# Confidence Metrics for Classification by deep Neural Networks

March 15, 2019

Adam Oberman
with Chris Finlay
Math and Stats, McGill

# Challenges for deep learning

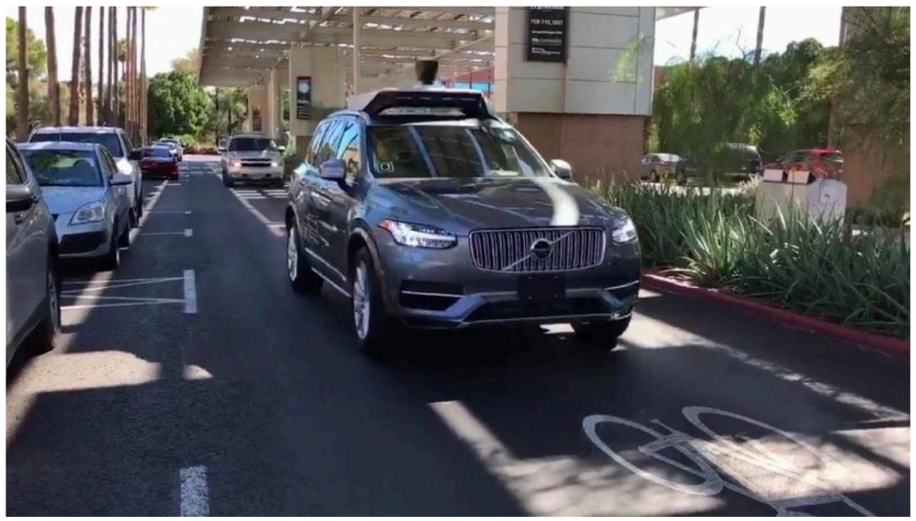
"It is not clear that the existing AI paradigm is immediately amenable to any sort of software engineering validation and verification. This is a serious issue, and is a potential roadblock to DoD's use of these modern AI systems, especially when considering the liability and accountability of using AI"

JASON report

## Self-Driving Uber Hits, Kills Pedestrian in Arizona

*The Uber vehicle was operating in autonomous mode with a human behind the wheel in Tempe, Arizona, when the incident occurred overnight.*

By Angela Moscaritolo March 19, 2018 2:07PM EST



Learning networks. Two things to make clear to the reader (1) We don't know how Deep Learning works and (2) when it makes a prediction, we don't have an explanation why it arrived at that prediction. That is just scratching the

Fact:
the output "probabilities" of neural networks for image classification are not the probabilities that the classification is correct.

this is correct. unlike other classifiers, e.g. Naive Bayes, there is no interpretation of the output of the network as a probability

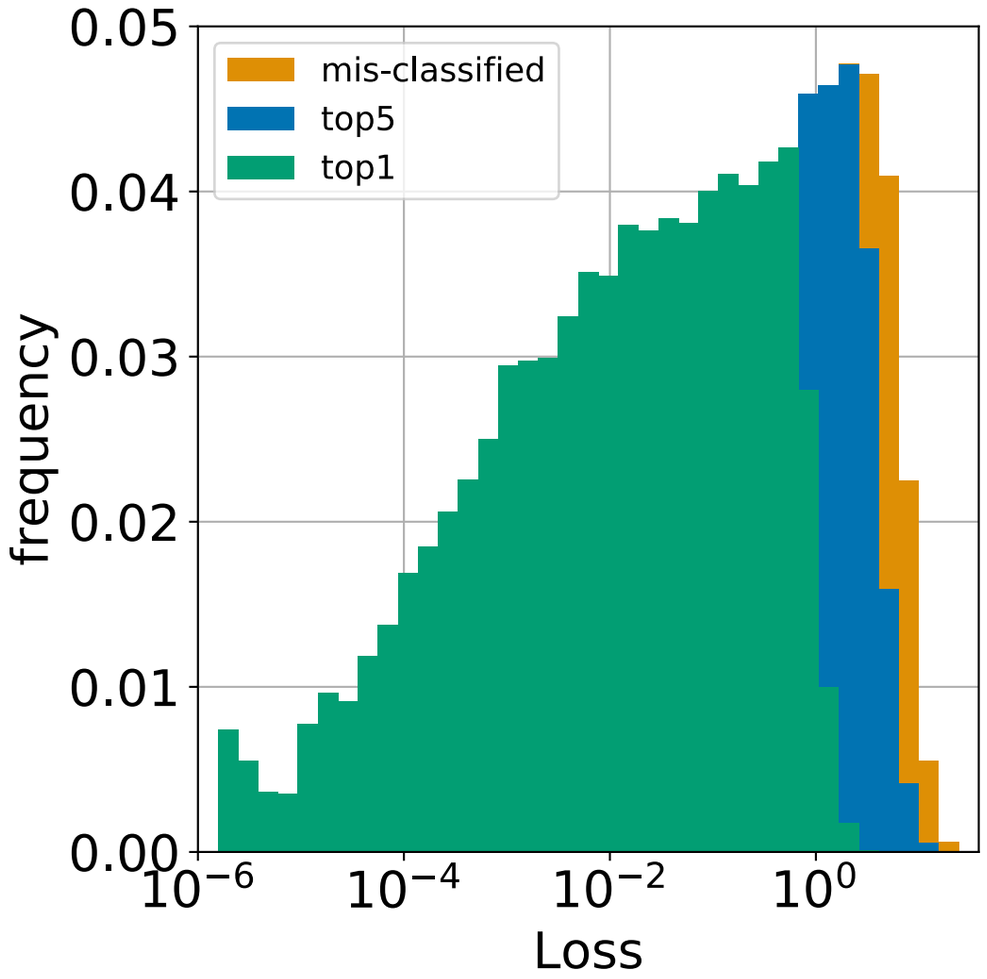
Misinterpretation:
the output probabilities are not meaningful predictors of classification error.

