Fraunhofer IKS

SAFE INTELLIGENCE

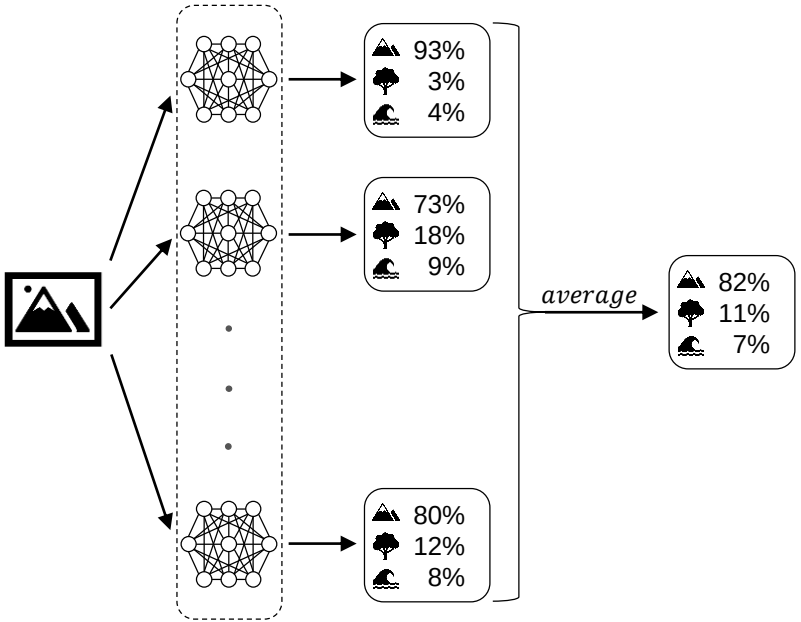
COGNITIVE SYSTEMS | ARTIFICIAL INTELLIGENCE & MACHINE LEARNING | AUTONOMOUS SYSTEMS | AUTONOMOUS DRIVING | INDUSTRY 4.0 | IOT

AI SAFETY 2021

MEASURING ENSEMBLE DIVERSITY AND ITS EFFECTS ON MODEL ROBUSTNESS

LENA HEIDEMANN, ADRIAN SCHWAIGER, KARSTEN ROSCHER

RECAP – DEEP ENSEMBLES



Deep Ensemble of M members

- Each ensemble member
  - has the same model architecture
  - is trained on the same dataset with random initialization
- For classification:
  - The ensemble output is the average of the predicted probabilities of all members

MOTIVATION – ENSEMBLE DIVERSITY

— Good performance of deep ensembles mostly explained by diversity among ensemble members

— There are several metrics aiming to quantify this ensemble diversity, but are they good indicators of ensemble robustness?

→ Main questions:

1. To what extent can ensembles with the same training conditions differ in their performance and robustness?
2. Are diversity metrics correlated with safety-relevant metrics and therefore suitable for selecting members to form a more robust ensemble?

MEASURING ENSEMBLE DIVERSITY

Disagreement

= fraction of samples for which two classifiers predict different labels

Normalized Disagreement

= Disagreement / (1 – Accuracy)

Double Fault Measure

= fraction of samples for which both classifiers make a wrong prediction

Output Correlation

= pairwise correlation of softmax outputs

Prediction space diversity

Cosine Similarity

= cosine similarity between the model parameters of two classifiers

Weight space diversity

EVALUATION SETUP

- *Architectures*: BasicCNN, MobileNetV2, 34-layer ResNet
- *Datasets*: CIFAR-10, CINIC-10, GTSRB
- For each architecture and dataset, 20 models were trained
- All possible ensembles of a combination of 5 out of the 20 trained networks were evaluated on In-Distribution (ID), corrupted, and Out-of-Distribution (OOD) data (CIFAR-100, SVHN)
- *Evaluation Metrics*:
  - ID/Corruptions: Accuracy, Acc@1%, ECE, NECE
  - OOD: AUROC, FPR95

