Out of Distribution Data Detection Using Dropout Bayesian Neural Networks

Andre T. Nguyen,[1,2,3] Fred Lu,[1,2,3] Gary Lopez Munoz,[1,2] Edward Raff,[1,2,3] Charles Nicholas,[3] James Holt[1]

[1]Laboratory for Physical Sciences
[2]Booz Allen Hamilton
[3]University of Maryland, Baltimore County
andre@lps.umd.edu, lu_fred@bah.com, dlmgary@lps.umd.edu, edraff@lps.umd.edu, nicholas@umbc.edu, holt@lps.umd.edu

Abstract

We explore the utility of information contained within a dropout based Bayesian neural network (BNN) for the task of detecting out of distribution (OOD) data. We first show how previous attempts to leverage the randomized embeddings induced by the intermediate layers of a dropout BNN can fail due to the distance metric used. We introduce an alternative approach to measuring embedding uncertainty, justify its use theoretically, and demonstrate how incorporating embedding uncertainty improves OOD data identification across three tasks: image classification, language classification, and malware detection.

1 Introduction

Detecting out of distribution (OOD) data at test time is critical in a variety of machine learning applications. For example, in the context of malware classification (Raff and Nicholas 2020), OOD data could correspond to the emergence of a new form of malicious attack. Gal and Ghahramani (2016b) developed an approach to variational inference in Bayesian neural networks (BNNs) that showed a neural network with dropout (Hinton et al. 2012; Srivastava et al. 2014), a technique commonly used to reduce overfitting in neural networks (NNs) by randomly dropping units during training, applied before every weight layer is equivalent to an approximation of a deep Gaussian process (Damianou and Lawrence 2013). Training with dropout effectively performs variational inference for the deep Gaussian process model, and the posterior distribution can be sampled from by leaving dropout on at test time. This approach to Bayesian deep learning has been popular in practice as it is easy to implement and scales well.

Measures of uncertainty usually are a function of the sampled softmax outputs of such a BNN, for example predictive entropy and mutual information. There is however useful information at every intermediate layer of a dropout BNN. The dropout based approach to Bayesian deep learning suffers, like most variational inference methods, from the tendency to fit an approximation to a local mode instead of to the full posterior because of a lack of representational capacity and because of the directionality of the KL divergence (Smith and Gal 2018; Wilson and Izmailov 2020). This behavior however allows us to expect the randomized intermediate representation samples in a dropout BNN to be meaningfully related as they are sampled from a local mode. In this paper, we explore how to leverage additional information generated at every layer of the network for the task of OOD data detection at test time. In particular, we interpret the intermediate representation of a data point at a particular layer as a randomized embedding. The embedding is randomized due to the use of dropout at test time.

The idea to use a randomized embedding induced by the intermediate layers of a dropout BNN has been attempted previously, but can fail due to the underlying Euclidean distance metric used in previous work. The use of Euclidean distance does not account for the confounding variability caused by changes in embedding magnitudes. We will theoretically justify and empirically show that by instead using a measure based on cosine distance, this problem can be rectified. We then leverage this improved uncertainty estimation to show better OOD data identification across three highly different tasks to demonstrate the robustness of our approach.

The objective of this paper is not to develop a state-of-the-art approach to OOD data detection, but rather in the context of dropout BNNs to: (1) show how to cheaply improve OOD data detection in systems where a dropout BNN is already deployed, by using intermediate computational results that are already being computed but not fully leveraged, and (2) provide theoretical and practical evidence to highlight why it is valuable to deconflate angular information about embedding dispersion from embedding norm information. Additionally, previous works have evaluated OOD detection by assuming access to a large OOD dataset of similar size to the in distribution dataset. This is an unrealistic assumption as in areas like cyber security where OOD examples are limited and expensive. So, we also examine the effect of small dataset sizes for OOD detection in our experiments.

2 Related Work

Two kinds of uncertainty can be distinguished (Kendall and Gal 2017). Aleatoric uncertainty is caused by inherent noise and stochasticity in the data. More training data will not help to reduce this kind of uncertainty. Epistemic uncertainty on the other hand is caused by a lack of similar training data. In regions lacking training data, different model parameter settings that produce diverse or potentially conflicting predic-

tions can be comparably likely under the posterior. OOD data is expected to have higher uncertainty, epistemic in particular. Mukhoti et al. (2021) prove that one cannot infer epistemic uncertainty from a deterministic model's softmax entropy, so additional information is needed to estimate epistemic uncertainty.

Uncertainty modeling using probabilistic embeddings has primarily been used for estimating aleatoric uncertainty (Oh et al. 2019; Shi and Jain 2019; Chun et al. 2021; Chang et al. 2020) in tasks such as determining the quality of a test input image. These methods do not easily translate to estimating epistemic uncertainty. For example, Oh et al. (2019) try to apply their method on an epistemic uncertainty estimation task and find that it did not work well for novel classes, and they leave the modeling of epistemic uncertainty as future work.

The only prior work we are aware of that looks at a randomized embedding approach similar to ours is by Terhörst et al. (2020), who use dropout at test time to generate a stochastic embedding. They estimate face image quality through the stability of the embedding as measured using Euclidean distance. As we will show, the use of Euclidean distance is problematic as it does not account for factors affecting embedding norms and more generally, the assumptions made by Terhörst et al. (2020) are not met in reality. We also note that they are actually estimating epistemic uncertainty (see (Oh et al. 2019) for an explanation) when test image quality is an inherently aleatoric uncertainty estimation problem. We will show both empirical evidence as well as mathematical grounding as to why our proposed approach, without the addition of any complexity, fixes these issues.