in fact, we can extract useful information from the output, combined with the statistics of the loss on the test set.

Fact:
if a neural network generalizes well , and is gives correct classifications 95% of the time (say), then we can estimate probability correct based on p_max

How?
Suppose, for the sake of argument, that we are given, along with a prediction, the value of the loss, (but not the correct label).

# Value of the loss




Then we could have an imperfect, but much better idea of the probability that the prediction is correct.

- For small values < 0.8 of the loss, always correct.
- For large values > 3 always incorrect
- For intermediate values, make a histogram, with probability correct in each bin.

We have easy to compute metrics which are almost as good as the loss

$$\mathsf{Entropy}(p) = -\sum_i p_i \log(p_i)$$

$$-\log p_{\mathsf{max}}$$

$$-\log \sum p_{1:5}$$

# How to measure predictive value?

Before presenting the results, we need to explain a way to measure the quality of a metric. This will allow us to compare the predictive value of different metrics on different data sets.

We will show that we can easily compute metrics which give great than10X improved confidence.

Moreover, we can define a simplified "green, yellow, red" zones, where the confidence is very high, moderate, and very low.

In the latter case, we value the fact that we have increased confidence in the probability of an error.

# Measuring the odds

Consider a Bernoulli random variable $X = B(p)$. The odds for $X$ are given by

$$O(p) = \frac{p}{1-p}$$

In the case that $p = 1$, we define, for small $\epsilon > 0$, the regularized odds to be

$$O^{\epsilon}(p) = \frac{p}{\max(1-p, \epsilon)}$$

Now consider a test, $Y = B(p_y)$, for which

$$P(X = 1|Y = 1) = p^+$$

Then the odds, given the test succeeds, are

$$O(p^+) = \frac{p^+}{1-p^+}$$

# Bayes Factor measures the value of information

Given a test $Y$, the Bayes factor for $Y$ is given by

$$BF(X \mid Y) = \max\left(\frac{O^+}{O}, \frac{O}{O^+}\right)$$

*Example* 1.2. For example, if $p = .95$ then $O(p) = \frac{p}{1-p} = 19$. If $p^+ = .99$ then $O(p^+) = 99$, and $BF(p, p^+) = 5.25$. On the other hand, if $p^+ = 2/3$ then $O(p^+) = 2$ and $BF(p, p^+) = 9.5$

Whether to bet for or against depends on the new odds.

The Bayes Factor tells us how much our expected winnings increase, if we know the value of the test (and bet correctly).

