Improved Robustness to Open Set Inputs via Tempered Mixup

Ryne Roady[1]   Tyler L. Hayes[1]   Christopher Kanan[1,2,3]

[1] Rochester Institute of Technology, Rochester, NY 14623 USA
[2] Paige, New York, NY 10044 USA
[3] Cornell Tech, New York, NY 10044 USA
{rpr3697,tlh6792,kanan}@rit.edu

Abstract. Supervised classification methods often assume that evaluation data is drawn from the same distribution as training data and that all classes are present for training. However, real-world classifiers must handle inputs that are far from the training distribution including samples from unknown classes. Open set robustness refers to the ability to properly label samples from previously unseen categories as novel and avoid high-confidence, incorrect predictions. Existing approaches have focused on either novel inference methods, unique training architectures, or supplementing the training data with additional background samples. Here, we propose a simple regularization technique easily applied to existing convolutional neural network architectures that improves open set robustness without a background dataset. Our method achieves state-of-the-art results on open set classification baselines and easily scales to large-scale open set classification problems.

1 Introduction

Modern supervised classification methods often assume train and test data are drawn from the same distribution and all classes in the test set are present for training. However, deployed models will undoubtedly be exposed to out-of-distribution inputs that do not resemble training samples and these models are expected to robustly handle these novel samples. Performance in this 'open-world' setting is often hidden by current computer vision benchmarks in which the train and test sets have the same classes and the data is sampled from the same underlying sources. One solution to this problem is to develop open set classifiers which have the ability to identify novel inputs that do not belong to any training classes so that they are not assigned an incorrect label [23]. This capability is especially important for the development of safety-critical systems (e.g., medical applications, self-driving cars) and lifelong learning agents that automatically learn during deployment [19].

There are two major paradigms for enabling open set classification in existing convolutional neural network (CNN) architectures. The first paradigm replaces the standard closed-set prediction scheme with a new inference mechanism [2,17,16,24]. The second paradigm is to regularize the classification model
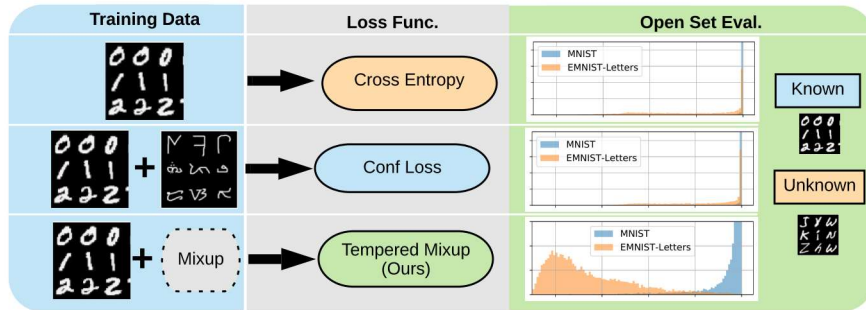
Fig. 1: TEMPERED MIXUP. Traditional methods for building model robustness to unknown classes involves training with a background set, which is problematic for large-scale datasets. Tempered Mixup provides state-of-the-art robustness to novel inputs without the requirement of a representative background set. The model confidence of a LeNet++ model trained to classify MNIST digits is shown for both known and unknown classes here represented by the Extended-MNIST-Letters dataset. Confidence is calculated using only the softmax of the model output and does not rely on a computationally expensive open set inference method to separate known from unknown classes.

during training to enable it to better separate known classes from potential unknowns [15,8]. For the latter approach, the most effective methods involve training with a large set of background images to penalize an overconfident prediction when a sample from a novel or unknown class is encountered. This approach has been shown to excel at open set classification, but existing results have been limited to small-scale datasets. In part, this is because it is increasingly difficult to construct an effective background dataset that does not semantically overlap with a large-scale training dataset.

In this paper, we overcome this limitation by proposing a new training approach that penalizes overconfident predictions on samples outside of the known training classes without access to a set of background/unknown inputs. Instead our method hallucinates invalid images using a novel form of the Mixup data augmentation technique [29] that we combine with a unique auxiliary loss function. **This paper makes the following contributions:**

1. We propose a novel end-to-end training algorithm for regularizing existing CNN architectures for open set classification that does not require the use of an explicit background dataset.
2. We propose a new loss function specifically designed to train a model to be less confident towards samples from unknown classes.
3. We show that our method exceeds methods that use explicit background datasets on standard small-scale benchmark settings for open set classification (e.g., MNIST and CIFAR).

