Uncertainty Quantification in Computer Vision and Robotics

Cologne AI & ML Meetup

May 18, 2021

Dr. Matias Valdenegro Toro
German Research Center for Artificial Intelligence, Bremen

matias.valdenegro@dfki.de
@mvaldenegro
http://github.com/mvaldenegro

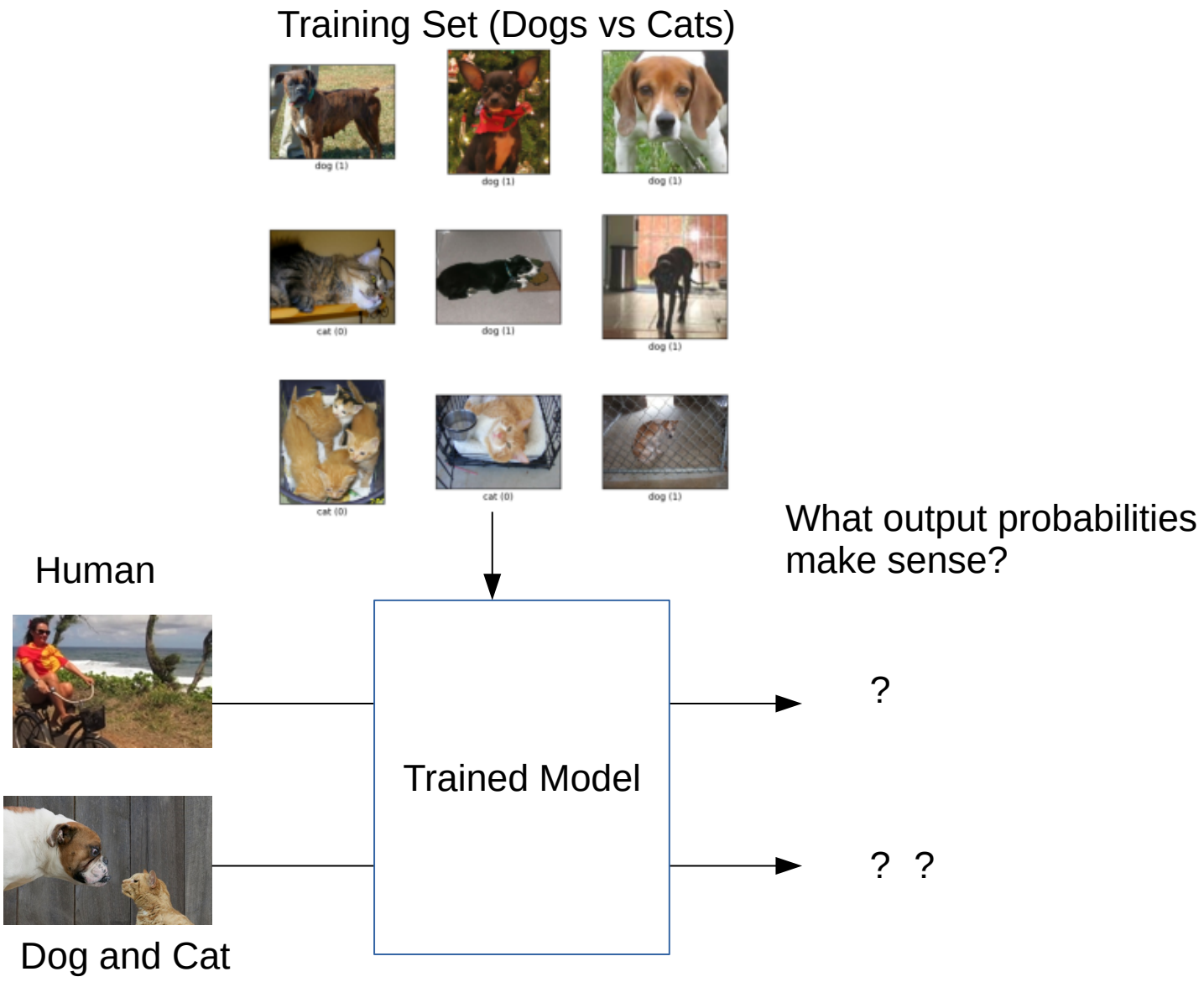
Contents

1. Intro to Uncertainty Quantification

2. Uncertainty in Robotics

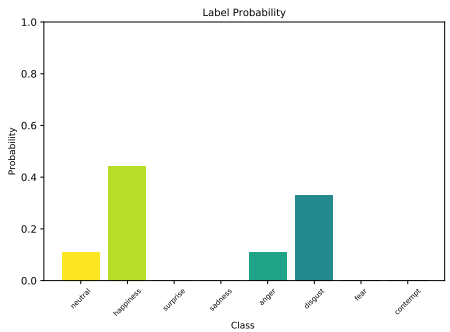
3. My Research in UQ

What is Uncertainty in Machine Learning?

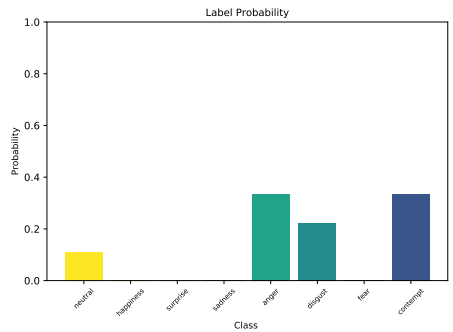


What is Uncertainty in Machine Learning?

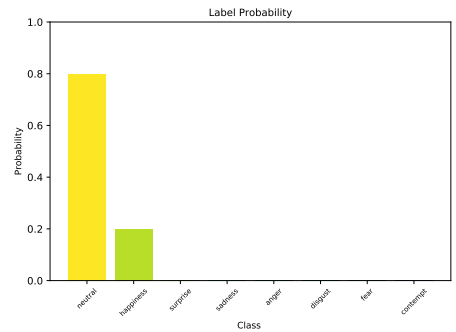
Happiness



Anger



Neutral



FER+ fataset, with crowd sourced labels for emotion recognition, over classes Neutral, Happinesss, Surprise, Sadness, Anger, Disgust, Fear, and Contempt.

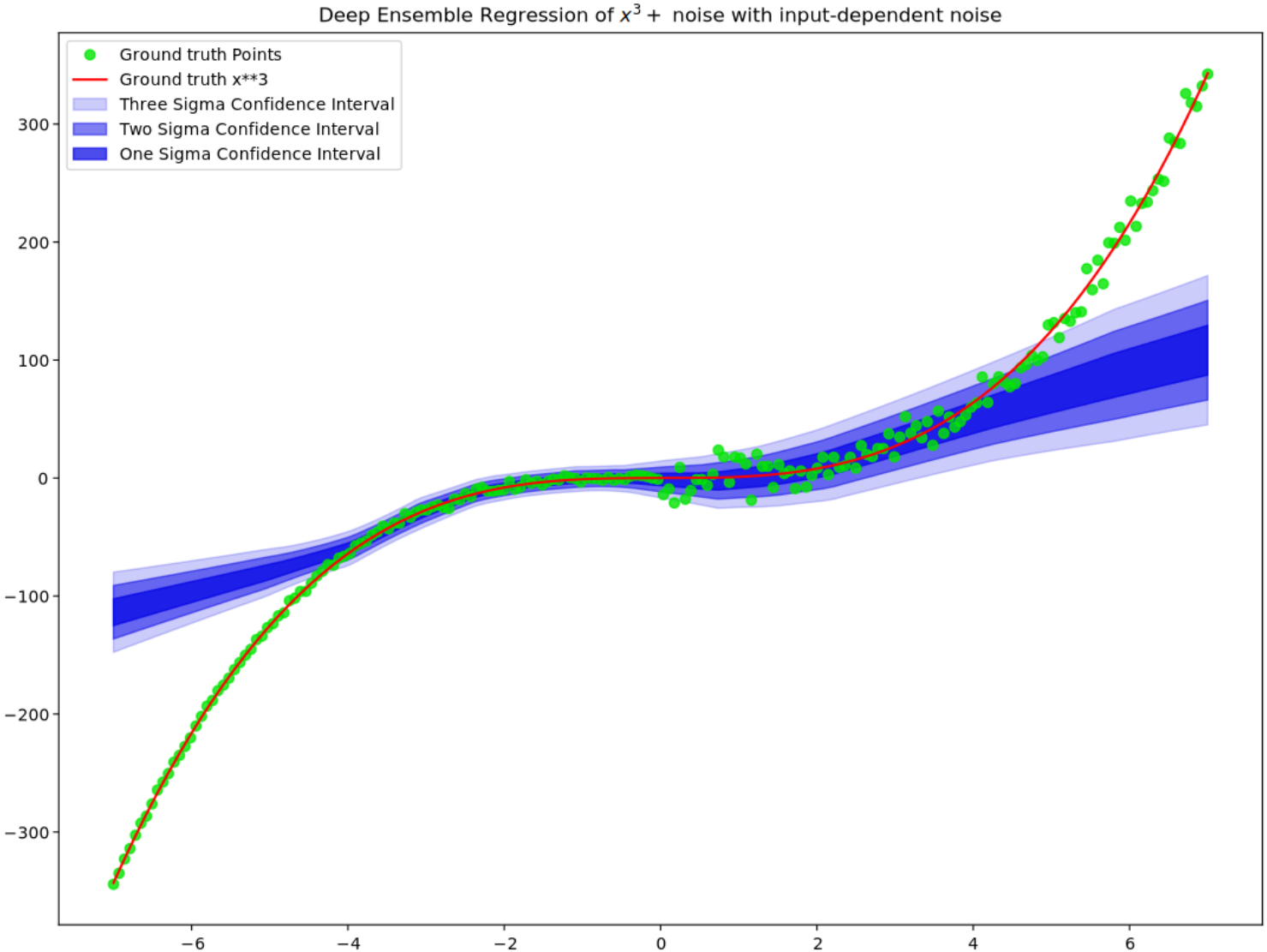
What is Uncertainty in ML?

