Improvement of Rejection for AI Safety Through Loss-Based Monitoring

Daniel Scholz[1, 3], Florian Hauer[2], Klaus Knobloch[1], Christian Mayr[3]
[1]Infineon Technologies Dresden
[2]Infineon Technologies München
[3]Technische Universität Dresden
24.07.2022

KI-ASIC

Bundesministerium für Bildung und Forschung

TECHNISCHE UNIVERSITÄT DRESDEN

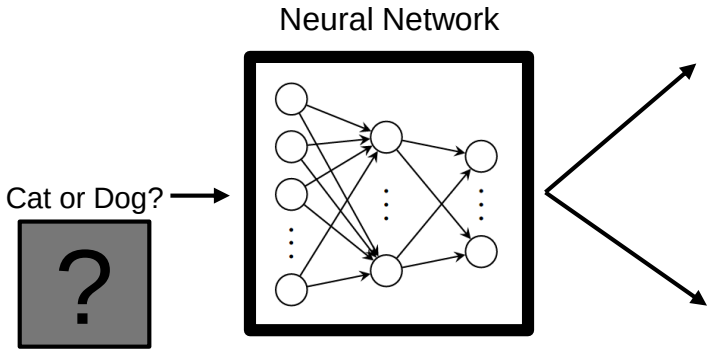
Infineon

Motivation

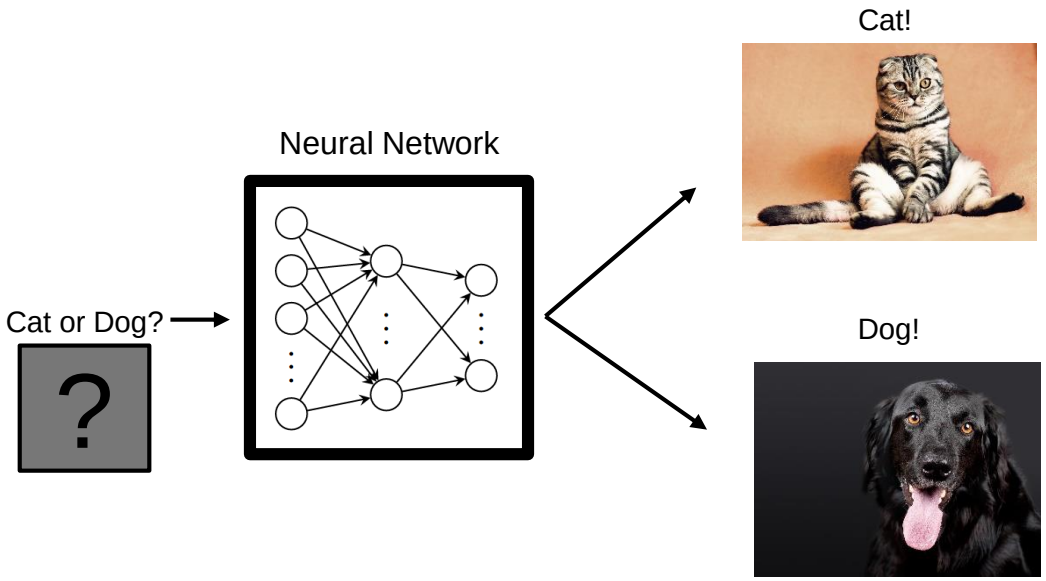
- AI increasingly included in safety critical domains
- Decisions based on output probabilities
- Rejection was proposed but often fails to catch high output probability errors

- Presented approach: Rejection performed by monitoring model which is trained on loss