# Predicting preference for Voice vs Text

Consider the situation where you have exchanged phone numbers with someone, and you wish to contact them. The question is whether to send a text message or phone their number. Approximately 95% of people prefer to message. Let $X$ be the probability that a person prefers to message. The expected value and odds for $X$ is given by

$$p_X = 0.95, \qquad O(p_X) = 19$$

Histogram for age:

$$\begin{cases} Y_1 = 1_{\{U<20\}}, & \mathbb{E}[Y_1] = .4 \\ Y_2 = 1_{\{20\leq U\leq 65\}}, & \mathbb{E}[Y_2] = .5 \\ Y_3 = 1_{\{65<U\}}, & \mathbb{E}[Y_2] = .1 \end{cases}$$

# Value of age information

$$\begin{cases} p(X \mid Y_1) = .999, & O(p_{X,Y_1}) = 999 \\ p(X \mid Y_2) = .94, & O(p_{X,Y_2}) = 15.7 \\ p(X \mid Y_3) = .9, & O(p_{X,Y_2}) = 9 \end{cases}$$

What is the expected value of knowing the age (without knowing in advance the range)? Expected Bayes Ratio.

$$\begin{cases} BR^1 = 999/18.84 = 53 \\ BR^2 = 18.84/15.6 = 1.2 \\ BR^3 = 18.84/4 = 4.7 \end{cases}$$

$$\mathbb{E}\left[BR(X|A)\right] = 53 \times .4 + 1.2 \times .5 + 4.7 \times .1 = 22.3$$

# Less valuable information: where they live

let $Y_1, Y_2, Y_3$ be the histogram random variables. Suppose

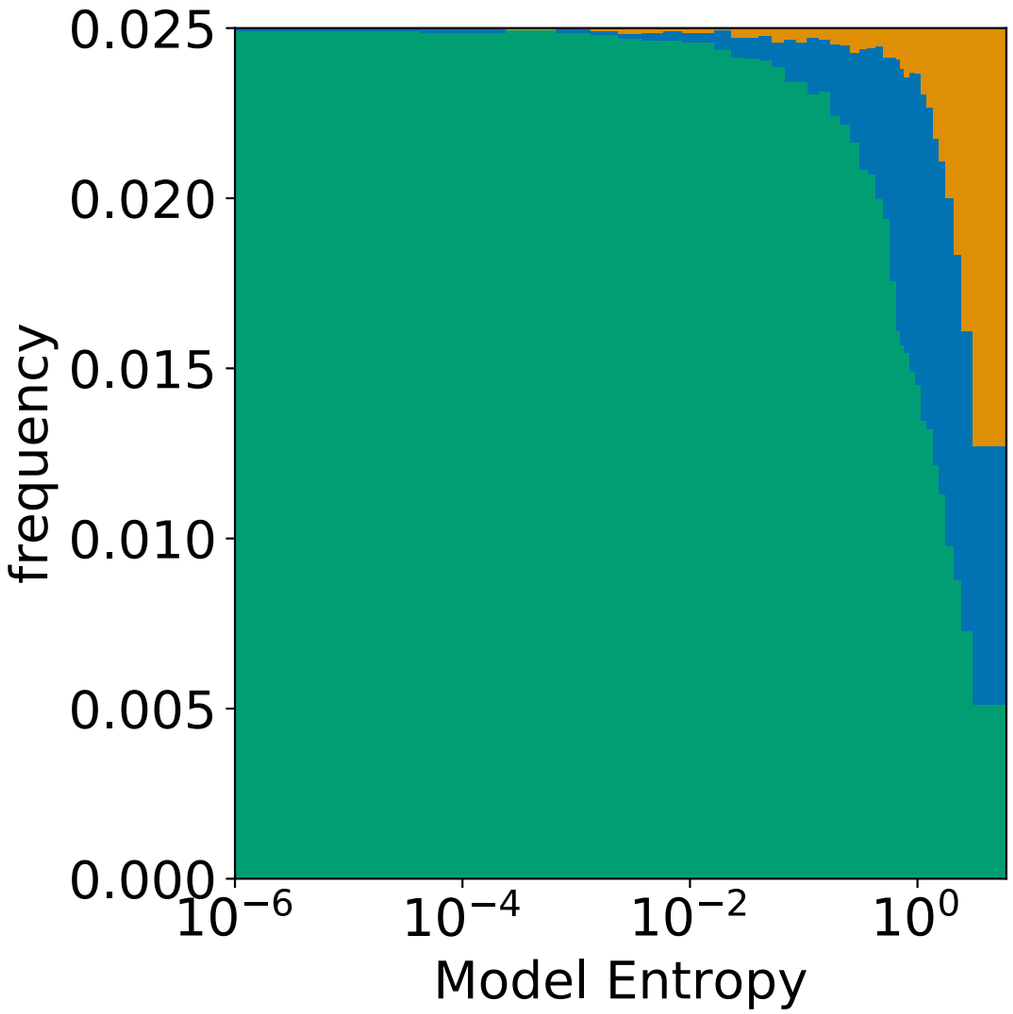
$$\begin{cases} p(X \mid Y_1) = .03 \\ p(X \mid Y_2) = .05 \\ p(X \mid Y_3) = .07 \end{cases} \qquad \begin{cases} \mathbb{E}[Y_1] = .3 \\ \mathbb{E}[Y_2] = .5 \\ \mathbb{E}[Y_3] = .3 \end{cases}$$