- Real-world datasets are typically unbalanced, so confidences on each class should be different, reflecting the training data and model inferences.
- Real-world datasets might contain noise, like imprecise labels, ambiguous measurements, or sensor noise. A model should be aware of this.
- Most neural networks are overconfident, meaning that softmax confidences do not have a good probabilistic interpretation and could be misleading.

What do Classical Models Lack?

- Most machine learning models do not explicitly model uncertainty at their outputs.
- They produce `point-wise` predictions. A model with uncertainty outputs a `distribution`.
- A distribution can usually include more information than a single point-wise prediction, for example, mean and variance for a regression output instead of just a point prediction.
- Neural networks are often overconfident, producing wrong predictions with high confidence.

What do Classical Models Lack?



Practical Applications of Uncertainty

- Reliable confidence estimates can be used to detect misclassified examples or when the model is extrapolating.
- A model can reject to produce an output if the uncertainty is too high, for example, to require human processing instead of automated. This is called out of distribution detection.
- The confidence or uncertainty of a prediction tells the human how much it should really trust the prediction.
- Additional decision making can be made with a realistic confidence score, which is very important for medical and human-interaction applications.

Types of Uncertainty

Aleatoric Uncertainty

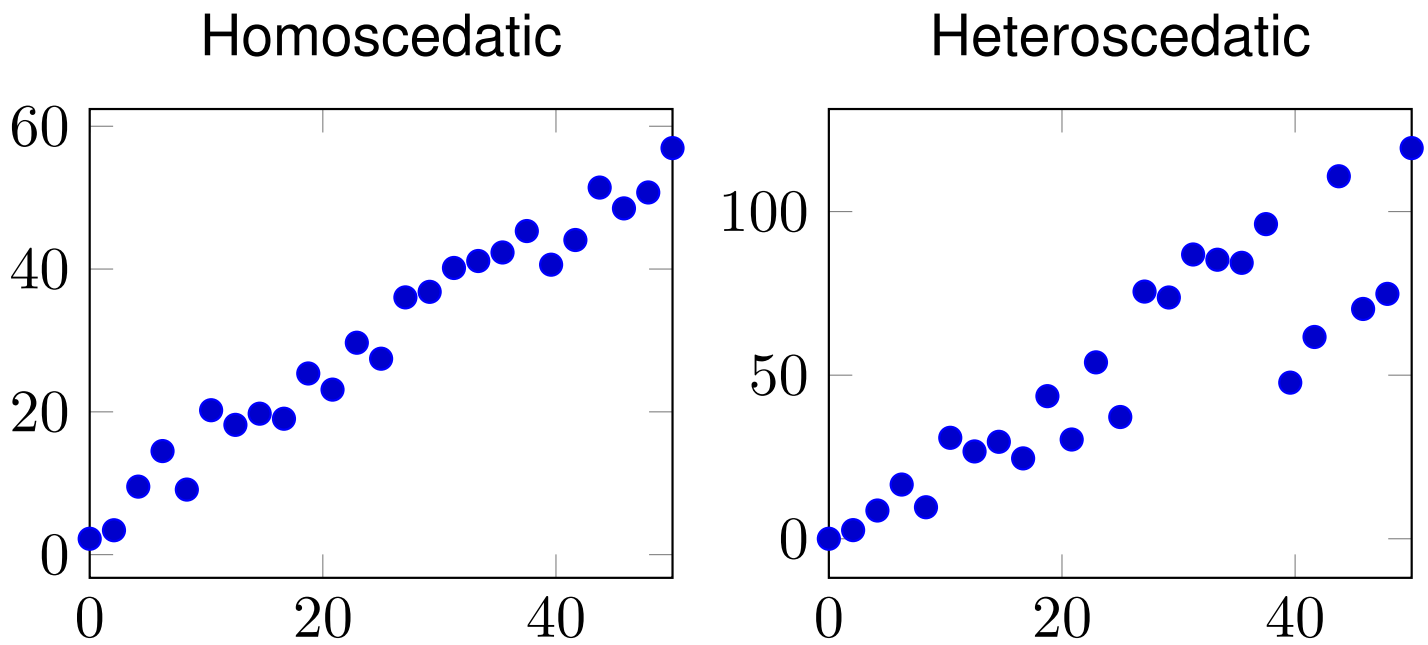
Uncertainty that is inherent to the data, for example, sensor noise, stochastic processes.
Cannot be reduced by adding more information.

Epistemic Uncertainty

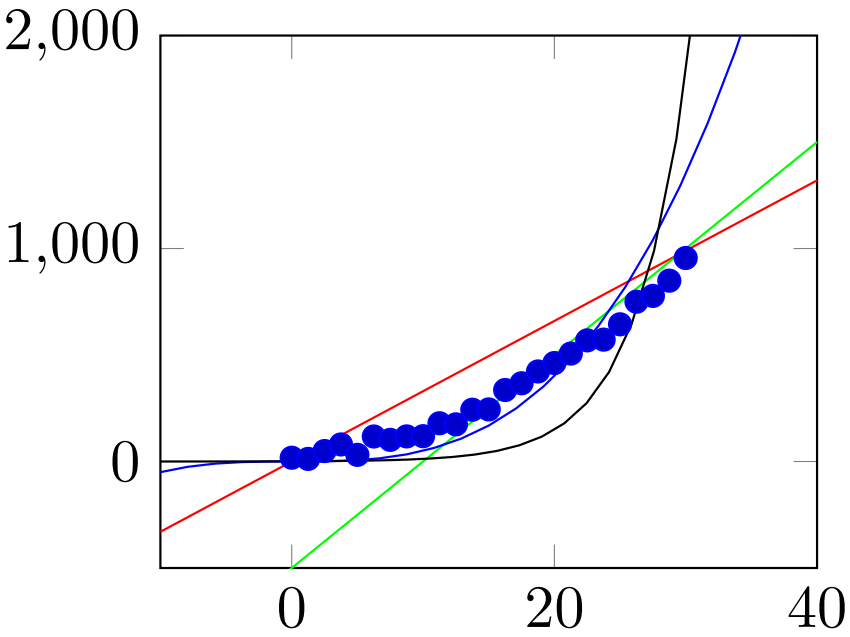
Uncertainty produced by the model, for example, model misspecification, class imbalance, lack of training data.
Can be reduced by adding more information to the training process.

Aleatoric Uncertainty

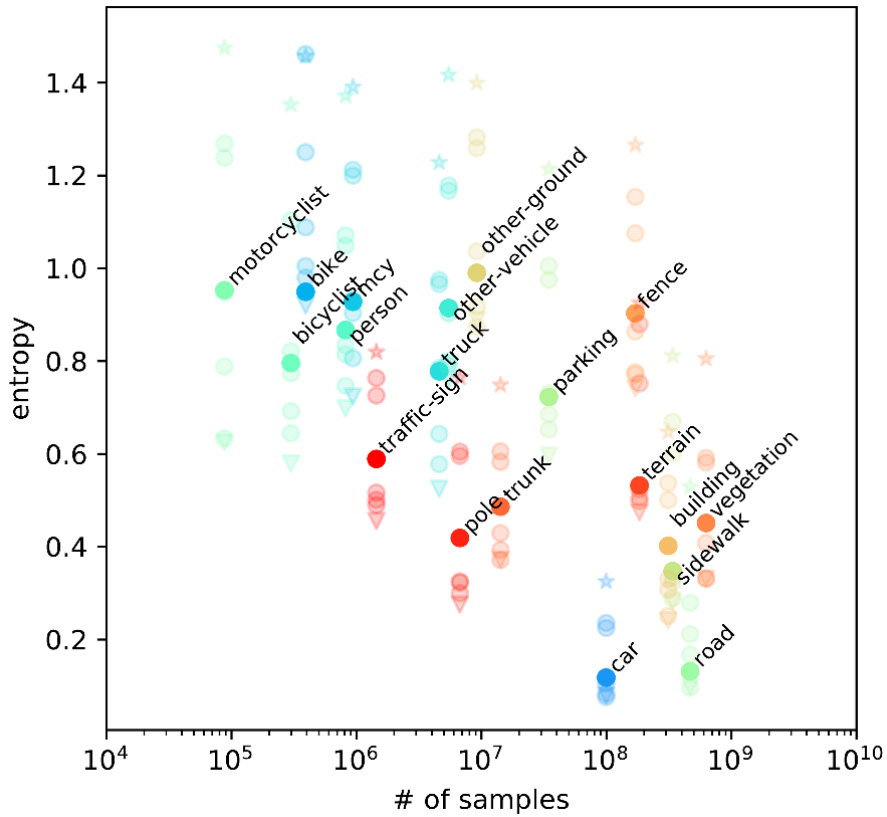
The simplest example of AU is measurements corrupted by additive noise, like $f(x) = x^3 + \epsilon$ Where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and x^3 would be the true function.
If σ^2 is constant, this is called homoscedatic noise, if σ^2 is a function of the input or variable, then it is called heteroscedatic noise.



