Tracking Change on the Edge

Enrique Nueve
NAISE
enrique.iv@northwestern.edu

Nicola J. Ferrier
Argonne National Laboratory
nferrier@anl.gov

Sean Shahkarami
Argonne National Laboratory
sshahkarami@anl.gov

Seongha Park
Argonne National Laboratory
seongha.park@anl.gov

Abstract

In this work, I will attempt to identify samples with high probability of a large loss that is being evaluated over an edge network. By assumption, samples with high OOD detection values have high probability of having a high loss value. Upon flagging a set of samples with high OOD detection values, I will send a particular subset of the data back to the cloud for human evaluation. Based on whether the model is predicting right or wrong on the sent back samples, I will either alert the user of drift or not. I desire to make my method capable of providing statistical guarantees of loss through the data sent back to the cloud for evaluation by a person.

1 Journal

Open ideas

1. Run a benchmark of different OOD methods on edge devices, could make a short paper out of it
2.
3.

Topics I need to look into

1. Benchmark methods and datasets
2. Planning code for test bed
3. Do a write up on different metrics

Preprint. Under review.

2 Open Questions

1. What statistical guarantees can I provide for a drop in loss through evaluating the data sent back to the cloud?
2. What data sets will I use for test

3 General Notes

3.1 Types of Concept Drift

1. **Type 1 (Virtual)**: $P_t(X) \neq P_{t+1}(X)$ yet $P_t(Y|X) = P_{t+1}(Y|X)$
2. **Type 2 (Real)**: $P_t(X) = P_{t+1}(X)$ yet $P_t(Y|X) \neq P_{t+1}(Y|X)$
3. **Type 3 (Real)**: $P_t(X) \neq P_{t+1}(X)$ and $P_t(Y|X) \neq P_{t+1}(Y|X)$

3.2 Methods that will be used

1. Maximum Softmax [HG18]
2. MC Dropout [GG16]
3. Deep Ensemble [LPB17]
4. Energy Based [Liu+20]
5. Variational Info Bottlneck [AFD18]

4 Lit Review

4.1 Confidence and Out of distribution detection

1. **Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift**
   - WILL USE!: A.9 Computational and Memory Complexity of Different methods
   - **Link**: https://arxiv.org/abs/1906.02530
   - **Usable**: YES
   - **Summary**: Confidence methods perform differently on in distribution data and out of distribution data. This paper states that most methods perform worse on OOD. The paper compares methods on how confidence changes on data from OOD (out-of-distribution) through experimentation. The abstract states that some methods that marginalize over models give good results for OOD detection. States that post-hoc calibration methods perform bad on OOD. In the section called takeaways, it states that deep ensembles perform the best across all metrics such as MC dropout with diminishing returns beyond size 5 of re-sampling.
   - **Terms**:
     (a) Aleatoric uncertainty: uncertainty due to inherit noise of data
     (b) Epistemic uncertainty: uncertainty due to lack of samples over whole sample space
     (c) Covariate shift: refers to the change in the distribution of the input variables present in the training and the test data
   - **Don't Understand**:
   - **Assumptions**:
   - **Limits**:
   - **Potential Improvements**:
   - **What ideas will I use**: I can use this in my related works to discuss the performance of different methods and why I choose the ones that I did.
2. **A statistical theory of out-of-distribution detection**
   - **Link**: https://arxiv.org/abs/2102.12959
   - **Usable**: NO
   - **Summary**: Defines train data to all be in distribution and then shows that by definition out of sample data could be calculated if you knew the prob of each labeler assign a class. Constantly refers to a paper by the name of *Deep anomaly detection with outlier exposure*. It is hard to understand the paper given that it constantly refers to it. I recommend first reading *Deep anomaly detection with outlier exposure* then this paper.
   - **Don't Understand**: To calculate the OOD prob or not OOD prob, equations 6 and 7 are needed. However, there is a sum within the equation. This probability is discussed in the background section yet, does not explain clearly how to calculate in practice.
   - **Assumptions**:
   - **Limits**: Assumes you can train the model on some proxy OOD dataset along with the main dataset.
   - **Potential Improvements**:
   - **What ideas will I use**:
3. **Energy-based Out-of-distribution Detection**
   - **Link**:https://arxiv.org/abs/2010.03759
   - **Usable**: YES
   - **Summary**: Proposed an OOD method that uses energy scores since energy scores are theoretically aligned with the probability density of the inputs and are less susceptible to the overconfidence issue. Proposed energy method can be used both as a scoring function for any pre-trained neural classifier (without re-training), and a trainable cost function to fine-tune the classification model.
   - **Terms**:
     (a) Energy: Scalar value EBM model maps an input too.

(b) Energy-based model (EBM): essence of the energy-based model (EBM) is to build a function $E(x) : \Re^D \rightarrow \Re$ that maps each point x of an input space to a single, non-probabilistic scalar called the energy

- **Don't Understand**: Not sure how they pick the tau value to determine in or out sample data
- **Assumptions**:
- **Limits**:
- **Potential Improvements**: Try to modify density estimation methods to use this idea of energy functions
- **What ideas will I use**: I will use their method that uses an energy function to test whether or not a sample is OOD or not.

