Unified Probabilistic Deep Continual Learning Through Generative Replay and Open Set Recognition

Martin Mundt, Sagnik Majumder, Iuliia Pliushch,
Yong Won Hong, and Visvanathan Ramesh

2020

Bardia Mojra
September 22, 2021
Robotic Vision Lab
University of Texas at Arlington

Introduction

- Based on Extreme Value Theory (EVT)
- Probabilistic open set recognition
- Allows for continual learning
- Prevents catastrophic forgetting
- Does not store entire data set
- Model:
  - Combines joint probabilistic encoder with a GAN
  - Tight classification bound on high-density learned parameter regions
  - Only use high-density data points in regenerative model training

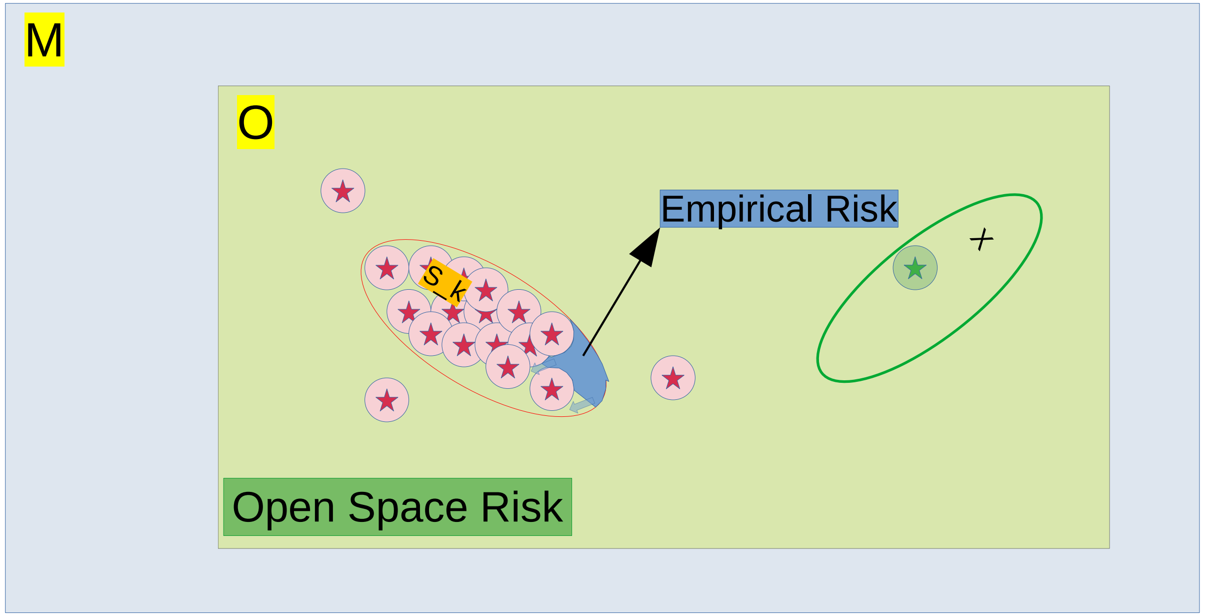
Variational Bayesian Inference

- Inference hidden variables
- Deterministic
- Easy to gauge convergence
- Requires dozens of iterations
- No conjugacy requirement
- More complicated math (slightly)

Open Set Recognition

› Definition per Scheirer et al. [31]:

- *“For any recognition function f over an input space* X*, the open set* O *is defines as* $O \subseteq X - S_K$*, where* S_K *is a union of balls of radius* r_o *including all of the training examples for known classes* $x \in K$*“.*