Epistemic Uncertainty



Model Misspecification



Variations on Training Data

Bayesian Formulation

A Bayesian Neural Network is one where weights are probability distributions, instead of point estimates. Weight distributions implicitly encode uncertainty in the network.
This requires radically different inference algorithms to learn these distributions from data, this talk does not cover this. The Bayesian predictive posterior for y from inputs x and weight distributions θ is:

$$p(y \mid x) = \int_{\Theta} p(y \mid x, \theta) P(\theta \mid x) d\theta$$

This is called Bayesian model averaging, as weights are sampled from the learned weight distributions, and used to produce output estimates, weighted by the probability of each weight. This makes estimating the full posterior computationally very expensive, so it is rarely used in practice.

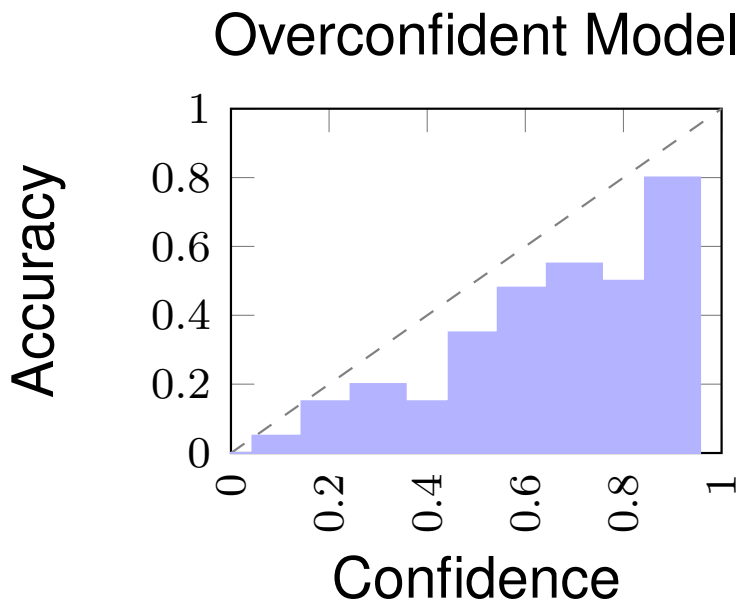
Calibration

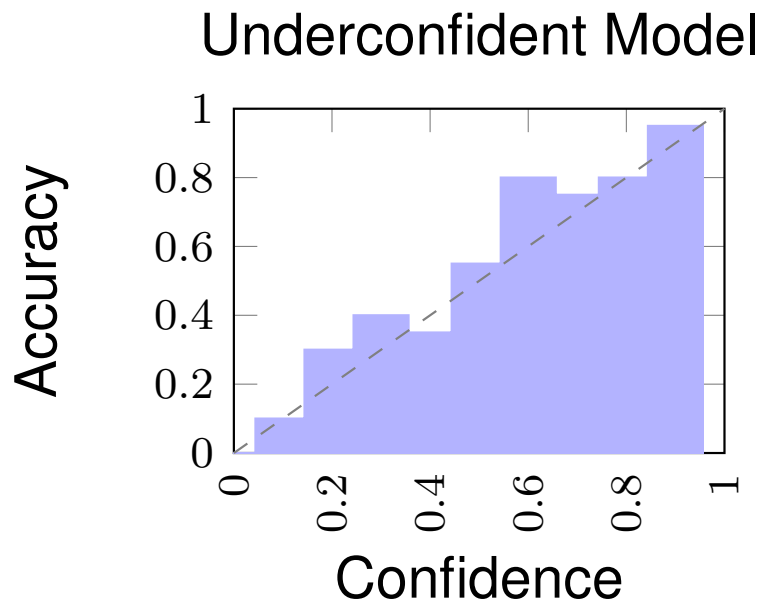
- We talked about a concept that indicates how much we can trust the confidence of a model.
- This can be formalized by comparing task performance (such as accuracy) as the confidence of predictions change.
- For example, if a prediction is made with 10% confidence, then we expect that such predictions will be correct 10% of the time.
- And correspondingly, if a prediction is made wit 90% confidence, then only 10% of such predictions will be incorrect.

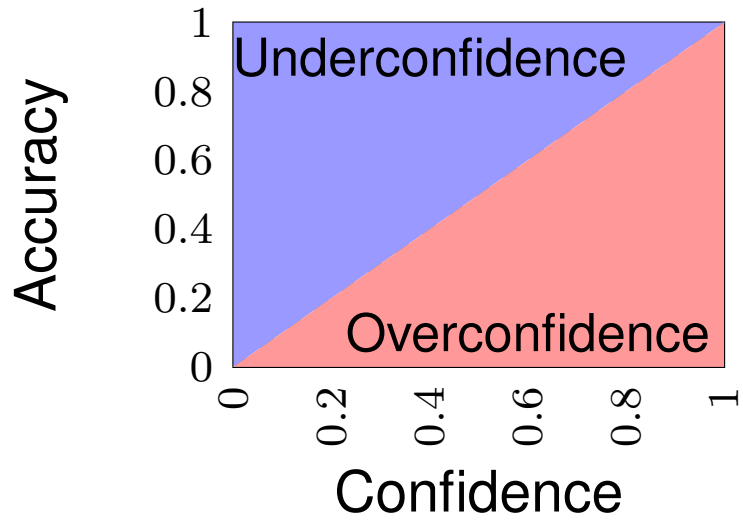
Calibration - Reliability Plots

- Calibration can be observed by making a Reliability plot.
- We take the predictions of a model over a dataset, divide the predictions by confidence values conf(B_i) into bins B_i, for each bin the accuracy acc(B_i) is computed, and then the values $(\text{conf}(B_i), \text{acc}(B_i))$ are plotted.
- Regions where conf$(B_i) <$ acc(B_i) indicate that the model is underconfident, while regions conf$(B_i) >$ acc(B_i) indicate overconfidence.
- The line conf$(B_i) =$ acc(B_i) indicates perfect calibration.

Calibration - Reliability Plots







1. Intro to Uncertainty Quantification

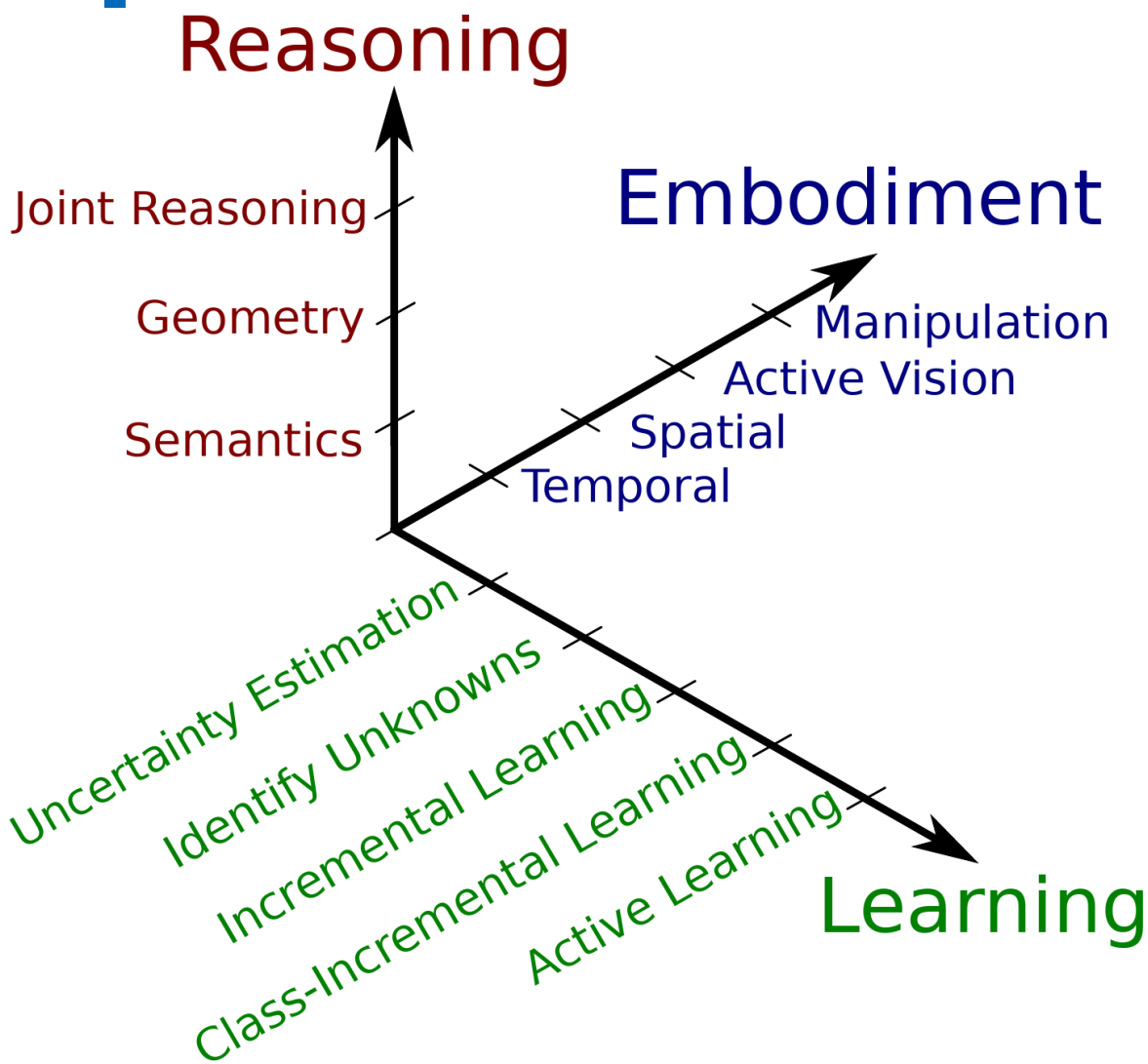
2. Uncertainty in Robotics

3. My Research in UQ

Challenges of DL in Robotics [Sünderhauf et al. 2018]

