Novel Ensemble Diversification Methods for Open-Set Scenarios

Miriam Farber
Amazon
Haifa, Israel
mffarber@amazon.com

Roman Goldenberg
Google Research *
Haifa, Israel
romanfg@google.com

George Leifman
Google Research *
Haifa, Israel
gleifman@google.com

Gal Novich
Amazon
Haifa, Israel
ganovich@amazon.com

Abstract

We revisit existing ensemble diversification approaches and present two novel diversification methods tailored for open-set scenarios. The first method uses a new loss, designed to encourage models disagreement on outliers only, thus alleviating the intrinsic accuracy-diversity trade-off. The second method achieves diversity via automated feature engineering, by training each model to disregard input features learned by previously trained ensemble models. We conduct an extensive evaluation and analysis of the proposed techniques on seven datasets that cover image classification, re-identification and recognition domains. We compare to and demonstrate accuracy improvements over the existing state-of-the-art ensemble diversification methods.

1. Introduction

The importance of diversity in ensembles of models has been recognized since late 90s [40, 28, 7, 34]. The topic was thoroughly discussed in the Multiple Classifier Systems (MCSs) community, while investigating (i) metrics for measuring the diversity [1, 2, 17, 46], (ii) the connection between diversity and the ensemble accuracy [22, 36], and (iii) methods for constructing diversified ensembles [42, 4].

The existing body of work primarily deals with the closed-set problems, where a correct output is expected for valid input only. For open set problems, on the other hand, the ensemble should yield a correct response for all known/learned classes, while rejecting "unknown" inputs.

Open-set problems gain an increasing attention in recent years [26, 45, 12]. Extreme open-set scenarios emerge naturally in real world tasks (e.g. autonomous driving, face recognition), where an object of unknown class or an unenrolled person needs to be correctly handled by the system.

In closed-set scenarios the ensemble diversity is beneficial to counterbalance biases of individual models to reduce the bias of the ensemble on valid input. That is, when ensemble models are dissimilar, an error of one model can potentially be compensated by others. Such diversity is usually imposed by optimizing for predictions diversity on training data, which comes with the inherent trade-off between the individual model accuracy and diversity.

In the open-set case, on the other hand, the additional role of diversity is to cause individual models to disagree on *unknown* input (outliers or un-enrolled participants). The disagreement between models can be leveraged as an outlier indicator. In the identification setup, the disagreement between models can be used as an indication that the participant is not enrolled into the system. Interestingly, this "open-set" diversity does not intrinsically contradict the individual model accuracy goal.

In this paper we improve the existing, well established ensemble-based recognition methods. While these ensemble-based methods are not necessarily the SOTA for all open-set tasks, they have their merits. E.g., for many practical biometric identification tasks, where the FPR must be very low (e.g. $< 1/10e6$), ensembles is probably the only way to go. Today's public datasets don't have nearly enough labeled identities even to measure such accuracy. That's where ensembles come to rescue: e.g. voting with two independent models with FPR$< 1/1000$ yields the overall FPR$< 1/10e6$.

We show that ensemble models generated using our proposed techniques are more diverse and yield better open-set recognition results compared to other SOTA ensemble diversification methods. We design ensemble training methods that encourage models to disagree on unknown input, while agreeing on known classes. We thus avoid or soften the accuracy-diversity trade-off and allow outliers detection, while maintaining high accuracy on inliers.

Our main contributions are:

1. We develop a new diversification loss for *simultaneous* ensemble models training, designed for open-set problems. The loss encourages diverse response to outliers, thus resolving, for open-set scenarios, the intrinsic diversification dilemma of trading the individual model

*This work was conducted under Amazon.

accuracy for ensemble diversity.

2. We propose a novel feature-based diversification method for *sequential* training of ensemble models. The method automatically selects features not yet utilized by the previously trained models, thus making the "diversification by complimentary features" approach practical.

3. We evaluate the proposed methods for open-set recognition, re-identification, and classification, using seven public datasets, the largest of which includes 85k identities and 5.8M images. We compare to other ensemble diversification approaches and demonstrate the advantage of our open-set specific techniques. We compare the two proposed methods and identify use-cases they are best suited for.

2. Notation

To facilitate the discussion, let us introduce some notation. For simplicity, we consider a K-class classification problem. The same setup can be applicable for identification/recognition problems, where classes correspond to identities in the training set.

Let ensemble E=$(\mathbf{f}, g)$ be defined by a collection of n models $\mathbf{f}$=$\langle f_1, f_2, \ldots, f_n\rangle$ and a combining function g. For an input sample x each model produces a class probability vector $f_i(x) = (f_i^1(x), f_i^2(x), \ldots, f_i^K(x))$. The combining function $g(x) = g \circ \mathbf{f} = g(f_1(x), f_2(x), \ldots, f_n(x))$ merges individual model predictions to produce the final ensemble output. Among typical examples of combining functions are averaging $g_{ave}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ and voting.