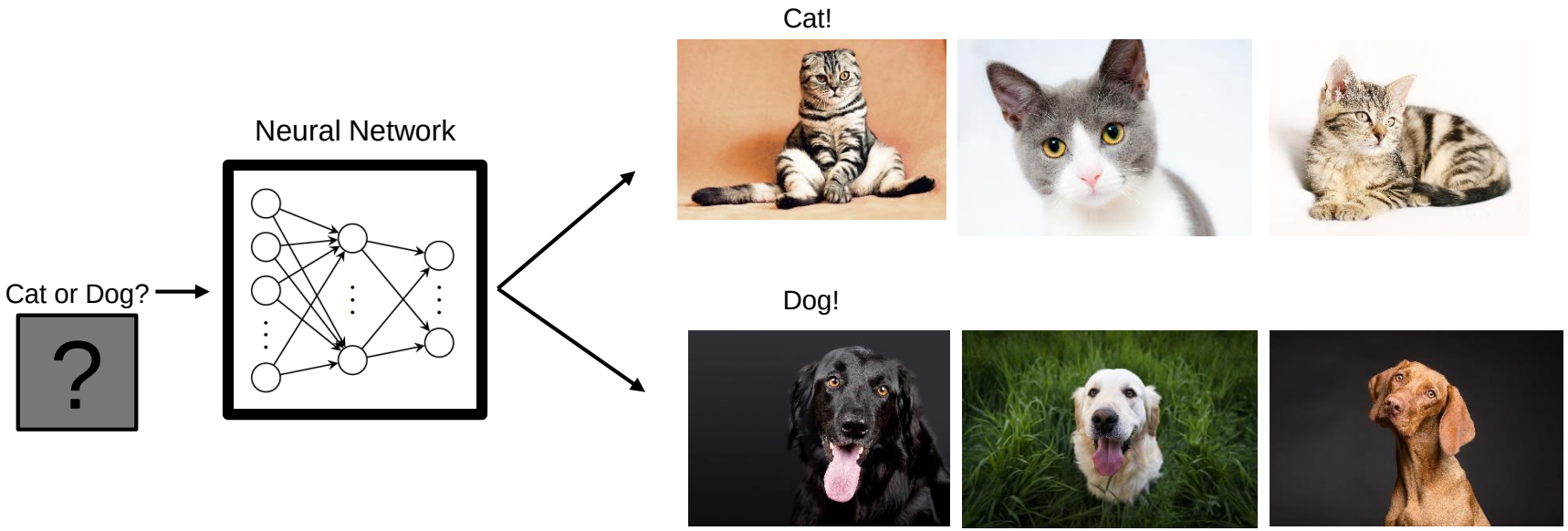
Thought Experiment



Thought Experiment

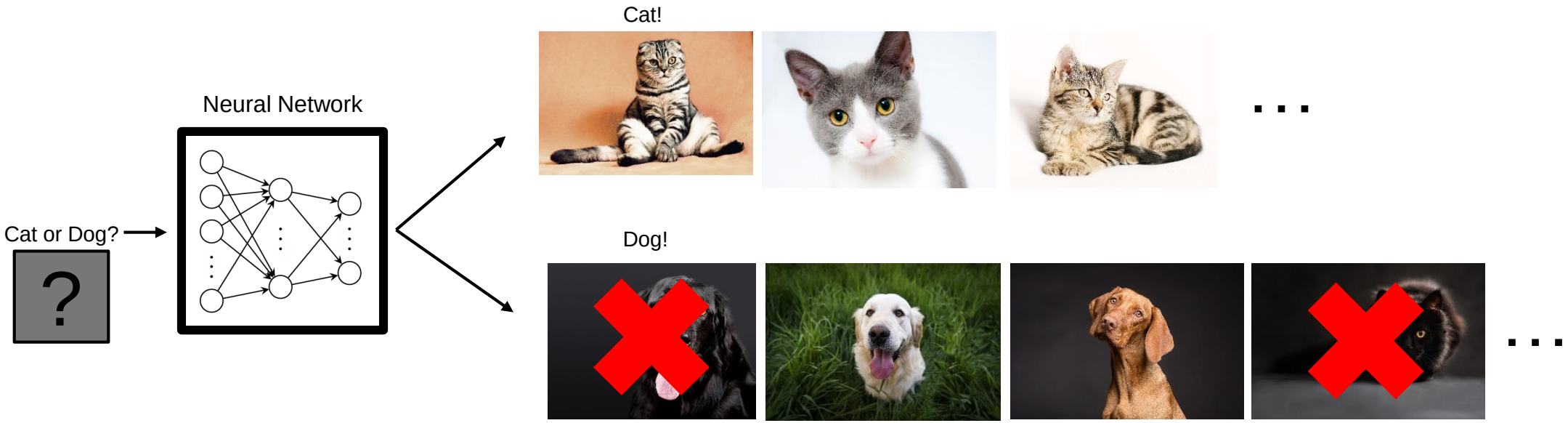


Thought Experiment



Thought Experiment

Thought Experiment



→ 0% error rate but rejected inputs

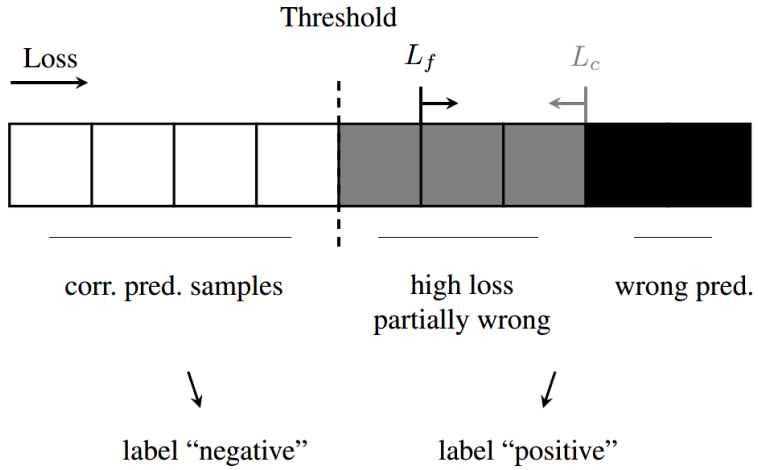
Loss-Based Rejection

- Cross-entropy loss with one-hot-encoding

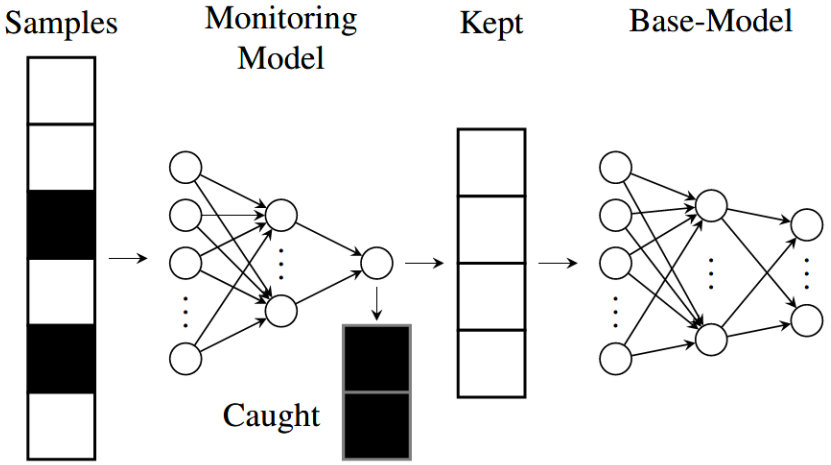
$$L = -\sum_{i=1}^{M} y_i log_e(p_i) \qquad L = -log_e(p_c)$$

- Bounds for correct and false predictions

$$L_c < -log_e(\frac{1}{n}) \qquad L_f > -log_e(0.5)$$



Ideal behavior:



Performance Evaluation

- Goal: Reduced error-rate, targeting safety critical applications

- Metric for monitoring model not sufficient
  - ROC curve will not capture effectiveness