Bayesian Probability

- Observations – x
- Hidden variable – z
- Fixed (learned) parameters – alpha

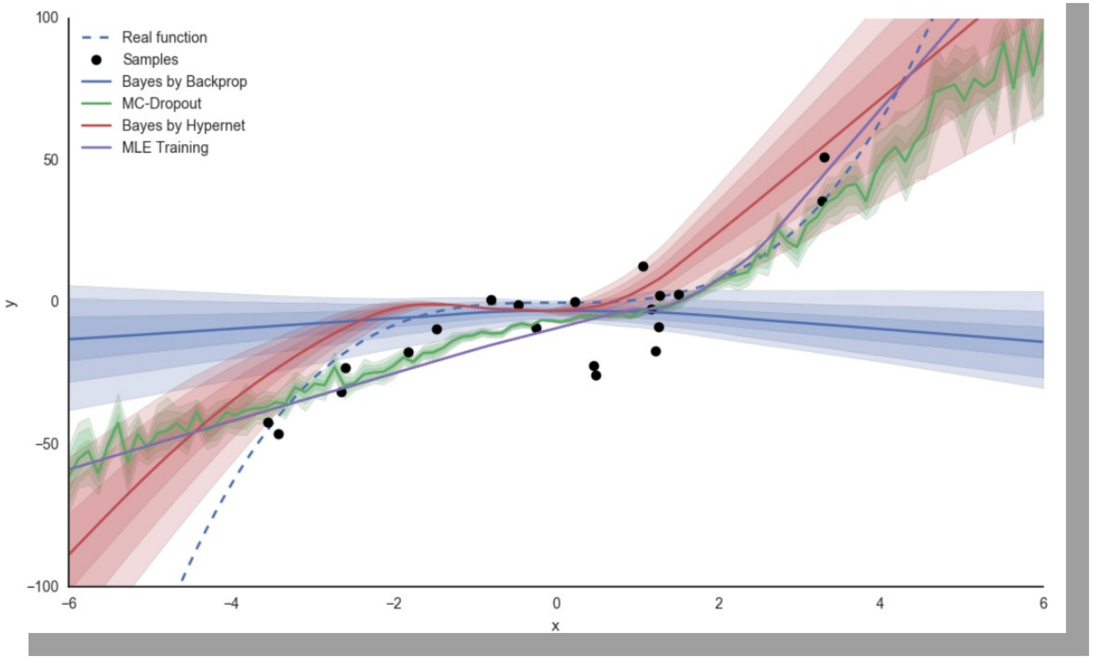
$$P(z|x,\alpha) = \frac{P(z,x|\alpha)}{\int_z P(z,x|\alpha)}$$

- Bayesian Inference:
  - Posterior requires calculating over all means and cluster assignments

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}$$

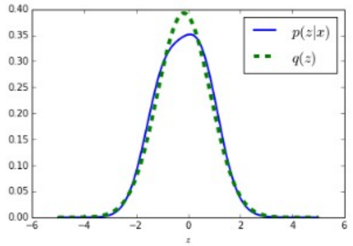
Bayesian Uncertainty

- Naturally rejects statistical outlier by dilution
- 100s to 1000s of forward passes with MCMC (Bayesian approx.)
- Fewer but many forward passes with MC Dropout (Bayesian approx.)
- Direct calculation not feasible
- Relies on the max entropy principle assumption
  - Directly contradicts with the Open Set Recognition definition



Source:Implicit Weight Uncertainty in Neural Networks - http://bayesiandeeplearning.org/2017/papers/36.pdf

Variational Inference



A variational Gaussian approximation to a scalar posterior.

A natural approach to fitting the approximation parameters λ is to minimize the KL divergence between our approximation $q(z;\lambda)$ and the posterior $p(z|x)$.[2] Writing this out,

$$KL\left[q(z;\lambda)\|p(z|x)\right] = \int q(z;\lambda)\log\frac{q(z;\lambda)}{p(z|x)}dz,$$

we see that it depends on the posterior density $p(z|x)$ which we don't know. However, we do have access to the joint distribution $p(x,z)$, which is proportional to the posterior, so we can just apply simple algebra to unpack the normalizing constant:

$$\begin{aligned}
KL\left[q(z;\lambda)\|p(z|x)\right] &= \int q(z;\lambda)\log\frac{q(z;\lambda)}{p(z|x)}dz \\
&= \int q(z;\lambda)\left[\log q(z;\lambda) - \log p(z|x)\right]dz \\
&= \int q(z;\lambda)\left[\log q(z;\lambda) - \log\frac{p(x,z)}{p(x)}\right]dz \\
&= \log p(x) + \int q(z;\lambda)\left[\log q(z;\lambda) - \log p(x,z)\right]dz \\
&= \log p(x) - \mathcal{F}(\lambda;x).
\end{aligned}$$

This shows that the KL divergence is equal to the model evidence $\log p(x)$, which is an (unknown) normalizing constant, minus a term $\mathcal{F}$ given by

$$\mathcal{F}(\lambda;x) = \int q(z;\lambda)\left[\log p(x,z) - \log q(z;\lambda)\right]dz.$$