Let $q(x) \in \{1, 2, \ldots, K\}$ be the ground-truth class for the sample x. Let us denote the individual loss, i.e. the loss used to train a stand-alone ensemble model f_i, by $L_i(x) = L(f_i(x), q(x))$. If ensemble and individual model predictions are of the same form, one can apply the same loss L to the ensemble output to yield a combined ensemble loss: $L_{combined}(x) = L(g(x), q(x))$. A popular individual loss for classification problems is the cross-entropy loss:

$$L_{CE}(f_i(x), q(x)) = -\sum_{j=1}^{K} 1_{q(x)=j} \log(f_i^j(x)), \quad (1)$$

where $1_{q(x)=j}$ is the indicator function, equal to 1 if $q(x)$=j and 0 otherwise. For identification problems the cross entropy loss is often combined with additional loss components [38].

3. Related Work

In this section we discuss some of the ensemble diversification techniques most relevant to our work. For comprehensive reviews of ensemble learning we refer the reader to [47, 22, 32, 25, 10]. As this work focuses on using ensembles for open-set problems, other non-ensemble open-set methods are beyond the scope of this paper. Extensive surveys on open set recognition methods can be found in [12, 3, 33].

Diversity has been shown to be critical for generating accurate and robust ensembles [21]. Yet, it is not obvious whether an explicit encouragement of such diversity through the objective function is required. As an alternative, one can train each model in a slightly different way to gain diversity - e.g. using different training sets, randomized training order [5], weights initialization [18], etc.

Another approach is to let the combined ensemble accuracy optimization to figure out the benefits of diversification in a "natural" way. That is, train the ensemble as one piece and expect the models to come out diverse, assuming the diversity indeed contributes to ensemble accuracy. Such joint training approach was proposed in [11] and further investigated in [41] where authors experiment with a weighted combination of individual model losses and the combined ensemble loss:

$$Joint(x) = (1-\alpha)\frac{1}{n}\sum_{i=1}^{n} L_i(x) + \alpha L_{combined}(x), \quad (2)$$

smoothly interpolating between independent ($\alpha = 0$) and combined ($\alpha = 1$) training of predictors.

Measuring the degree of diversity requires a metric. Different diversity metrics have been proposed [1, 2, 17, 46]. Without loss of generality, we use the average pair-wise correlation between model predictions as the diversity metric.

Although the joint loss in Eq. 2 does not include an explicit correlation reduction term, it can be shown (see Supplementary) that the correlation between models decreases due to the joint loss (stronger than due to training randomization). It was demonstrated in [41] that the joint training shows promise for resource limited scenarios, but does not generalize well to the test set and requires putting more weight on individual models performance. These studies suggest that optimizing for the combined ensemble loss is sub-optimal and call for a loss that includes individual model losses and a term that explicitly encourages diversity.

Ideally, such loss should be derived from a decomposition of a combined ensemble loss. For the MSE regression problems and ensembles that use the averaging combining function g_{ave}, an analytical justification for diversity was derived by decomposing the combined MSE ensemble loss using the *ambiguity* decomposition [20]:

$$\text{MSE}(x) = (q(x) - g_{ave}(x))^2 = \\ \frac{1}{n}\sum_{i=1}^{n}(q(x) - f_i(x))^2 - \frac{1}{n}\sum_{i=1}^{n}(f_i(x) - g_{ave}(x))^2, \quad (3)$$

where, for regression problems, $q(x)$ is the ground truth

value for x. The first term in Eq. 3 is the average MSE of ensemble models, while the second term encourages the diversity between different models. This decomposition was heuristically derived in [28] as the negative correlation loss (NCL) for regression tasks:

$$\mathrm{NCL}(x) = (1-\alpha)\frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i(x) + \alpha\frac{1}{n}\sum_{i=1}^{n}p_i(x), \quad (4)$$

where the penalty function p_i

$$p_i(x) = (f_i(x) - g_{ave}(x))\sum_{j\neq i}(f_j(x) - g_{ave}(x)), \quad (5)$$

decreases the correlation between the error of model f_i and the errors of other ensemble models. Same as in Eq. 2, α controls the relative weight between the individual model performance and the diversity. Putting more weight on individual model accuracy can be viewed as regularization.

The NCL was studied in [7, 34], recently used with deep models [35], and generalized in [37] for binary classification with cross-entropy loss and averaging combining function. We can extend the binary classification loss proposed in [37] to multi-class (MC) classification as follows:

$$\begin{aligned}\mathrm{NCL}_{MC}(x) =& \frac{1-\alpha}{n}\sum_{i=1}^{n}L_{CE}(q(x), f_i(x)) - \\ & \frac{\alpha}{n(n-1)}\sum_{i}\sum_{j\neq i}L_{CE}(f_i(x), f_j(x)).\end{aligned} \quad (6)$$

The first term is the average of individual cross-entropy losses and the second term encourages negative correlation between all pairs of class probability vectors $(f_i(x), f_j(x))$.

It can be clearly seen from Eqs. (3)-(6) that all ensemble losses above "suffer" from the intrinsic trade-off between the individual model accuracy and the diversity. The trade-off stems from the fundamental contradiction between the two objectives encoded therein: (a) all f_i's are to be as close as possible to the ground truth q and, (b) at the same time, they need to be as different as possible from each other. In this work we present a diversification method for the open set scenario that overcomes or softens this trade-off.