- Machine/Deep Learning and Computer Vision by itself is quite different from Robotics. The main difference is that a robot has a "body".
- A good description paper about this topic is "The Limits and Potentials of Deep Learning for Robotics" by Sünderhauf et al. 2018.
- Embodiment is the main difference between Robot Learning/Perception and their more theoretical fields of Machine/Deep Learning and Computer Vision.

Challenges of DL in Robotics [Sünderhauf et al. 2018]



Challenges of DL in Robotics - Learning [Sünderhauf et al. 2018]

Level	Name	Description
5	Active Learning	The system is able to select the most informative samples for incremental learning on its own in a data-efficient way. It can ask the user to provide labels.
4	Class-Incremental Learning	The system can learn *new* classes, preferably using low-shot or one-shot learning techniques, without catastrophic forgetting. The system requires the user to provide these new training samples along with correct class labels.
3	Incremental Learning	The system can learn off new instances of known classes to address domain adaptation or label shift. It requires the user to select these new training samples.
2	Identify Unknowns	In an open-set scenario, the robot can reliably identify instances of unknown classes and is not fooled by out-of distribution data.
1	Uncertainty Estimation	The system can correctly estimate its uncertainty and returns calibrated confidence scores that can be used as probabilities in a Bayesian data fusion framework. Current work on Bayesian Deep Learning falls into this category.
0	Closed-Set Assumptions	The system can detect and classify objects of classes known during training. It provides uncalibrated confidence scores.

Challenges and Applications

Medical Systems and Decision Making

Practically all medical applications require correct (epistemic) uncertainty estimates to be used with humans/animals, receive regulatory approval, and be useful for practitioning medical doctors to make decisions.

Robotics

Generally in Robotics, useful uncertainties are not modeled, for example uncertainty in dynamical systems (parameters), perception (object detection), or estimate when robot capabilities are being extrapolated. The best example is autonomous driving.

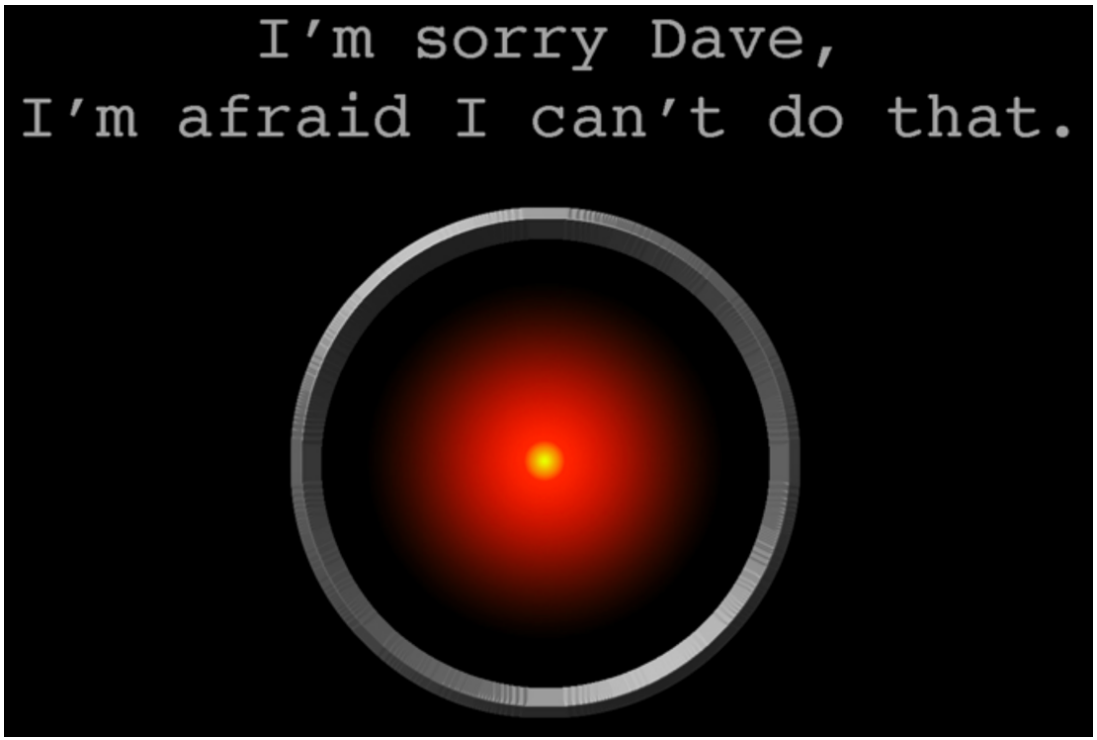
Challenges and Applications

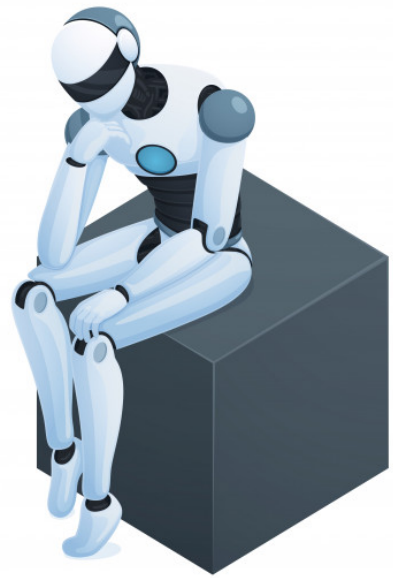
Reinforcement Learning

In the same way, it is very important to have RL-learned policies that can estimate their own epistemic uncertainty, and not take an action when the environment is too different from the training one.

- RL in robots or real mechanisms, with safety constraints (Safe RL).
- RL in non-stationary environments (for example, dynamic or unpredictable obstacles).
- Reduce the sample complexity required for training through Active Learning and Exploration.

Objective - Safe and Trustable Robots





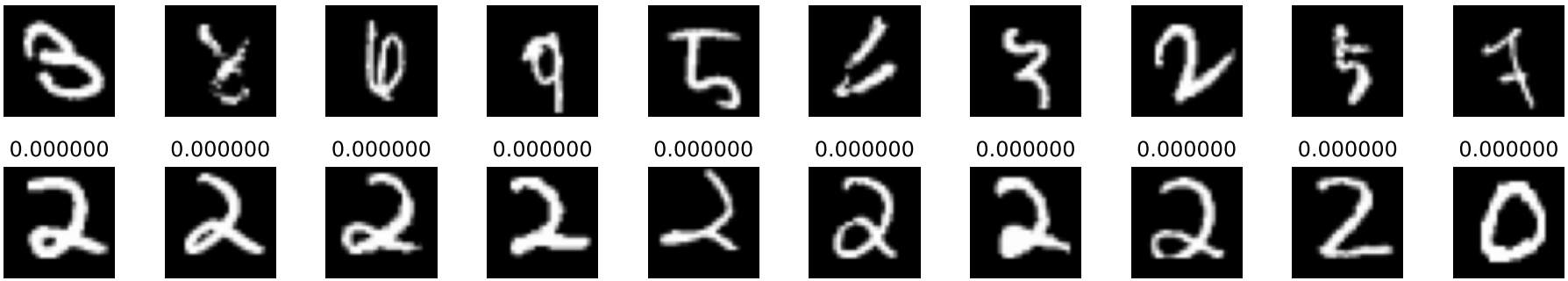
Objective - Safe and Trustable Robots

Examples

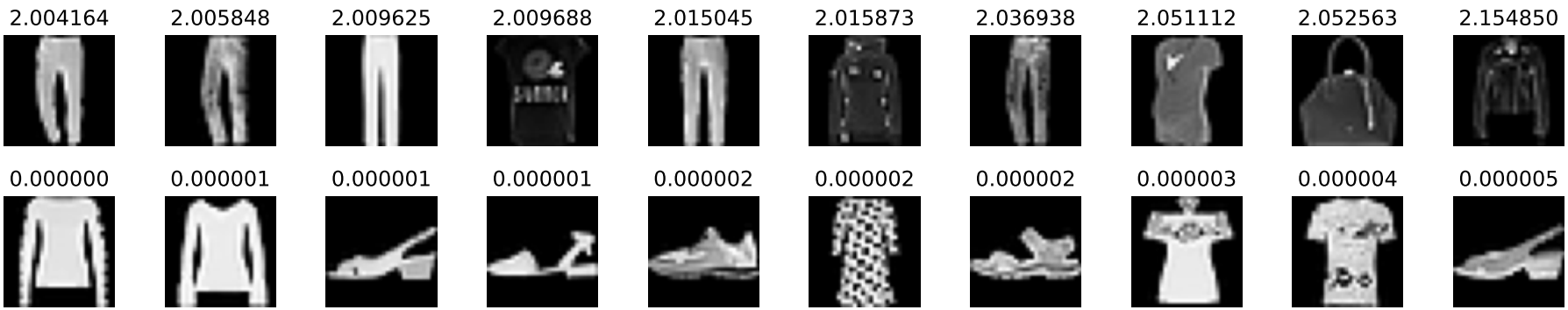
- Multiple incidents of experimental Autonomous Vehicles hitting human pedestrians and producing accidents, due to conditions not considered in development/training (similar to Kidnapped Robot Problem).
- Possible issues with Robots at care homes for the elderly. Algorithms should be tuned for maximum safety.
- Well known examples of face recognition being biased against some skin colors, OOD detection can help in preventing or alleviate these.
- AI/Robotics should be done for the social good.