4. We demonstrate that our method can easily scale to open set classification on large-scale datasets and produce comparable results to background set regularization without having to build an additional dataset for training.

2 Background

2.1 Open Set Classification

The goal of open set classification is to explicitly label points that are 'far' from known training classes as an unknown class instead of arbitrarily assigning one of the known categories. Open set classification methods discriminate among K known categories seen during training and reject samples that do not belong to the known categories. Formally, given a training set $D_{train} = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, where X_i is the i-th training input tensor and $y_i \in C_{train} = \{1, 2, \ldots, K\}$ is its corresponding class label, the goal is to learn a classifier $F(X) = (f_1, ..., f_k)$, that correctly identifies the label of a known class and separates known from unknown examples:

$$\hat{y} = \begin{cases} \operatorname{argmax}_k F(X) \text{ if } S(X) \geq \delta \\ K+1 \qquad\qquad \text{ if } S(X) < \delta \end{cases} \tag{1}$$

where $S(X)$ is an acceptance score function that determines whether the input belongs to the training data distribution, δ is a user-defined threshold, and $K + 1$ indicates 'unknown class.' For proper open set testing, the evaluation set contains samples from both the set of classes seen during training and additional unseen classes, i.e., $D_{test} = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, where $y_i \in (C_{train} \bigcup C_{unk})$ and C_{unk} contains classes that are not observed during training.

2.2 Inference Methods for Open Set Classification

Inference methods incorporate open set classification abilities into a pre-trained CNN by creating a unique acceptance score function and threshold that rejects novel inputs [23,2,3]. Current state-of-the-art inference methods often rely on features from multiple layers of the CNN [16,1] and many methods use multiple forward and backward passes through the CNN to improve performance [17,16]. These approaches significantly increase computational and memory requirements during inference [21], which may be sub-optimal for deployed models. We instead focus on building better open set classification performance through model regularization so that, during inference time, a much simpler and computationally efficient method such as confidence thresholding can be used to detect unknown classes.

2.3 Confidence Loss Training for Open Set Classification

One approach for improving open set robustness is training with a confidence loss penalty [15,8], which improves detection of unknown classes by penalizing

overconfident predictions on samples that are outside the training classes. As previously observed [8,21], inputs from unknown classes tend to be centered around the origin and have a smaller magnitude than samples from known classes in the deep feature space of a well-trained CNN .

Given an effective background dataset that is representative of unknowns expected to be seen during deployment, confidence loss training [8] collapses novel samples toward the origin of the deep feature space, resulting in lower confidence model outputs. The regularization penalty is a simple addition to standard cross-entropy loss during training, i.e.,

$$\mathcal{L}_{conf} = \begin{cases} -\log S_k(F(X)) & \text{if } X \in D_{train} \\ -\frac{1}{K} \sum_{k=1}^{K} \log S_k(F(X)) & \text{if } X \in D_{bkg} \end{cases} \qquad (2)$$

where the first term is a standard cross-entropy loss for known classes (D_{train}) and the second loss term forces the model to push samples from a representative background class (D_{bkg}) toward a uniform posterior distribution.

Multiple approaches have been developed for producing open set images including using alternative datasets which are distinct from the training set [8,20] and even using generative methods to produce images outside of the manifold defined by the training set [15,18]. Naturally, the performance of the open set classifier trained with open set images is tied to the "representativeness" of the background dataset and its similarity to the known training set. The difficulty for even small-scale classification is that the effectiveness of certain background sets cannot be known a priori and seemingly representative background datasets can sometimes fail to produce better open set robustness. We demonstrate this phenomenon using the CIFAR-10 dataset in Fig. 2. To overcome this limitation, we designed
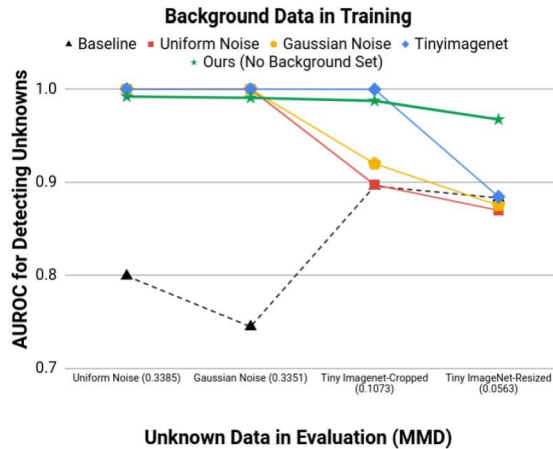

Fig. 2: Novelty Detection vs Background Class Selection The performance of confidence loss for improving open set performance is dependent on the selected background class. As the background class becomes more dissimilar from the in-distribution data, it becomes less useful for improving model performance on difficult open set classification problems. Our method does not require a background class for training and thus is robust across a wide-range of different input types.

our method around drawing novel samples for confidence loss using data augmentation instead of relying on an explicit background set.