The approaches discussed so far encourage diverse predictions on the training set as a means to get a diverse ensemble. In [31], the authors suggest an alternative approach to ensemble diversification through encouraging diverse extrapolation, via extrapolating differently on local patches of the data manifold. To make f_i dissimilar to f_j, they add an approximation of $\mathbb{E}[(f_i(x_{f_j}^{max}) - f_i(x))^2]$ to the loss function, where $x_{f_j}^{max} = \underset{x'}{\mathrm{argmax}} f_j(x')$, and x' ranges over the neighborhood of x. The authors showed promising results, experimenting with 256-unit single hidden layer fully connected network. The limitations of their approach include unstable behaviour for activations that are not from the rectifier family, as well as the need to compute second derivative during the optimization process.

Another alternative to diversifying model predictions is diversification by feature selection. One example is Random Forests [6], where subsets of input features are randomly chosen to compose an ensemble. This and other random or ad-hoc feature selection methods typically result in rather weak individual models, which is less suitable for ensembles with only a few computationally heavy deep learning models.

The method we present in this work performs an implicit feature selection that does not degrade the individual model accuracy. Similarly to [31], it encourages diverse extrapolation, but does not suffer from the same limitations, and can be easily applied to state-of-the-art networks.

4. Methods - Diversification for Open-Set

In this section we develop two ensemble diversification methods for open-set problems. In open-set scenarios, both in- and out-of-distribution (OOD) data are fed into the model at the test stage. The goal is to correctly classify the in-distribution data and reject the OOD samples. The K-class open-set classification can be viewed as a classification into $K+1$ classes, where an unseen outliers class is added at test time to the K classes used in training. Using ensembles, a straightforward approach would be to leverage the disagreement between models as an indication for outliers. For example, using a majority voting combining function, we can decide to reject an input sample if the size of the largest consensus group is below a threshold. Otherwise, the class chosen by the majority is returned as the classification result.

The question is "how to train an ensemble to disagree on unknown data if this data is unavailable during training?"

4.1. Open-Set Correlation Reduction

Our first solution uses the "wrong class" probabilities generated by models on valid input as a proxy for model response to outliers. That is, we train on known-class data, but request the inter-model disagreement on wrong class probabilities only. Formally, let x be a valid training sample of class $q(x) \in [1, 2, \ldots, K]$. Let us denote by

$$f_i \backslash q(x) = (f_i^1(x), \ldots, f_i^{q(x)-1}(x), f_i^{q(x)+1}(x), \ldots, f_i^n(x))$$

the vector obtained from $f_i(x)$ by omitting the $q(x)^{th}$ component. We then seek to achieve a disagreement between $f_i \setminus q(x)$ and $f_j \setminus q(x)$ for every pair of models i and j.

In principle, any diversity measure could be used to drive the inter-model disagreement. For example, we could borrow the negative cross-entropy term used in NCL (Eq. 6).

We decided against the cross-entropy because of its numerical instability: unlike the closed-set case, where we expect at least one of $f_i(x)$ components to be large enough, it is likely that all $f_i \setminus q(x)$ components are close to zero, causing the cross-entropy to explode. Instead, we chose to use the correlation $\mathrm{Corr}(f_i \setminus q(x), f_j \setminus q(x))$ as a loss term to encourage the disagreement. For the rank-based voting functions g it seems natural to use Spearman's rank order correlation. Taking a second look, the majority voting g only cares about the highest rank classes, i.e. to reject an unknown sample, models have to disagree on the rank1. Low Spearman correlation does not necessarily imply the disagreement on the rank1. Hence we decided to use Pearson that gives more weight to the higher probability vector components, and define the open-set diversification loss term as

$$L_{\mathrm{Corr}}(x) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \mathrm{Corr}(f_i \setminus q(x), f_j \setminus q(x)) \quad (7)$$

For the accuracy term we took the average cross entropy loss, as in NCL (Eq. 6). Finally, the open-set correlation reduction loss (OSCRL) function we propose is given by

$$OSCRL(x) = \frac{1-\alpha}{n} \sum_{i=1}^{n} L_{CE}(q(x), f_i(x)) + \alpha L_{\mathrm{Corr}}(x),$$

where, as before, α controls the relative weight between the individual model accuracy and the diversity.

The proposed OSCRL method behaves very well for problems with low number of classes (see *Experiments*). For large number of classes, the correlation between the very long class probability vectors is expected to be always low and the diversification loss term in Eq. 7 becomes ineffective. To address the open-set problems with high number of classes (millions of classes in identification/recognition problems) we came up with the following method:

4.2. Feature-Based Diversification

Here we chose to achieve the diversity by implicit feature selection. An explicit or random feature selection, as in Random Forest [6], is problematic both because for non-tabular data the features are not readily available, and because it usually results in relatively weak individual classifiers. Instead, following the motivation similar to [31], we propose an approach for feature-based diversification that encourages diverse extrapolation. For open-set scenarios the diverse extrapolation is especially important as the model has to deal with the OOD data. As opposed to [31], our approach does not suffer from architectural limitations, and can be applied to a wide range of SOTA models.

We explain the algorithm for building a two-model ensemble, and it extends naturally for larger ensembles. The first ensemble model f_1 is trained with no diversity constraints. We then train the second model f_2 to rely on features "orthogonal" to those learned by f_1. The process consists of two steps:

1. **Feature distillation:** Distil the features learned by f_1.
2. **Feature-based diversification training:** Train f_2 while encouraging it to disregard features learned by f_1.