Out of Distribution Detection (OOD) - MNIST vs Fashion MNIST

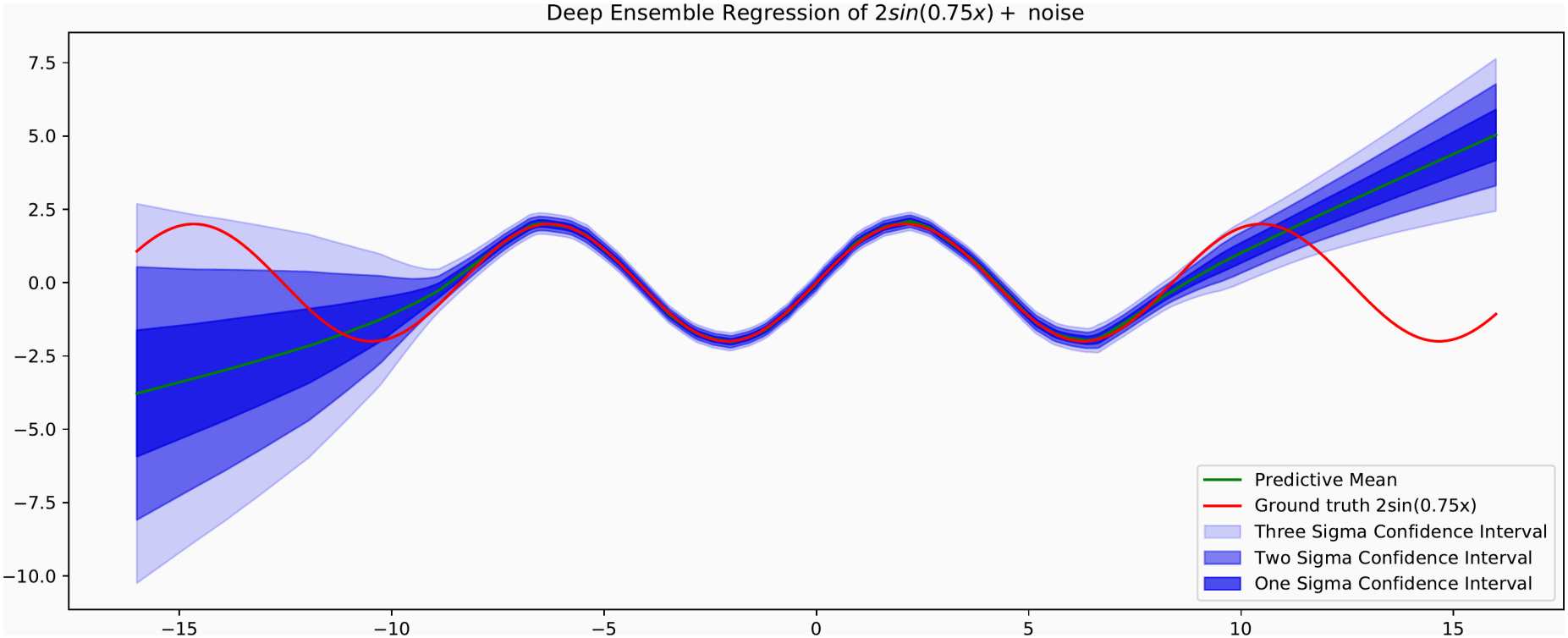
MNIST



Fashion MNIST



Out of Distribution Detection (OOD) - Sinusoid Regression with an Ensemble



Here the training set is $x \in [-8, 8]$. You can observe that outside of this range the standard deviation (uncertainty) increases considerably, and increases with the distance to the training set.

1. Intro to Uncertainty Quantification

2. Uncertainty in Robotics

3. My Research in UQ

Sub-Ensembles [Valdenegro. 2019]

- A great problem with Ensembles is that computational costs increase linearly with the number of members in the ensemble.
- A basic question is: Is it necessary that all ensemble members be independent? Can weights be shared across ensemble members?
- Turns out the answer is no and yes, weights on layers from the input can be shared, and last layers in the network ensembled, and this works as an approximation of the full ensemble.

Sub-Ensembles [Valdenegro. 2019]