2.4   Mixup Training

Mixup is a regularization approach that combines two separate images from the training set into a single example by forming an elementwise convex combination of the two input samples [29]. Mixup can improve model accuracy [29], model calibration [26], and model robustness to certain types of image corruptions [5]. However, Mixup has not been shown to be beneficial for open set classification, and in [5] Mixup resulted in a 50% reduction in detecting unknown classes versus baseline cross-entropy training.

Mixup is based on the principle of Vicinal Risk Minimization (VRM) [4] where a classification model is trained not only on the direct samples in the training set, but also in the *vicinity* of each training sample to better sample the training distribution. In Mixup, these vicinal training samples $(\tilde{x}, \tilde{y})$ are generated as a simple convex combination of two randomly selected input samples, x_i and x_j:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \tag{3}$$

where y_i and y_j are the associated targets for the selected input samples. The linear interpolation factor $\lambda \in [0, 1]$ is drawn from a symmetric $Beta(\alpha, \alpha)$ distribution where the shape of the distribution is determined by the hyper-parameter, α which trades off training with mostly unmixed examples versus training with averaged inputs and labels. By training with standard cross-entropy loss on these vicinal examples, the model learns to vary the strength of its output between class manifolds. The effect of this training is a substantial improvement in model calibration and accuracy on large-scale image classification tasks [30,26,5]. Our approach uses a variant of Mixup to overcome the need for an explicit background set with confidence loss training.

3   Mixup and Open Set Classification

3.1   Re-balancing Class Targets

As originally proposed, Mixup is a form of model regularization that trains on linear combinations of inputs and targets, and encourages a model to learn smoother decision boundaries between class manifolds. While smoother decision boundaries promote better generalization, they also benefit open set classification. This is because smoother boundaries reduce the likelihood of a model producing a confident but wrong prediction with an input that does not lie on the class manifolds learned from the training set. Mixup essentially turns a single label classification problem into a multi-label problem by creating additional samples through the linear mixture of inputs and feature space embeddings. The question then becomes: is a linear combination of the targets appropriate for a linear combination of features in training a model to produce accurate uncertainty estimates in the space between class manifolds?
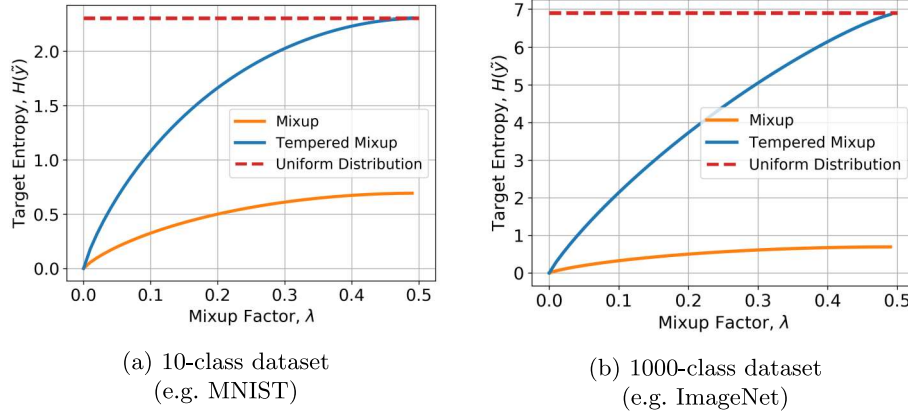
(a) 10-class dataset
(e.g. MNIST)

(b) 1000-class dataset
(e.g. ImageNet)

Fig. 3: TARGET REBALANCING: Ideally, for unknown inputs a classifier's output probabilities will approach a uniform distribution. Using Mixup, samples that are more mixed should have higher entropy, but this does not occur to the extent necessary in the original formulation. Instead, Tempered Mixup uses a novel formulation that ensures the target entropy for mixed examples approaches that of a uniform distribution. This figure demonstrates this in both the 10 class and 500 class setting.