(a) (b) (c)

Figure 1. **Feature distillation process:** Given image (a), we start from a blank image (b) and iteratively modify it in the direction which shortens the distance between its embedding vector and that of (a), resulting in image (c)

Without loss of generality, we assume that models f_i generate an embedding vector $e_i(x)$ as internal representation of the input x (a layer before the softmax), which is then translated into the model output $f_i(x)$. The notion of embedding is widespread in search, clustering and recognition problems, and can also be applicable to classification.

4.2.1 Feature Distillation

The distillation process is visualized in Figure 1. For a given image x, we would like to generate an image x_{f_1} such that its embedding vector $e_1(x_{f_1})$ is close to the embedding vector $e_1(x)$ of the original image. To achieve this we use an iterative process inspired by the "deep-dream" [29] approach. We start with a blank image $x_{f_1}^{<0>} = \text{const}$ and proceed iteratively with the back propagation (ξ is the step size):

$$x_{f_1}^{<t+1>} = x_{f_1}^{<t>} - \xi \nabla_{x_{f_1}} \|e_1(x_{f_1}) - e_1(x)\|, \quad (8)$$

The process stops once the distance between the embeddings is small enough. In practice we performed 100 iterations and used $\xi = 0.1$. The distilled x_{f_1} (Fig. 1(c)) includes the features that, from the point of view of f_1, capture the essence of the image x (Fig. 1(a)).

4.2.2 Feature-Based Diversification Training

To train f_2 we use a loss that consists of two parts: the individual model accuracy loss (e.g. the cross-entropy loss L_{CE}) and the feature-based diversity term. To encourage diversity, we penalize the model f_2 for learning features that are useful for f_1, forcing f_2 to be agnostic to those features.

To achieve the desired diversity we would like the distilled features encoded in x_{f_1} to be non-discriminative for

f_2, that is f_2 should fail to accurately classify x_{f_1} images. We enforce the non-discriminability for x_{f_1} images by a loss term that minimizes the gap between the correct class probability and the average probability for the incorrect classes:

$$L_{FD}(x_{f_1}) = \left\| \frac{1}{K} \sum_{i=1}^{K} f_2^i(x_{f_1}) - f_2^{q(x_{f_1})}(x_{f_1}) \right\|, \quad (9)$$

where $q(x_{f_1})$ is the correct class of x_{f_1}. Alternatively, one can require a uniform logit distribution, e.g. using the Entropic Open-Set Loss proposed in [9]. The final loss function is a weighed sum of the two terms:

$$FDL(x, x_{f_1}) = (1-\alpha)L_{CE}(x) + \alpha L_{FD}(x_{f_1}). \quad (10)$$

Practically, the advantage of the proposed approach is that it can be easily applied to existing architectures with an additional loss term based on the distilled input images. As the diversity loss term operates on the distilled images, which are OOD w.r.t. the training set, the resulting models are expected to exhibit diverse response on open sets. The downside is the fact that the distilled input images need to be generated prior to or during the training, which can be time consuming. The proposed technique can be easily extended to any number of models, by adding the diversity loss terms (Eq. 9) for all the models trained before.

To demonstrate the diverse extrapolation property of the proposed method, let us compare two ensembles trained on the MNIST dataset. One two-model ensemble is trained using the proposed FDL method. The two models of the other ensemble are trained independently, using training randomization. Given an OOD image (Fig. 2(a)), which is not a MNIST digit, we show the features distilled from this image by all four models (Fig. 2(b-e)). As expected, since the image is OOD, the distilled features do not make much sense. Notably, features distilled by the independently trained models are very similar (Fig. 2(b,c)), while the features for the FDL models are different (Fig. 2(d,e)). Moreover, both independently trained models predict the same class '7' for the Fig. 2(a) image, while the two FDL models predict different classes - '7' and '2', which triggers the outlier rejection. In supplementary we discuss how the proposed method affects the embedding space.

5. Experiments

In this section we apply the proposed methods to various open-set recognition and classification problems, and compare them with other diversification methods: NCL (Eq. 4), Joint training (Eq. 2), and the baseline - an ensemble of independently trained models. As the scope of the paper is ensemble diversification, we do not compare to non-ensemble open-set recognition or classification methods. Moreover, the explored ensemble diversification methods are invariant to the underlying single model architecture and loss. In every experiment below, for a fair comparison, we use the same, not necessarily state-of-the-art, single model type for all diversification methods we compare to. We also evaluate the impact of the proposed techniques on adversarial robustness (see supplementary).

Figure 2. **Distilled features:** Given an OOD image (a), the distilled features learned by a pair of independently trained models are similar (b,c), while the distilled features learned by "orthogonal" models differ (d,e).

5.1. Open-Set Recognition

In this section we apply the proposed methods to various open-set recognition problems and present quantitative results on publicly available benchmarks. In open-set recognition scenarios, each input probe either has a match in the gallery (target), or has not (non-target). The targets are to be recognized (correct match established), whereas non-targets are to be rejected. We use ensemble consensus to distinguish between targets and non-targets in the following way: a probe is matched if all ensemble models agree on its identity. Otherwise, the probe is rejected. Following [49]) we use TTR (true target rate) at different FTRs (false target rate) as the accuracy metric. Considering recognition as a target/non-target classification, TTR is equivalent to recall and FTR to false positive rate. For evaluation, all test IDs are enrolled in the gallery. To compute the TTR, we count the probes correctly matched to their gallery images. To compute the FTR, for every probe we exclude all its corresponding images from the gallery, and count the probes that are not rejected.

