university of groningen

faculty of science and engineering

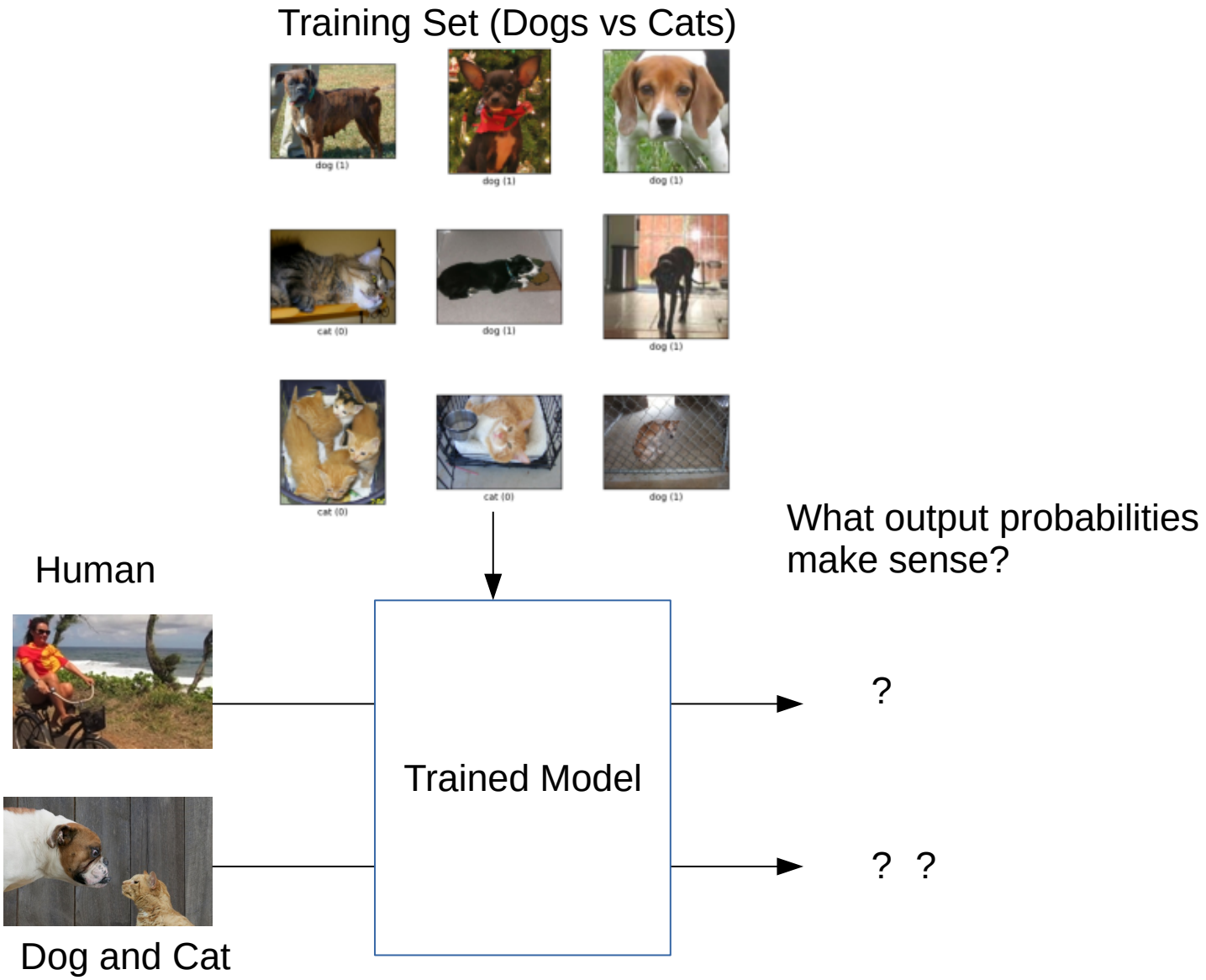
Introduction to Uncertainty in Machine Learning

Dr. Matias Valdenegro
Department of Artificial Intelligence
University of Groningen

January 24, 2024

Outline

Motivation


Training Set (Dogs vs Cats)
dog (1)
dog (1)
dog (1)
cat (0)
dog (1)
dog (1)
cat (0)
cat (0)
dog (1)
What output probabilities make sense?
Human
Trained Model
?
? ?
Dog and Cat

Motivation

- In the previous example, the model cannot correctly predict if an input is a human or cat+dog image. It was just not trained for those classes.
- But a standard machine learning model will make predictions with softmax confidences that will mislead the user.
- The question is: how can the ML model tell the user that is should not trust this prediction?
- This specific example is called Out of Distribution Detection.

