Novelty detection using ensembles with regularized disagreement
Alexandru Tifrea, 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 ⇒ difficult to distinguish ID and OOD data.

Existing OOD detection methods (assuming different access to OOD data) perform poorly on novel-class detection.

**OUR SETTING**

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

Unknown OOD setting:

Our approach makes use of two key ingredients:
1. regularization
2. a suitable score for OOD detection.

Previous methods that employ the Unknown OOD setting (e.g. nnPU, MCD) fail to leverage unlabeled data effectively.

**OUR APPROACH**

Idea: Train an Ensemble w/ Regularized Disagreement.

**Algorithm 1: Fine-tuning the ERD ensemble**

**Input:** Train set $S$, Validation set $V$, Unlabeled set $U$.
Weights $W$ pretrained on $S$. Ensemble size $K$

**Result:** ERD ensemble $\{f_k\}_{k=1}^K$.

Sample $K$ different labels $\{y_1, ..., y_K\}$ from $Y$.

for $c \leftarrow \{y_1, ..., y_K\}$ do // fine-tune $K$ models

$\langle U, c \rangle \leftarrow \{(x, z) : z \in U\}$

$\hat{f}_c \leftarrow \text{Fine-tuneWithEarlyStopping}(f, S \cup \langle U, c \rangle)$;

return $\hat{f}_c$.

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 $\tau$.

**EXPERIMENTS**

**Easy OOD:** SVHN vs CIFAR10, CIFAR10 vs SVHN etc

**Novel class OOD:** CIFAR100[0-49] vs CIFAR100[50-99] etc

**Evaluation metric:** TNR at a TPR of 95%.

→ TNR = correctly identified ID; TPR = correctly flagged OOD.

**Prior work:** Entropy of average predictor $(H \circ \text{Avg})$.

Our average pairwise disagreement score:

$\frac{2}{K(K-1)} \sum_{i \neq j} \rho(f_i(x), f_j(x))$

→ e.g. $\rho = \text{total variation distance}$

Unlike $(H \circ \text{Avg})$, our score exploits ensemble diversity. ⇒ lower FPR at the same TPR.

**Key 1: ROLE OF REGULARIZATION**

**Goal:** Prevent complex models from interpolating on $S \cup \langle U, c \rangle$.

**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 $(H \circ \text{Avg})$.

Our average pairwise disagreement score:

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

→ e.g. $\rho = \text{total variation distance}$

Unlike $(H \circ \text{Avg})$, our score exploits ensemble diversity. ⇒ lower FPR at the same TPR.