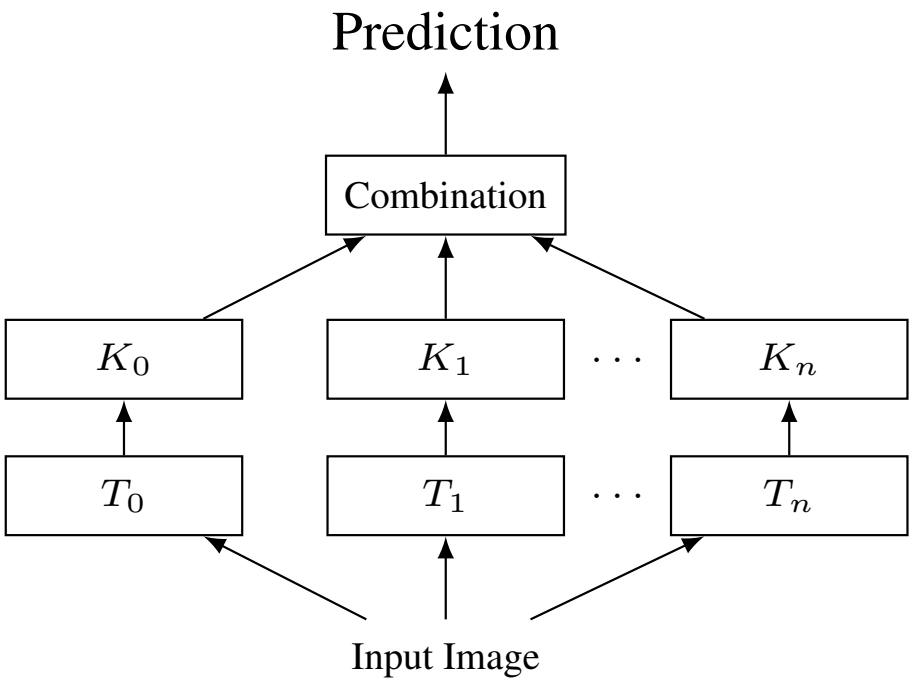


Figure 1: Ensemble

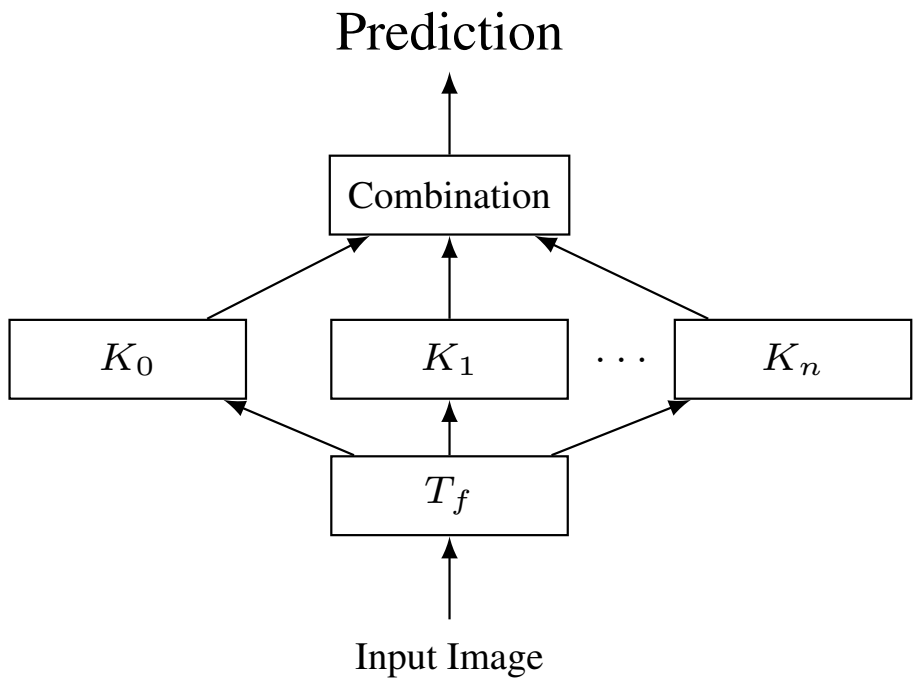
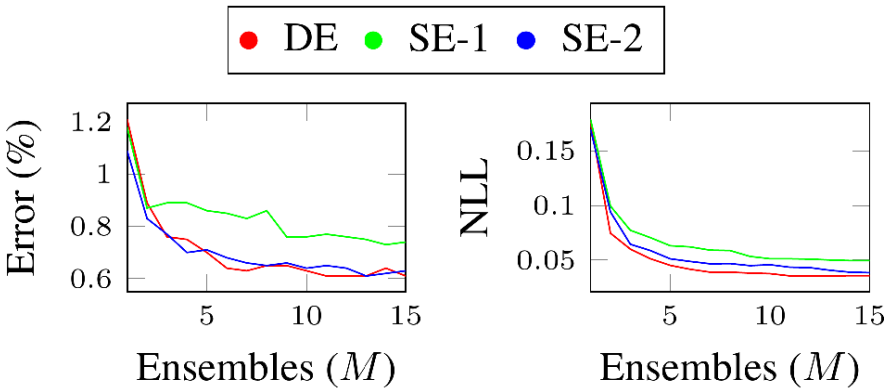
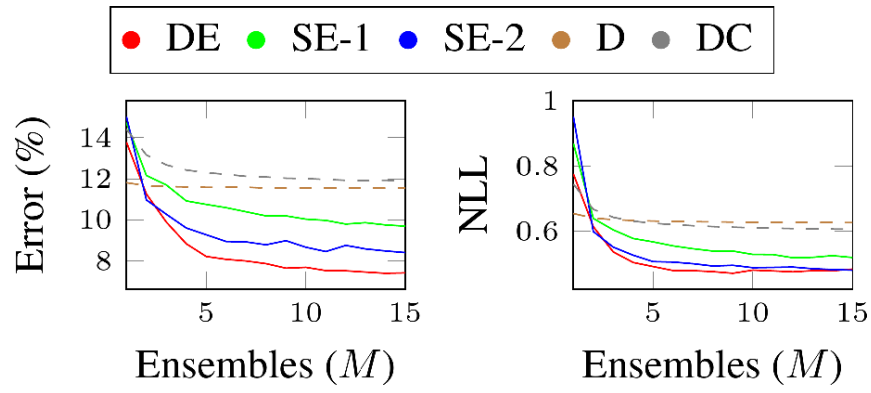


Figure 2: Sub-Ensemble

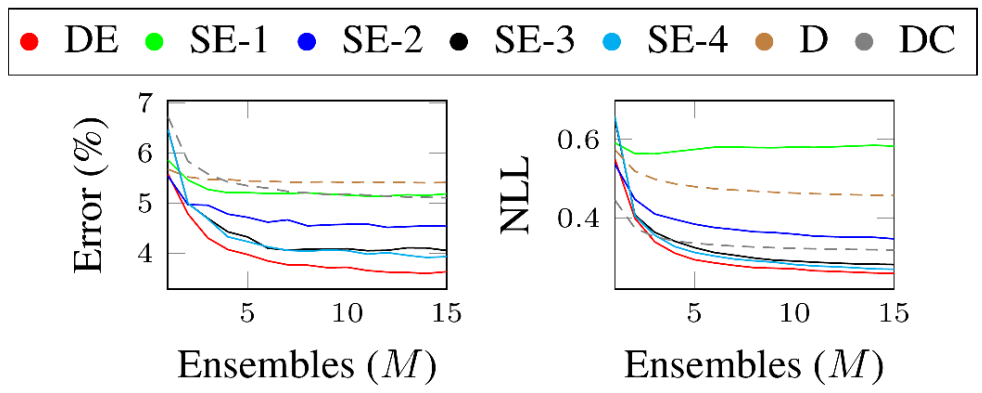
Sub-Ensembles - Performance



(a) MNIST



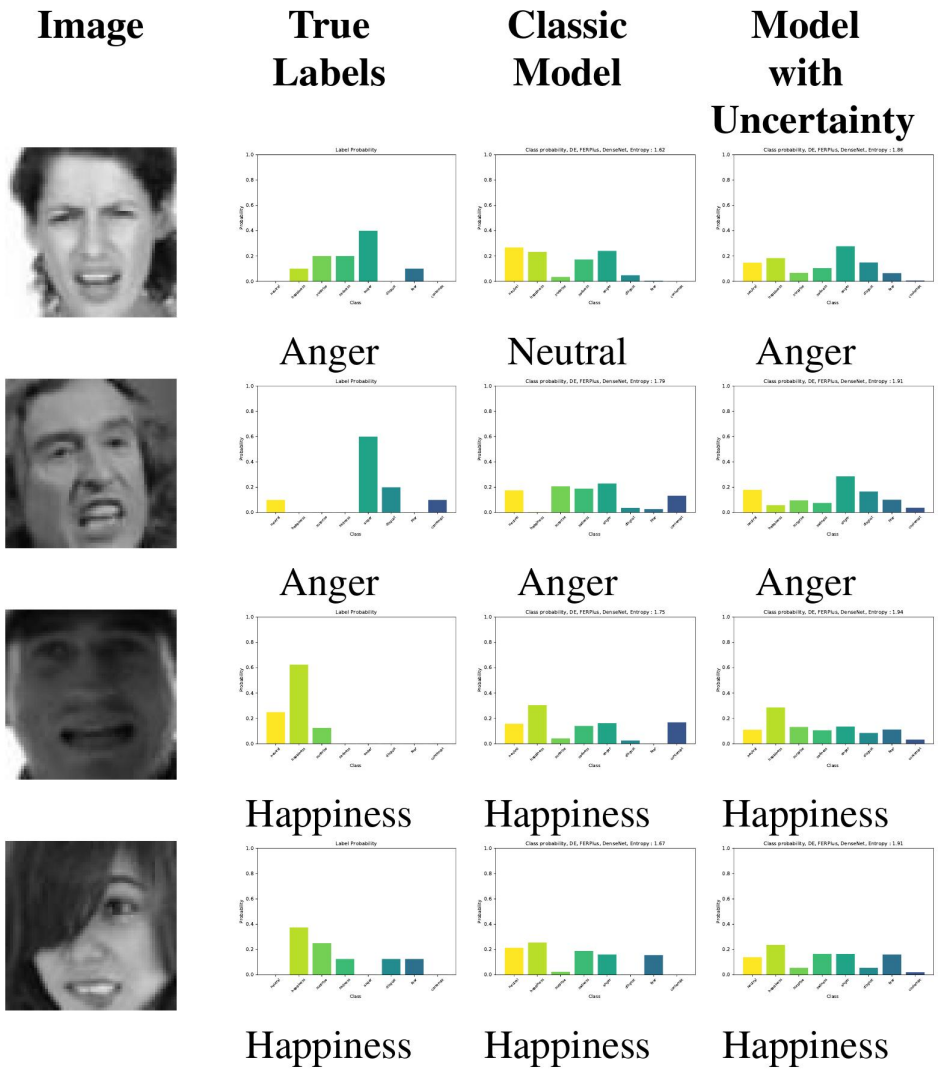
(b) CIFAR10



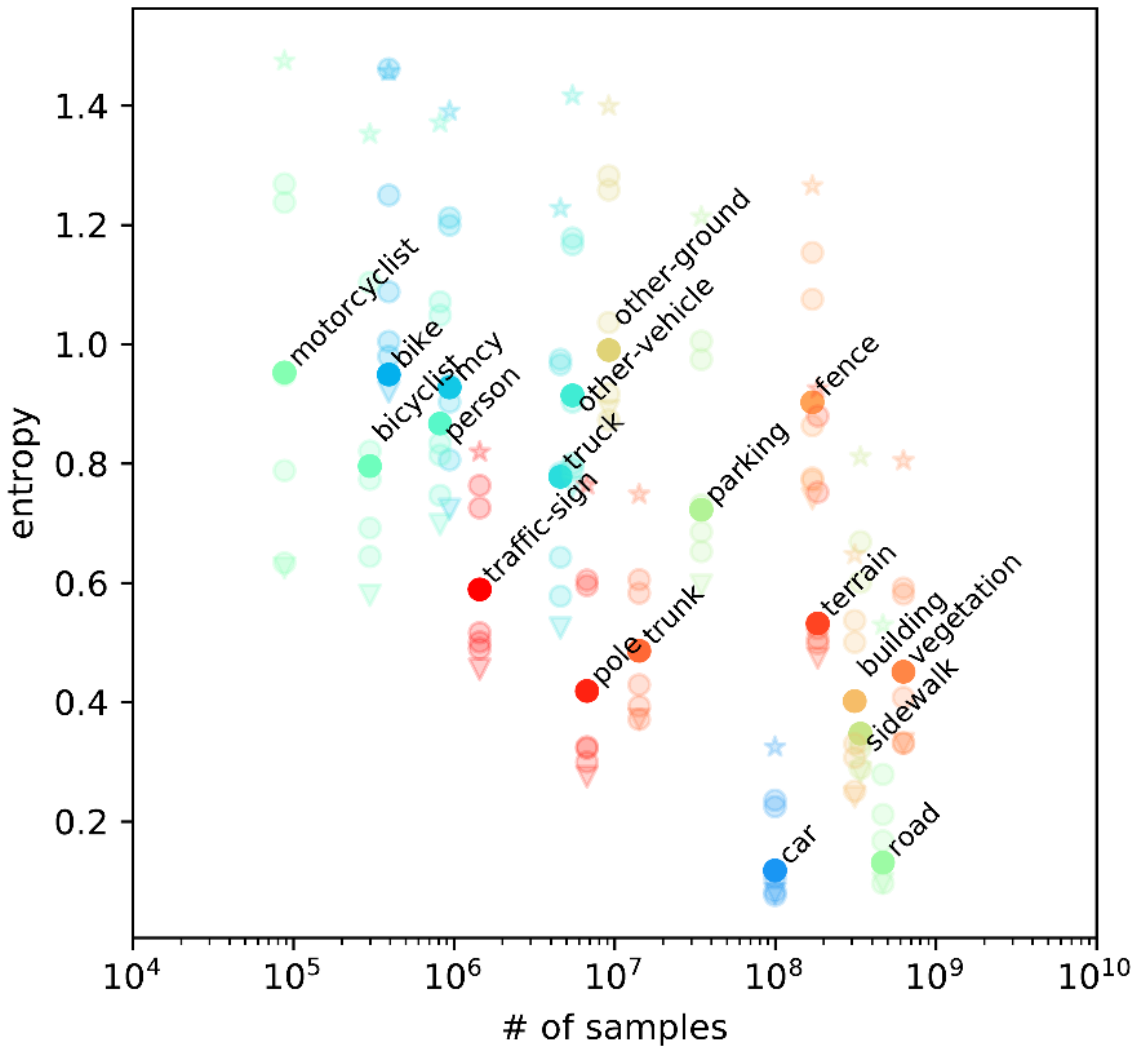
Results on SVHN using a batch normalized VGG-like network