There is evidence that intermediate layers of a neural network contain information useful for epistemic uncertainty estimation and out of distribution detection. Postels et al. (2020) establish a connection between the density of hidden representations and the information-theoretic surprise of observing a specific sample in the setting of a deterministic neural network. In particular, they suggest that the first layers of a neural network should be used to estimate epistemic uncertainty due to feature collapse, a phenomena where out-of-distribution data is mapped to in-distribution feature representations in later layers of a network (van Amersfoort et al. 2020; Mukhoti et al. 2021), though they also suggest that OOD data detection can benefit from aggregating uncertainty information from several layers. Our work differs from their work as we are not fitting a density to representations of the training data, increasing the applicability of our approach to situations where fitting and storing a density is not an option for computational or regulatory reasons.

Other recent work has also looked at uncertainty estimation using a single forward pass of a neural network that has had its intermediate representations regularized to produce good uncertainty estimates (van Amersfoort et al. 2020; Liu et al. 2020). We note that many single forward pass based methods like (Mukhoti et al. 2021; Liu et al. 2020) require residual based networks in combination with spectral normalization to enforce a bi-Lipschitz inductive bias (Bartlett, Evans, and Long 2018). While the method of (van Amersfoort et al. 2020) is not residual network constrained, it requires significant changes to the model and training procedure. While our approach requires multiple forward passes (as is the case with all dropout BNNs), it does not require any modifications to existing dropout BNNs, by only using information that is already being computed within a dropout BNN.

(Mandelbaum and Weinshall 2017) propose a confidence score that uses a data embedding derived from the penultimate layer of a neural network. The embedding is achieved using either a distance-based loss or adversarial training. Similarly to other methods, this method requires density estimation, and our work differs as our method does not involve a comparison to nearest neighbors from the training set, which may be difficult to deploy in practice due to both storage and regulatory constraints.

Many works have investigated OOD data detection in probabilistic contexts. Ovadia et al. (2019) benchmarks Bayesian deep learning methods in the context of dataset shift and OOD data at test time. Xiao, Gomez, and Gal (2020) use epistemic uncertainty to detect OOD language data. Ren et al. (2019) detect OOD data using likelihood ratios in the context of deep generative models and evaluate on OOD genomic sequences. Our work makes a contribution to probabilistic OOD identification by being the first work to systematically investigate the appropriate use of the randomized embeddings induced by the intermediate layers of a dropout BNN.

3 Methods

In a supervised setting, suppose a neural network structure with N (non-linearity included) layers $f_i, i \in [1, N]$ where x_1 is the input and x_{N+1} is the prediction: $x_{i+1} = f_i(x_i)$. Gal and Ghahramani (2016b) showed that a neural network with dropout (Hinton et al. 2012; Srivastava et al. 2014) applied before every weight layer is equivalent to an approximation of a deep Gaussian process (Damianou and Lawrence 2013), and that training with dropout effectively performs variational inference for the deep Gaussian process model. At test time, the posterior distribution can be sampled from by leaving dropout on. This gives us the network structure:

$$x_{i+1} = f_i(\text{dropout}(x_i)) \tag{1}$$

3.1 Randomized Embeddings

Computing an Embedding In the context of a trained dropout Bayesian neural network, we can use the intermediate representations from the various layers (the x_{i+1} in eq. (1)) as a randomized embedding of a data point. The embedding is randomized as multiple forward passes with dropout on will yield different embedding values. The variation in the embedding values could be used to measure epistemic uncertainty (Oh et al. 2019), allowing for the detection of OOD data and dataset shift.

Measuring Uncertainty A datum is embedded to a set of randomized embedding values at each layer. We can compute the maximum pairwise distance between the embeddings for a specific datum at a specific layer. This can be done at each layer in the BNN, giving us a feature for each layer that can then be used for tasks such as OOD identification. *All previous work has used Euclidean distance to compute the*

Algorithm 1: Computing Randomized Embedding Based Features for OOD Data Detection

Input: A datum x, a N layer NN trained with dropout $\{f_1, ..., f_N\}$, and number of samples T.
Output: N randomized embedding based features $z_1, ..., z_N$, each corresponding to a layer in the network, for a OOD data detection task.

for $t \leftarrow$ *1 to T* **do**
  for $i \leftarrow$ *1 to N* **do**
    $x_{i+1,t} \leftarrow f_i(\text{dropout}(x_{i,t}))$
for $i \leftarrow$ *1 to N* **do**
  $z_i \leftarrow \max(\text{PairwiseCosineDistances}(x_{i,:}))$
return $z_1, ..., z_N$ `// Return features.`

pairwise distances, without examining the appropriateness of Euclidean distance for the task. Part of our contribution is an analysis in section 3.3 of why Euclidean distance is in fact not appropriate, and we introduce a preferable cosine distance based approach which we use in all of our experiments. A small value of 1e-6 was added to the embeddings to avoid numerical issues caused by corner-case zero normed embedding vectors.[1] In our experiments, embeddings from non-linear layers (such as convolutions) are flattened prior to computing this metric. A summary of our approach can be found in algorithm 1. The intuition behind this approach is that if measured appropriately, the "spread" or maximal variation in a datum's embedding contains uncertainty information. If all embedding samples are realized to a same point in the embedding space, then there is less uncertainty than if the embedding samples are realized to wildly different parts of the embedding space.