4. **A BASELINE FOR DETECTING MISCLASSIFIED AND OUT-OF-DISTRIBUTION EXAMPLES IN NEURAL NETWORKS**
   - **Link**:https://arxiv.org/pdf/1610.02136.pdf
   - **Usable**: YES
   - **Summary**: We consider the two related problems of detecting if an example is misclassified or out-of-distribution.However, in this work we also show the prediction probability of incorrect and out-of-distribution examples tends to be lower than the prediction probability for correct examples. In summary, while softmax classifier probabilities are not directly useful as confidence estimates, estimating model confidence is not as bleak as previously believed. Simple statistics derived from softmax distributions provide a surprisingly effective way to determine whether an example is mis- classified or from a different distribution from the training data, as demonstrated by our experimental results spanning computer vision, natural language processing, and speech recognition tasks. This creates a strong baseline for detecting errors and out-of-distribution examples which we hope future research surpasses.
   - **Terms**:
     (a) AUROC: AUROC can be interpreted as the prob- ability that a positive example has a greater detector score/value than a negative example. Consequently, a random positive example detector corresponds to a 50 percent AUROC, and a "perfect" classifier corresponds to 100 percent.
     (b) ROC: ROC curve is a graph showing the true positive rate (tpr = tp/(tp + fn)) and the false positive rate (fpr = fp/(fp + tn)) against each other
     (c) AUPR: AUPR adjusts for these different positive and negative base rates. The PR curve plots the precision (tp/(tp + fp)) and recall (tp/(tp + fn)) against each other. The baseline detector has an AUPR approximately equal to the precision (Saito Rehmsmeier, 2015), and a "perfect" classifier has an AUPR of 100 percent
     (d) abnormality module: decoder trained post model to estimate OOD
   - **Don't Understand**:
   - **Assumptions**:
   - **Limits**:
   - **Potential Improvements**:
   - **What ideas will I use**: Use Wilcoxon rank-sum test to see if methods are significant, I want to look into this. Try using abnormality module in my paper as baseline.

5. **DEEP ANOMALY DETECTION WITH OUTLIER EXPOSURE**
   - **Link**:https://arxiv.org/pdf/1812.04606.pdf
   - **Usable**: reading
   - **Summary**:
   - **Terms**:
     (a) AUROC:
   - **Don't Understand**:
   - **Assumptions**:
   - **Limits**:
   - **Potential Improvements**:

- **What ideas will I use**:

6. **Uncertainty in the Variational Information Bottleneck**
   - **Link**:https://arxiv.org/pdf/1807.00906.pdf
   - **Usable**: reading
   - **Summary**:
   - **Terms**:
     (a) term 1
   - **Don't Understand**:
   - **Assumptions**:
   - **Limits**:
   - **Potential Improvements**:
   - **What ideas will I use**:

5 Metrics

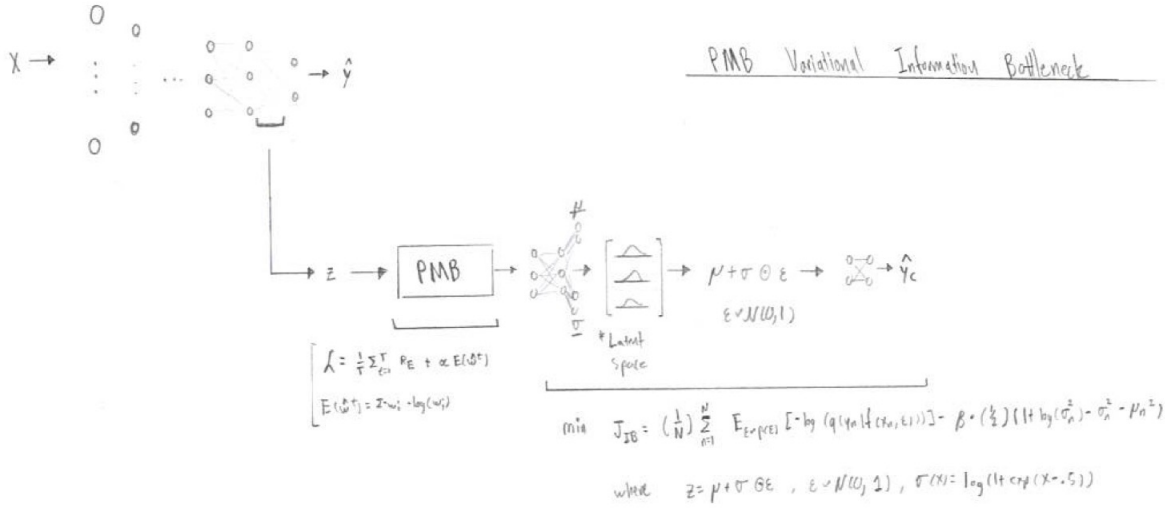
If you treat OOD detection as a binary classifier problem, the following metrics can be used:

1. **AUROC**: how well model can discriminate between positive or negative class
2. **FPRN**: computes prob of a labeled sample being misclassified as an OOD sample (false positive) when at least N percent of the true OOD samples are correctly detected.
3. **Detection Error**: computes the misclassification prob
4. **AUPR**: like AUROC but takes into account precision

Metrics proposed by the paper *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty UnderDataset Shift*