We answer this question by looking at how the target entropy of a model trained via cross-entropy loss changes as a function of the mixing factor λ. As shown in Fig. 3, using a linear combination of targets, Eq. 3, to mix the labels does not capture the increase in uncertainty that we desire for examples that are off of the class manifold, e.g., the highly mixed examples. Instead we can re-balance the target labels with an additional label smoothing term modulated by the interpolation factor, λ, with an adjusted target mixing scheme as follows:

$$\tilde{y}_{tm} = |2\lambda - 1|\tilde{y} + \frac{1 - |2\lambda - 1|}{K} \ , \qquad (4)$$

where $\tilde{y}$ is the normal linear mixing from Eq. 3 and K is the number of known target classes. Using this novel re-balancing approach, we can temper the model confidence for highly mixed samples such that they approach a uniform distribution prediction. As shown in Fig. 3, this approach assigns a much higher target entropy to highly mixed up samples, including when they are mixed from the same class. This tempering effect is magnified as the number of known classes increases.

3.2 Tempered Mixup

Tempered Mixup is an open set classification training method that overcomes the need for a background training set by using a modified form of Mixup with a novel variant of confidence loss regularization. Using Mixup enables the creation

of *off-manifold* samples based on the training input distribution, and it enables control over how similar the simulated outliers are to the known classes. This allows the CNN to learn features that are robust to open set classes through a number of different specialized inference methods, including baseline confidence thresholding.

Instead of training with standard cross-entropy (softmax) loss with labels drawn from the convex combination of two randomly selected images as prescribed in the standard Mixup algorithm, Tempered Mixup uses the same mixed up input but a modified version of the auxiliary confidence loss function to regularize how the model maps these between-class inputs in deep feature space. To do this, we apply the Mixup coefficient drawn per sample from a symmetric Beta distribution to a modified confidence loss equation. This allows us to simultaneously minimize the loss for misclassifying samples from the known classes and map unknown samples that are far from the known classes to the origin. The Tempered Mixup loss is given by:

$$\mathcal{L}_{TempMix} = -|2\lambda - 1| \sum_{k=1}^{K} \tilde{y}_k \log \sigma_S(F(\tilde{X}))_k - \zeta \frac{1 - |2\lambda - 1|}{K} \sum_{k=1}^{K} \log \sigma_S(F(\tilde{X}))_k,$$

(5)

where σ_S is the softmax function applied to the vector $F(\tilde{X})$, λ is the sampled mixing interpolation factor, $\tilde{y}$ denotes the linearly mixed targets, K is the number of known classes, and ζ weights the amount of confidence loss applied to highly mixed up samples. Tempered Mixup is a straight-forward extension of traditional mini-batch stochastic gradient descent with cross-entropy loss training for deep neural network models.

3.3 Visualizations of Deep CNN Feature Space to Unknowns

To visually illustrate the benefit of Tempered Mixup in separating known and unknown samples, we trained a simple CNN model (LeNet++ architecture [14]) to classify MNIST digits as known classes and Extended MNIST Letters as unknown classes (overlaid as black points). The CNN architecture has a bottlenecked two dimensional feature space to allow the visualization of the resulting embeddings.

As shown in Fig. 4, the Tempered Mixup model collapses the embedding of samples from the unknown classes towards the origin, thus reducing the overlap (and confusion with) known classes. This is a dramatic improvement over common supervised training methods that improve model robustness including Label Smoothing [25] and Center Loss [27]. Tempered Mixup even improves on methods trained with an explicit background set such as Entropic Open Set and Objectosphere [8].

4   Experiments

To evaluate the open set robustness of our method against current state-of-the-art techniques, we first compare against baselines established using the MNIST [14]

(a) Cross-Entropy   (b) Label Smoothing   (c) One-vs-Rest

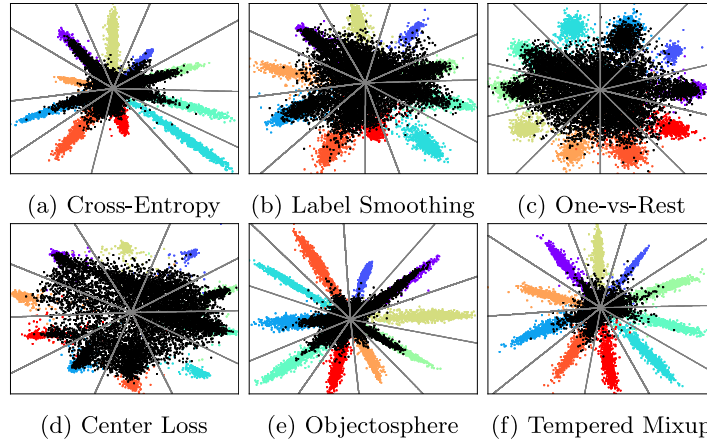(d) Center Loss   (e) Objectosphere   (f) Tempered Mixup