Motivation

- As ML is used for real processes, the testing set is effectively infinite, and there are needs to evaluate how certain are the predictions in completely unseen samples, as this additional information can be used to assert if the predictions are useful.
- **Autonomous Driving**: The decision making process while driving needs to know if the predictions from a perception model are reliable, and to *know when the model does not know*.
- **Medical Applications**: Low confidence predictions might indicate that additional tests might be required, instead of taking a decision over faulty data.

Motivation

- We need ML models that can answer the following questions:

Do I know that I do not know?
Can I refuse to provide an answer?

- This is related to out of distribution detection, a model can provide information to say that the input is dissimilar to the training set, or the task is very different, and refuse to answer.
- Classic example is to train a cat/dog model, and test with a bird image. The answer will always be wrong for a typical model.

Why Uncertainty? [Leibig et al., 2017]

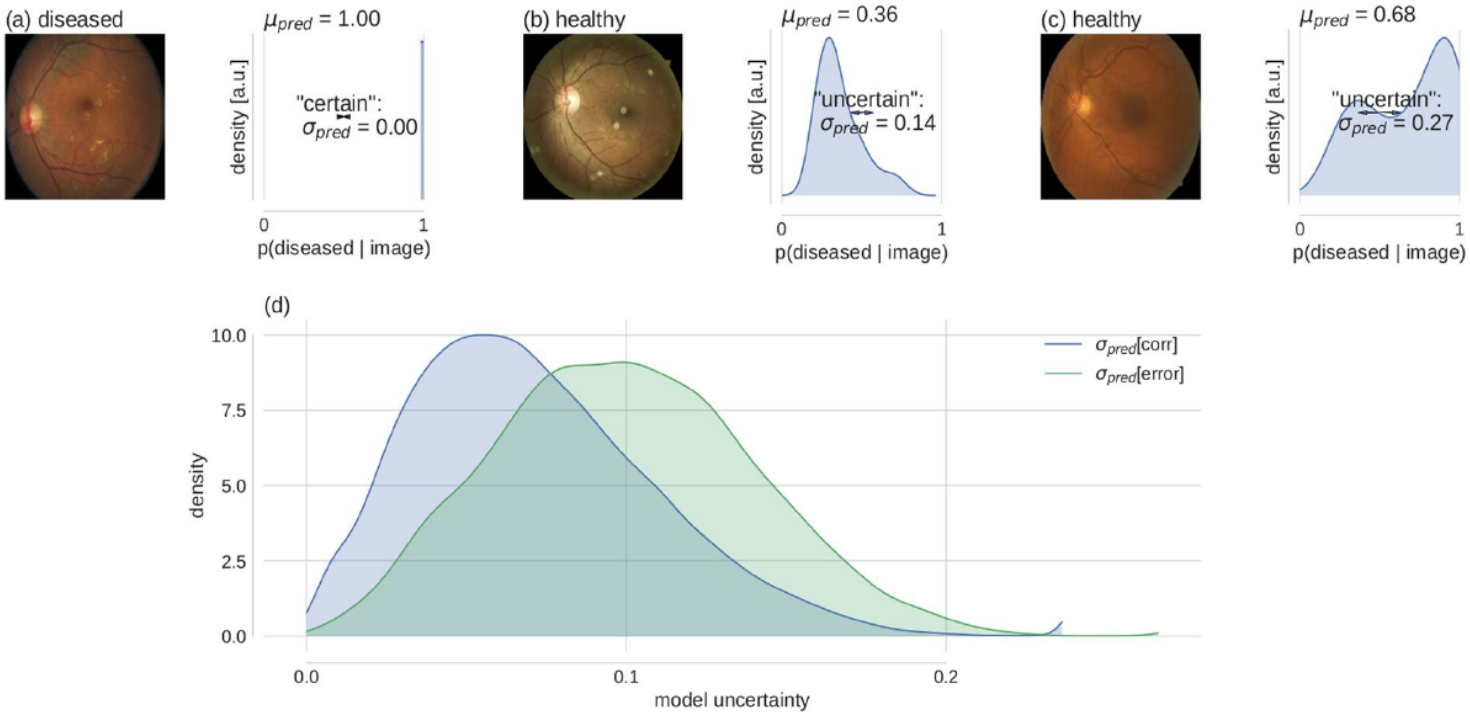


Figure 1. Bayesian model uncertainty for diabetic retinopathy detection. (**a**–**c**) **left:** Exemplary fundus images with human label assignments in the titles. (**a**–**c**) **right**: Corresponding approximate predictive posteriors (Eq. 6) over the softmax output values p(diseased | image) (Eq. 