3.2 Baseline Features

We compare the addition of our randomized embedding based features to a set of common baseline features. For classification tasks, uncertainty estimates in dropout BNNs are usually a function of the sampled softmax outputs. In particular, overall uncertainty can be measured using predictive distribution entropy: $H[\mathbb{P}(y|x, D)] = -\sum_{y \in C} \mathbb{P}(y|x, D) \log \mathbb{P}(y|x, D)$. To isolate and measure epistemic uncertainty mutual information can be used: $I(\theta, y|D, x) = H[\mathbb{P}(y|x, D)] - \mathbb{E}_{\mathbb{P}(\theta|D)} H[\mathbb{P}(y|x, \theta)]$.

The terms of these equations can be approximated using Monte Carlo estimates obtained by sampling from the dropout BNN posterior (Smith and Gal 2018). In particular, $\mathbb{P}(y|x, D) \approx \frac{1}{T} \sum_{i=1}^{T} \mathbb{P}(y|x, \theta_i)$ and $\mathbb{E}_{\mathbb{P}(\theta|D)} H[\mathbb{P}(y|x, \theta)] \approx \frac{1}{T} \sum_{i=1}^{T} H[\mathbb{P}(y|x, \theta_i)]$ where the θ_i are samples from the posterior over models and T is the number of samples. In addition to predictive distribution entropy and mutual information, we also use maximum softmax probability (the value of the largest element of $\mathbb{P}(y|x, D)$) as a feature, shown by Hendrycks and Gimpel (2017) to be an effective baseline for the OOD data detection task.

[1] We also note that normalized Euclidean distance, where embedding vectors are normalized to unit length prior to computing Euclidean distance, could also be used in place of cosine distance as its square can be shown to be proportional to cosine distance.

3.3 How to Measure Embedding Dispersion

We will now explore why Euclidean distance as used by previous works is not appropriate to measure randomized embedding dispersion. We illustrate using a LeNet5 (Yann LeCun et al. 1998) model with added dropout before each layer trained on MNIST, with MNIST variants as OOD data. Further data, model, and experimental details correspond to those expanded upon in section 4.1.

The Problem With Euclidean Distance Terhörst et al. (2020) suggest the Euclidean distance to measure when a data point is suitable for a downstream task, where lower variability in the stochastic embedding induced by a dropout neural network suggests higher suitability for a data point. In particular, they use the sigmoid of the negative mean Euclidean distance between all stochastic embedding pairs for a data point as the measure of suitability. In other words, their hypothesis is that a form of uncertainty can be measured using the Euclidean distance between embedding samples.

We find that if Euclidean distance is used as the metric to measure distance between samples, their hypothesis holds only with excessive training and likely over-fitting. fig. 1a shows that with enough training to get to the accuracy plateau (10 epochs of training with a batch size of 64, with a test accuracy of 0.9885), we actually see the opposite effect. Embeddings for OOD data are actually less spread out than embeddings for in distribution data. fig. 1b shows that with excessive training (100 epochs of training, with a lower test accuracy of 0.9882), we see that the hypothesis holds better but note that there is still a good amount of overlap between the histograms, limiting the usefulness for OOD detection (and adding a difficult to select stopping criteria). We note that what we are observing is *not* feature collapse.

This points to two issues that we need to resolve. First, how can we get consistent behavior regardless of over/under-training? Second, how can we more usefully measure spread in a way that matches intuition?

Spectral Normalization Stabilizes Behavior Spectral normalization rescales the weights during training with the spectral norm of the weight matrix, enforcing a Lipschitz constraint that bounds the derivative of the learned function (Miyato et al. 2018). This helps to preserve distance as a data point makes its way through the network. fig. 1c shows that a spectral normalized version of the network results in consistent behavior even with longer training (100 epochs of training, with a test accuracy of 0.9927). So, there is a solution to the first problem. However, we still see that the spread for OOD data is lower than for in distribution data.

Why Cosine Distance Is Needed To Properly Measure Embedding Dispersion Previous research around OOD detection has noted that a lower maximal softmax output value is correlated with a data point being OOD (Hendrycks and Gimpel 2017). One possible explanation could be logits (softmax inputs) of smaller norm. This would make intuitive sense as potentially, less neurons would activate for OOD

data since OOD data would lack the in distribution features the network is looking for.

The squared Euclidean distance between vectors $\mathbf{u}$ and $\mathbf{v}$ can be written as, where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$:

$$||\mathbf{u}-\mathbf{v}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 - 2\,||\mathbf{u}||\,||\mathbf{v}||\,\cos\theta \quad (2)$$

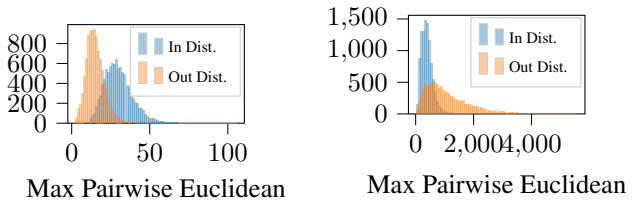
If embedding norms are inherently smaller for OOD data, then Euclidean distance which is norm dependent cannot be used to compare embedding spread across OOD and in distribution datasets, due to confounding. As shown in eq. (2), angular information is affected by norm in both an additive and multiplicative manner with Euclidean distance. So, assuming confounding caused by systematic norm differences, cosine distance should be used to isolate the angular information when measuring embedding dispersion. If Euclidean distance mostly captures information already captured by the norm, then the benefit of being Bayesian for this task is not fully leveraged as norm can be estimated with a single point estimate. *To take full advantage of a dropout BNN, angular information about embedding dispersion needs to be deconflated from embedding norm information.*

We explored this hypothesis and found it to be empirically true and formally justifiable. In fig. 2a, Euclidean distance is used to measure embedding dispersion, we see that dispersion is correlated with the logits norm and that the relationship is nearly identical for OOD and in distribution data. This means that measuring the spread of the embeddings using Euclidean distance conveys little extra information than just looking at the norm of the logits. In Appendix **??**, we perform a simulation to further illustrate this problem in the case of a two layer ReLU activated network.