Fig. 4: 2-D visualization of the effect of the different feature space regularization strategies on separating in-distribution and outlier inputs. The in-distribution training set is MNIST, while the unknown samples are from the **Extended-MNIST-Letters** dataset [6]. For Entropic Open Set and Objectosphere, the Omniglot dataset is used as a source for background samples.

and CIFAR-10 [12] datasets as known classes and samples drawn from similar, but distinct, datasets as unknown classes. We then compare the performance of Tempered Mixup to standard Mixup training and other forms of VRM data augmentation. Finally, we extend our small-scale experiments to show how our method scales to large-scale open set classification problems.

4.1   Open Set Performance Assessment

An open set classifier needs to correctly classify samples from known classes and identify samples from unknown classes. This makes evaluation more complex than out-of-distribution detection, which simplifies the detection task to a binary in/out classification problem.

Our primary metric is the area under the open set classification (AUOSC) curve. It measures the correct classification rate among known classes versus the false positive rate for accepting an open set sample, and has been used as a standard metric for open set classification [8]. The correct classification rate can be viewed as the difference between normal model accuracy and the false negative rate for rejection. Intuitively, AUOSC takes into account whether true positive samples are actually classified as the correct class and thus rewards methods which reject incorrectly classified positive samples before rejecting samples that are correctly classified. In addition to reporting the area under the curve, we also calculate the correct classification rate at a specific false positive rate of 10^{-1}. Finally, we also report the area under the Receiver Operating Characteristic (AU-

ROC) for identifying unknown classes as a measure of pure out-of-distribution detection performance.

4.2   Comparison Methods

We compare our method against the following approaches:

1. **Cross-entropy**: As a baseline, we train each network with standard cross-entropy loss to represent a common feature space for CNN-based models.
2. **One-vs-Rest** [24]: The one-vs-rest training strategy was implemented by substituting a sigmoid activation layer for the typical softmax activation and using a binary cross-entropy loss function. In this paradigm, every image is a negative example for every category it is not assigned to. This creates a much larger number of negative training examples for each class than positive examples.
3. **Label Smoothing** [25]: By smoothing target predictions during cross-entropy training, the model learns to regularize overconfident predictions and produce less confident and more calibrated predictions.
4. **CenterLoss** [27]: A form of model regularization that increases the robustness of the class-conditional feature representation by encouraging tightly grouped class clusters. This is achieved by penalizing the Euclidean distance between samples and their class-mean (inter-class variance). By reducing the inter-class variance, a more precise rejection threshold can theoretically be established for separating known from unknown samples.
5. **Entropic Open Set** [8]: The Entropic Open Set method applies the confidence loss formulation (Eq. 2) using a background class to train the model to reduce model confidence on unknown classes.
6. **Objectosphere** [8]: Objectosphere separates known from unknown classes by training with a background class and reducing the magnitude of learned features for unknown classes. Instead of using the confidence loss formulation, it uses a margin based hinge-loss centered around the origin in deep feature space. This loss reduces the magnitude of features from the background set and increases the magnitude of features for samples from known classes to be larger than a user defined margin. This method was shown to better separate known from unknown samples in network architectures with higher dimensional feature spaces.

For all methods, identification of unknown classes is done by thresholding the maximum class posterior found by passing the model's output through a softmax activation.

4.3   Open Set Baselines

We first study Tempered Mixup on common open set benchmarks. Following the protocol established in [8], the first baseline uses the LeNet++ CNN architecture [14] with MNIST [14] for known classes and a subset of the Extended-MNIST-Letters dataset [6] for unknown classes.

Table 1: Open-Set Classification Baselines. The correct classification rate at a false positive rate for open set classification of 10^{-1} and the areas under the resulting OSC and ROC curves. In these experiments, the positive class used samples from the known set of classes seen during training and the negative class used samples from the unknown classes. Best performance for each experiment and metric is in **bold**.