Presented at the Bayesian Deep Learning Workshop @ NeurIPS 2019.

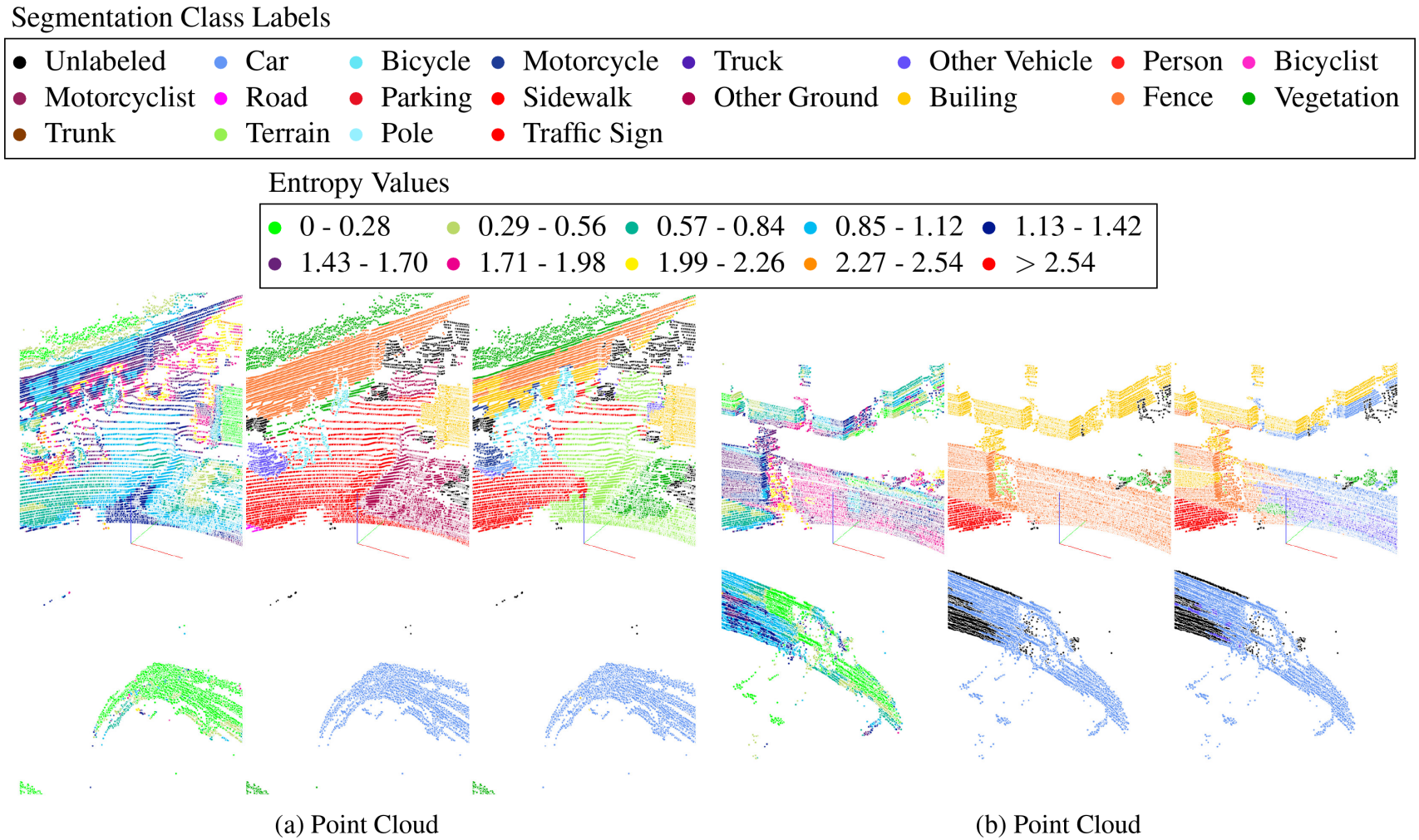
Uncertainty in Emotion Classification [Matin et al. 2020.]



Uncertainty in Point Cloud Segmentation [Bhandary et al. 2020]



Uncertainty in Point Cloud Segmentation [Bhandary et al. 2020]



(a) Point Cloud (b) Point Cloud

Entropy (Right) Ground truth (Center), Predictions (Right).

Unsupervised Difficulty Estimation [Arriaga & Valdenegro. 2020]

Idea. Look how the loss evolves for each sample on the train/val set, accumulating loss for each sample (as a metric).

Hypothesis. Difficult examples accumulate more loss than easy ones.

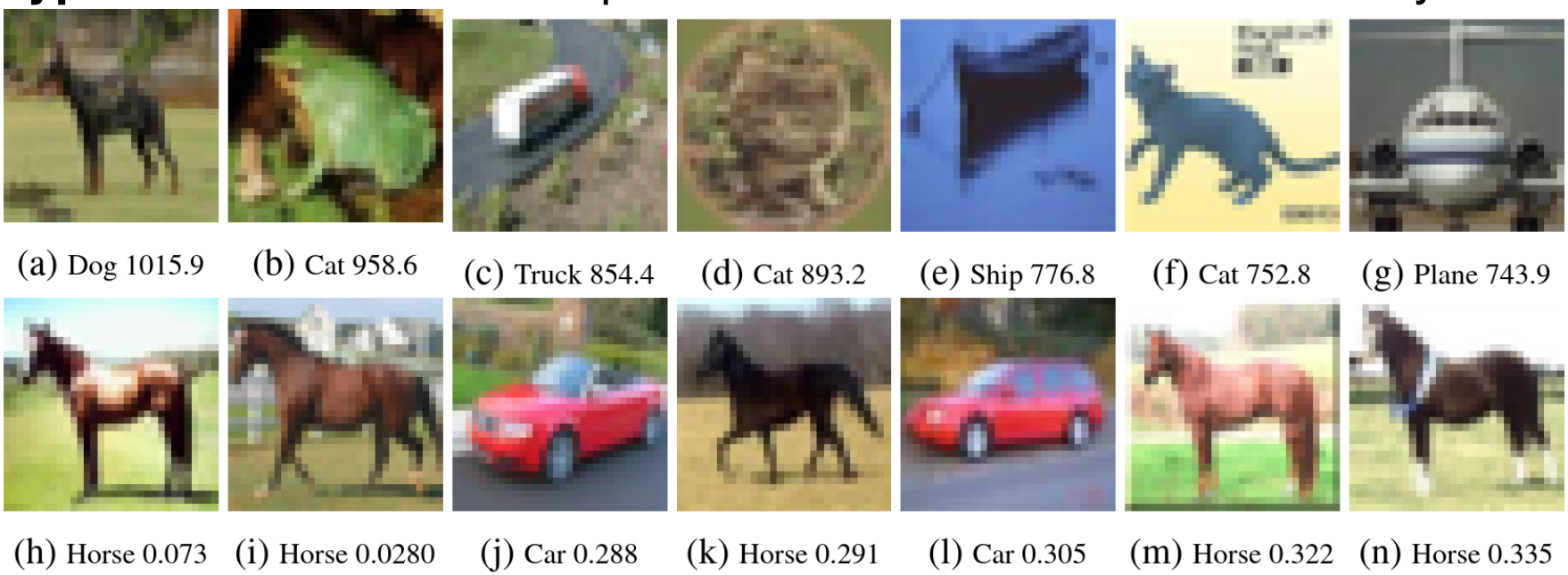
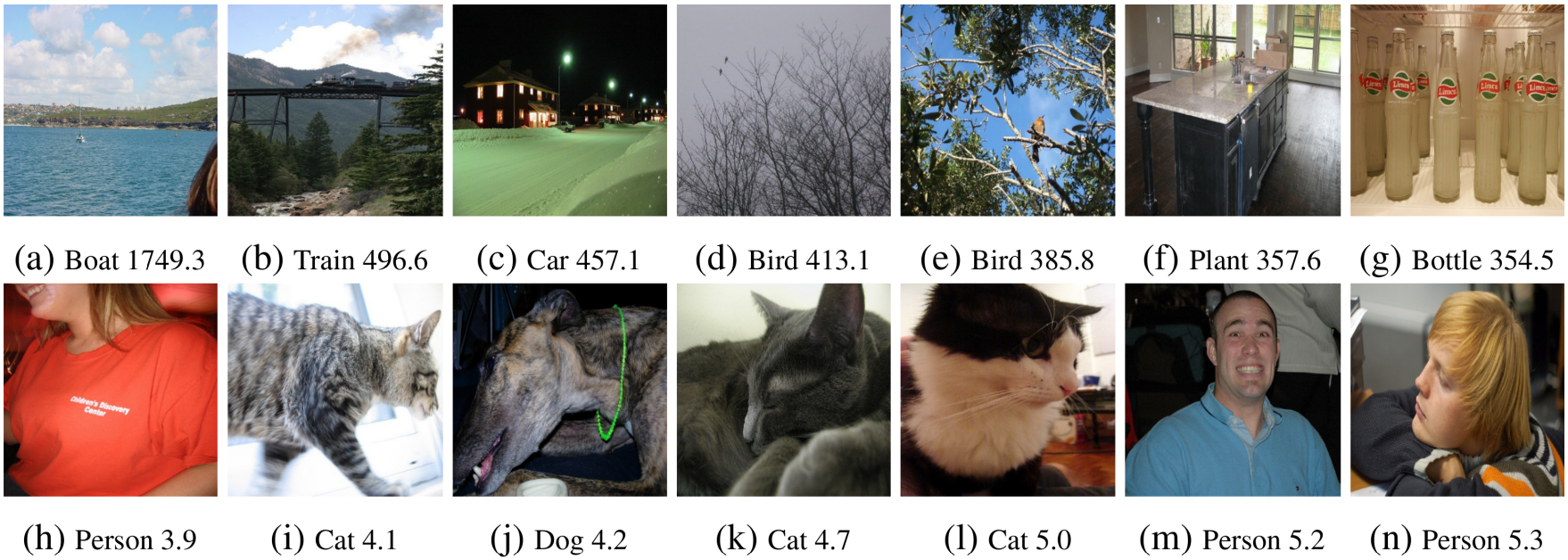