We want to measure spread in a way that is independent of the embedding norm. This can be done a couple of different ways. For example, a simple switch to cosine distance could be used, or the embeddings could be normalized prior to using Euclidean distance (which can be shown to be related to cosine distance). As illustrated in fig. 2b, using cosine distance results in OOD and in distribution data having behaviors that are no longer identical. Appendix **??** shows similar results in an unsupervised setting involving a stacked denoising autoencoder variant.

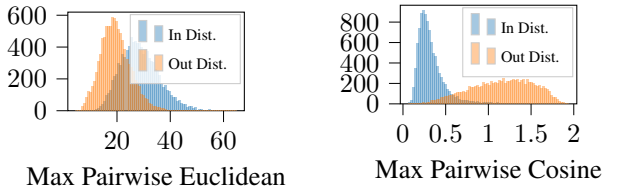
fig. 1d shows the same information as fig. 1a, except a cosine distance based measure of spread is used instead of a Euclidean based one. With cosine distance, we now see the expected behavior of OOD having more spread than in distribution, and we see a better separation as well which is good for OOD detection. We have shown results for the last layer of a network but note that a similar analysis can be done for each layer. Having shown empirical evidence for why angular information needs to be isolated from norm information when measuring embedding dispersion, we next provide a formal analysis for why cosine distance allows for an additional source of information.

Formal Analysis of Cosine Embedding Dispersion We aim to compute a metric that is invariant to the relative magnitudes among embedding samples, and also accurately represents the dispersion of the embedding samples. In the following, we argue that the mutual information score is



(a) LeNet5 with 10 epochs of training, Euclidean based measure of embedding dispersion.

(b) LeNet5 with 100 epochs of training, Euclidean based measure of embedding dispersion.



(c) Spectral Normalized LeNet5 with 100 epochs of training, Euclidean measure of dispersion.

(d) LeNet5 with 10 epochs of training, cosine based measure of embedding dispersion.

Figure 1: Comparison of last layer randomized embedding dispersion distributions for in distribution data (MNIST) and OOD data (Not-MNIST). The y-axis is "Count" in all cases.

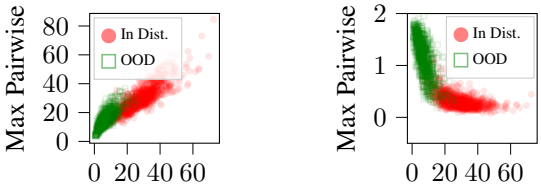
not satisfactory for these two objectives. Our goal is not to replace the mutual information as an uncertainty measure, but rather to demonstrate that our pairwise cosine similarity yields an additional source of information that is not captured otherwise.

Let $\{z_i\}_{i=1}^m$ denote m embedding vectors sampled through dropout. The mutual information score is defined as

$$I(w, y|D, x) = H[p(y|x, D)] - \mathbb{E}_{p(w|D)}H[p(y|x, w)]$$

and is approximated by $\hat{I}(w, y|D, x) = H\left[\frac{1}{m}\sum_{i=1}^m \text{softmax}(z_i)\right] - \frac{1}{m}\sum_{i=1}^m H\left[\text{softmax}(z_i)\right]$ where $H(\cdot)$ is the entropy function $H(y) = -\sum_i y_i \log y_i$.

We first introduce a theorem from Amos (2019) that clarifies the geometric properties of the softmax function. The proof is readily shown using Lagrange multipliers.



(a) Euclidean distance.

(b) Cosine distance.

Figure 2: A comparison of the relationships between last layer randomized embedding mean norm and the maximum pairwise distance for Euclidean and cosine distances respectively, for in distribution data (MNIST) and OOD data (Not-MNIST). Both models using LeNet5 trained for 10 epochs.

Theorem 3.1. *The softmax function* $\text{softmax}(x)_j = \frac{\exp(x_j)}{\sum_i \exp(x_i)}$ *is a map from* $\mathbb{R}^d$ *to the* $(d-1)$*-simplex that satisfies*

$$\text{softmax}(x) = \arg\min_{0<y<1} -x^\top y - H(y) \text{ s.t. } 1^\top y = 1$$

From this we see that the softmax solution is a balance between two competing objectives: maximizing $x^\top y$ which aims to place all weight on the coordinate with the largest x_i value, and maximizing the entropy of y which steers toward the uniform vector with value $1/d$. In addition, the softmax temperature changes the relative weighting, which allows us to evaluate the effect of the magnitude of the embedding vector. We leverage this for a further Lemma and Theorem:

Lemma 3.2. *The softmax function with temperature* α*, defined by* $\text{softmax}(x/\alpha)$*, satisfies*

$$\text{softmax}(x/\alpha) = \arg\max_{0<y<1} x^\top y + \alpha H(y) \text{ s.t. } 1^\top y = 1$$

Proof. From the previous theorem we get $\text{softmax}(x/\alpha) = \arg\min_{0<y<1} -(x/\alpha)^\top y - H(y) \text{ s.t. } 1^\top y = 1$. Multiplying by scalar α and switching the optimization to maximizing the negative does not change the optimal solution, yielding the statement above. □

These facts help indicate that softmax-based metrics are not suited for assessing the angular dispersion among vectors. We note that the mapped vector is α-dependent and hence dependent on the L_2 magnitude of the input vector. Furthermore, arbitrary translations of the vector, which can completely change the direction of the vector, do not impact the softmax. These observations are formalized below.

