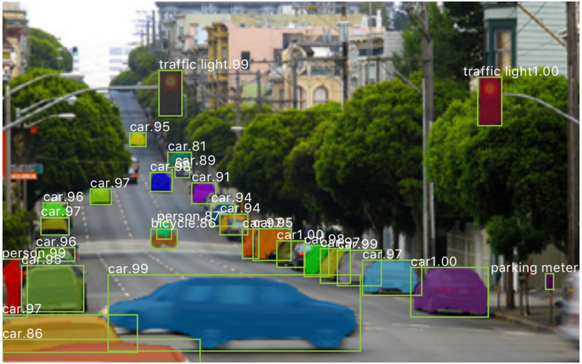
PAC Confidence Predictions for Deep Neural Network Classifier

Sangdon Park, Shuo Li, Insup Lee, and Osbert Bastani

PRECISE Center
University of Pennsylvania

ICLR 2021

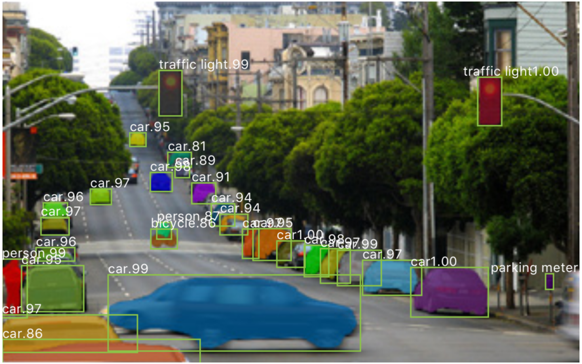
Quantifying Uncertainty of Deep Neural Network Predictions



Source: Mask RCNN

✓ Accurate label predictions

Quantifying Uncertainty of Deep Neural Network Predictions

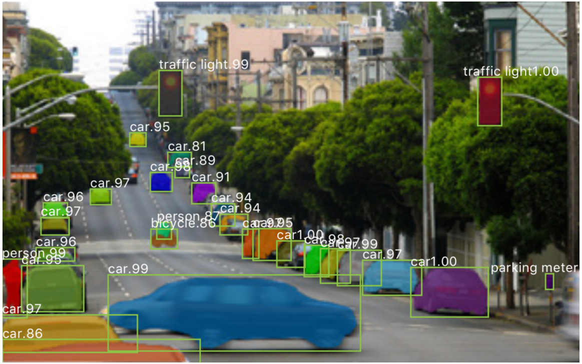


Source: Mask RCNN

✓ Accurate label predictions

✗ No finite-sample correctness guarantees on confidence prediction

Quantifying Uncertainty of Deep Neural Network Predictions



Source: Mask RCNN

- ✓ Accurate label predictions
- ✗ No finite-sample correctness guarantees on confidence prediction

How to predict confidences with finite sample guarantees?

PAC Calibration

PAC calibration: The goal of PAC calibration is to find a confidence coverage predictor $\hat{C}$ such that it contains true confidence with high probability—*i.e.*,

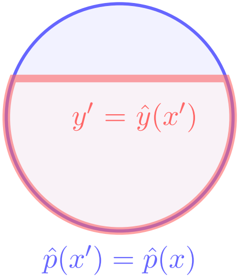
$$(\text{true confidence}) \in \hat{C}(x; \hat{f})$$

Here, a pretrained predictor $\hat{f}$ is given.

True Confidence

True confidence on x associated with $\hat{f}$ (due to the known calibration definition [DeGroot and Fienberg, 1983, Zadrozny and Elkan, 2002, Park et al., 2020]):

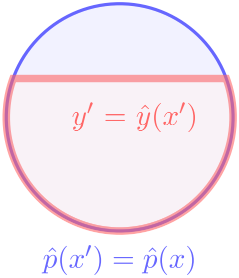
$$c^*_{\hat{f}}(x) := \mathbb{P}_{(x',y')\sim D}\left[y' = \hat{y}(x') \mid \hat{p}(x') = \hat{p}(x)\right]$$