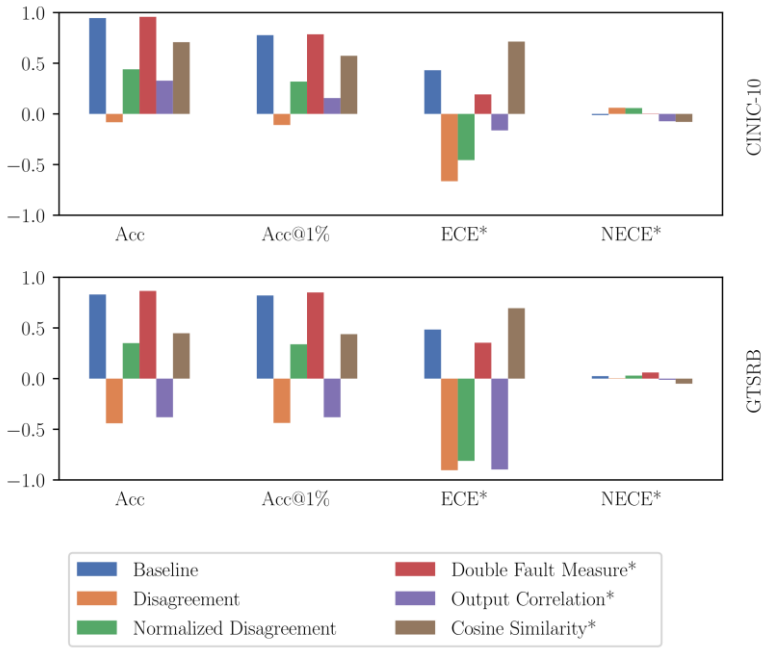
ID DATA

Variance of Metrics

- Accuracy: up to 2 pp
- Acc@1%: up to 8 pp
- ECE: up to 4 pp
- NECE: << 1 pp

Correlation with Diversity Metrics

- *Baseline*, *Double Fault Measure*, and *Cosine Similarity* correlate positively with Accuracy, Acc@1%, and ECE
- NECE: almost no correlation, but also small variance
- *Disagreement* and *Output Correlation* correlate negatively, especially for GTSRB



*reversed sign

CORRUPTED DATA

- 3 types of corruptions: brightness, contrast, cutout
- **Variance of Metrics**: higher than on ID data
- **Correlation with Diversity Metrics**: generally weaker than on ID data

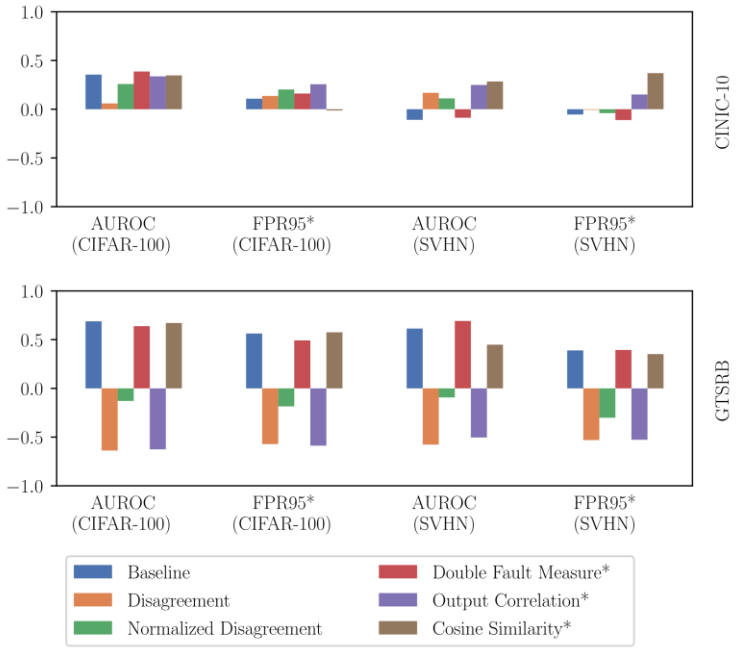
OOD DATA

Variance of Metrics

- AUROC (CIFAR-100): up to 4 pp
- FPR95 (CIFAR-100): up to 25 pp
- AUROC (SVHN): up to 15 pp
- FPR95 (SVHN): up to 43 pp

Correlation with Diversity Metrics

- *Baseline*, *Double Fault Measure*, and *Cosine Similarity* mostly correlate positively with AUROC and FPR95
- ResNets trained on GTSRB:
  - Highest *Cosine Similarity*: 15.8% FPR95 (CIFAR-100)
  - Lowest *Cosine Similarity*: 0.8% FPR95 (CIFAR-100)



*reversed sign

CONCLUSIONS AND FUTURE WORK

- Ensembles trained under the same conditions may vary significantly in performance and robustness metrics
- *Cosine Similarity* and *Double Fault Measure* show a high correlation with the evaluation metrics, but rarely exceed the baseline of selecting based on accuracy
- Future work:
  - New diversity metrics based on the specifics of DNNs
  - Increase ensemble diversity with the help of these metrics