  - Rejection based on loss does not specifically divide correct/wrong samples

- Two meaningful metrics [1]
  - Remaining accuracy rate (rar)
  - Remaining error rate (rer)

$$rar = \hat{\phi} \cdot (1 - \hat{r}) = \hat{\phi} - rer$$

$$rer = \hat{r} \cdot \hat{\phi}$$

  - Defined via risk and coverage [2]

$$\hat{r}(f, g|S_m) \triangleq \frac{\frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i) g(x_i)}{\hat{\phi}(g|S_m)} \qquad \hat{\phi}(g|S_m) \triangleq \frac{1}{m}\sum_{i=1}^{m} g(x_i)$$

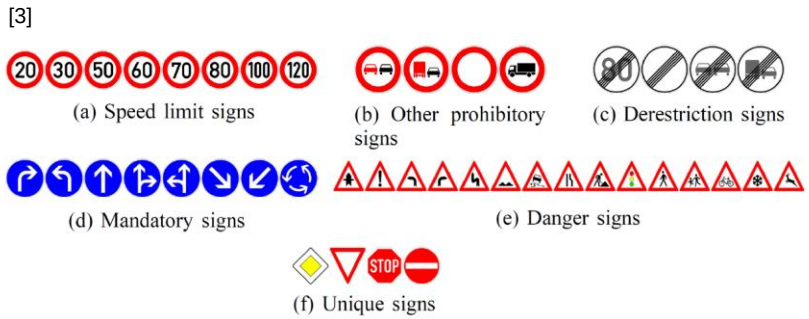
- Trivial approaches:
  - Aim for high coverage: keep all → rar will be same as base acc. But error rate unchanged
  - Aim for low error rate: reject all → rer will be 0 but rar, too