Experiment	Algorithm	CCR @ FPR 10^{-1}	AUOSC	AUROC
MNIST Unknown: EMNIST-Letters Arch: LeNet++	Baseline	0.7259	0.9066	0.9103
	One-vs-rest	0.9556	0.9654	0.9814
	Label Smoothing	0.8543	0.9315	0.9443
	CenterLoss	0.9633	0.9695	0.9877
	Entropic Open-Set	0.9712	0.9797	0.9892
	Objectosphere	0.9570	0.9739	0.9801
	Tempered Mixup (Ours)	**0.9761**	**0.9821**	**0.9924**
CIFAR-10 Unknown: TinyImageNet Arch: Pre-ResNet-32	Baseline	0.5211	0.7694	0.8105
	One-vs-rest	0.2363	0.7064	0.7559
	Label Smoothing	0.0920	0.6841	0.7283
	CenterLoss	0.4930	0.7613	0.8038
	Entropic Open-Set	0.6766	0.7880	0.8344
	Objectosphere	0.6720	0.8045	**0.8584**
	Tempered Mixup (Ours)	**0.6923**	**0.8099**	0.8503

The second benchmark uses a 32-layer Pre-Activation ResNet architecture [10] with CIFAR-10 [12] as known classes and 178 classes from TinyImageNet [13] as unknown classes. TinyImageNet images were all resized to 32×32 images. We removed 22 classes from the original TinyImageNet dataset because they contained semantic overlap with CIFAR-10 classes based on hypernym or hyponyms, which were determined using the Wordnet lexical database [9].

For both benchmarks, we train all open set classification methods on the known classes. Methods that use a background class (i.e., Entropic Open Set and Objectosphere) are additionally trained on the first 13 classes in the Extended-MNIST-Letters dataset. For the CIFAR-10 baseline, a background class training set is drawn from non-overlapping classes in the CIFAR-100 dataset. All methods are then evaluated on an even split of samples drawn from known and unknown classes. For MNIST, 10000 samples are used as a source of unknowns drawn from the final 13 classes in the Extended-MNIST-Letters dataset. For CIFAR-10, 10000 samples are randomly selected from the TinyImageNet dataset as a source of unknowns.

Results. Tempered Mixup achieves state-of-the-art-results in open set classification without the use of an explicit background class (see Table 1). For the MNIST experiment, all methods show an improvement over the cross-entropy baseline. Tempered Mixup surpasses cross-entropy by 8% in terms of AUOSC and even surpasses confidence loss methods that use a background set that is
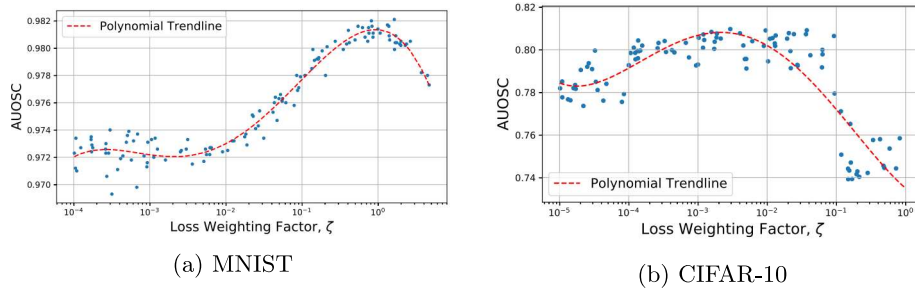
(a) MNIST

(b) CIFAR-10

Fig. 5: EFFECT OF TEMPERED LABELS: The effect of the tempering label smoothing loss term is varied to demonstrate the overall benefit on MNIST (left) and CIFAR-10 (right) Open Set Robustness.

derived from the same dataset as the unknowns. For CIFAR-10, which uses a modern ResNet v2 architecture, Tempered Mixup achieves more than a 5% improvement in terms of AUOSC over all methods that do not require a background set for training and is state-of-the-art in terms of AUOSC over all evaluated methods including those that train with an additional background set.

4.4 Additional Evaluations

Performance Improvement from Target Rebalancing. We seek to understand the benefit from target rebalancing towards improving the open set robustness gained from normal Mixup training. To model this effect, we varied the weight applied to the label smoothing term using the loss weighting factor (ζ in Eq. 5). This allowed us to see the performance difference as the confidence loss portion of our formulation is emphasized over standard Mixup training (which is equivalent as $\zeta \to 0$). As Fig. 5 shows, performance improves as we increase the weight of the confidence loss until a point when the confidence loss prevents the model from achieving a high closed set accuracy, thus reducing the overall open set performance.

Alternate Data Augmentation Schemes. Mixup is part of a family of data augmentation approaches which work on the VRM principle. As an alternative to the standard Empirical Risk Minimization formulation, VRM attempts to enlarge the support of the empirical training distribution by creating *virtual* examples through various data augmentation schemes. Other VRM data augmentation schemes have been proposed recently which have demonstrated increased robustness to certain forms of input corruption [5]; however, few have explicitly tested for open set robustness.