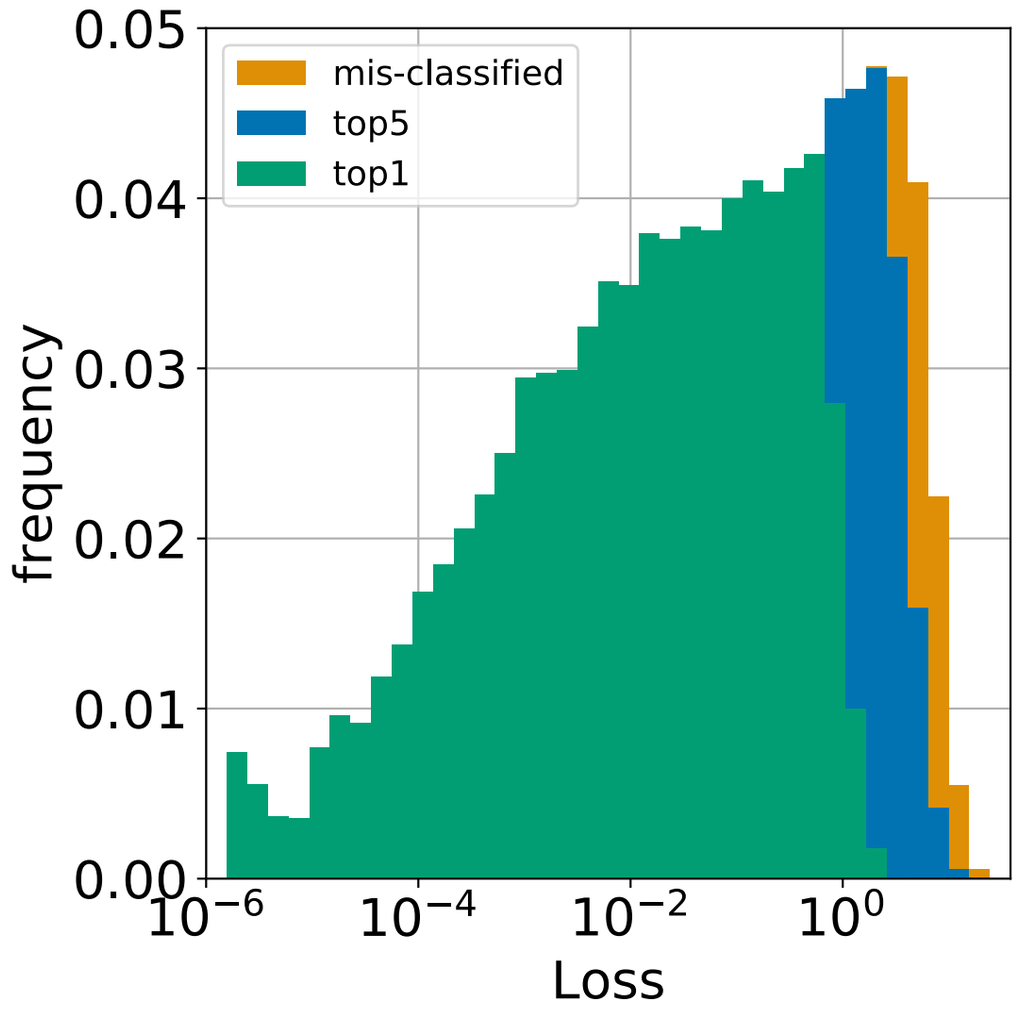
Since $\mathbb{E}[X] = .95$,

$$\begin{cases} BF(X \mid Y_1) = 1.9 \\ BF(X \mid Y_1) = 1.1 \\ BF(X \mid Y_1) = 1.3 \end{cases} \qquad \mathbb{E}[BF(X|Y_i)] = 1.5$$

So compared to expected value of knowing the age of 22, the expected value of knowing location is low, 1.5.