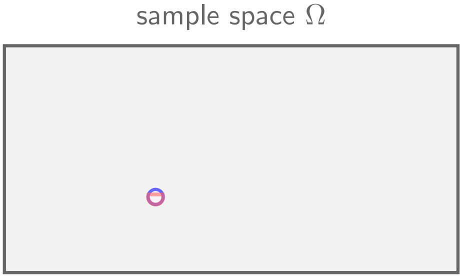


True Confidence

True confidence on x associated with $\hat{f}$ (due to the known calibration definition [DeGroot and Fienberg, 1983, Zadrozny and Elkan, 2002, Park et al., 2020]):

$$c^*_{\hat{f}}(x) := \mathbb{P}_{(x',y') \sim D} \left[y' = \hat{y}(x') \mid \hat{p}(x') = \hat{p}(x) \right]$$



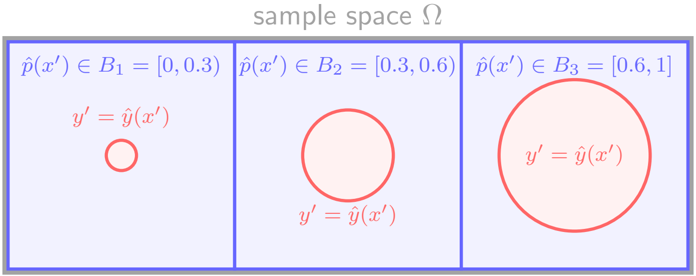


Estimating the true confidence with finite samples is challenging

"Coarsen" True Confidence

$$c_{\hat{f}}(x) := \mathbb{P}_{(x',y')\sim D}\left[y' = \hat{y}(x') \,\middle|\, \hat{p}(x') \in B_{\kappa_{\hat{f}}(x)}\right]$$

- $\kappa_{\hat{f}} : \mathcal{X} \to \{1, 2, \ldots, K\}$: the index of the bin for x—*i.e.*, $\hat{p}(x) \in B_{\kappa_{\hat{f}}(x)}$



PAC Calibration

Definition

Given $\delta \in \mathbb{R}_{>0}$ and $n \in \mathbb{N}$, $\hat{C}$ is *probably approximately correct (PAC)* if for all D

$$\overbrace{c_{\hat{f}}(x) \in \hat{C}(x; \hat{f}, Z_n)}^{\text{approximately correct}}$$

For a formal connection to the PAC learning theory, see Appendix A.

PAC Calibration

Definition

Given $\delta \in \mathbb{R}_{>0}$ and $n \in \mathbb{N}$, $\hat{C}$ is *probably approximately correct (PAC)* if for all D

$$\bigwedge_{x \in \mathcal{X}} \overbrace{c_{\hat{f}}(x) \in \hat{C}(x; \hat{f}, Z_n)}^{\text{approximately correct}}$$

For a formal connection to the PAC learning theory, see Appendix A.

PAC Calibration

Definition

Given $\delta \in \mathbb{R}_{>0}$ and $n \in \mathbb{N}$, $\hat{C}$ is *probably approximately correct (PAC)* if for all D

$$\underbrace{\mathbb{P}_{Z_n \sim D^n}\left[\bigwedge_{x \in \mathcal{X}} \overbrace{c_{\hat{f}}(x) \in \hat{C}(x; \hat{f}, Z_n)}^{\text{approximately correct}}\right] \geq 1 - \delta.}_{\text{probably approximately correct}}$$

For a formal connection to the PAC learning theory, see Appendix A.

PAC Calibration

Definition

Given $\delta \in \mathbb{R}_{>0}$ and $n \in \mathbb{N}$, $\hat{C}$ is *probably approximately correct (PAC)* if for all D

$$\underbrace{\mathbb{P}_{Z_n \sim D^n}\left[\bigwedge_{x \in \mathcal{X}} \overbrace{c_{\hat{f}}(x) \in \hat{C}(x; \hat{f}, Z_n)}^{\text{approximately correct}}\right] \geq 1 - \delta.}_{\text{probably approximately correct}}$$

For a formal connection to the PAC learning theory, see Appendix A.

