Confidence Scoring Using Whitebox Meta-models with Linear Classifier Probes

Tongfei Chen*, Jiří Navrátil, Vijay Iyengar, Karthikeyan Shanmugam

IBM Thomas J. Watson Research Center; * Johns Hopkins University

Summary

- A learnable confidence scorer (**meta-model**) observes an existing neural classifier (**base model**) succeeding / failing at its task
- Using ***linear classifier probes*** to collect features from the base model (*whitebox*) to predict success or failure of the base model

Base vs Meta

- **Base model:** Prediction $\hat{\mathbf{y}} = F(\mathbf{x})$;
- **Meta-model:** Confidence score $z = G(\mathbf{x}, \mathbf{\Theta}_F)$. Trained as a binary classifier where G predicts whether F is correct or not.

Blackbox vs Whitebox

Blackbox: Intermediate computations of F not accessible — G can only observe the prediction $\hat{\mathbf{y}}$.

$$z = G_{\blacksquare}(\hat{\mathbf{y}}).$$

Softmax response (Geifman and El-Yaniv, 2017):

$$z = P(y^*|\mathbf{x}, \mathbf{\Theta}_F) = \max_i \hat{\mathbf{y}}_{(i)}.$$

Whitebox: Meta-model G assumes full access to the internals of F:

$$z = G_{\square}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n).$$

Linear Classifier Probes

For each intermediate result $\mathbf{x}_i$, train a ***linear classifier probe*** (Alain and Bengio, 2016) F_i to predict the correct class y using only that result:

$$\hat{\mathbf{y}}_i = F_i(\mathbf{x}_i) = \text{softmax}(\mathbf{W}_i\mathbf{x}_i + \mathbf{b}_i).$$

$$z = G(\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_n).$$

Structure of meta-model:

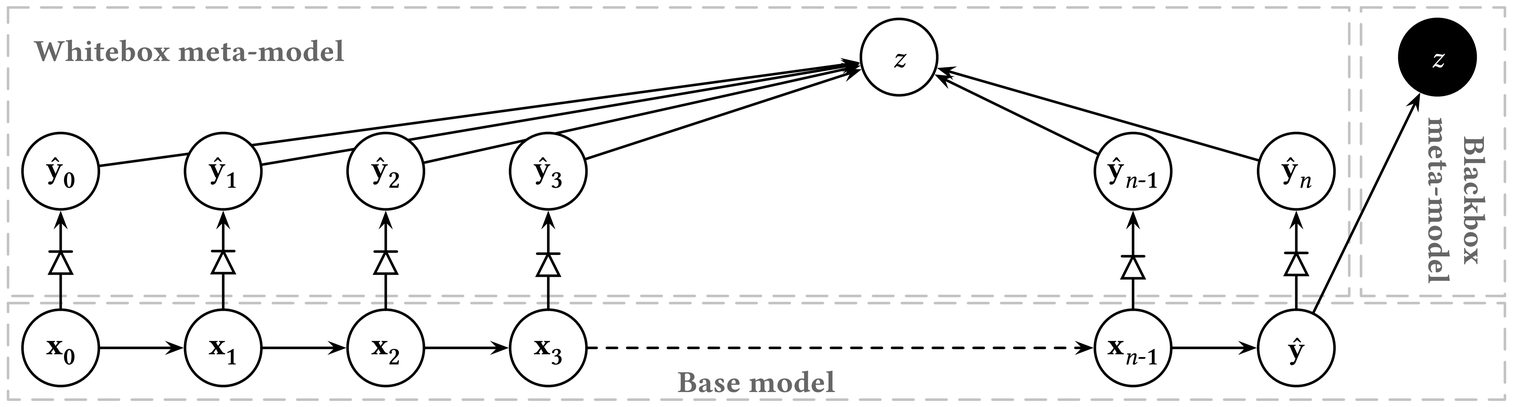
1. Logistic regression;
2. Gradient boosting machine (Friedman, 2001).