Theorem 3.3. *The softmax function is invariant to translation of input vector* x*. It is not invariant to scaling* x *except in the special case when* $x_1 = x_2 = \ldots = x_d$*. Furthermore, as the magnitude of* x *increases (without changing direction), the softmax shifts weight to the vertex of the simplex corresponding to the largest coordinate in* x*.*

Proof. Invariance to translation follows from observing that $\text{softmax}(x + K) = \exp(x_j + K)/\sum_i \exp(x_j + K) = \exp(x_j)/\sum_i \exp(x_j) = \text{softmax}(x)$.

The dependence on scaling follows from Lemma 1.2. Consider two vectors x, x' such that $x' = x/\alpha$. The value of α adjusts the scale of the $H(y)$ term. Since the $\max x^\top y$ objective aims to shift weight in y to the largest x coordinate and the $\max H(y)$ objective aims to distribute weight evenly, their solutions do not coincide, giving $\text{softmax}(x)$ and $\text{softmax}(x')$ different solutions. In the special case that $x_1 = x_2 = \ldots = x_d$ then $x^\top y$ is constant, so the optimization of $H(y)$ gives the uniform distribution vector. Otherwise, increasing the magnitude of x' is equivalent to sending $\alpha \to 0$, which decreases the contribution of $H(y)$. This causes the solution vector to shift weight to the element with largest value in x. □

We confirm this analysis by simulation in Appendix **??**, where we find that our new cosine-based feature adds an orthogonal measure of information that is not captured in previously used measures of uncertainty.

4 Experiments and Results

In this section, we evaluate the value of randomized embedding based features across three different OOD data detection tasks in the vision, language, and malware domains. All experiments were implemented in PyTorch (Paszke et al. 2019), and neural networks were optimized using Adam with the default recommended settings (Kingma and Ba 2015). A dropout probability of $p = 0.1$ was used, and when sampling from the base neural network models to compute features for OOD detection, 32 samples are used. Experiments were run on an 80 CPU core machine with 512GB of RAM using a single 16GB Tesla P100 GPU. Experiment specific details are described in their respective sections.

We explore the use of two model classes for the OOD detection algorithms. The first model is an L2-regularized logistic regression (LR) with the regularization strength chosen using 3-fold cross-validation. We min-max scaled the input features for the LR model to the range $[0, 1]$ based on the training data. The second model is a 500 tree random forest (RF) classifier. We choose these two models to assess linear vs. non-linear behavior in the OOD detection task. We also explore the effect of varied, small training set sizes for the OOD task in all of our experiments. In many production contexts such as cyber security, examples of OOD data are limited and usually expensive to obtain.

4.1 Image Classification

For our vision experiments, similarly to the evaluation protocol from (van Amersfoort et al. 2020; Ren et al. 2019; Postels et al. 2020; Mukhoti et al. 2021) we explore MNIST variants as OOD data. In particular, we train our base model, a LeNet5 (Yann LeCun et al. 1998) with added dropout before each layer, on MNIST and use Kuzushiji-MNIST (Clanuwat et al. 2018), notMNIST (Bulatov 2011), and Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017) as OOD data. When training the downstream OOD data detection algorithms, we train the OOD detector on one of the OOD datasets and test on the other two. For example, we first train a digit classifier on MNIST. Then, we train an OOD data detector that uses randomized embedding based features from the digit classifier to classify MNIST vs. notMNIST. Then we test the OOD data detector on MNIST vs. Kuzushiji-MNIST and Fashion-MNIST.

Due to its importance in practical use, we will test the sample efficiency of the OOD tasks (i.e., how few samples of OOD are needed to detect future OOD data). In particular, we evaluate performance, as measured by area under the receiver operating characteristic curve (captures desired data ordering performance) and accuracy (captures desired decision making value), using training datasets consisting of n=1000, 100, and just 10 data points from each class (in distribution and OOD). We note that this differs from most previous works which have evaluated by assuming access to a large OOD dataset of similar size to the in distribution dataset, an often unrealistic assumption. Each experiment was run 100 times with random training set samples, where all appropriate data not in the training set is included in the test set, and we report a mean and standard deviation for each. In all of our experiments, the

standard deviations are much smaller than effect sizes, so we report only the means in this section, and standard deviations can be found in appendix **??**.

Detecting OOD Data table 1 compares performance with and without the cosine embedding spread features for various experimental configurations and OOD detection models for a dropout LeNet5 trained for 100 epochs. Features labeled as "Last" consist of common baseline features computed using softmax output samples from the network (predictive entropy, mutual information, and maximum softmax probability). Features labeled as "Last+Spread" consist of these baseline features plus our additional randomized embedding maximum cosine spread features for each layer.