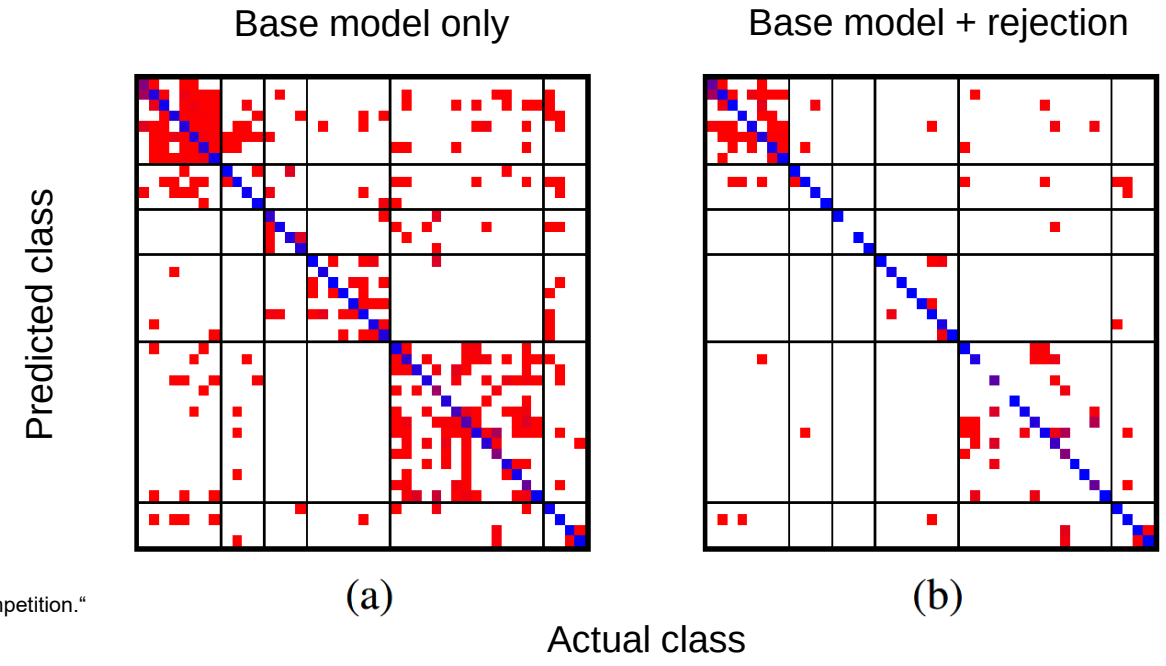
[1] Henne, Maximilian, et al. "Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics." SafeAI@ AAAI. 2020.
[2] Geifman, Yonatan, and Ran El-Yaniv. "Selective classification for deep neural networks." Advances in neural information processing systems 30 (2017).

GTSRB Confusion Matrix

- LeNet-5 architecture tested with German Traffic Sign Recognition Benchmark (GTSRB)

- Relative color code of given class
  - White equal to zero
  - Blue equal to one
  - Increasing from red to blue





[3] Stallkamp, Johannes, et al. "The German traffic sign recognition benchmark: a multi-class classification competition." *The 2011 international joint conference on neural networks*. IEEE, 2011.

Fashion-MNIST Confusion Matrix

Base model only

Base model + rejection

Predicted \ Actual	Negative	Positive
Negative	5809 85%	482 15%
Positive	1038 15%	2671 85%

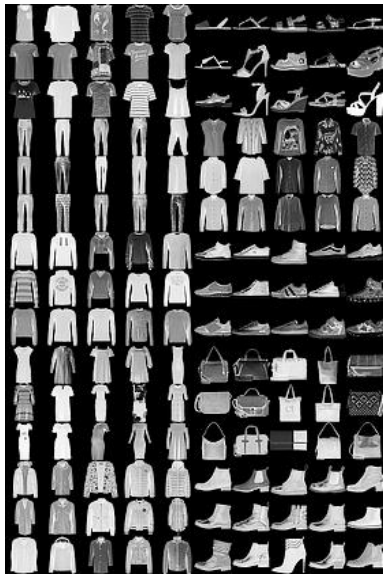
(a)