This term is alternately referred to as (negative) variational free energy or the evidence lower bound (ELBO). It is a lower bound on $\log p(x)$

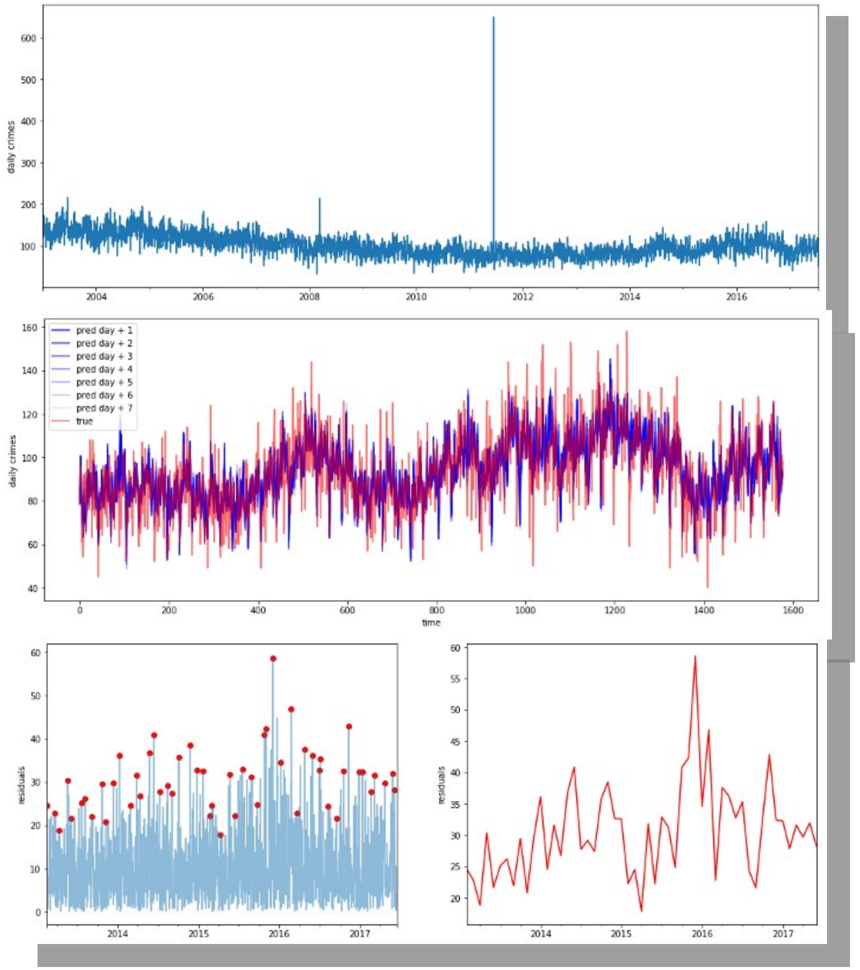
Source: Variational Inference in 5 Minutes - http://davmre.github.io/blog/inference/2015/11/13/elbo-in-5min

Calibration

- What to do with outliers?
  - [38] uses perturbation and temperature scaling
  - [39] uses a separate GAN with additional loss term for outliers
  - [40] uses unknown class label and true negative samples
- They use a GAN and outlier rejection using EVT (ELBO)

Extreme Value Theory (EVT)