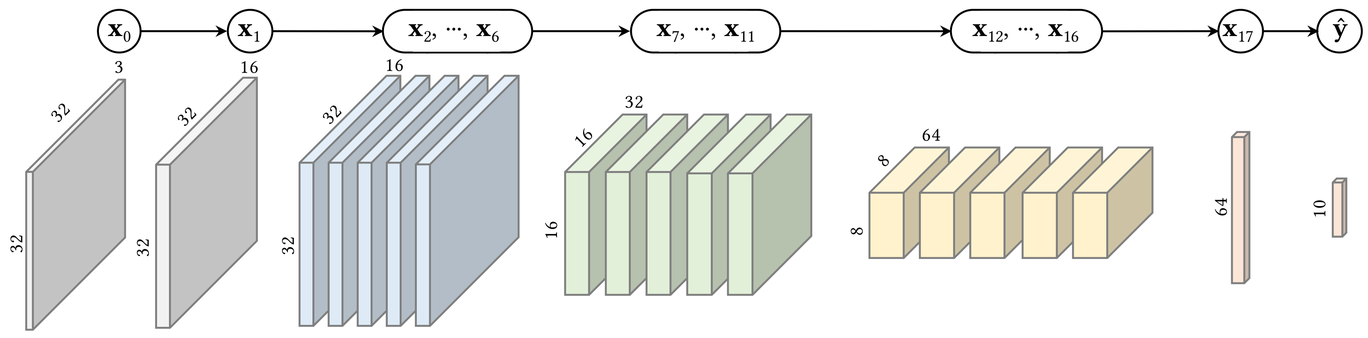
Experiments

CIFAR-10: 50k training samples partitioned into (Train-base: 30k; Train-meta: 10k; Dev: 10k)

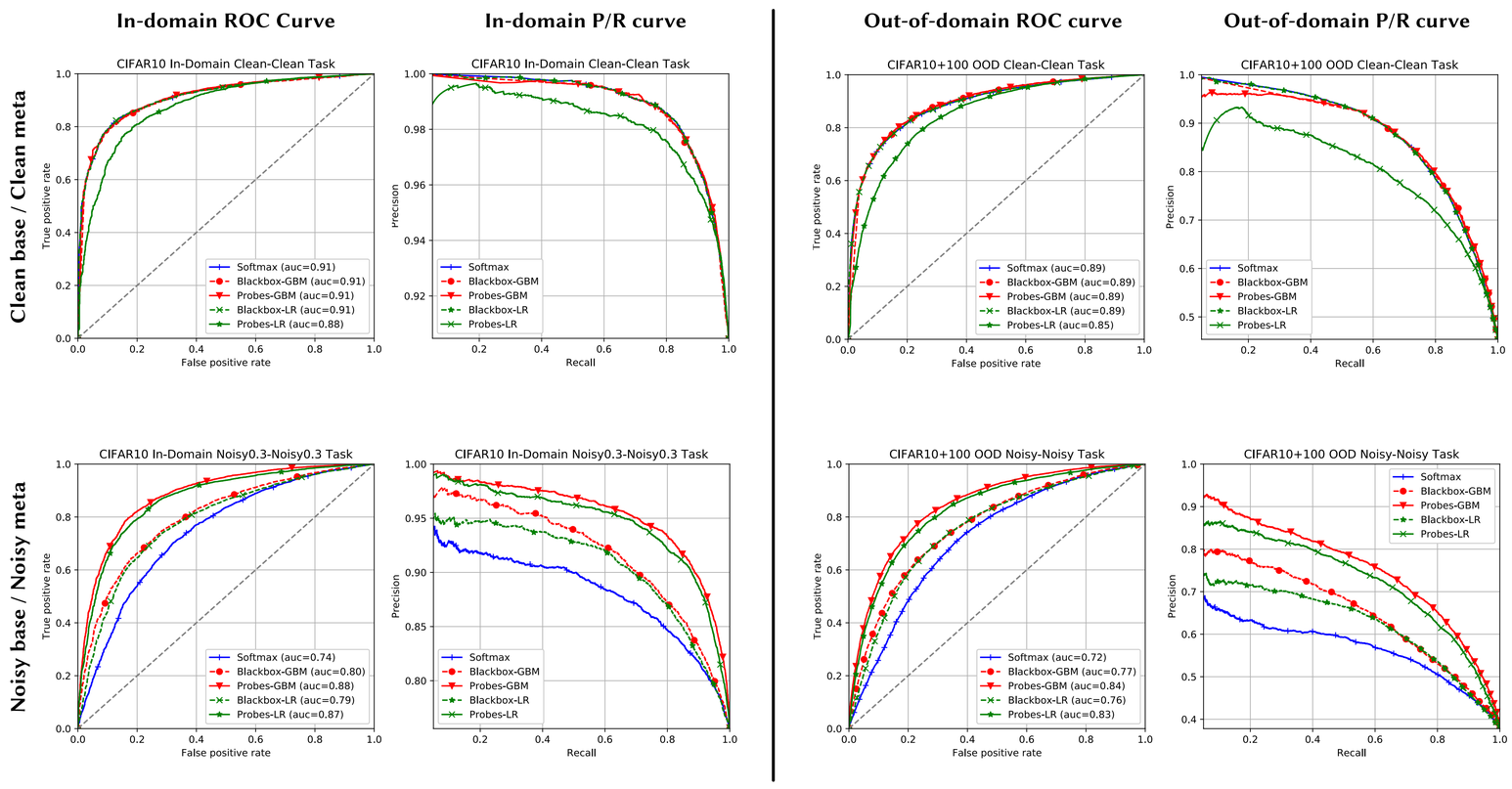
1. **In-domain task:** Filter out predictions considered uncertain
2. **Out-of-domain task:** Filter out out-of-domain samples (**CIFAR-100**)