1). Predictions are based on μ_{pred} (Eq. 7) and uncertainty is quantified by σ_{pred} (Eq. 8). Examples are ordered by increasing uncertainty from left to right. (**d**) Distribution of uncertainty values for all Kaggle DR test images, grouped by correct and erroneous predictions. Label assignment for "diseased" was based on thresholding μ_{pred} at 0.5. Given a prediction is incorrect, there is a strong likelihood that the prediction uncertainty is also high.

Why Uncertainty? [Leibig et al., 2017]

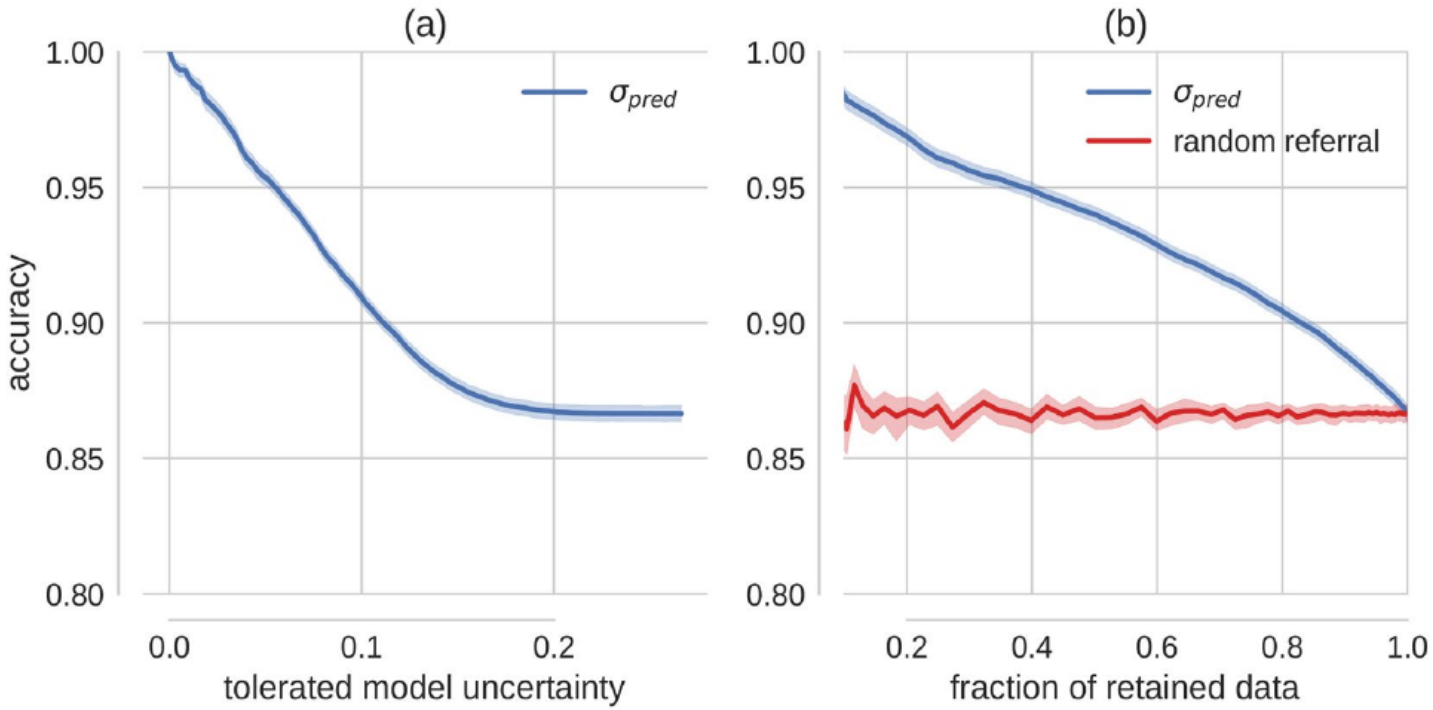


Figure 3. Improvement in accuracy via uncertainty-informed decision referral. (**a**) The prediction accuracy as a function of the tolerated amount of model uncertainty. (**b**) Accuracy is plotted over the retained data set size (test data set size - referral data set size). The red curve shows the effect of rejecting the same number of samples randomly, that is without taking into account information about uncertainty. Exemplarily, if 20% of the data would be referred for further inspection, 80% of the data would be retained for automatic diagnostics. This results in a better test performance (accuracy $\geq$ 90%, point on blue curve) on the retained data than on 80% of the test data sampled uniformly (accuracy $\approx$ 87%, point on red curve). Uncertainty informed decision referral derived from the conventional softmax output cannot achieve consistent performance improvements (Fig. 4).