Predicted \ Actual	T-shirt/top	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
T-shirt/top	858 86%	4 0%	13 1%	21 2%	0 0%	0 0%	160 16%	0 0%	3 0%	0 0%
Trouser	1 0%	965 96%	1 0%	3 0%	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
Pullover	12 1%	1 0%	815 81%	17 2%	150 15%	0 0%	84 8%	0 0%	1 0%	0 0%
Dress	27 3%	22 2%	12 1%	918 92%	46 5%	0 0%	31 3%	0 0%	7 1%	0 0%
Coat	4 0%	3 0%	63 6%	15 1%	742 74%	0 0%	66 7%	0 0%	1 0%	0 0%
Sandal	2 0%	0 0%	1 0%	0 0%	0 0%	972 97%	0 0%	21 2%	5 0%	11 1%
Shirt	83 8%	4 0%	91 9%	19 2%	55 5%	0 0%	645 64%	0 0%	2 0%	1 0%
Sneaker	0 0%	0 0%	0 0%	0 0%	0 0%	16 2%	0 0%	959 96%	2 0%	37 4%
Bag	13 1%	1 0%	4 0%	5 0%	7 1%	0 0%	14 1%	0 0%	979 98%	0 0%
Ankle boot	0 0%	0 0%	0 0%	2 0%	0 0%	12 1%	0 0%	20 2%	0 0%	951 95%

(b)

Predicted \ Actual	T-shirt/top	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
T-shirt/top	482 97%	0 0%	3 1%	5 1%	0 0%	0 0%	32 14%	0 0%	1 0%	0 0%
Trouser	1 0%	942 99%	1 0%	1 0%	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
Pullover	3 1%	0 0%	334 96%	1 0%	25 23%	0 0%	8 4%	0 0%	0 0%	0 0%
Dress	4 1%	11 1%	1 0%	638 99%	3 3%	0 0%	6 3%	0 0%	1 0%	0 0%
Coat	0 0%	0 0%	5 1%	0 0%	79 71%	0 0%	5 2%	0 0%	0 0%	0 0%
Sandal	1 0%	0 0%	0 0%	0 0%	0 0%	910 98%	0 0%	12 2%	5 1%	9 1%
Shirt	1 0%	0 0%	3 1%	0 0%	1 1%	0 0%	171 75%	0 0%	1 0%	1 0%
Sneaker	0 0%	0 0%	0 0%	0 0%	0 0%	11 1%	0 0%	757 97%	2 0%	10 1%
Bag	7 1%	0 0%	0 0%	1 0%	3 3%	0 0%	5 2%	0 0%	906 99%	0 0%
Ankle boot	0 0%	0 0%	0 0%	0 0%	0 0%	9 1%	0 0%	9 1%	0 0%	864 98%

(c)



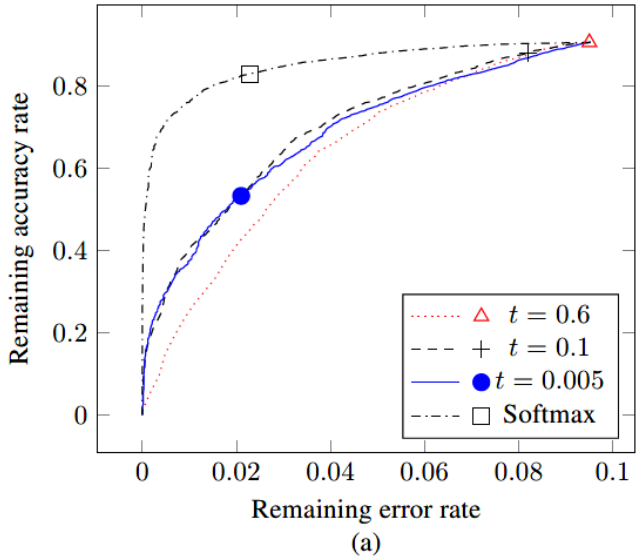
[4]