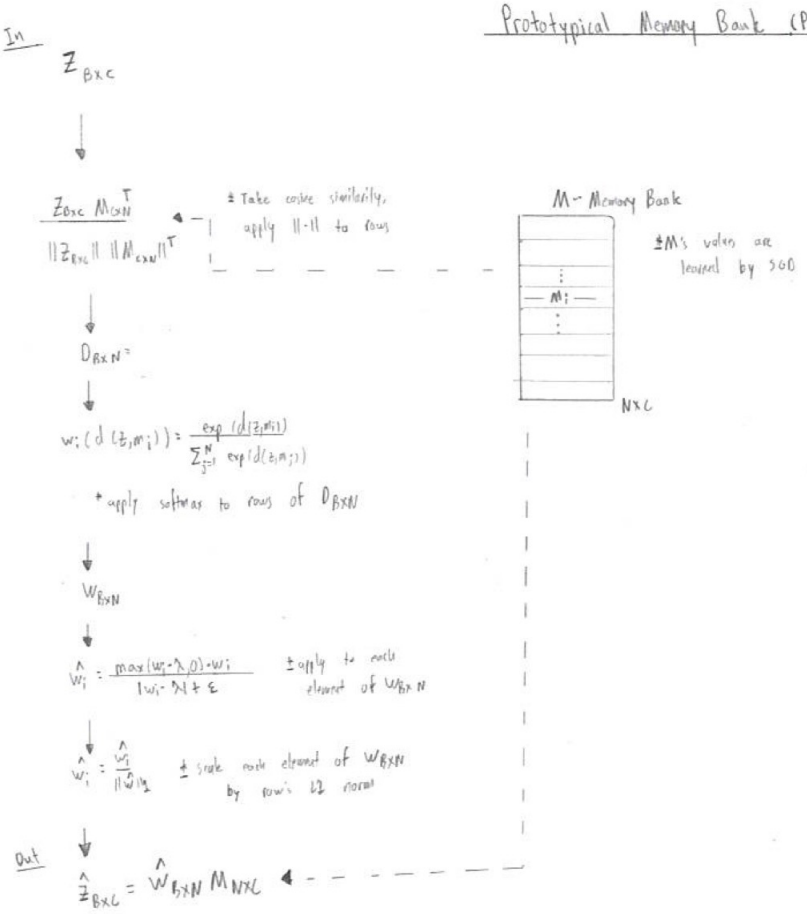
1. **Negative Log-Likelihood**:
2. **Brier Score**: The Brier score has a convenient interpretation as BS = uncertainty resolution + reliability, where uncertainty is the marginal uncertainty over labels, resolution measures the deviation of individual predictions against the marginal, and reliability measures calibration as the average violation of long-term true label frequencies
3. **Expected Calibration Error**: Measures the correspondence between predicted probabilities and empirical accuracy. Refer to *Obtaining Well Calibrated Probabilities Using Bayesian Binning*.
4. **Maximum Calibration Error:**
5. **Entropy:** $\sum_i p(y_i|x) \log p(y_i|x)$

6 VIB Memory Bank Test

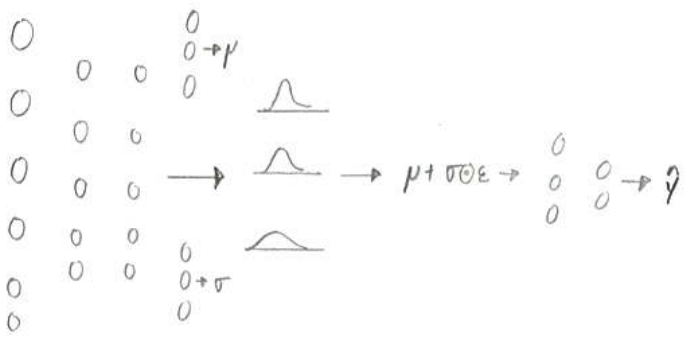


B = Batch Size
C = # - features
N = # - memory cells

Prototypical Memory Bank (PMB)