5.1.1 Re-ID

We apply our methods to the open-set person re-identification (re-ID) problem. A typical scenario is a multi-camera setup where a person seen by one camera is to be recognized by another camera based on the whole body appearance. The problem is further complicated as the images are taken from different angles and the views are often partially occluded.

Dataset and Architecture: We use Market-1501 benchmark [48], which contains 32,668 images of 1,501 identities (750 train and 751 test). The setup is similar to the multi-query setting reported in [50]. We use ResNet50 [13] as the backbone for this task.

	Baseline	NCL	Joint	FDL	OSCRL
FTR=1%	16.9%	20.1%	19.6%	**22.2%**	20.3%
FTR=10%	56.3%	54.7%	53.9%	**61.3%**	60.9%
FTR=20%	71.6%	69.0%	70.6%	**76.8%**	75.6%
FTR=30%	78.3%	76.4%	78.0%	**83.5%**	80.1%

Table 1. **Re-ID on Market-1501:** Re-identification accuracy (TTR) at different FTR targets for the 5 ensemble diversification approaches. The best results are in green.

	Baseline	NCL	Joint	FDL	OSCRL
FTR=1%	98.26	98.42	98.27	**98.57**	98.50
FTR=.5%	83.33	85.71	84.11	85.04	**86.41**

Table 2. **Face recognition on LFW:** Face recognition accuracy (TTR) at FTR=$\{0.5\%, 1\%\}$ for 5 ensemble diversification approaches. Best results in green.

DeepID3 [39]	LightCNN29 [44]	IDL Sngl.[27]	IDL Ensm.[27]	**FDL**	**OSCRL**
81.40	93.62	94.12	95.80	**99.33**	**99.16**

Table 3. **Face recognition on LFW - comparison to non-ensemble SOTA:** DIR@FAR=1% following the protocol in [16].

Table 1 shows TTR at different FTRs, for various ensemble diversification methods. It can be seen that FDL yields the best result, outperforming the baseline by up to 30%. OSCRL is the 2^{nd} best approach. Here we report the accuracy for optimal α's. The results for other α's are discussed in supplementary and "Sensitivity to α" section below.

5.1.2 Face Recognition

Almost any real-world face recognition system operates in a open-set environment, where only a fraction of probe identities are enrolled in the gallery.

Datasets: We conduct the face recognition experiments by training on MS1MV2 dataset [8] (version of the MS-Celeb-1M [43]), containing 85k identities and 5.8M images. For testing we use the Labeled Faces in the Wild (LFW) dataset [15] with 13,233 images of 5,749 identities.

Architecture: We use the IR-SE50 ArcFace [8] architecture, with the same training regime as in [8].

Table 2 compares the five ensemble diversification methods at FTR=1% (the standard evaluation point [24]) and FTR=0.5%. For FTR=1% the FDL-based ensemble outperforms all other methods, reducing the error rate by almost 20% (from 1.74% to 1.43%, when compared to the baseline). For FTR=0.5% the OSCRL-trained ensemble outperforms all other methods (more results in supplementary).

Finally, even though the scope of this paper is ensemble methods, to provide a comparable reference point, we compare to several non-ensemble SOTA methods in Table 3. The table shows DIR(Rank-1 Detection and Identification rate)@FAR=1% (equivalent to TTR@FTR) measured on LFW using the protocol from [16]. This protocol enrolls only 1% of the IDs to the gallery, which is less challenging than $\sim$100% enrollment used for Table 2. This explains the higher FDL/OSCRL accuracy in Table 3. Both FDL and OSCRL outperform other methods.

5.2. Open-Set Classification

We use the same data setup as in [26]: The training set includes the in-distribution data only. The test set is formed by adding an increasing proportion, from 2% to 100% (denoted by γ), of OOD data to the in-distribution test set.

Datasets: We use CIFAR-100 and CIFAR-10 [19], consisting of 50K training and 10K test images, drawn from 100 and 10 classes, respectively. For OOD data we use the TinyImageNet(crop) dataset [26] (randomly cropped 32-by-32 images from ImageNet) of 10K images.

Architectures: We train 3 configurations of ensembles of DenseNet-BC networks [14]: DN-100-12, DN-82-8 and DN-64-6 with depths of 100, 82 and 64, and growth rates of 12, 8 and 6, respectively. We follow the configurations and hyper-parameters used in [41]. Each network is trained 3 times, and results are averaged over the training instances.

In the experiments below, we train two-model ensembles ($n = 2$) using five methods (OSCRL, FDL, Joint, NCL and the baseline) and compare their performance. Note that while using the same data as in [26], we do not compare to their results since they perform OOD detection instead of measuring the overall K+1-class classification accuracy (like we do). In addition, they use OOD data as validation to tune the model, while we avoid using such data during the training process. Figure 3 summarizes the results of the experiments, where each sub-figure corresponds to a data set (CIFAR-10/CIFAR-100) and a network-architecture (DN-64-6/DN-82-8/DN-100-12). The x-axis represents γ, and y-axis represents the accuracy of the model on the test set that includes both in-distribution and OOD data. For each method we depict the graph for the best performing α (we chose the one that optimizes the average accuracy over all possible OOD rates - the area under curve). Table 4 compares ensemble accuracy for different diversification methods on a test set with 1:1 mix ($\gamma = 100\%$) of in- and out- of distribution data (the setup used in [45, 26]).