Why Uncertainty? [Leibig et al., 2017]

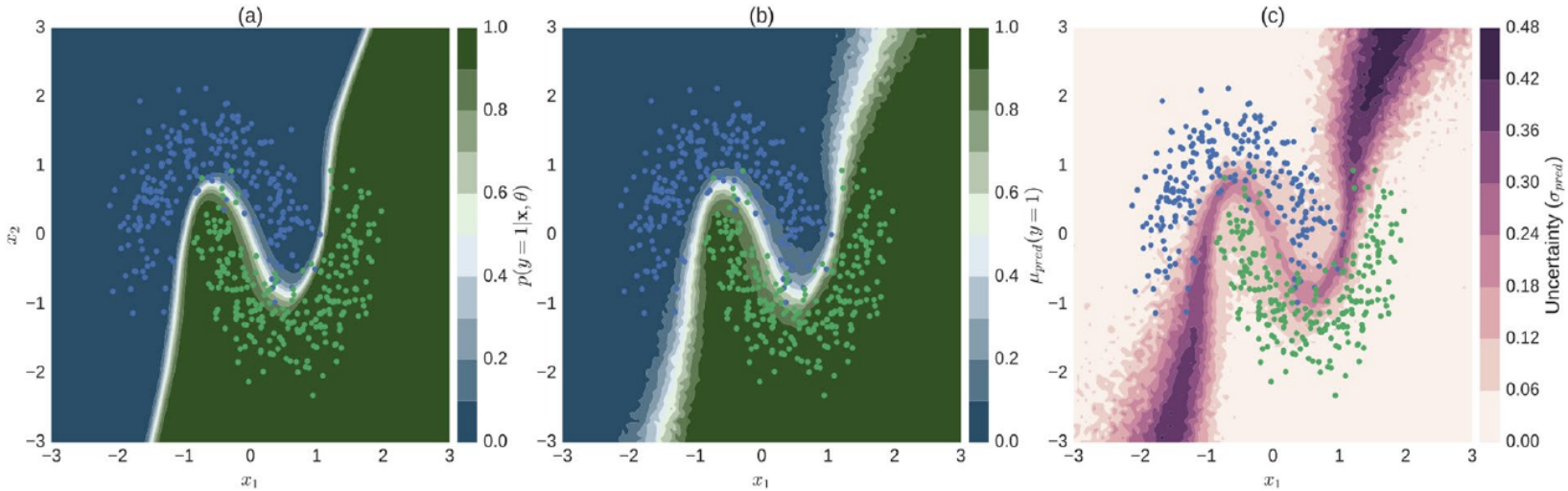
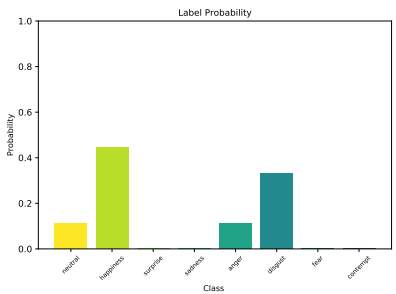


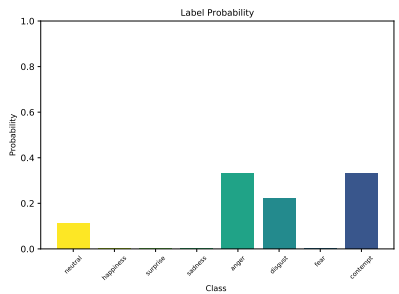
Figure 5. Illustration of uncertainty for a 2D binary classification problem. (**a**) Conventional softmax output obtained by turning off dropout at test time (eq. 1). (**b**) Predictive mean of approximate predictive posterior (eq. 7). (**c**) Uncertainty, measured by predictive standard deviation of approximate predictive posterior (eq. 8). The softmax output (**a**) is overly confident (only a narrow region in input space assumes values other than 0 or 1) when compared to the Bayesian approach (**b**,**c**). Uncertainty (**c**) tends to be high for regions in input space through which alternative separating hyperplanes could pass. Colour-coded dots in all subplots correspond to test data the network has not seen during training.