Figure 1: Most difficult (top-row) and easiest examples (bottom-row) in CIFAR10. Our proposed *action score* is displayed below each image as well as the true label.

Unsupervised Difficulty Estimation [Arriaga & Valdenegro. 2020]



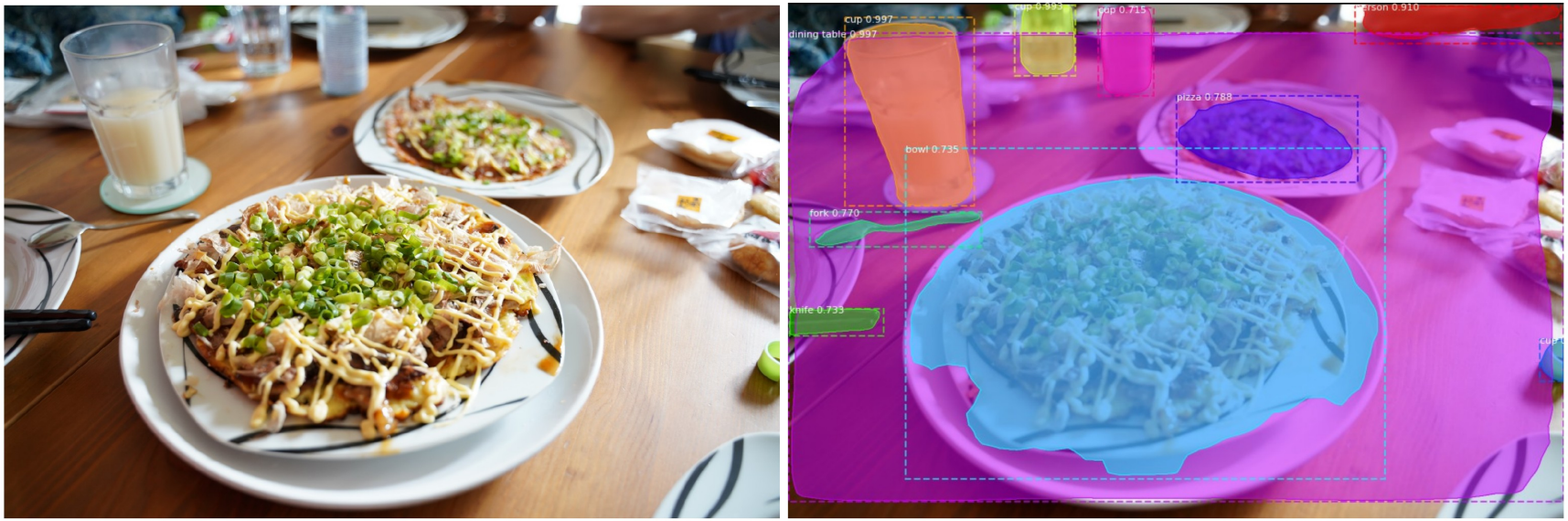
(a) Boat 1749.3 (b) Train 496.6 (c) Car 457.1 (d) Bird 413.1 (e) Bird 385.8 (f) Plant 357.6 (g) Bottle 354.5

(h) Person 3.9 (i) Cat 4.1 (j) Dog 4.2 (k) Cat 4.7 (l) Cat 5.0 (m) Person 5.2 (n) Person 5.3

Figure 2: Most difficult (top-row) and easiest examples (bottom-row) in the VOC 2007-VAL with the SSD localization loss. The *action scores* are displayed below each image as well as the true label.

We are also looking at relationship between action score and uncertainty (entropy), and possible predictions of model and data biases.

Lack of Uncertainty in Computer Vision



Incorrect detections with Mask R-CNN trained on COCO. Left presents the input image, and right the predictions. The Okonomiyaki is misclassified as a Bowl and a Pizza, and a Spoon is misclassified as a Fork, all with high confidence of $\sim 75\%$. The Bowl-Okonomiyaki has high variability in the predicted mask boundaries, signaling some amount of uncertainty.

M. Valdenegro, "I Find your Lack of Uncertainty in Computer Vision Disturbing.", Accepted at CVPR 2021 workshops.

Lack of Uncertainty in Computer Vision

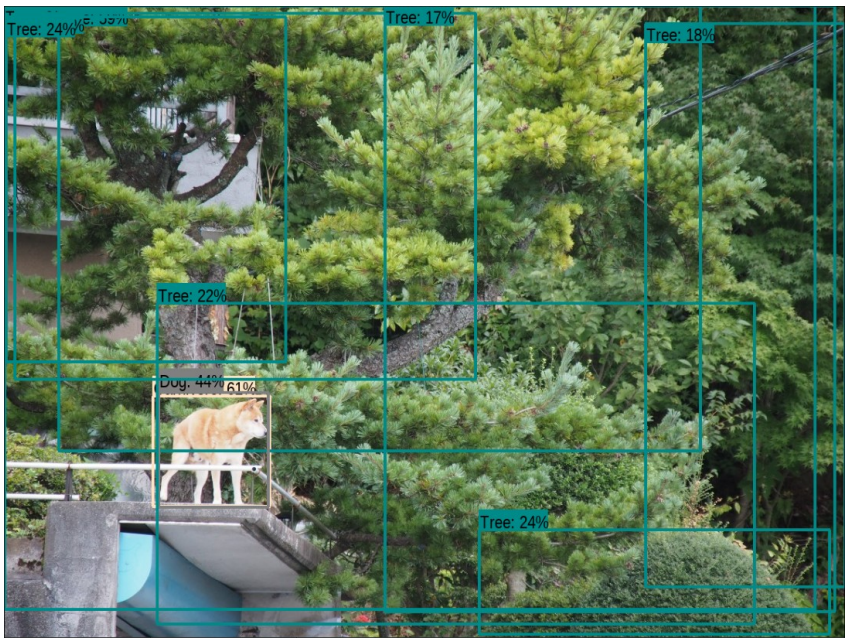


Figure 3: Multiple incorrect detections with low confidence. Shiba Dog is detected as 44% dog and 61% carnivorous, which is counterintuitive for humans.

Figure 4: Multiple incorrect detections with relatively high confidence, including detecting persons and bowls.

Conclusions and Future Thoughts

- Uncertainty is a useful measure to detect misclassified and out of distribution examples.
- Bayesian neural networks are not often used in practice, and many applications would benefit from them. Computational performance is a big reason.
- It is important to spread these techniques and their possible applications, specially now that ML is used in real-world applications that require to estimate model limits.
- Robotics in particular is a great application field, for example with Bayesian Reinforcement Learning, Probabilistic Object Detection, etc.
- I expect increase use of these techniques in practice.