The inclusion of the additional cosine spread features improves OOD detection performance consistently across datasets, training set sizes, and model types. In limited cases where the "Spread" features do not improve the LR model, the RF model with "Spread" features performs the best overall, suggesting that the relationship is not necessarily linear. **??** in the Appendix summarizes results from a similar experiment where the base model is a spectral normalized dropout LeNet5 trained for 100 epochs. A comparison of table 1 and Appendix **??** suggests that, while spectral normalization is not required to see an improvement from the inclusion of cosine spread features, spectral normalization does improve OOD detection performance consistently.

In Appendix **??**, we further examine the need for a small amount of OOD training data, evaluate Euclidean based spread features, and investigate the feature importances associated with our cosine spread features.

Train/Test	Feat.	n=1000		n=100		n=10	
Fashion/	Last	97	**91**	97	91	96	88
Kuzushiji	+Spread	**98**	91	**97**	**91**	**97**	**90**
RF version	Last	96	92	95	91	94	88
	+Spread	**98**	**92**	**97**	**92**	**97**	**91**
Fashion/	Last	97	91	97	91	96	88
notMNIST	+Spread	**98**	**93**	**98**	**93**	**97**	**89**
RF version	Last	96	92	95	90	94	88
	+Spread	**99**	**94**	**98**	**92**	**96**	**90**
Kuzushiji/	Last	97	92	97	92	97	90
Fashion	+Spread	**99**	**95**	**98**	**94**	**98**	**92**
RF version	Last	96	92	96	91	95	90
	+Spread	**99**	**94**	**98**	**93**	**97**	**91**
Kuzushiji/	Last	97	91	97	91	96	89
notMNIST	+Spread	**98**	**93**	**98**	**91**	**97**	**89**
RF version	Last	96	92	95	90	94	89
	+Spread	**98**	**94**	**97**	**92**	**95**	**90**
notMNIST	Last	97	91	96	91	96	89
/Fashion	+Spread	**98**	**94**	**97**	**93**	**98**	**93**
RF version	Last	96	91	96	90	95	89
	+Spread	**99**	**94**	**98**	**94**	**98**	**92**
notMNIST/	Last	96	**90**	95	**89**	95	88
Kuzushiji	+Spread	**97**	89	**95**	89	**97**	**90**
RF version	Last	96	**91**	95	90	94	88
	+Spread	**98**	91	**97**	**92**	**97**	**91**

Table 1: Performance (AUC, accuracy) with and without the cosine randomized embedding spread features for various experimental configurations for a dropout LeNet5 trained on MNIST. Features labeled as "Last" consist of common baseline features computed using softmax output samples from the network (predictive entropy, mutual information, and maximum softmax probability). Features labeled as "Last+Spread" consist of these baseline features plus our additional randomized embedding maximum cosine spread features for each layer. Each experiment was repeated multiple times, and the mean is reported here while the standard deviation is reported in **??**. Best results are shown in bold.

4.2 Language Classification

Out of distribution data detection is also of interest in natural language processing, where systems are trained to work on specific languages, and inputs from other languages are considered OOD (Xiao, Gomez, and Gal 2020). For these experiments, we train a Char-CNN (Zhang, Zhao, and LeCun 2015) with dropout added before every layer to classify languages using the WiLI dataset (Thoma 2018). Training consisted of 50 epochs with a batch size of 128, where the 100 most common characters in the training set (after stripping accents) were used as the vocabulary and each datum was truncated/padded to a length of 200 characters. We train the language classification model to distinguish between French, Spanish, German, English, Italian, and Portuguese text. We use Basque, Polish, Luganda, Finnish, Tongan, and Xhosa as out of distribution languages. All of our in and out of distribution languages are chosen to use the Latin writing system. For the OOD task, training sets consisted of n=100, 50, 25, and 10 data points from each class (in distribution and OOD). Each experiment was run 100 times with random training data subsamples, where all languages not trained on are tested on. table 2 shows that the inclusion of our randomized embedding based features consistently improves OOD detection across experimental settings, with average and maximal AUC improvements of 0.06 and 0.15.

We note that while OOD data detection is usually treated as a purely binary classification task by most previous work, OOD versus in distribution is a false binary. There are different levels and degrees of how OOD data can be. In the context of language, we can examine the nuances between different flavors of OOD data. While Basque is a language isolate that linguistically does not share any significant similarities to any other languages, Catalan is a Romance language with many linguistic similarities to French and Italian (and Spanish to a lesser extent). While both Basque and Catalan are considered OOD in our setting, we expect good estimates of epistemic uncertainty to capture the property that Catalan is "less OOD" than Basque is. fig. 3 shows that this desired property is captured by the norm of our randomized embedding features, while the mutual information distributions for Basque and