What is Uncertainty in Machine Learning?

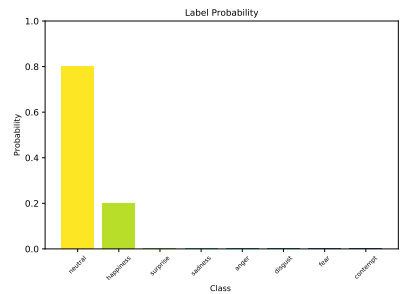
Happiness



Anger



Neutral



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

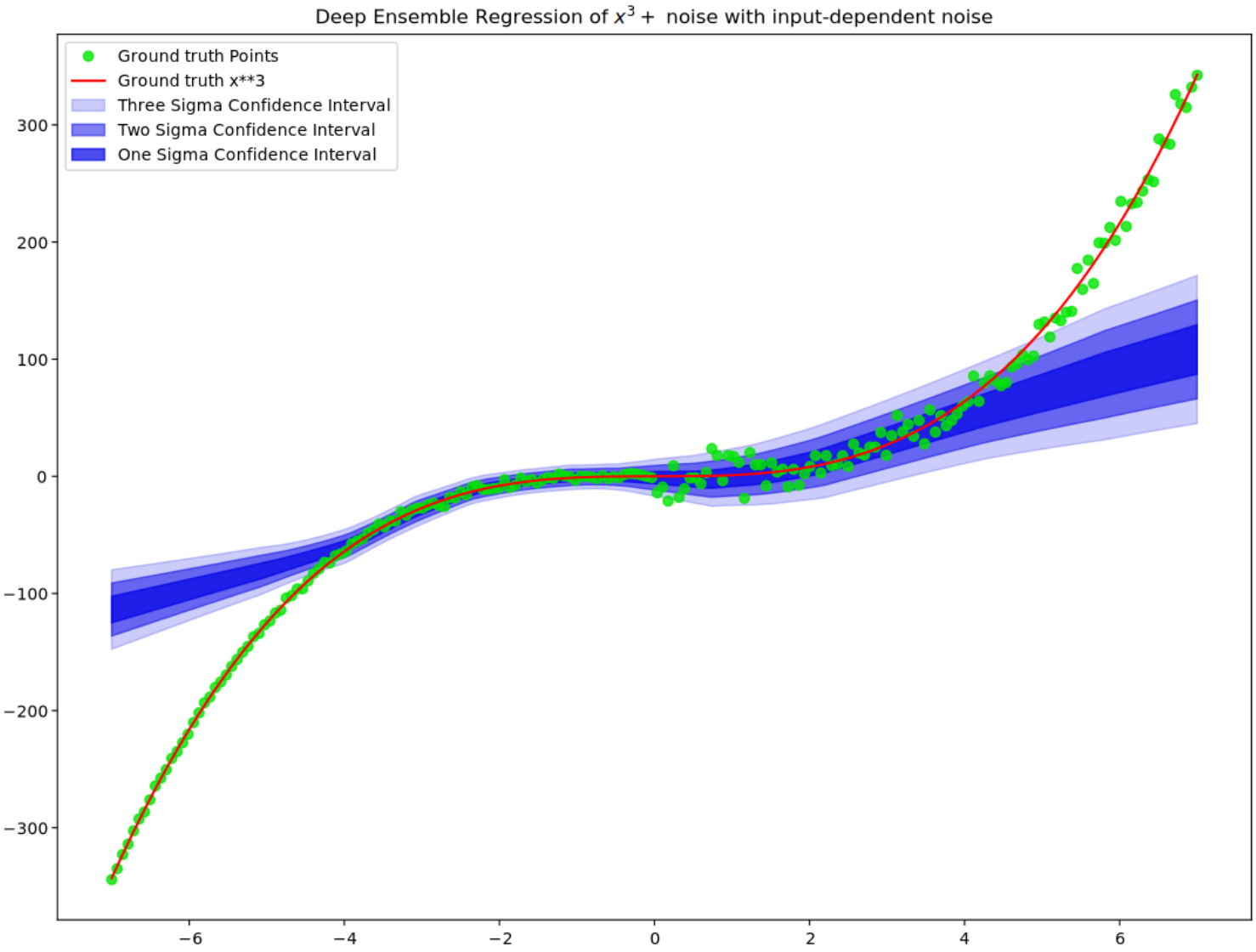
What is Uncertainty in ML?