On CIFAR-10, the OSCRL significantly outperforms other methods by a large margin for most γ values, showing up to 10% accuracy increase when comparing to the baseline. Joint training is the 2nd best method, while NCL and FDL are comparable to the baseline (or a bit better). On CIFAR-100, the FDL outperforms the baseline by a large margin. A closer look at CIFAR-100 results reveals that the relative ensemble behaviour is network-dependent: on DN-64-6, the FDL outperforms the rest of the methods, with OSCRL being in the 2nd place for mid-large range γ's. On DN-82-8, the FDL outperforms the rest of the methods for small-mid range γ values, while OSCRL wins for mid-large range γ values. For DN-100-12, NCL outperforms the rest

CIFAR-100					
Architecture	Baseline	Joint	NCL	FDL	OSCRL
DN-64-6	60.9	62.8	62.5	**64.2**	**64.2**
DN-82-8	65.1	67.3	66.3	67.1	**69.4**
DN-100-12	65.4	69.6	**69.9**	67.7	69.0
CIFAR-10					
Architecture	Baseline	Joint	NCL	FDL	OSCRL
DN-64-6	73.8	75.8	73.7	73.5	**77.9**
DN-82-8	72.3	75.6	72.3	72.4	**77.0**
DN-100-12	71.2	75.5	72.6	72.2	**78.4**

Table 4. **Open-set classification on CIFAR** Accuracy of ensembles trained using 5 different diversification approaches: independently trained models - baseline, Joint training, NCL, OSCRL, and FDL, on a 1:1 in-/outlier mix ($\gamma = 100\%$) test set. Rows correspond to different network architectures - three types of DenseNets. Best results in green.

of the methods. To summarize, in 5 out of the 6 experiments (datasets and architectures), our diversification methods, designed specifically to tackle the open-set problems, outperform other ensemble diversification methods.

Note that, as stated earlier, the α values used for the discussion above are those that optimize the average accuracy over all possible γ's (in the range of 2%-100%). If γ is expected to be outside of this range (very small or very large), α should be adjusted accordingly.

5.2.1 SVHN and MNIST

We repeated the experiments depicted in Table 4 on SVHN [30] and MNIST [23] datasets using DN-100-12. Both datsets contain 10 classes representing digits (MNIST contains handwritten digits, while SVHN consists of street-view digits). MNIST dataset includes 60K training images and 10K test images. SVHN dataet includes 73,257 training images, and we used randomly sampled 10K test images, out of the standard 26032 test images. In both cases, for OOD data, we added 10K images from the TinyImageNet(crop) dataset (same as for the CIFAR experiments).

Once again, OSCRL significantly outperformed the competing methods, yielding 86% and 80% accuracy on SVHN and MNIST respectively (vs 78% and 77% obtained by Joint and NCL respectively).

5.3. Sensitivity to α

In the experiments above we present the accuracy of models for the optimal α. To be practical, a method cannot be too sensitive to tunable parameters. To test this sensitivity, we evaluated the proposed methods in a range of α's around the optimum. Table 5 presents the accuracy of FDL and OSCRL in a range of α's for classification (top) and recognition (bottom), respectively. One can see that both

CIFAR-10								
Architecture	FDL				OSCRL			
α=	.02	.05	.1	.2	.05	.1	.15	.2
DN-64-6	72.7	72.6	73.0	73.5	75.4	75.8	**77.9**	77.5
DN-82-8	72.4	72.3	72.3	72.4	76.9	**77.0**	76.6	75.8
DN-100-12	71.9	71.6	71.5	72.2	76.9	78.2	78.2	**78.4**

Face Recognition								
Architecture	FDL			OSCRL				
α=	.01	.05	.1	.1	.25	.5	.75	.9
FTR=1%	**98.57**	98.48	98.49	98.50	98.47	98.46	97.96	97.48
FTR=.5%	85.04	84.64	84.48	83.06	85.61	84.57	83.98	**86.41**

Table 5. Classification accuracy (top) and recognition TTR (bottom) of FDL and OSCRL ensembles for a range of α's. Compare to Tables 4 and 2. The best results are in green.

methods are robust to α variations and keep outperforming other methods in a wide range around the optimum (compare to accuracy of other methods in Tables 4 and 2).

6. Discussion

FDL vs OSCRL: The experiments above show that the two proposed methods outperform other diversification techniques. The question is when FDL is a better fit than OSCRL? Our experiments suggest that OSCRL may be better for classification, usually with a small number of classes, while FDL for identification, especially with many identities, e.g. biometric identification. For large number of classes, OSCRL becomes less effective, as de-correlating a long tail of irrelevant low probability classes doesn't necessarily decrease the correlation for the relevant top classes. Another possible reason is the nature of outliers. In identification, the outliers are of the same structure as inliers (e.g. unenrolled vs. enrolled faces), whereas in classification the outliers can be anything (e.g. cats vs. digits). FDL can better exploit the diversity of features when the learned features are still observable in outliers. See supplementary for insights into embedding spaces learned by the two methods.