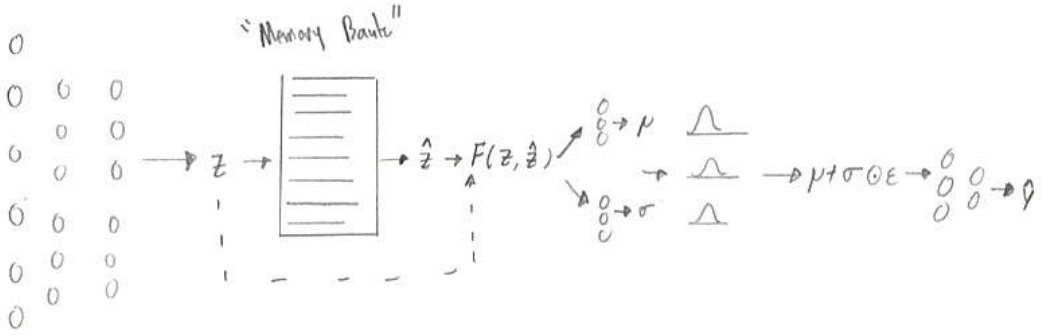


VIB

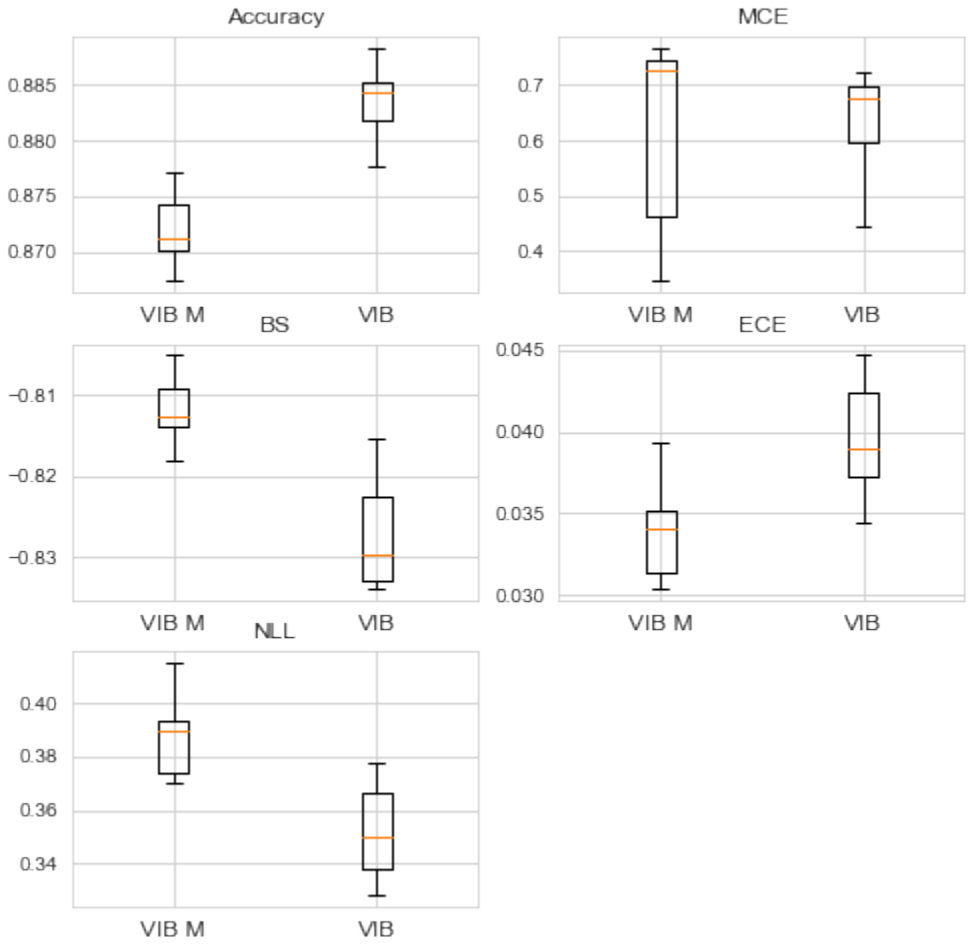


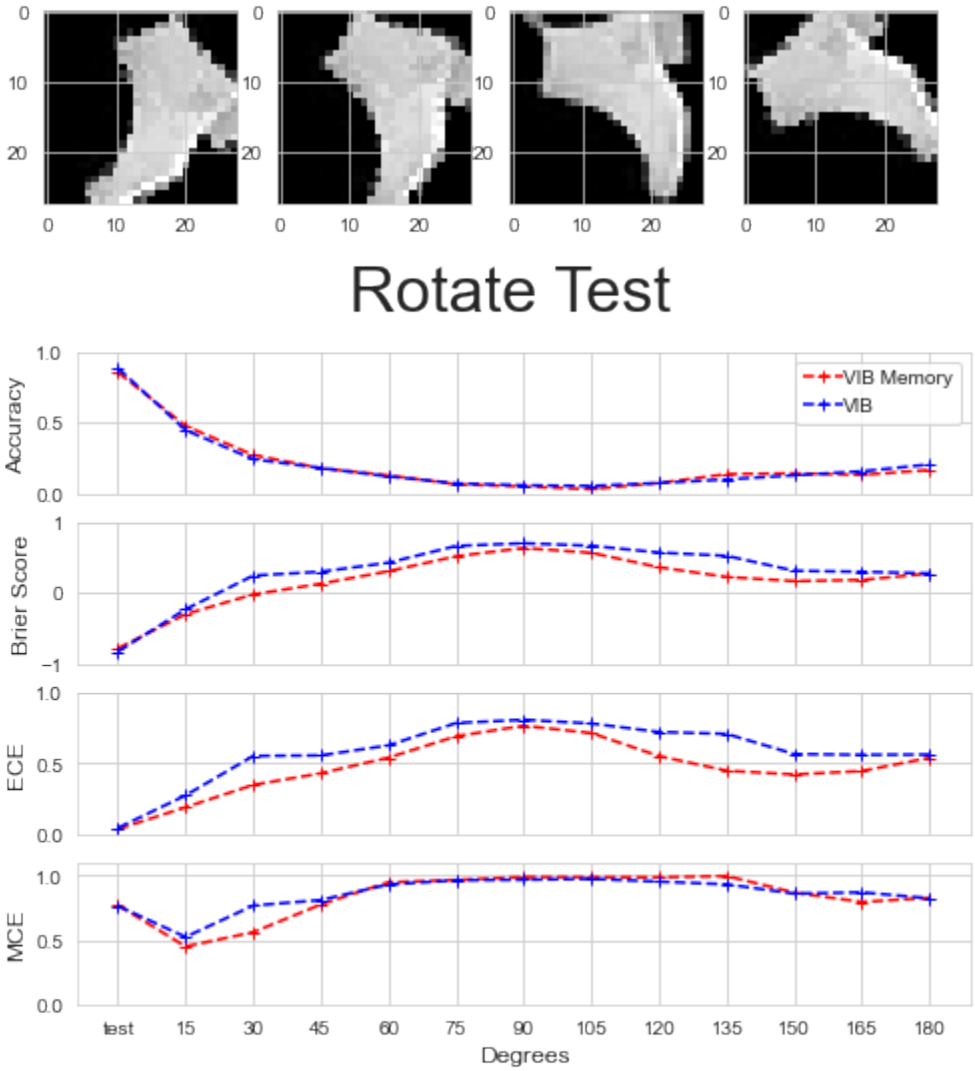
