER 2018 Test?



- Included only data above quantile Q.
- Also ran this experiment for the ratio of KLD:MSE which performed slightly worse.
- Small but negligible gains in AUC for certain thresholds.

"Unknown Unknown" Detection by thresholding KL Divergence on EMBER 2018 Test?



- Included only data below quantile Q.
- Also ran this experiment for the ratio of KLD:MSE which performed slightly worse.
- Small but negligible gains in AUC for certain thresholds.