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 \( \hat{y} = F(x) \);
- **Meta-model**: Confidence score \( z = G(x, \Theta_F) \). 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 \( \hat{y} \).

**Softmax response** (Geifman and El-Yaniv, 2017):
\[
z = P(y^*|x, \Theta_F) = \max_{i} \hat{y}_i.
\]

**Whitebox**: Meta-model \( G \) assumes full access to the internals of \( F \):
\[
z = G(x_1, x_2, \ldots, x_n).
\]

Linear Classifier Probes

For each intermediate result \( x_k \), train a linear classifier probe (Alain and Bengio, 2016) \( F_k \) to predict the correct class \( y \) using only that result:
\[
\hat{y}_k = F_k(x_k) = \text{softmax}(W_k x_k + b_k).
\]

Structure of meta-model:
- Logistic regression;
- Gradient boosting machine (Friedman, 2001).

Experiments

**CIFAR-10**: 50k training samples partitioned into (Train-base: 30k; Train-meta: 10k; Dev: 10k).
- **In-domain task**: Filter out predictions considered uncertain.
- **Out-of-domain task**: Filter out out-of-domain samples (CIFAR-100).
- **Clean base / clean meta**: Original data
- **Noisy base / noisy meta**: 30% of labels corrupted.

References

- Y. Geifman, R. El-Yaniv (2017). *NeurIPS.*

Contact

Jiří Navrátil: jiri@us.ibm.com

IBM Research