1. **Clean base / clean meta:** Original data
2. **Noisy base / noisy meta:** 30% of labels corrupted.

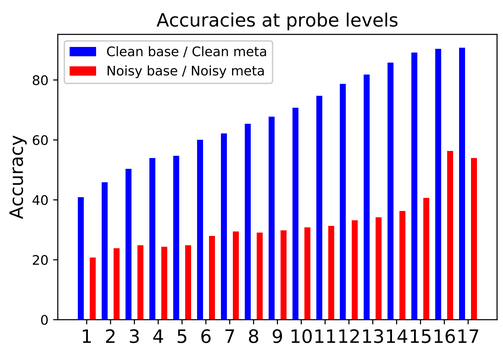




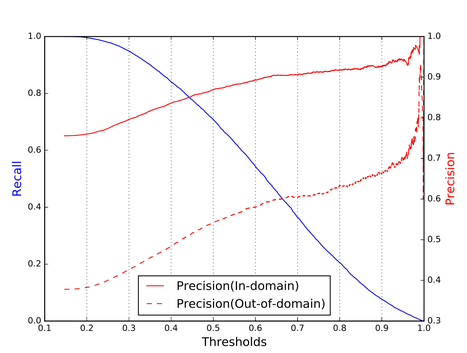
Base model: ResNet for image classification (CIFAR-10 dataset).



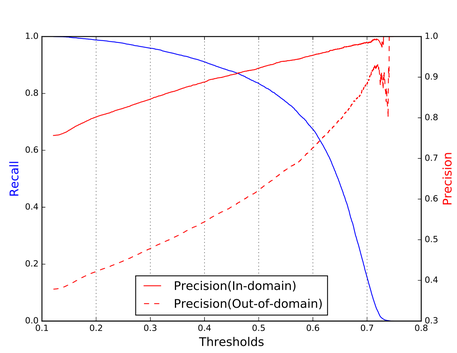
ROC and Precision/Recall curves under various settings.



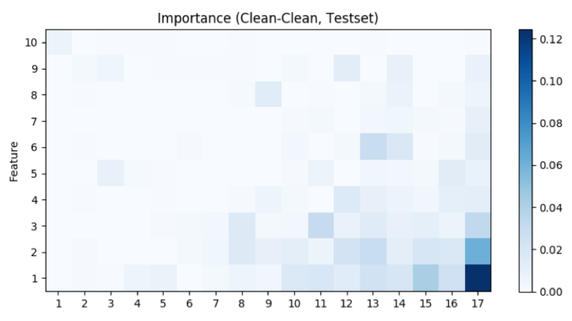
Accuracies at various probe levels (1 – 17)



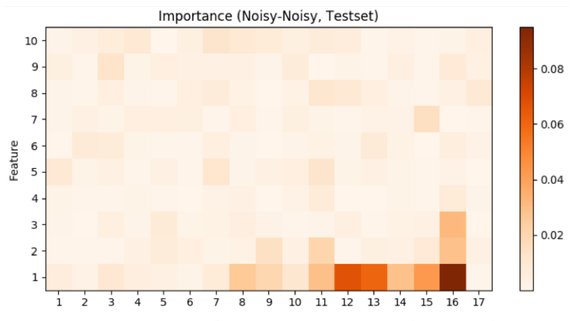
P/R w.r.t. threshold (Blackbox)



P/R w.r.t. threshold (Whitebox)



Feature informativeness (Clean)



Feature informativeness (Noisy)

References

- G. Alain, Y. Bengio (2016). *arXiv.*
- J. H. Friedman (2001). *Ann. Stat.*
- Y. Geifman, R. El-Yaniv (2017). *NeurIPS.*

Contact

Jiří Navrátil: jiri@us.ibm.com

IBM **Research**