Model	Ft	n=100		n=50		n=25		n=10	
Basque	Last	89	80	88	79	88	79	88	79
LR	+S	**93**	**84**	**92**	**84**	**92**	**84**	**93**	**83**
Basque	Last	86	80	86	79	85	79	84	79
RF	+S	**92**	**85**	**92**	**84**	**92**	**84**	**91**	**82**
Finnish	Last	89	**90**	89	79	88	79	88	79
LR	+S	**91**	82	**91**	**82**	**91**	**82**	**91**	**82**
Finnish	Last	86	79	86	79	85	79	84	78
RF	+S	**91**	**82**	**91**	**82**	**91**	**82**	**90**	**81**
Luganda	Last	89	81	89	80	89	80	88	79
LR	+S	**94**	**86**	**94**	**85**	**94**	**85**	**93**	**84**
Luganda	Last	87	80	86	80	85	80	84	79
RF	+S	**94**	**86**	**93**	**85**	**93**	**84**	**92**	**83**
Polish	Last	90	82	90	82	89	82	89	81
LR	+S	**94**	**87**	**94**	**86**	**94**	**86**	**93**	**85**
Polish	Last	87	79	86	79	85	78	85	78
RF	+S	**94**	**87**	**93**	**86**	**93**	**86**	**92**	**84**
Tongan	Last	86	**82**	84	81	82	80	79	77
LR	+S	**89**	81	**88**	81	**88**	**82**	**88**	**81**
Tongan	Last	77	70	77	70	77	68	79	70
RF	+S	**92**	**85**	**91**	**85**	**91**	**84**	**90**	**82**
Xhosa	Last	89	81	89	80	89	80	88	79
LR	+S	**94**	**87**	**94**	**86**	**93**	**85**	**93**	**84**
Xhosa	Last	86	79	86	78	85	78	85	77
RF	+S	**94**	**87**	**93**	**86**	**93**	**85**	**92**	**84**

Table 2: Performance (AUC, accuracy) with and without the cosine randomized embedding spread features for a Char-CNN with dropout added before every layer trained to classify languages using the WiLI dataset. Standard deviations are reported in **??**, and best results are shown in bold.

Catalan are nearly indistinguishable.

4.3 Malware Detection

Finally, we evaluate the usefulness of our randomized embedding based features in the context of malware detection. Uncovering new or significantly different malware is of particular interest in the quickly evolving cyber security space. We use a dropout variant of the MalConv model (Raff et al. 2017), a convolutional NN for malware detection that operates on raw byte sequences. We apply dropout before each fully connected layer of MalConv. Applying dropout to only the last layers of a NN corresponds to using maximum a posteriori (MAP) estimates for the initial layers and Bayesian estimates for the later layers (Gal and Ghahramani 2016a). We train the dropout MalConv model for 5 epochs with a batch size of 32 on the EMBER2018 dataset which consists of portable executable files (PE files) scanned by VirusTotal in or before 2018 (Anderson and Roth 2018).

We run two experiments on the Bayesian MalConv model. First, of the 200000 files in the EMBER test set, 363 have

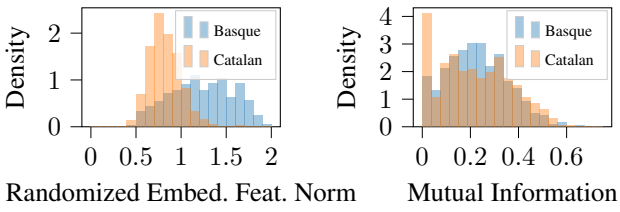


Figure 3: Basque and Catalan are linguistically similar but different languages. Our cosine based embeddings (left) show that they have high overlap but are more OOD than normal data. Prior work using MI (right) is unable to meaningfully distinguish any difference between the related languages.

Model	Feat.	n=100		n=50		n=25	
Ember	Last	78.9	70.4	**78.6**	68.2	**77.8**	65.0
LR	+Sprd	**79.3**	**71.8**	78.3	**68.9**	76.6	**65.8**
Ember	Last	75.7	73.5	75.2	72.7	74.8	71.4
RF	+Sprd	**79.1**	**78.4**	**78.2**	**76.4**	**77.0**	**74.3**
Brazil	Last	68.5	**64.5**	68.0	60.7	66.8	58.4
LR	+Sprd	**74.1**	62.0	**73.4**	**61.7**	**71.2**	**60.5**
Brazil	Last	72.4	69.3	70.5	67.4	67.9	65.2
RF	+Sprd	**83.9**	**79.7**	**81.3**	**77.2**	**77.6**	**73.6**

Table 3: Performance with and without the cosine randomized embedding spread features for a MalConv model with dropout added before each fully connected layer trained to detect malware using EMBER2018. Standard devs. are reported in **??**, and best results are bolded.

as their top most likely malware family label (as labeled by AVClass (Sebastián et al. 2016)) a family that was not present in the train set. We evaluate OOD detection performance first on these unseen malware families. Second, we evaluate OOD detection performance on a different malware dataset containing malware samples obtained from a Brazilian financial entity (Ceschin et al. 2019). The malware from this dataset could be considered as OOD due to differing geographical specificity and intent, leading to the use of malware tactics, techniques, and procedures likely specific to a Brazilian banking target. There are also temporal differences as the Brazilian samples were all collected before the EMBER dataset, and we additionally only used malware first seen by VirusTotal before 2012. OOD task training sets consisted of n=100, 50, and just 25 data points from each class (in distribution and OOD). Each experiment was run 100 times with random train/test splits, where all of the data not in the training set is included in the test set. Results are summarized in table 3, showing that the inclusion of our randomized embedding based features consistently improves OOD detection across experimental settings. Because of the high class imbalance in this use case, as access to good OOD data is more limited in the malware domain, we reported the ROC AUC and the recall for the OOD class in table 3, noting that recall is often

the primary metric of interest in practice for cyber security.