VIB Memory Bank

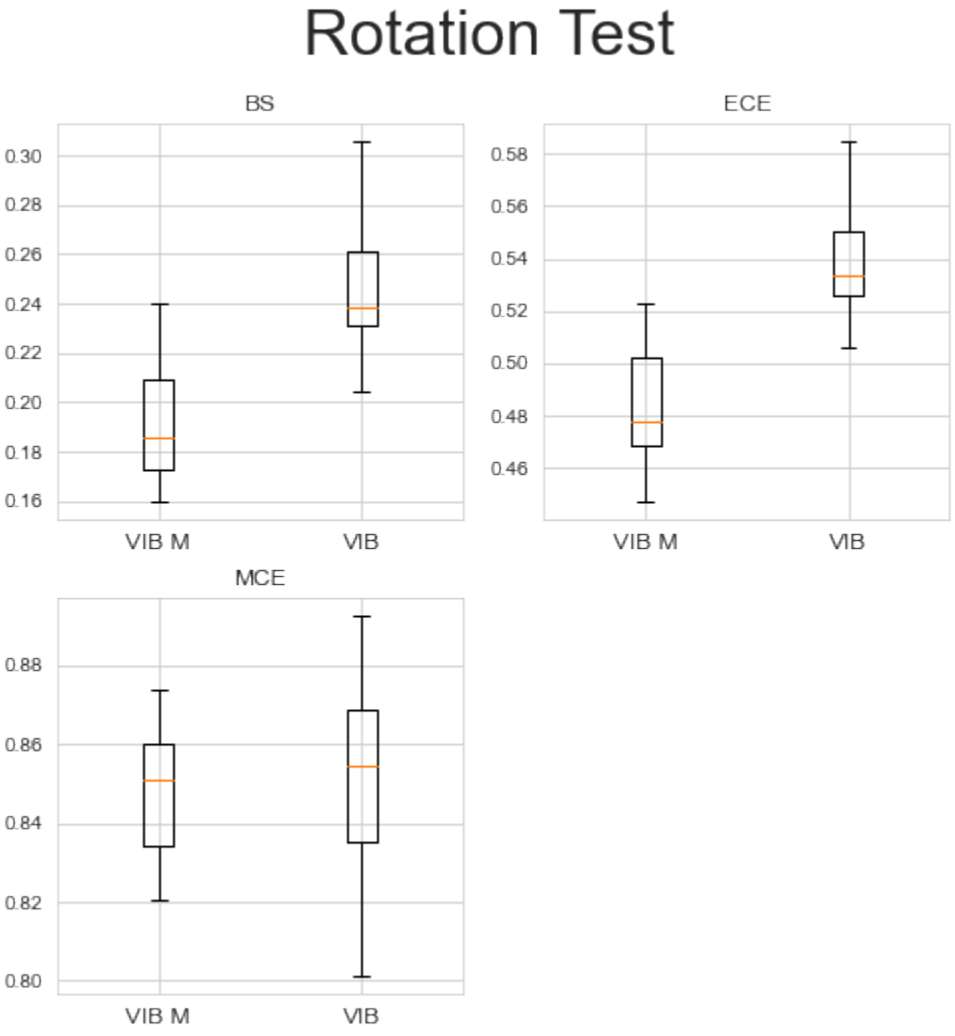
$$F(z, \hat{z}) = \hat{z} + U(-1,1) \odot (z - \hat{z})^2$$