The Cutmix strategy [28] overlays a random path from a separate training image and adjusts the target labels during training based on the ratio of the patch area to the original image. Cutout [7] is a similar variation, however the

Table 2: ALTERNATE VRM COMPARISON. AUOSC performance of our method versus other VRM data augmentation techniques and their tempered variants. Experiments use either MNIST or CIFAR-10 as known dataset and three different unknown datasets that vary in similarity to the known dataset.

Experiment	Algorithm	Gaussian Noise	FMNIST	EMNIST-Letters
	Baseline	**0.9878**	0.9848	0.9066
	Cutmix	0.8837	0.8751	0.8172
MNIST	Cutout	0.9830	0.9813	0.9028
	Mixup	0.9874	**0.9875**	0.9737
	Tempered Cutmix	0.9805	0.9780	0.9249
	Tempered Cutout	0.9844	0.9829	0.9121
	Tempered Mixup (Ours)	0.9846	**0.9875**	**0.9821**
Experiment	Algorithm	SVHN	LSUN	Tiny ImageNet
	Baseline	0.8271	0.7934	0.7694
	Cutmix	0.6249	0.7956	0.7697
CIFAR-10	Cutout	0.8174	0.7425	0.7414
	Mixup	0.8193	0.7966	0.7886
	Tempered Cutmix	0.7274	0.7803	0.7572
	Tempered Cutout	0.7783	0.6963	0.7150
	Tempered Mixup (Ours)	**0.8340**	**0.8062**	**0.8099**

patch is made up of black (zero-valued) pixels. We evaluated our Tempered Mixup formulation against these competing VRM schemes both in their normal formulation and with a tempered target label set where the entropy of the target distribution is adjusted based on the interpolation factor, λ. As results show in Table 2, Tempered Mixup is superior to both Cutmix and Cutout and their tempered variants.

4.5   Large-Scale Open Set Classification

Deployed systems typically operate on images with far higher resolution and many more categories than the open set baselines previously established for model regularization techniques. It is necessary to understand how well these systems work with higher resolution images and when the number of categories exceeds 100. As the number of categories increases, it can become increasingly difficult to identify a suitable set of background images for background regularization methods.

To study open set classification for large-scale problems, we use the ImageNet Large-scale Visual Recognition Challenge 2012 dataset (ImageNet) [22]. ImageNet has 1.28 million training images (732-1300 per class) and 50000 labeled validation images (50 per class), which we use for evaluation. All methods use an 18-layer ResNet CNN for classification [11] with an input image resolution of 224×224. In our experimental setup, the known set of classes consists of 500 classes from ImageNet. Following [2], unknown images for open set evaluation are drawn from categories of the 2010 ImageNet challenge that were not

subsequently used and do not have semantic overlap with the 2012 ImageNet dataset. In total the open set dataset consisted of 16950 images drawn from the 339 categories.

We compare Tempered Mixup against the baseline cross-entropy method, Objectosphere, and a combination that incorporates both Objectosphere and Tempered Mixup training. Objectosphere was chosen because it is the best method for using a background dataset based on our previous experiments. To our knowledge, Objectosphere has not been previously evaluated on large-scale problems, thus for a background training set, we use 1300 images from the Places scene understanding validation dataset [31]. We again ensure that all classes do not have semantic overlap with any ImageNet category in either the known or unknown evaluation set, as verified by hypernym and hyponym relationship lookup in the Wordnet lexical database. All models are trained using SGD with a mini-batch size of 256, momentum weighting of 0.9, and weight decay penalty factor of 0.0001 for 90 epochs, starting with a learning rate of 0.1 that is decayed by a factor of 10 every 30 epochs. The baseline cross-entropy trained model for the 500 class partition achieves 78.04% top-1 (94.10% top-5) accuracy.

Results. The results from our ImageNet experiments are shown in Fig. 6. We compute the AUOSC metric using the top-1 correct classification rate and report AUROC as a measure of OOD detection capability. For this large-scale experiment, Tempered Mixup shows roughly the same open set robustness as compared to Objectosphere without having to train with an additional dataset of background samples. To try and gain even better open set performance we augmented our Tempered Mixup formulation with the same background samples used in the Objectosphere training and a uniform distribution target among the known classes for these samples. In this way, our model trains on multiple combinations of mixed up samples, including combinations of known and unknown classes. The resulting hybrid model achieved the best open set performance over either the Tempered Mixup or the Objectosphere methods alone.