- 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 or Data Uncertainty

Uncertainty that is inherent to the data, for example, sensor noise, stochastic processes.

Cannot be reduced by adding more information.

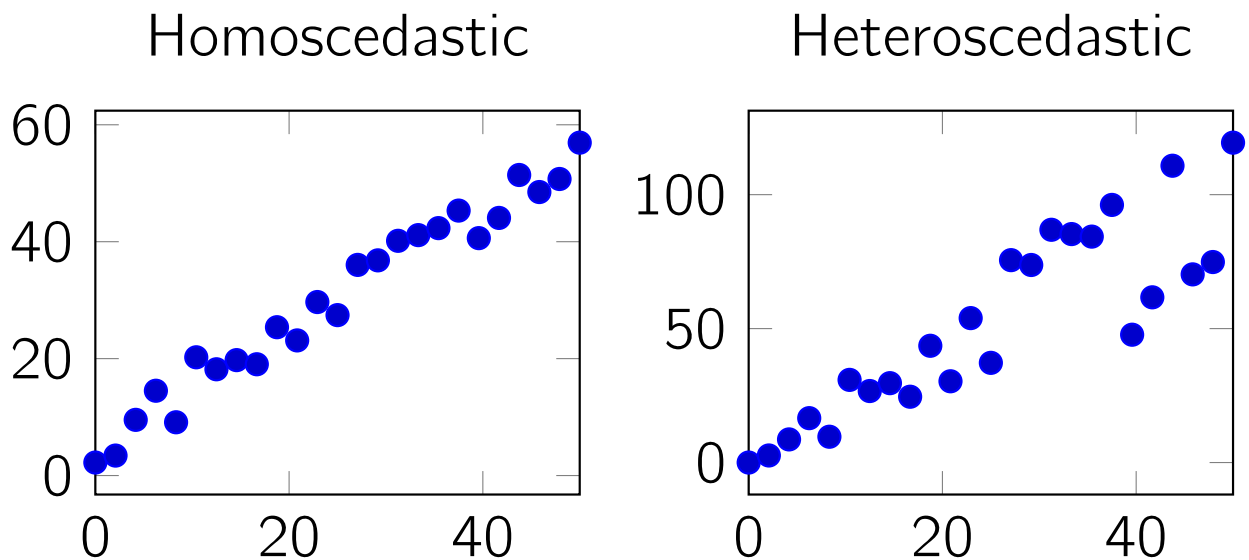
Epistemic or Model Uncertainty

Uncertainty produced by the model, for example, model misspecification, class imbalance, lack of training data.

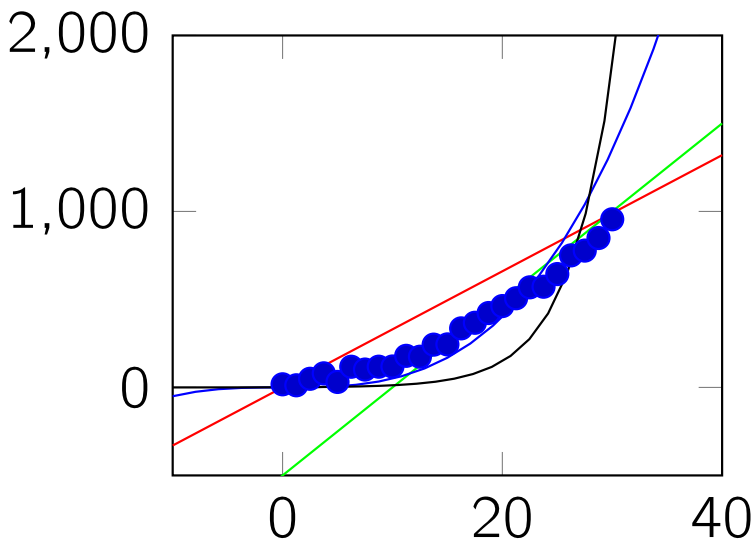
Can be reduced by adding more information to the training process.