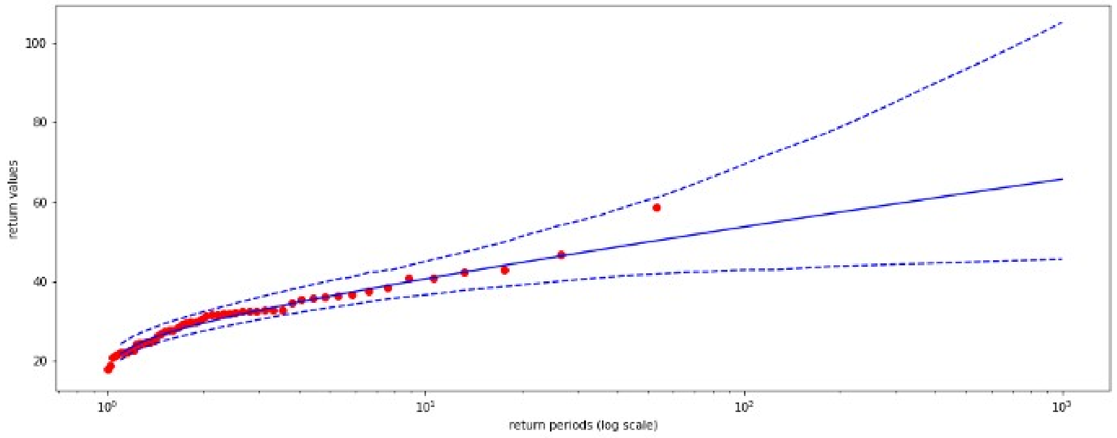
- Consider training residual absolute values, Extreme Values
- Model Extreme Values and assign distribution, Dirichlet distribution
- Compare empirical and estimated distribution
- Expect similar number of anomalies

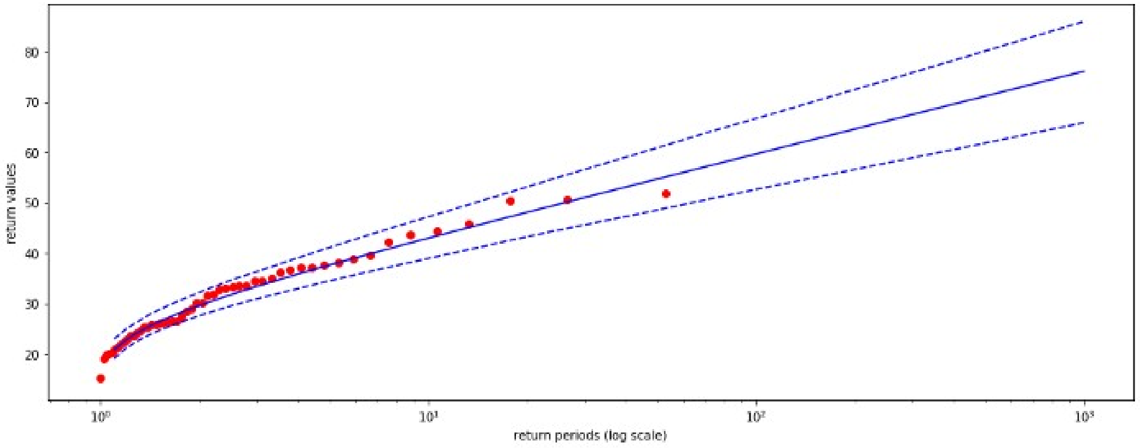


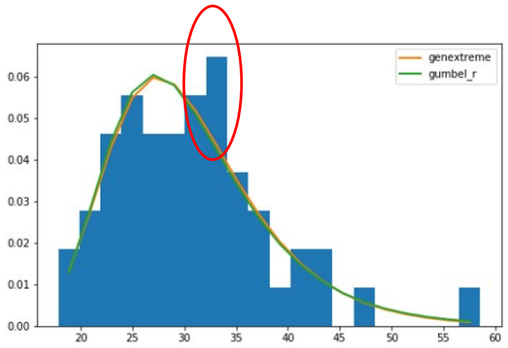
Source: https://towardsdatascience.com/anomaly-detection-with-extreme-value-analysis-b11ad19b601f

Extreme Value Theory (EVT)

- Consider training residual absolute values, Extreme Values
- Model Extreme Values and assign distribution, Dirichlet distribution
- Compare empirical and estimated distribution
- Expect similar number of anomalies