5 Conclusions

We have demonstrated why previous attempts at measuring randomized embedding dispersion using Euclidean distance are inherently flawed. Then we introduced and theoretically justified a cosine distance based, lightweight approach to test time OOD data detection in the context of dropout Bayesian neural networks. Information that is already computed is used as randomized embeddings, training dataset information does not need to be stored, additional regularization methods are not needed (though do help), and auxiliary neural networks do not need to be trained to take advantage of this additional information. While we note that our approach is limited to dropout BNNs, the popularity of the dropout approximation to BNNs and the existence of previous works exploring the use of stochastic embeddings based on dropout BNNs suggests the applicability of our approach to practice. Our approach can be deployed anywhere a dropout BNN is already deployed with minimal additional overhead. Future work includes the investigation of more elaborate features based off of the randomized embeddings.

References

Amos, B. 2019. *Differentiable optimization-based modeling for machine learning*. Ph.D. thesis, Carnegie Mellon University.

Anderson, H. S.; and Roth, P. 2018. EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models. *arXiv preprint arXiv:1804.04637*.

Bartlett, P. L.; Evans, S. N.; and Long, P. M. 2018. Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. *arXiv preprint arXiv:1804.05012*.

Bulatov, Y. 2011. notMNIST dataset. http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html. Machine Learning, etc. Accessed: 2022-05-05.

Ceschin, F.; Pinage, F.; Castilho, M.; Menotti, D.; Oliveira, L. S.; and Gregio, A. 2019. The Need for Speed: An Analysis of Brazilian Malware Classifers. *IEEE Security and Privacy*, 16(6): 31–41.

Chang, J.; Lan, Z.; Cheng, C.; and Wei, Y. 2020. Data Uncertainty Learning in Face Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5710–5719.

Chun, S.; Oh, S. J.; de Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8415–8424.

Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; and Ha, D. 2018. Deep Learning for Classical Japanese Literature. *arXiv preprint arXiv:1812.01718*.

Damianou, A. C.; and Lawrence, N. D. 2013. Deep Gaussian Processes. *Artificial intelligence and statistics, PMLR*, 31: 207–215.

Gal, Y.; and Ghahramani, Z. 2016a. Dropout as a Bayesian Approximation: Appendix. *33rd International Conference on Machine Learning, ICML 2016*, 3: 1661–1680.

Gal, Y.; and Ghahramani, Z. 2016b. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *33rd International Conference on Machine Learning, ICML 2016*, 3: 1651–1660.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *International Conference for Learning Representations*, 1–12.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 1–18.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 5575–5585.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *International Conference for Learning Representations*, 1–15.

Liu, J. Z.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax-Weiss, T.; and Lakshminarayanan, B. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 7498–7512.

Mandelbaum, A.; and Weinshall, D. 2017. Distance-based Confidence Score for Neural Network Classifiers. *arXiv preprint arXiv:1709.09844*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. *International Conference on Learning Representations*.

Mukhoti, J.; Kirsch, A.; van Amersfoort, J.; Torr, P. H. S.; and Gal, Y. 2021. Deterministic Neural Networks with Appropriate Inductive Biases Capture Epistemic and Aleatoric Uncertainty. *arXiv e-prints*.

Oh, S. J.; Murphy, K.; Pan, J.; Roth, J.; Schroff, F.; and Gallagher, A. 2019. Modeling Uncertainty with Hedged Instance Embedding. *International Conference on Learning Representations*.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; and Snoek, J. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *NeurIPS*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Postels, J.; Blum, H.; Strümpler, Y.; Cadena, C.; Siegwart, R.; Van Gool, L.; and Tombari, F. 2020. The Hidden Uncertainty in a Neural Networks Activations. *arXiv preprint arXiv:2012.03082*.

Raff, E.; Barker, J.; Sylvester, J.; Brandon, R.; Catanzaro, B.; and Nicholas, C. 2017. Malware Detection by Eating a Whole EXE. *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Raff, E.; and Nicholas, C. 2020. A Survey of Machine Learning Methods and Challenges for Windows Malware Classification. *arXiv preprint arXiv:2006.09271*, 1–48.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; DePristo, M. A.; Dillon, J. V.; and Lakshminarayanan, B. 2019. Likelihood Ratios for Out-of-Distribution Detection. *Advances in Neural Information Processing Systems*.

Sebastián, M.; Rivera, R.; Kotzias, P.; and Caballero, J. 2016. AVCLASS: A Tool for Massive Malware Labeling. In *International symposium on research in attacks, intrusions, and defenses.*

Shi, Y.; and Jain, A. K. 2019. Probabilistic Face Embeddings. *IEEE/CVF International Conference on Computer Vision*, 6902–6911.

Smith, L.; and Gal, Y. 2018. Understanding measures of uncertainty for adversarial example detection. *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2: 560–569.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15: 1929–1958.

Terhörst, P.; Niklas Kolf, J.; Damer, N.; Kirchbuchner, F.; and Kuijper, A. 2020. SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5651–5660.

Thoma, M. 2018. The WiLI benchmark dataset for written language identification. *arXiv preprint arXiv:1801.07779*.

van Amersfoort, J.; Smith, L.; Teh, Y. W.; and Gal, Y. 2020. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. *International conference on machine learning, PMLR*, 9690–9700.

Wilson, A. G.; and Izmailov, P. 2020. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. *Advances in neural information processing systems*, 4697–4708.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*.

Xiao, T. Z.; Gomez, A. N.; and Gal, Y. 2020. Wat zei je? Detecting Out-of-Distribution Translations with Variational Transformers. *arXiv preprint arXiv:2006.08344*.

Yann LeCun; Léeon Bottou; Yoshua Bengio; and Patrick Haffner. 1998. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 2278–2324.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. *Advances in neural information processing systems*.