Aleatoric Uncertainty

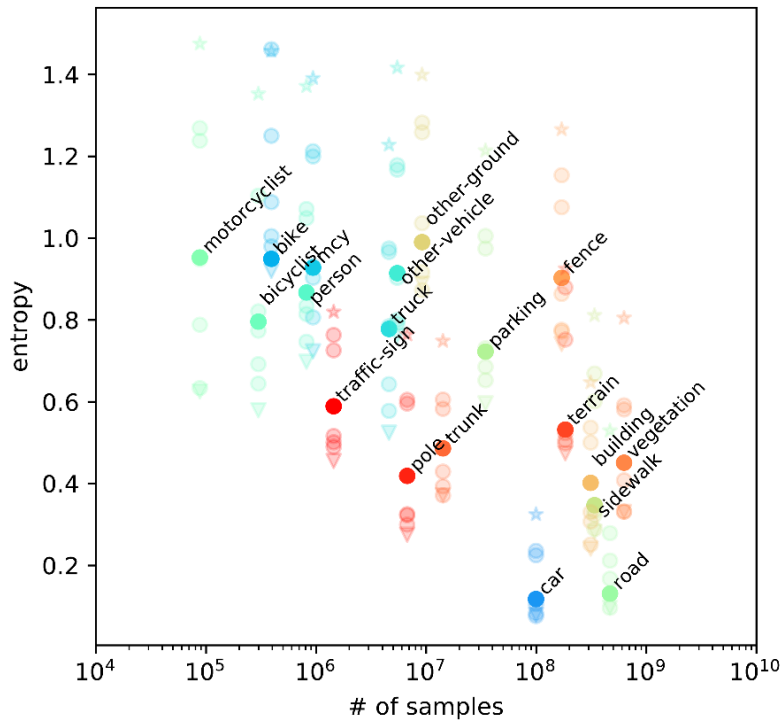
The simplest example of AU is measurements corrupted by additive noise, like $f(x) = x^3 + \epsilon$ Where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and x^3 would be the true function.
If σ^2 is constant, this is called homoscedastic noise, if σ^2 is a function of the input or variable, then it is called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



Epistemic Uncertainty



Model Misspecification



Variations of # Samples in Training Data

Data/Model Uncertainty in Regression

Data Uncertainty

Additive noise added to labels.

Noisy inputs, for example, values produced by sensors.

Model Uncertainty

Missing data points, less density in some areas, or inputs outside of training set range.

Uncertainty in model parameters, leading to uncertainty in the predicted value.

Data/Model Uncertainty in Classification

Data Uncertainty

Incorrect class labels (happens more than you think in most datasets).

Noisy inputs, for example, values produced by sensors.

Model Uncertainty

Missing data points, less density in some areas, or inputs outside of training set range.

Uncertainty in the model parameters, for example, unclear or uncertain decision boundary.

Predictive and Distributional Uncertainty

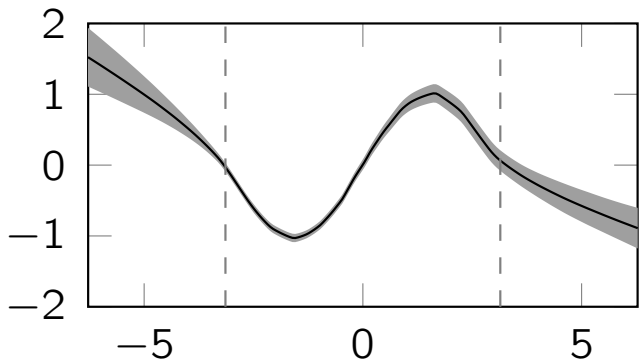
Predictive uncertainty is what models usually output, a combination of aleatoric and epistemic uncertainty.

$$\text{Predictive Uncertainty} = \underbrace{\text{Aleatoric}}_{\text{Data Uncertainty}} + \underbrace{\text{Epistemic}}_{\text{Model Uncertainty}} \tag{1}$$

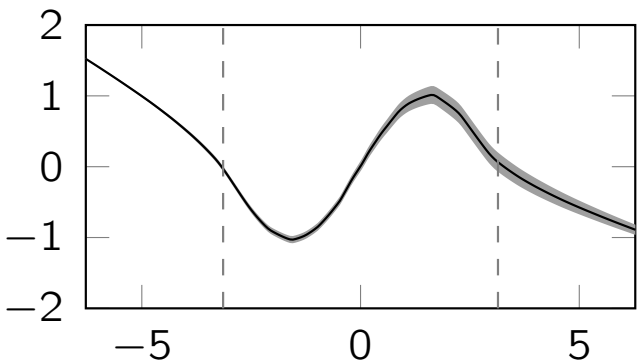
There is also distributional uncertainty, corresponding to lack of knowledge about the correct output distribution.