Sequential vs simultaneous training: We said that FDL ensemble models are trained sequentially, while OSCRL - jointly. In fact, the OSCRL can be slightly altered to support sequential training: by freezing the already trained models and applying the gradients to the model being trained.

Accuracy as a function of γ (percentage of outliers): Figure 3 shows the graphs of ensemble accuracy (acc) for open-set classification as a function of γ. Peculiarly, for CIFAR-10 the $acc(\gamma)$ functions are decreasing, while for CIFAR-100 they are increasing. By decomposing the accuracy acc into the accuracy on inliers (acc_{in}) and outliers (acc_{out}) one can express the derivative of acc w.r.t. γ as follows (see the derivation in supplementary):

$$\partial acc/\partial\gamma = (acc_{out} - acc_{in})/(1+\gamma)^2 \qquad (11)$$

Eq. 11 explains the above $acc(\gamma)$ slope differences. Indeed, acc_{in} is expected to be lower for CIFAR-100 (100-class

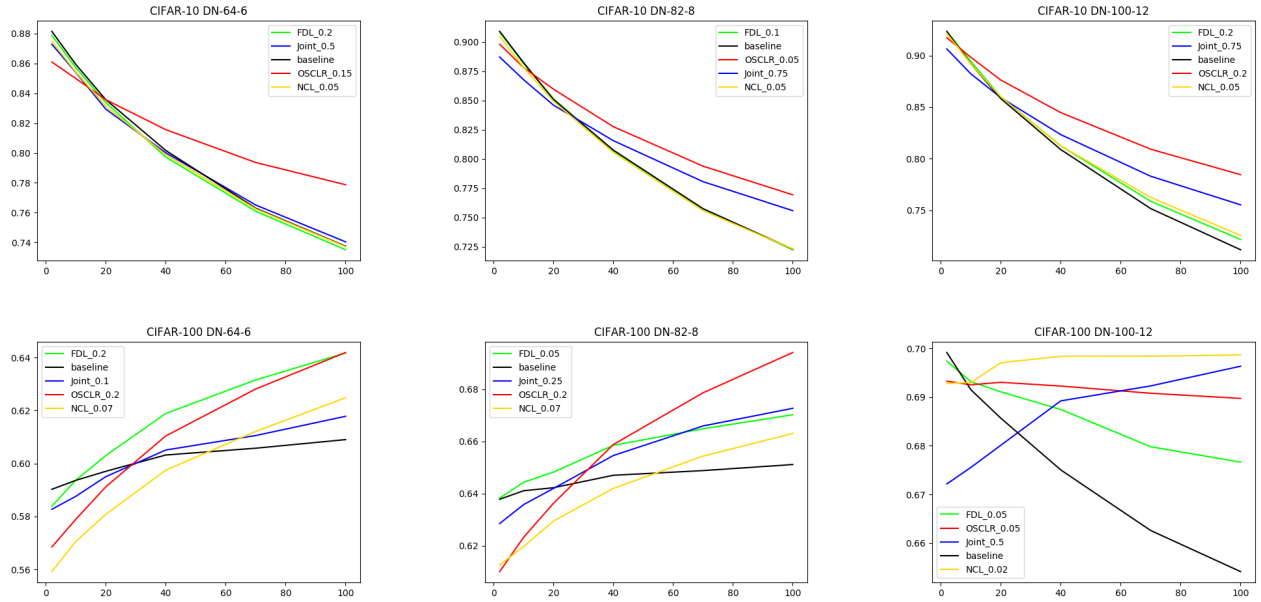

Figure 3. **Results on CIFAR**: Accuracy (Y-axis) of ensembles trained using 5 different diversification approaches: independently trained models - baseline (black), Joint training (blue), NCL(yellow), OSCRL(red) and FDL(green), as a function of outliers proportion γ (X-axis) in the test set. Top row - CIFAR-10. Bottom row - CIFAR-100. Columns correspond to different network architectures - three types of DenseNets.

problem is harder than 10-class). On the other hand, acc_{out} is expected to be higher for CIFAR-100, as it is less likely that ensemble models agree on the same class (out of 100) for an outlier. Hence, for CIFAR-100 it is more likely getting $acc_{out} > acc_{in}$, and, thus, an increasing $acc(\gamma)$.

Optimal α as a function of γ: By design, α controls the tradeoff between the individual model performance and the diversification. For open-set problems it is equivalent to controlling the tradeoff between the acc_{in} and acc_{out}. Thus, to achieve the highest overall accuracy, one should increase α for larger expected γ (proportion of outliers). E.g., for CIFAR with $\gamma < 50\%$, we use $\alpha < 0.2$, whereas for face recognition, with $\gamma > 70\%$, we increased it to 0.9. Similarly, our results on adversarial attacks (supplementary) indicate that harsher attack require higher α's. See supplementary on the relation of α to ensemble interpretability.

7. Conclusion

We proposed two approaches for training diverse ensemble of models for open-set scenarios.

The first approach performs a simultaneous training of multiple models using the specially designed open-set correlation reduction loss (OSCRL). OSCRL alleviates the accuracy-diversity trade-off by requesting diversity on non-valid (outliers) inputs only. This type of diversity is particularly beneficial for outliers detection by majority voting.