Problem

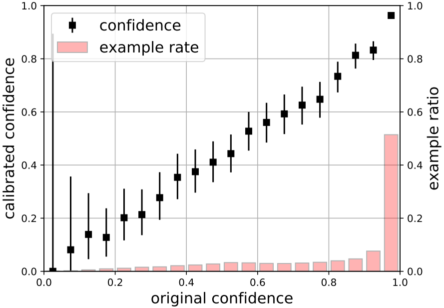
Find a PAC confidence coverage predictor $\hat{C}$, while ensuring its size is small.

Our Approach

Main idea: Coarsened true confidence is the parameter of a Binomial distribution

Our Approach

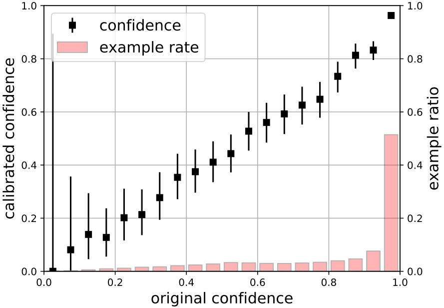
Main idea: Coarsened true confidence is the parameter of a Binomial distribution



Pictorial representation of $\hat{C}$

Our Approach

Main idea: Coarsened true confidence is the parameter of a Binomial distribution



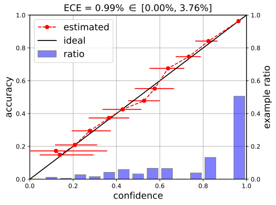
Pictorial representation of $\hat{C}$

Theorem

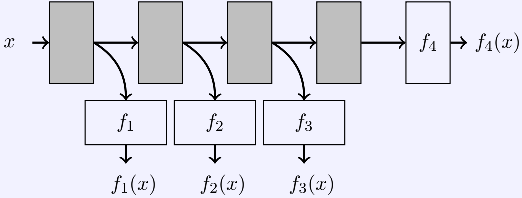
$\hat{C}$ satisfies the PAC property.

Applications

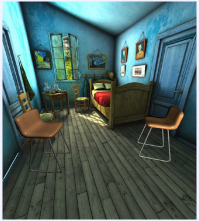
Classifier calibration



Fast DNN Inference



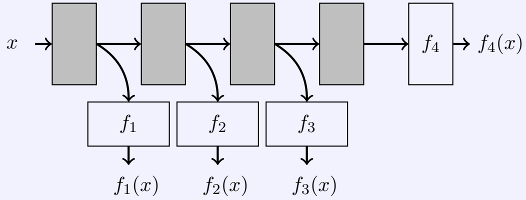
Safe Planning



Applications

Classifier calibration

Fast DNN Inference



Safe Planning

Application: Fast DNN Inference

Motivation:

- AlexNet: fast but inaccurate (top-1 error: 43.45%)
- ResNet152: slow but accurate (top-1 error: 21.69%)

Application: Fast DNN Inference

Motivation:

- AlexNet: fast but inaccurate (top-1 error: 43.45%)
- ResNet152: slow but accurate (top-1 error: 21.69%)

Can we combine the two models to improve inference speed
while maintaining high accuracy?

Model Composition

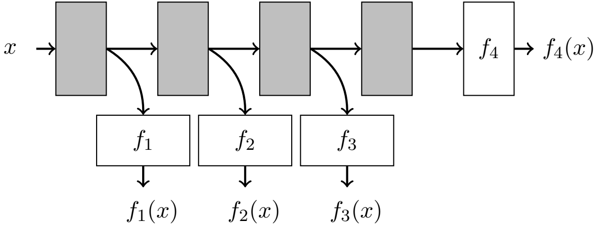


Figure: A composed model in a cascading way for $M = 4$.