# Value of Model Entropy

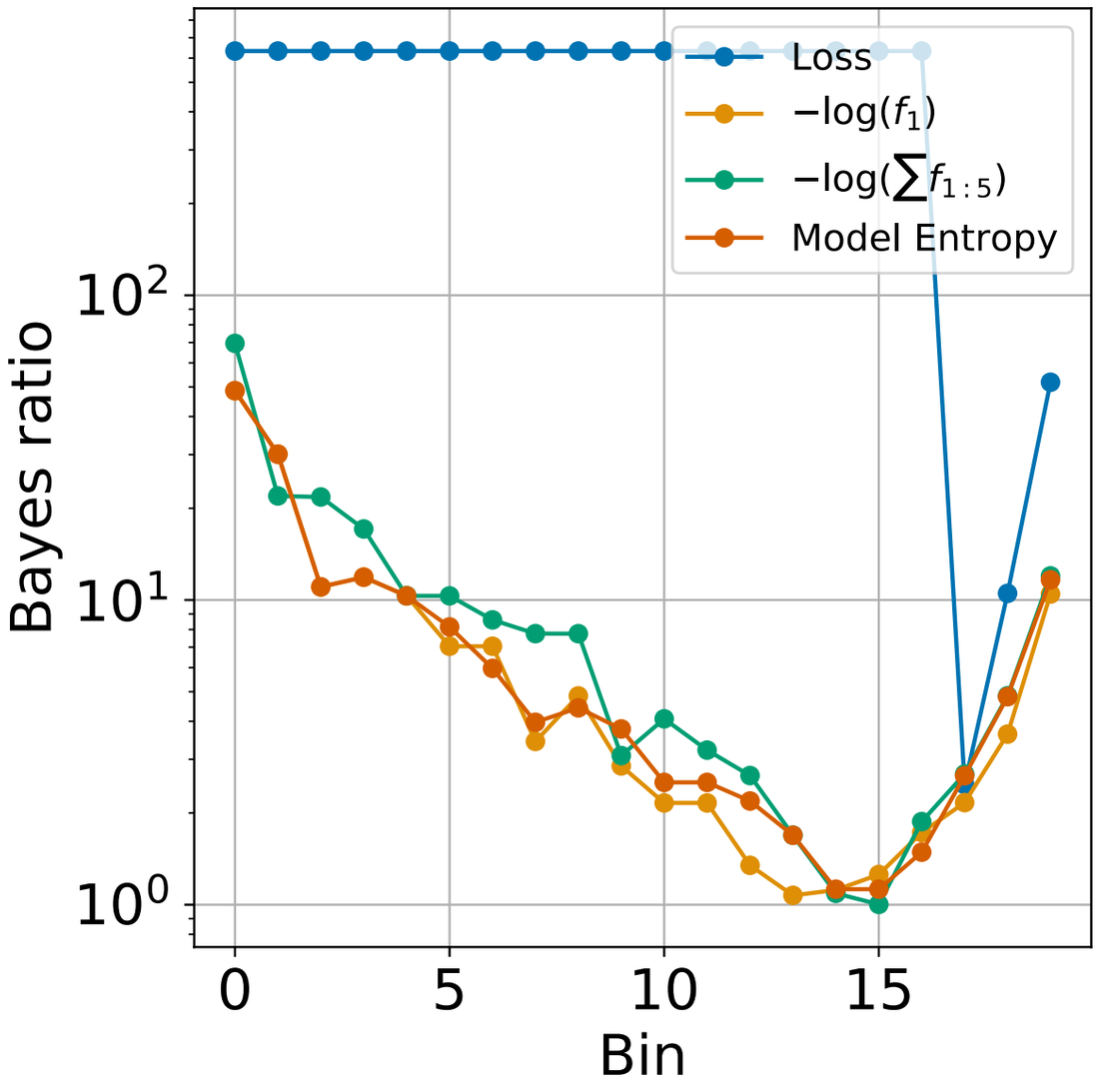


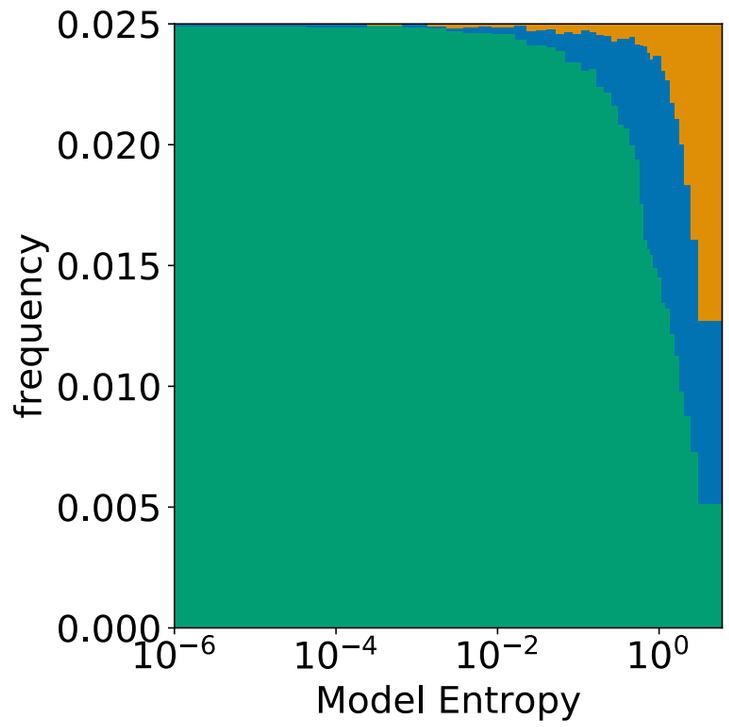




Still can estimate when prediction is correct.

- For small values < .001 of the function, always correct.
- For large values > 1 correct less than 20% of the time
- For intermediate values, make a histogram, with probability correct in each bin.

# Bayes Ratio for Loss, Entropy, U1, U2




Bayes ratio over equal 20 quantile bins for:
loss, entropy, U1, U5. Very large Bayes ratio in the first 10 and last 3 bins. On the other hand, bin 15 for U5 provides very little value.

# Expected Bayes Ratio Tables

TABLE 2. Expected odds ratio $\mathbb{E}[BR]$ against various measures of confidence. For CIFAR-10 the odds ratio is computed for the probability of $\mathrm{top1}$ correct; for CIFAR-100 and ImageNet-1K, with probability of $\mathrm{top5}$ correct. The mean is reported over 10-fold cross-validation of the test set. We also report the expected odds ratio against the loss for comparison.

| Confidence measure | CIFAR-10 | CIFAR-100 | ImageNet-1K |
|---|---|---|---|
| Model Entropy | 150.03 | 120.40 | 37.53 |
| $-\log p_{\mathsf{max}}$ | 141.56 | 131.38 | 44.97 |
| $-\log \sum p_{1:5}$ | - | 106.69 | 52.39 |
| $\|\nabla_x \|p\|\|$ | 264.85 | 132.37 | 45.89 |
| Dropout variance $(p = 0.002)$ | 239.92 | 99.36 | 24.15 |
| Dropout variance $(p = 0.01)$ | 172.86 | 31.09 | 34.76 |
| Dropout variance $(p = 0.05)$ | 19.77 | 1.39 | 1.57 |
| Loss | $\infty$ | 782.45 | 549.94 |

High Cost

# Comparison with other works

- Another confidence metric comes from Bayesian Dropout (Gal and Gharamani). In this case run inference 30 times with different random dropout. Confidence based on model variance.
  - less accurate
  - Very costly (30 X inference cost)

- Can also train an auxiliary neural network to predict whether another network is correct or wrong.
  - Have not compared accuracy for this method
  - But this is costly (1X inference)