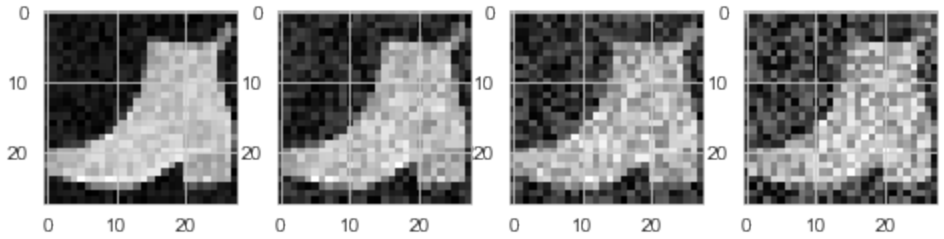


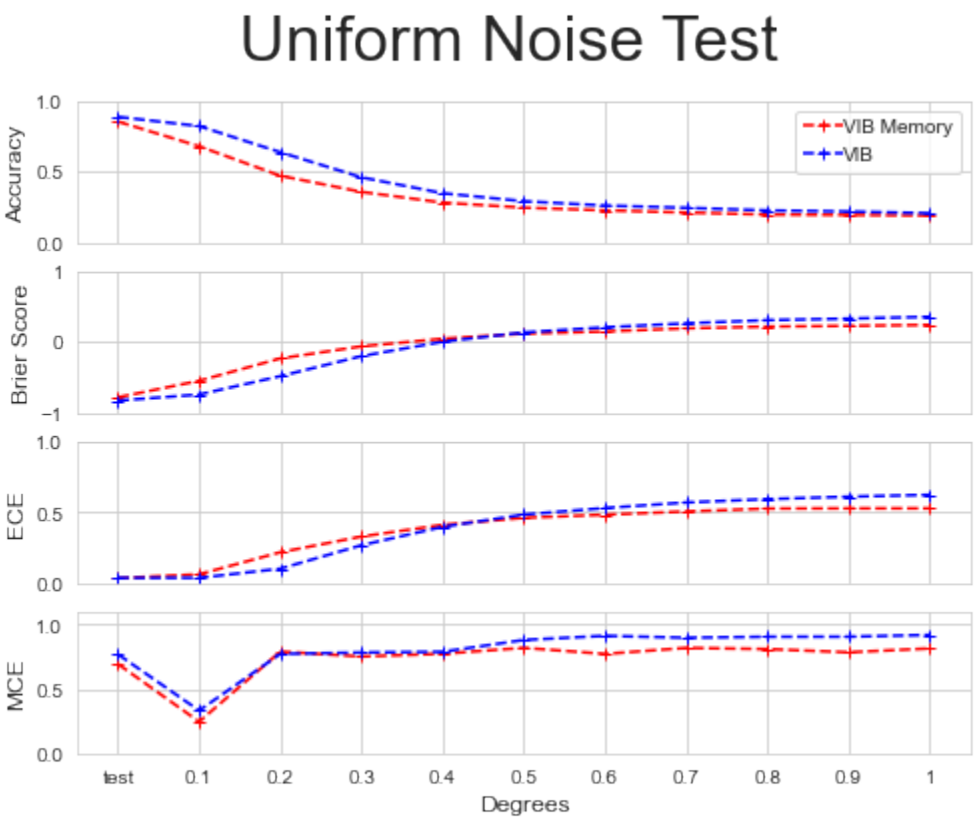
In Distribution

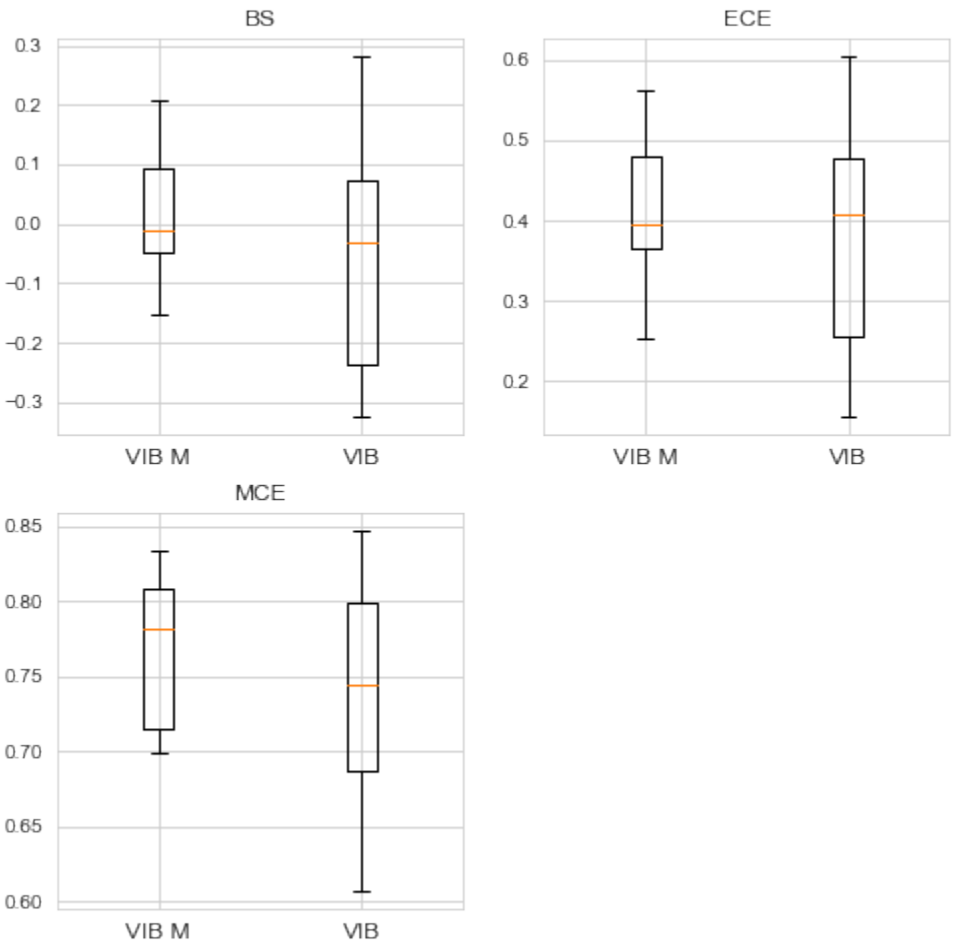


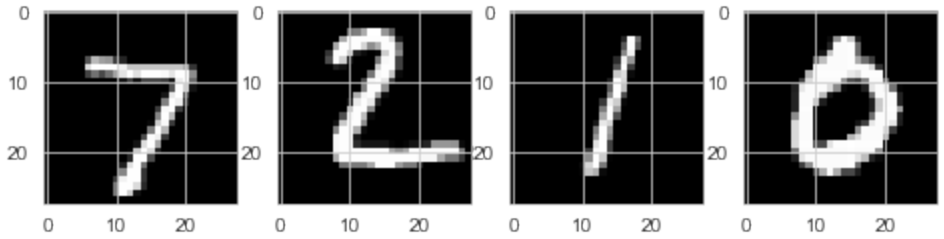

Rotate Test
VIB Memory
VIB
Accuracy
Brier Score
ECE
MCE
Degrees
test
15
30
45
60
75
90
105
120
135
150
165
180

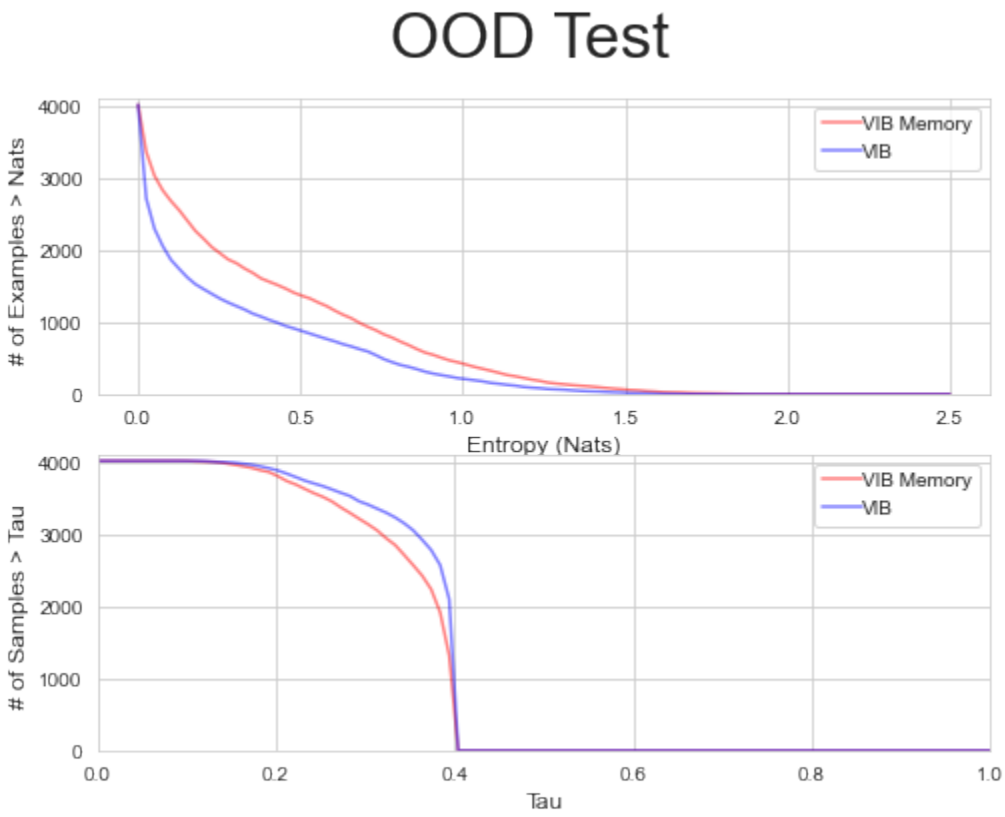

Rotation Test
BS
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
VIB M
VIB
ECE
0.58
0.56
0.54
0.52
0.50
0.48
0.46
VIB M
VIB
MCE
0.88
0.86
0.84
0.82
0.80
VIB M
VIB