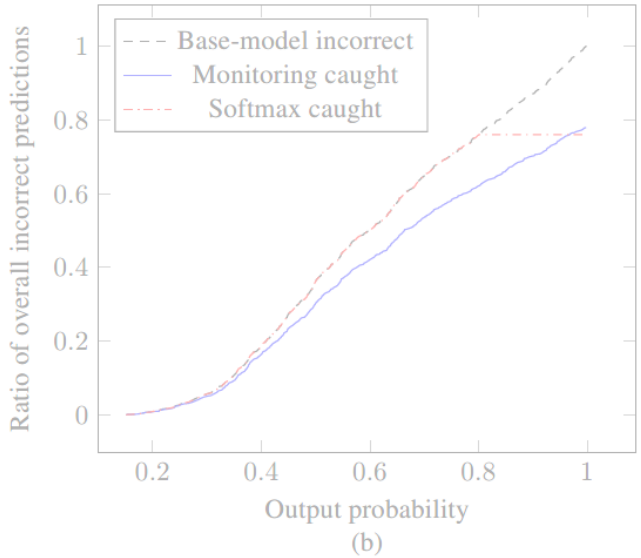
[4] Xiao, Han, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms." arXiv preprint arXiv:1708.07747 (2017).

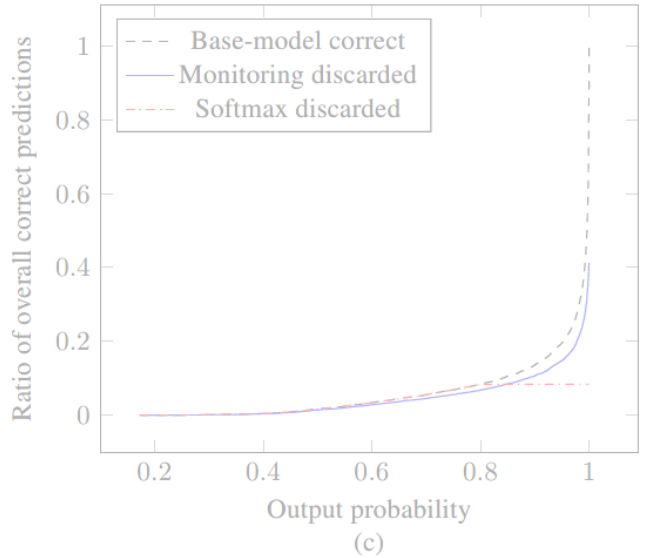
Results GTSRB

› Comparison of monitoring models trained with different thresholds t

› Rejection based on softmax misses high output probabilities
› Trade-off is harsher for loss-based rejection but whole range of output probabilities is caught



