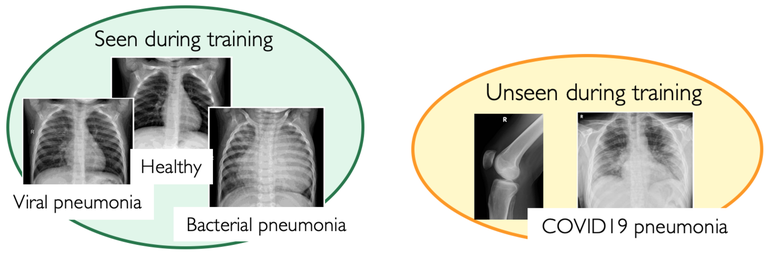
Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science, ETH Zurich

Novel classes as OOD data

Problem: Classifier predictions are incorrect on novel classes.

- Flag data from unseen classes as out-of-distribution (OOD).



- Novel classes are often similar to in-distribution (ID) classes
  $\Rightarrow$ difficult to distinguish ID and OOD data.

Current SOTA (unsupervised) novelty detection approaches **perform poorly on novel-class detection**.

Semi-supervised novelty detection

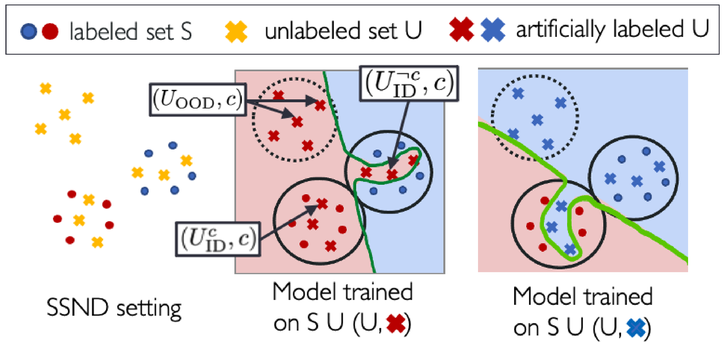
Available data:

1. Labeled set with ID samples.
   e.g. the training set for the prediction task.
2. Unlabeled set with unknown mixture of ID and OOD data.
   e.g. hospital collects all X-rays performed during the day.

Prior SSND methods fail to use unlabeled data effectively.
e.g. nnPU, MCD

Intuition for new method: Train ensemble on $S \cup (U, c)$.

- Each model sees set U labeled with different $c \in \mathcal{Y}$.



Our approach

At training time:

- Obtain an **Ensemble w/ Regularized Disagreement**.

Algorithm 1: Obtaining ERD ensemble via early stopping

Input : Train set S, ID Validation set V, Unlabeled set U, Model $\tilde{f}$ pretrained on S, Ensemble size K
Result: ERD ensemble $\{f_{y_i}\}_{i=1}^K$
Sample K different labels $\{y_1, ..., y_K\}$ from $\mathcal{Y}$
for $c \leftarrow \{y_1, ..., y_K\}$ **do** `// fine-tune K models`
  $f_c \leftarrow \mathit{Initialize}(\tilde{f})$
  $(U, c) \leftarrow \{(x, c) : x \in U\}$
  $f_c \leftarrow \mathit{EarlyStoppedFineTuning}\,(f_c, S \cup (U, c); V)$
return $\{f_{y_i}\}_{i=1}^K$

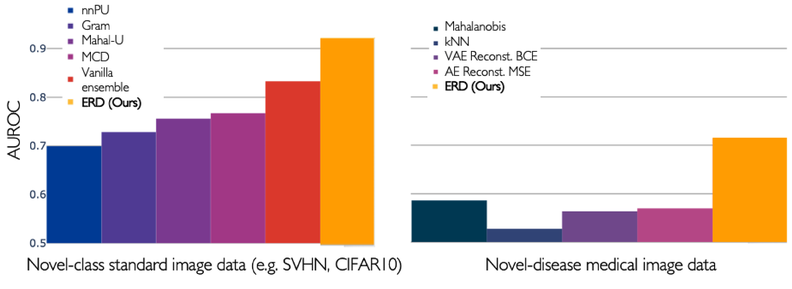
At test time:

- For a test sample x, use outputs $f_1(x), ..., f_k(x)$ to compute the **average pairwise disagreement score** (details later).
- Flag as OOD samples with score larger than threshold τ.

Experiments

Easy OOD: SVHN vs CIFAR10, CIFAR10 vs SVHN etc
Novel class OOD: CIFAR100[0-49] vs CIFAR100[50-99] etc

Evaluation metric: Area under ROC curve (AUROC)



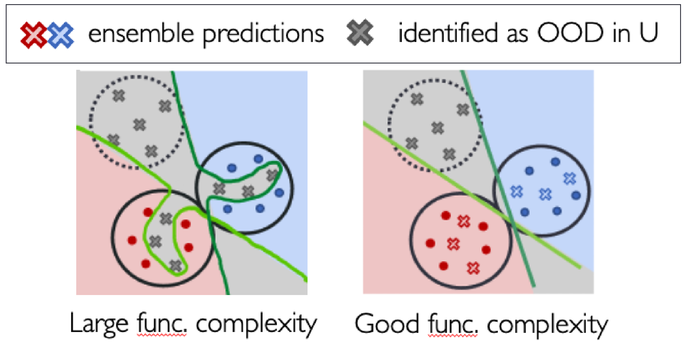
New SOTA SSND performance on near-OOD data.

Remark: Our approach makes use of two key ingredients:

1. regularization
2. a suitable score for OOD detection.

Key 1: Role of regularization

Goal: Prevent complex models from interpolating on $S \cup (U, c)$.



Advantages of early stopping:

- We prove that there exists an **optimal stopping time** at which every model predicts: (1) the correct label on ID data; and (2) the arbitrary label on the OOD unlabeled data.
- Efficient model selection (requires only one training run).

Key 2: Ensemble disagreement score

Prior work: Entropy of average predictor $(\mathrm{H} \circ \mathrm{Avg})$.
Our average pairwise disagreement score:

$$(\mathrm{Avg} \circ \rho)(f_1(x), ..., f_K(x)) := \frac{2}{K(K-1)} \sum_{i \neq j} \rho\left(f_i(x), f_j(x)\right)$$

$\rightarrow$ e.g. ρ = total variation distance

Unlike $(\mathrm{H} \circ \mathrm{Avg})$, our score exploits ensemble diversity.

- $(\mathrm{H} \circ \mathrm{Avg})$ obtains lower FPR at the same TPR.
- TP = correctly flagged OOD point; FP = ID flagged as OOD.