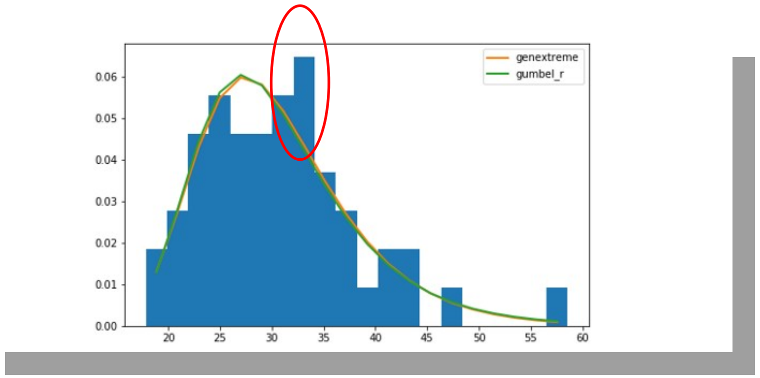


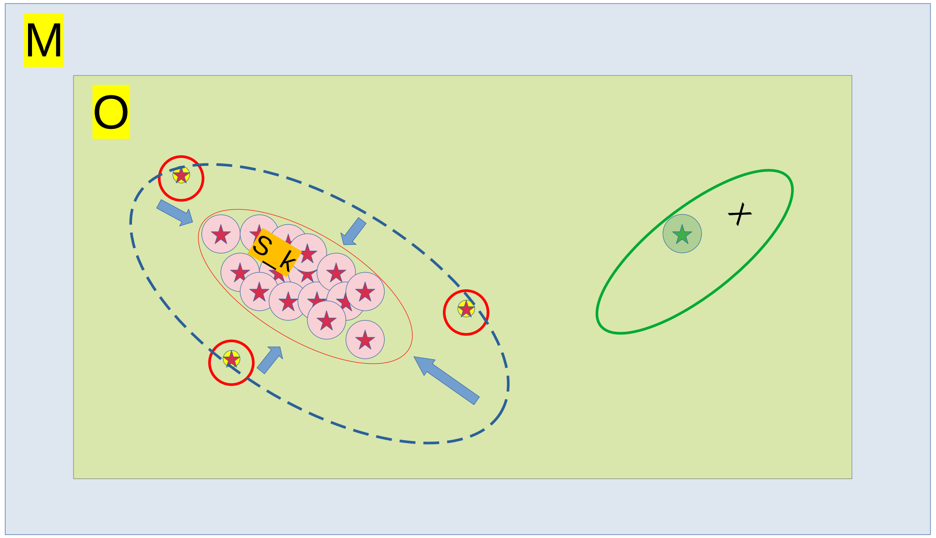
Source: https://towardsdatascience.com/anomaly-detection-with-extreme-value-analysis-b11ad19b601f

Extreme Value Theory (EVT)

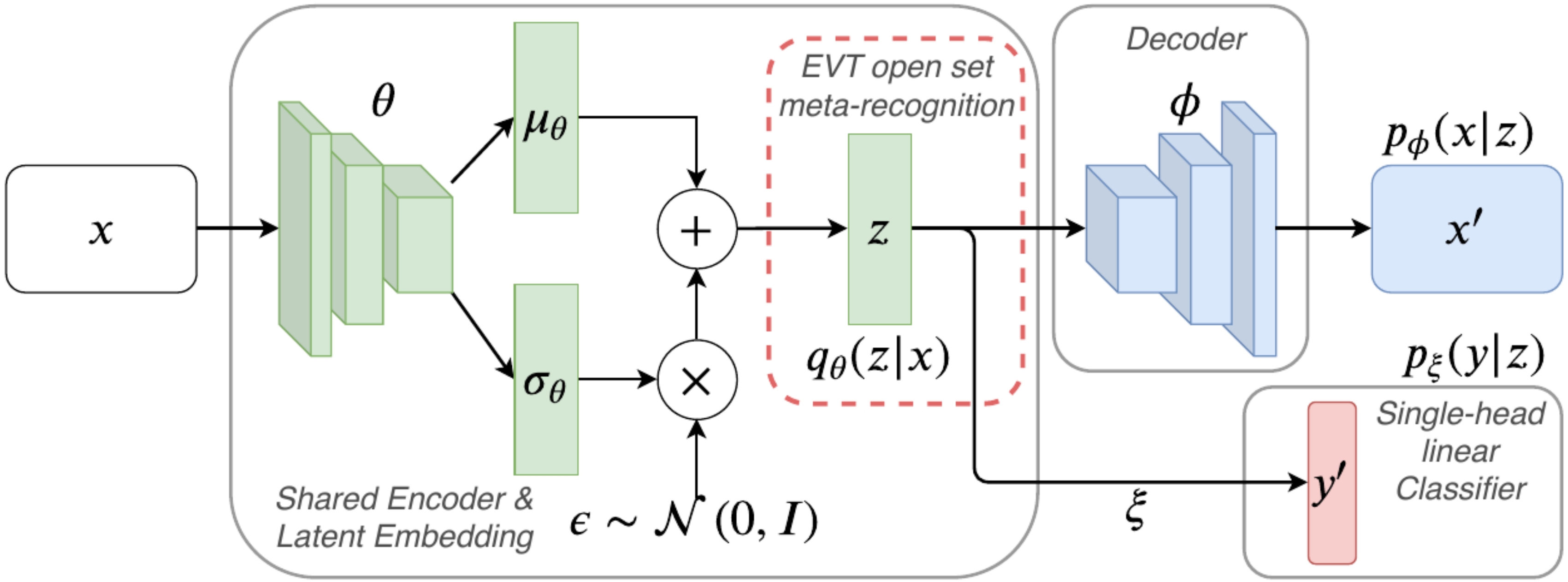
› In this paper:

- Use EVT to bound approximate posterior, instead of assigning confidence
- Use it to bound high density class distributions
- Used on each class separately





Model



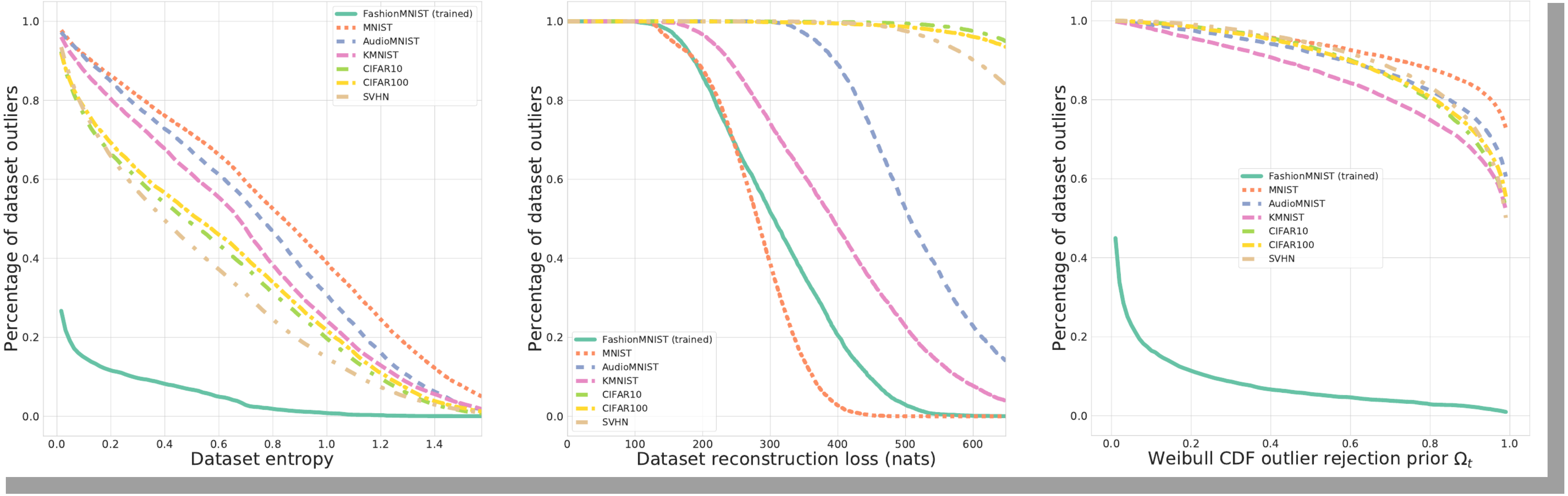
x
θ
μ_θ
σ_θ
Shared Encoder & Latent Embedding
$\epsilon \sim \mathcal{N}(0, I)$
EVT open set meta-recognition
z
$q_\theta(z|x)$
Decoder
ϕ
$p_\phi(x|z)$
x'
ξ
$p_\xi(y|z)$
y'
Single-head linear Classifier

Loss Function

$$\mathcal{L}\left(x^{(n)}, y^{(n)}; \theta\phi, \xi\right) = -\beta KL\left(q_\theta(z|x^{(n)})||p(z)\right) + \mathbb{E}_{q_\theta(z|x^{(n)})}\left[\log p_\phi(x^{(n)}|z) + \log p_\xi(y^{(n)}|z)\right] \tag{1}$$

- θ – shared encoder parameters
- φ – decoder parameters
- ξ – linear classifier parameters

Performance



- Thank you!