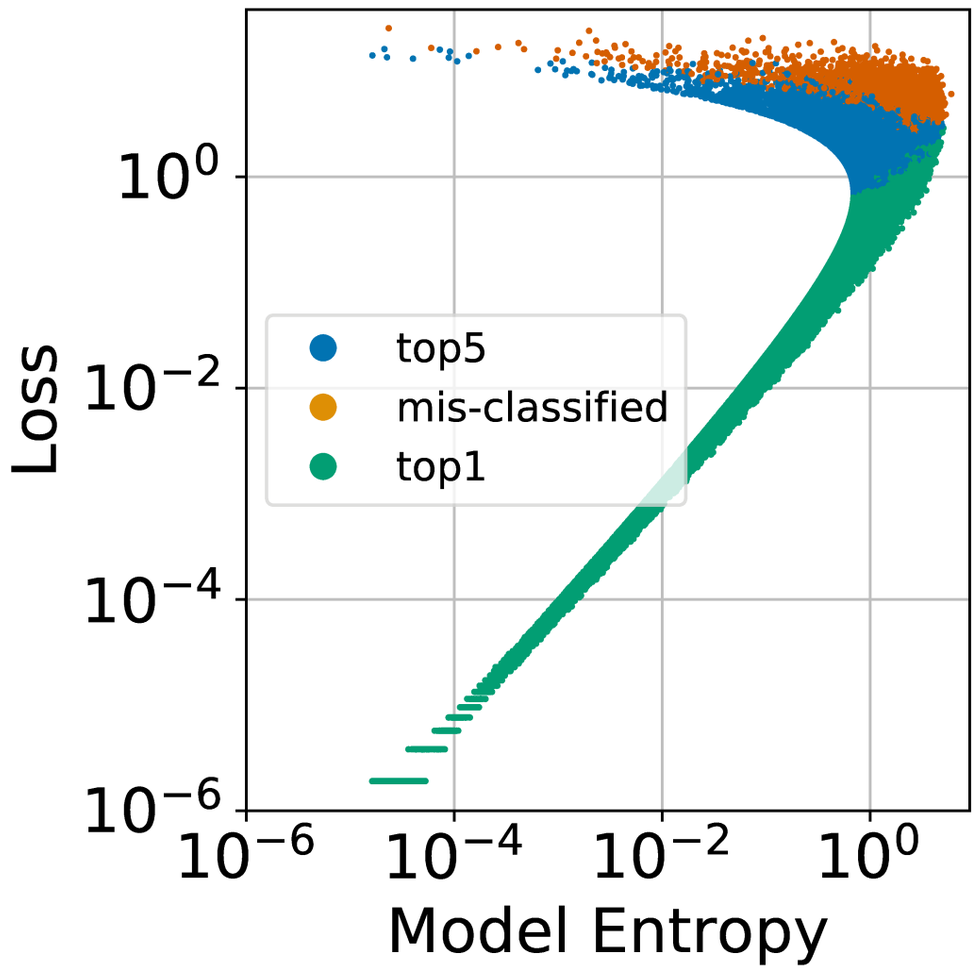
- Compared to these methods, our method is less costly, and we also can provide theoretical guarrantees that it works (for small values of U under the assumptions of generalization).

# Conclusions so far

- A confidence metric which is easy to compute (basically free) and which gives increased confidence in the probability that a prediction is correct.
- This measure can be adapted to Top 5, and can be used to detect increased probability of errors as well.
- On the larger dataset, ImageNet, the metric performs better relative to CIFAR-10.

# Detecting Incorrect Labels

- When we evaluate the uncertainty metric, we find some outliers.
- These turn our to be ambiguous images in the test set.




| | | | | | | |
|---|---|---|---|---|---|---|
| *Prediction:* | Bearskin | Wallaby | School bus | Pot | Ping pong ball | Baseball |
| *Label:* | Assault Rifle | Wombat | Minibus | Paintbrush | Beaker | Bucket |

# What about off-manifold data?

TABLE 6. Discarding out-of-distribution images from ImageNet-1K. For each confidence measure $Y$, the value of $a$ is chosen such that $P(Y \leq a \mid \text{image is from ImageNet-1k}) = 0.9$. Evasive attack penalizes both model entropy and gradient with Lagrange multiplier 0.1.

| Image source | Confidence measure | $a$ | $P$(image discarded) |
|---|---|---|---|
| COCO | Model Entropy | 1.75 | **0.38** |
| | $-\log p_{\max}$ | 0.77 | 0.34 |
| | $-\log \sum p_{1:5}$ | 0.13 | 0.37 |
| | $\|\nabla_x \|p\|\|$ | 1.06 | 0.23 |
| | Dropout variance ($p = 0.002$) | 0.024 | 0. |
| adversarially perturbed ($L_2$) | Model Entropy | 1.75 | 0.28 |
| | $-\log p_{\max}$ | 0.77 | 0.25 |
| | $-\log \sum p_{1:5}$ | 0.13 | 0.28 |
| | $\|\nabla_x \|p\|\|$ | 1.06 | **0.58** |
| | Dropout variance ($p = 0.002$) | 0.024 | 0.39 |

# Conclusions

- This tool for giving confidence (or uncertainty) to the classifications of neural networks has immediate applications to fields where confidence is valuable.
- Can also be used for
  - detecting errors in labels
  - detecting off manifold data, or adversarially perturbed data