The second approach achieves diversification by feature engineering. Ensemble training is done sequentially, when each new model is trained to "ignore" features already exploited by the previously trained models. To discover the set of features learned by a model we use a feature distillation process inspired by the "deep-dream" [29] approach.

The intrinsic dilemma of ensemble diversification is sacrificing individual model accuracy in favor of diversity. OSCRL resolves this dilemma for open set problems, which, we believe, is a significant differentiator from the existing ensemble diversification approaches. With regard to FDL - it takes the "diversification by complimentary features" technique to the next level.

Instead of hand-crafted feature selection, as was done before, FDL does it in a fully automatic way, thus making this powerful method practical (Best accuracy in all 7 identification categories except of FTR=0.5%, which permits only few false positives, resulting in higher noise). We propose the "anti-teaching" to make the trained model non-discriminative to previously learned distilled features, which, to the best of our knowledge, is an original approach.

We demonstrate the effectiveness of both approaches in several open-set recognition and classification domains. We show that in open-set scenarios the proposed methods consistently outperform existing ensemble diversification methods (Independent Models Training, Joint Training, Negative Correlation Loss), leading to an accuracy improvement of up to 10% in open-set classification and up to 20%-30% improvement on open-set recognition problems.

References

[1] Robert E Banfield, Lawrence O Hall, Kevin W Bowyer, and W Philip Kegelmeyer. A new ensemble diversity measure applied to thinning ensembles. In *International Workshop on Multiple Classifier Systems*, pages 306–316. Springer, 2003.

[2] Robert E Banfield, Lawrence O Hall, Kevin W Bowyer, and W Philip Kegelmeyer. Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1):49–62, 2005.

[3] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[4] Richard A Berk. An introduction to ensemble methods for data analysis. *Sociological methods & research*, 34(3):263–295, 2006.

[5] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[6] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[7] Gavin Brown, Jeremy L Wyatt, and Peter Tiňo. Managing diversity in regression ensembles. *Journal of machine learning research*, 6(Sep):1621–1650, 2005.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[9] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[10] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, pages 1–18, 2020.

[11] Anuvabh Dutt, Georges Quénot, and Denis Pellerin. Coupled Ensembles of Neural Networks. *Neurocomputing*, Apr. 2019.

[12] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[15] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2008.

[16] Chi Jin, Ruochun Jin, Kai Chen, and Yong Dou. A community detection approach to cleaning extremely large face database. *Computational intelligence and neuroscience*, 2018, 2018.

[17] Albert Hung-Ren Ko, Robert Sabourin, and Alceu de Souza Britto Jr. Compound diversity functions for ensemble selection. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):659–686, 2009.

[18] John F. Kolen, John F. Kolen, Jordan B. Pollack, and Jordan B. Pollack. Back propagation is sensitive to initial conditions. In *Complex Systems*, pages 860–867. Morgan Kaufmann, 1990.

[19] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[20] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *NIPS*, 1994.

[21] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, pages 231–238. MIT Press, 1995.

[22] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[23] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[24] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[25] Hui Li, Xuesong Wang, and Shifei Ding. Research and development of neural network ensembles: a survey. *Artificial Intelligence Review*, 49(4):455–479, 2018.

[26] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[27] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[28] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Networks*, 12:1399–1404, 1999.

[29] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015.

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

[31] Andrew Slavin Ross, Weiwei Pan, Leo Anthony Celi, and Finale Doshi-Velez. Ensembles of locally independent prediction models. *arXiv preprint arXiv:1911.01291*, 2019.

[32] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[33] W Scheirer, A de Rezende Rocha, A Sapkota, and T Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[34] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5382–5390, 2018.

[35] Zenglin Shi, L. Zhang, Yun Liu, X. Cao, Y. Ye, Ming-Ming Cheng, and G. Zheng. Crowd counting with deep negative correlation learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5382–5390, 2018.

[36] Catherine A Shipp and Ludmila I Kuncheva. Relationships between combination methods and measures of diversity in combining classifiers. *Information fusion*, 3(2):135–148, 2002.

[37] Ron Shoham and Haim Permuter. Amended cross-entropy cost: an approach for encouraging diversity in classification ensemble (brief announcement). In *International Symposium on Cyber Security Cryptography and Machine Learning*, pages 202–207. Springer, 2019.

[38] Yash Srivastava, Vaishnav Murali, and Shiv Ram Dubey. A performance comparison of loss functions for deep face recognition. *arXiv preprint arXiv:1901.05903*, 2019.

[39] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[40] Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, pages 90–95. IEEE, 1996.

[41] Andrew M Webb, Charles Reynolds, Dan-Andrei Iliescu, Henry Reeve, Mikel Luján, and Gavin Brown. Joint training of neural network ensembles. *arXiv preprint arXiv:1902.04422*, 2019.

[42] K Tyler Wilcox. An introduction to ensemble methods for machine learning. 2016.

[43] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 529–534, 2011.

[44] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[45] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019.

[46] Yang Yu, Yu-Feng Li, and Zhi-Hua Zhou. Diversity regularized machine. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[47] Cha Zhang and Yunqian Ma. *Ensemble Machine Learning: Methods and Applications*. Springer Publishing Company, Incorporated, 2012.

[48] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[49] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):591–606, 2015.

[50] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.