Composed classifier:

$$\hat{y}_C(x; \gamma_{1:M-1}) := \begin{cases} \hat{y}_1(x) & \text{if } \hat{p}_1(x) \geq \gamma_1 \\ \hat{y}_2(x) & \text{if } \hat{p}_1(x) < \gamma_1 \land \hat{p}_2(x) \geq \gamma_2 \\ & \vdots \\ \hat{y}_{M-1}(x) & \text{if } \bigwedge_{m=1}^{M-2} (\hat{p}_m(x) < \gamma_m) \land \hat{p}_{M-1}(x) \geq \gamma_{M-1} \\ \hat{y}_M(x) & \text{otherwise,} \end{cases}$$

Our Approach

$$\min_{\gamma_{1:M-1}} \quad \text{(inference time)}$$

$$\text{subj. to} \quad p_{\text{error}} := \underbrace{\mathbb{P}_{(x,y)\sim D}\left[\hat{y}_C(x;\gamma_{1:M-1}) \neq y\right]}_{\text{error}_{\text{composed model}}} - \underbrace{\mathbb{P}_{(x,y)\sim D}\left[\hat{y}_M(x) \neq y\right]}_{\text{error}_{\text{slow model}}} \leq \xi.$$

Our Approach

$$\min_{\gamma_{1:M-1}} \quad \text{(inference time)}$$

$$\text{subj. to} \quad p_{\text{error}} := \underbrace{\mathbb{P}_{(x,y)\sim D}\left[\hat{y}_C(x;\gamma_{1:M-1}) \neq y\right]}_{\text{error}_{\text{composed model}}} - \underbrace{\mathbb{P}_{(x,y)\sim D}\left[\hat{y}_M(x) \neq y\right]}_{\text{error}_{\text{slow model}}} \leq \xi.$$

Theorem

We have $p_{\text{error}} \leq \xi$ with probability at least $1-\delta$ over $Z \sim D^n$.

Experiment: Comparison

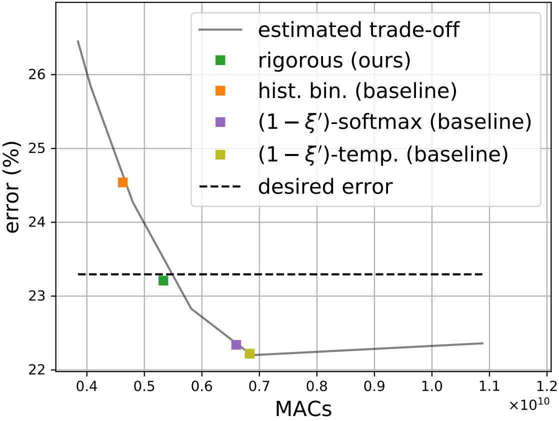


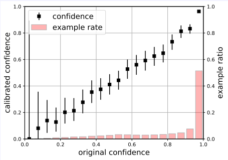
Figure: $M = 2,\ N = 20,000,\ \xi = 0.02,\ \delta = 10^{-2}$

- MACs: Multiplication ACcumulation operationS

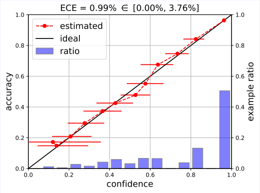
Conclusion

PAC Calibration Problem

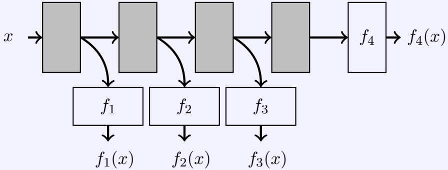
Approach: CP interval + hist. binning



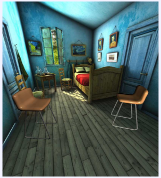
Application 1: Classifier Calibration



Application 2: Fast DNN Inference



Application 3: Safe Planning



Feel free to visit to our poster session if you have more questions!

References I

[DeGroot and Fienberg, 1983] DeGroot, M. H. and Fienberg, S. E. (1983).
The comparison and evaluation of forecasters.
Journal of the Royal Statistical Society: Series D (The Statistician), 32(1-2):12–22.

[Park et al., 2020] Park, S., Bastani, O., Weimer, J., and Lee, I. (2020).
Calibrated prediction with covariate shift via unsupervised domain adaptation.
In *The 23rd International Conference on Artificial Intelligence and Statistics*.

[Zadrozny and Elkan, 2002] Zadrozny, B. and Elkan, C. (2002).
Transforming classifier scores into accurate multiclass probability estimates.
In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699. ACM.