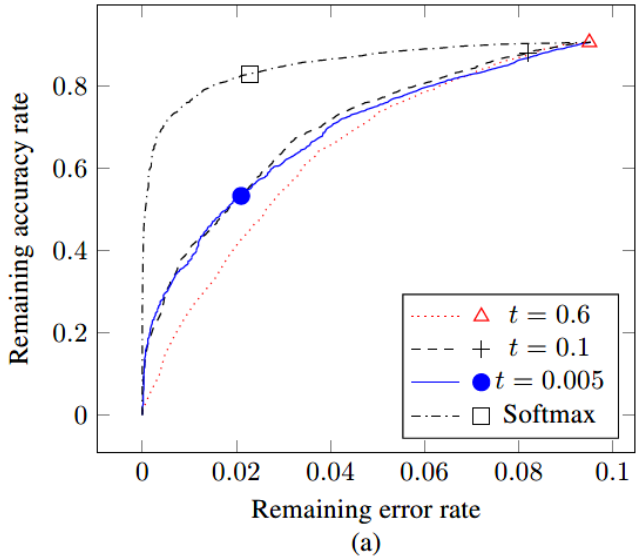


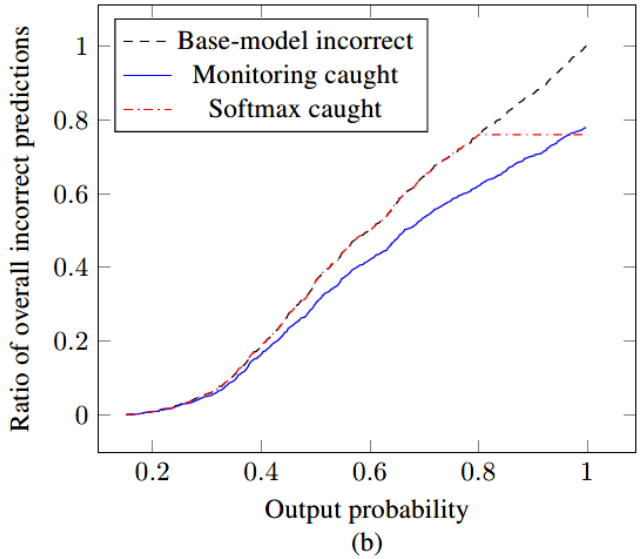


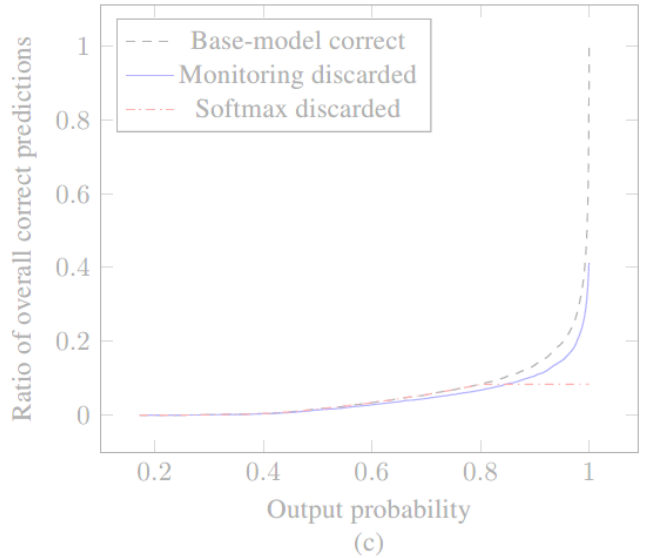
Results GTSRB

› Comparison of monitoring models trained with different thresholds t

› Rejection based on softmax misses high output probabilities
› Trade-off is harsher for loss-based rejection but whole range of output probabilities is caught



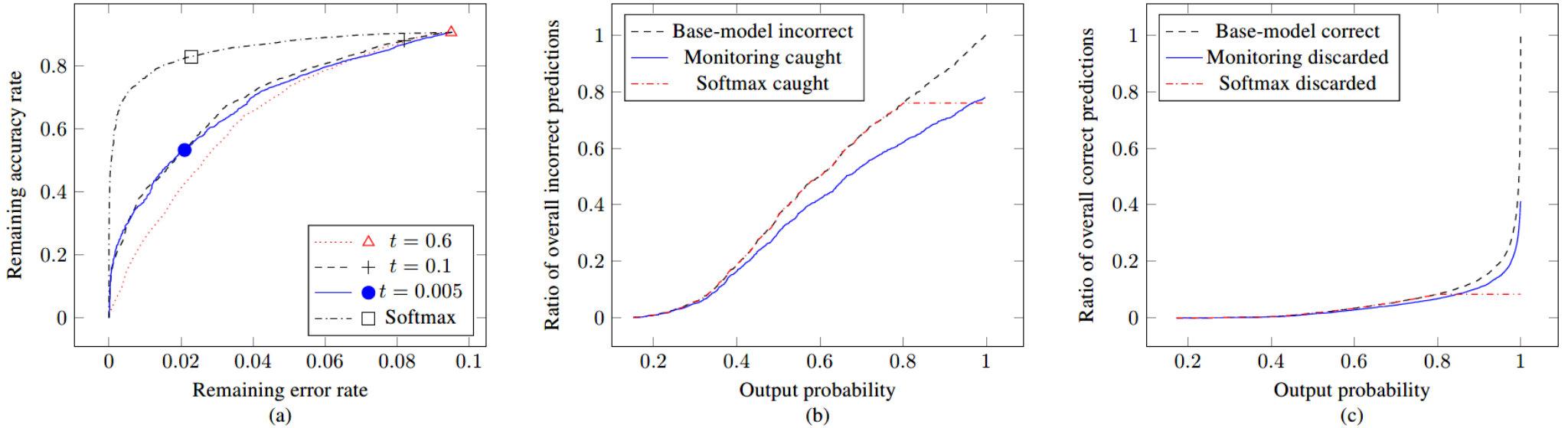




Results GTSRB

› Comparison of monitoring models trained with different thresholds t

› Rejection based on softmax misses high output probabilities
› Trade-off is harsher for loss-based rejection but whole range of output probabilities is caught



Summary

- Softmax threshold has better rer/rar trade-off but inherently fails in high output probability errors
- Data that base model has not seen needed to train monitoring model (excluding test data)

- Apart from harsher trade-off, loss based rejection superior for safety critical domains
  - Errors with all output probabilities are caught
  - Very high output probabilities are successfully caught
  - Tunable performance without missing probability sectors

- Applicable to black-box model, no re-training of base model

- Chosable application specific loss threshold to reach required coverage and error-rate

Part of your life. Part of tomorrow.