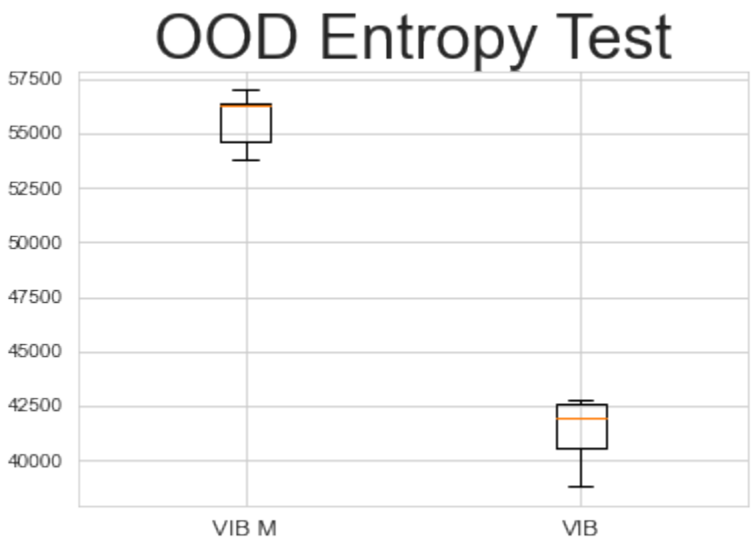



Uniform Noise Test
VIB Memory
VIB
Accuracy
Brier Score
ECE
MCE
Degrees
test
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

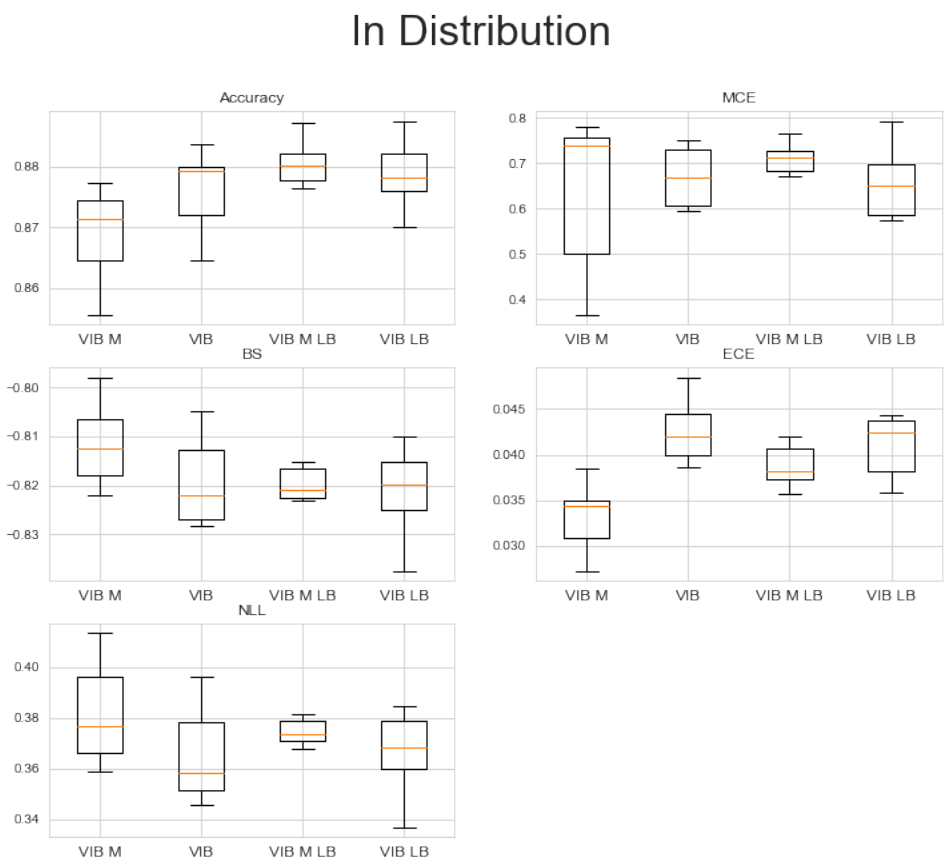

Uniform Noise Test
BS
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
VIB M
VIB
ECE
0.6
0.5
0.4
0.3
0.2
VIB M
VIB
MCE
0.85
0.80
0.75
0.70
0.65
0.60
VIB M
VIB