5   Discussion

The results from both small-scale and large-scale open set classification problems are evidence that Tempered Mixup is an effective means of improving open set robustness through feature space regularization without having to train with a source of representative unknown samples. We additionally have shown that when a representative background class is available, samples can easily be added into the training pipeline to gain additional robustness.

Our ImageNet open set classification results show that our Tempered Mixup formulation is as effective alone as training with a background class without the additional overhead. When an effective background class is available, we have also demonstrated that our formulation can take advantage of this additional training data to further improve open set robustness. In this case when the nature of the unknown classes to be rejected is known to a degree that an effective

	Training	Unknowns from ImageNet-Open	AUOSC / AUROC
Baseline			0.631 / 0.777
Objectosphere			0.688 / 0.801
Tempered Mixup (Ours)			0.688 / 0.806
Tempered Mixup (Ours)			0.704 / 0.857

Fig. 6: LARGE-SCALE OPEN SET CLASSIFICATION. Training data is made up of either ImageNet only or ImageNet plus background images from the Places dataset. Unknowns for evaluation are drawn from the ImageNet-Open dataset.

background class can be procured, then this is equivalent to hard negative mining for training with a confidence loss framework to reduce the network activation towards these unknown samples.

While we have demonstrated that our training paradigm builds a robust feature representation that improves model robustness in detecting novel classes unseen during training, this property is only tested in our work with a baseline confidence thresholding inference method. More advanced inference methods could easily be applied to our models to yield even better open set performance.

6   Conclusion

In this paper we developed a novel technique for improving the feature space of a deep CNN classifier to enable better robustness towards samples from unknown classes. We combined the concept of Mixup augmentation with a novel formulation of confidence loss to train a CNN to produce less confident predictions for samples off of the input distribution defined by the training set. Experimental evidence shows that this formulation performs favorably against current state-of-the-art methods including confidence loss regularization with a background class. This strategy could be especially useful when an appropriate background set is not available in large-scale, real-world classification environments.

References

1. Abdelzad, V., Czarnecki, K., Salay, R., Denounden, T., Vernekar, S., Phan, B.: Detecting out-of-distribution inputs in deep neural networks using an early-layer output. arXiv preprint arXiv:1910.10307 (2019)
2. Bendale, A., Boult, T.: Towards open world recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1893–1902 (2015)
3. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1563–1572 (2016)
4. Chapelle, O., Weston, J., Bottou, L., Vapnik, V.: Vicinal risk minimization. In: Advances in neural information processing systems. pp. 416–422 (2001)
5. Chun, S., Oh, S.J., Yun, S., Han, D., Choe, J., Yoo, Y.: An empirical evaluation on robustness and uncertainty of regularization methods. In: ICML Workshop on Uncertainty and Robustness in Deep Learning (2019)
6. Cohen, G., Afshar, S., Tapson, J., van Schaik, A.: Emnist: an extension of mnist to handwritten letters. arXiv preprint arXiv:1702.05373 (2017)
7. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
8. Dhamija, A.R., Günther, M., Boult, T.: Reducing network agnostophobia. In: Advances in Neural Information Processing Systems. pp. 9157–9168 (2018)
9. Fellbaum, C.: Wordnet. In: Theory and applications of ontology: computer applications, pp. 231–243. Springer (2010)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
12. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
13. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N (2015)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
15. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv preprint arXiv:1711.09325 (2017)
16. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems. pp. 7167–7177 (2018)
17. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018), `https://openreview.net/forum?id=H1VGkIxRZ`
18. Neal, L., Olson, M., Fern, X., Wong, W.K., Li, F.: Open set learning with counterfactual images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 613–628 (2018)
19. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks (2019)
20. Perera, P., Patel, V.M.: Learning deep features for one-class classification. IEEE Transactions on Image Processing (2019)

21. Roady, R., Hayes, T.L., Kemker, R., Gonzales, A., Kanan, C.: Are out-of-distribution detection methods effective on large-scale datasets? arXiv preprint arXiv:1910.14034 (2019)
22. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
23. Scheirer, W.J., Rocha, A., Sapkota, A., Boult, T.E.: Towards open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) **35** (July 2013)
24. Shu, L., Xu, H., Liu, B.: Doc: Deep open classification of text documents. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2911–2916 (2017)
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
26. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In: Advances in Neural Information Processing Systems. pp. 13888–13899 (2019)
27. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. pp. 499–515. Springer (2016)
28. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6023–6032 (2019)
29. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
30. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (ICLR) (2018), https://openreview.net/forum?id=r1Ddp1-Rb
31. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)