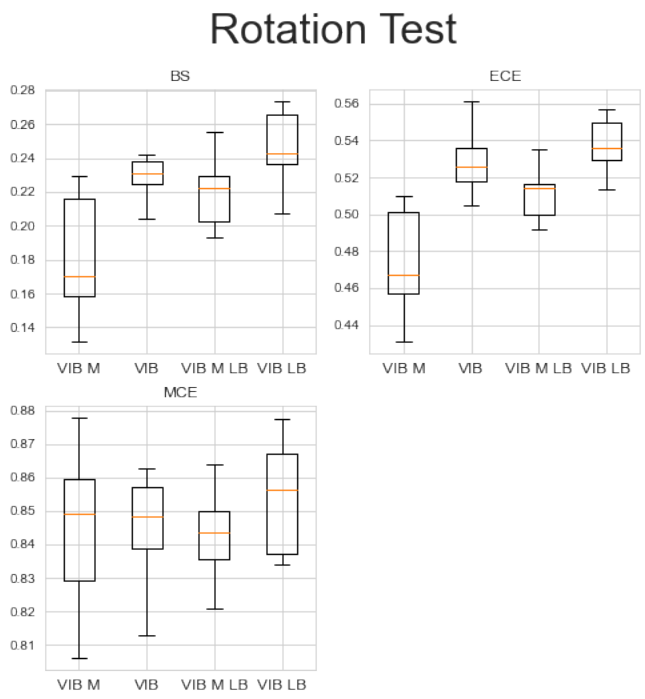


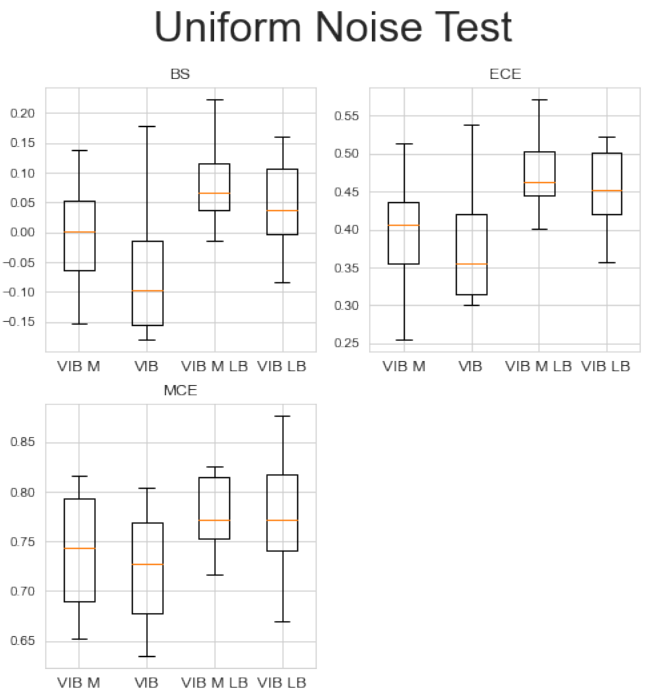

OOD Test
VIB Memory
VIB
# of Examples > Nats
Entropy (Nats)
# of Samples > Tau
Tau

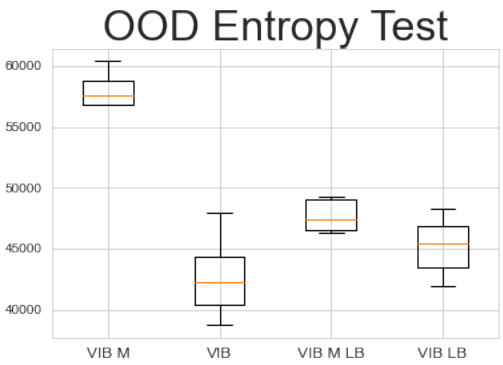

OOD Entropy Test
57500
55000
52500
50000
47500
45000
42500
40000
VIB M
VIB

7 VIB Memory Bank with Linear Block (LB) Test

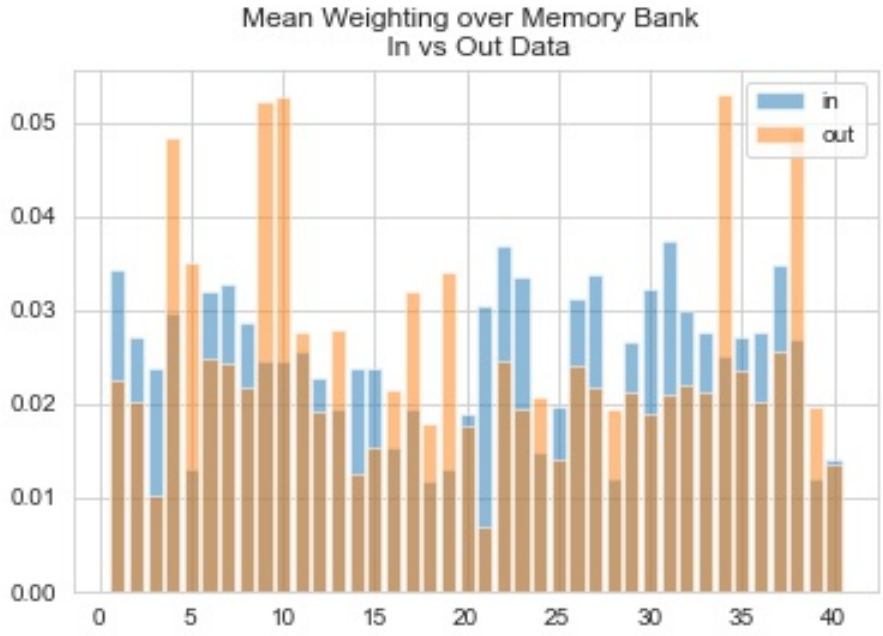



Rotation Test
BS
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
VIB M
VIB
VIB M LB
VIB LB
ECE
0.56
0.54
0.52
0.50
0.48
0.46
0.44
VIB M
VIB
VIB M LB
VIB LB
MCE
0.88
0.87
0.86
0.85
0.84
0.83
0.82
0.81
VIB M
VIB
VIB M LB
VIB LB


Uniform Noise Test
BS
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
VIB M
VIB
VIB M LB
VIB LB
ECE
0.55
0.50
0.45
0.40
0.35
0.30
0.25
VIB M
VIB
VIB M LB
VIB LB
MCE
0.85
0.80
0.75
0.70
0.65
VIB M
VIB
VIB M LB
VIB LB


OOD Entropy Test
60000
55000
50000
45000
40000
VIB M
VIB
VIB M LB
VIB LB

8 Weight over memory for in vs out of dist data



9 Methodology

1.

10 Experiments

11 Results

12 Conclusion

References

[GG16] Yarin Gal and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2016. arXiv: 1506.02142 [stat.ML].

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*. 2017. arXiv: 1612.01474 [stat.ML].

[AFD18] Alexander A. Alemi, Ian Fischer, and Joshua V. Dillon. *Uncertainty in the Variational Information Bottleneck*. 2018. arXiv: 1807.00906 [cs.LG].

[HG18] Dan Hendrycks and Kevin Gimpel. *A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks*. 2018. arXiv: 1610.02136 [cs.NE].

[Liu+20] Weitang Liu et al. *Energy-based Out-of-distribution Detection*. 2020. arXiv: 2010.03759 [cs.LG].