$$\text{Predictive Uncertainty} = \text{Aleatoric} + \text{Epistemic} + \text{Distributional} \tag{2}$$

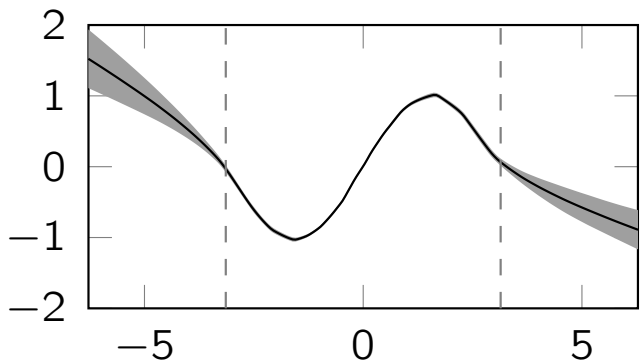
Uncertainty Disentanglement



(a) Predictive Uncertainty



(b) Aleatoric Uncertainty

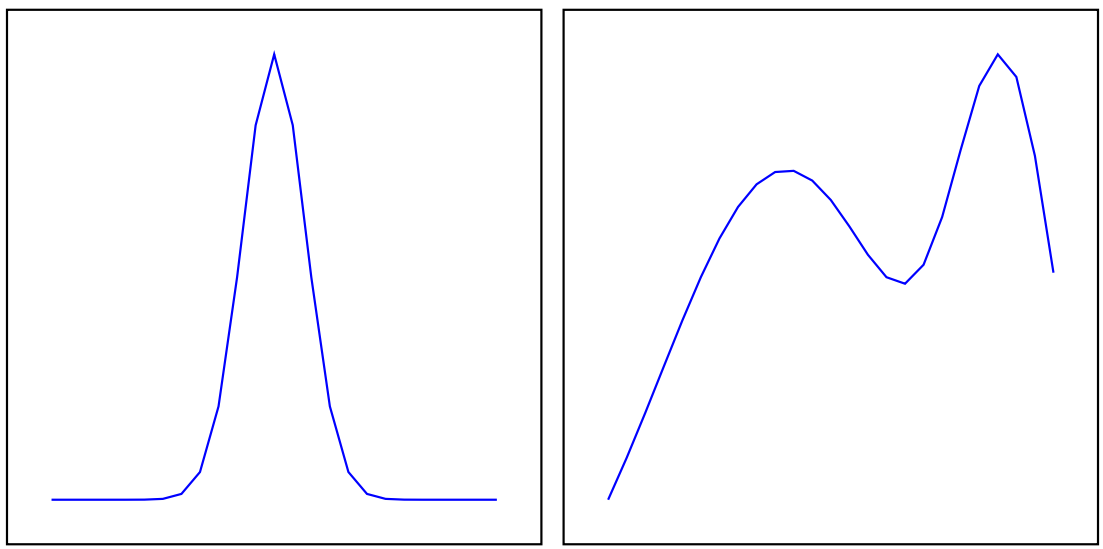


(c) Epistemic Uncertainty

Figure: Sample of uncertainty disentanglement in a toy regression example, produced using an ensemble of 15 neural networks.

Uncertainty Representations

- Before starting with techniques, we need to discuss how uncertainty information is represented.
- In general, this is not so trivial as there are multiple representations, and it depends on the kind of task.
- The most generic representation is to use a probability distribution on the output. This distribution indirectly encodes uncertainty.



Uncertainty Representation - Regression

Assuming that the output is represented by $f(x)$, then there are two principal ways to represent uncertainty.

Confidence Intervals

The output is within some defined interval $[a, b]$:

$$f(x) \in [a, b]$$

Usually some methods give a probability that the output belongs to the interval.

Uncertainty Representation - Regression

Assuming that the output is represented by $f(x)$, then there are two principal ways to represent uncertainty.

Mean and Variance

Uncertainty is represented as the variation of the output from the mean ($f(x)$):

$$f(x) \pm \sigma$$

A equivalent interval would be $f(x) \in [f(x) - \sigma, f(x) + \sigma]$

Uncertainty Representation - Classification

- For classification, representing uncertainty is a bit more hard.
- The only robust representation is to use a discrete probability distribution.
- The easiest way to implement it is to use a softmax activation (for multi-class) or a sigmoid activation (for binary classification).

Overconfidence and Calibration

- Many models produce probabilities or confidence intervals that are not good, mostly are overconfident.
- Probabilities or confidence intervals must represent likelihood of correct prediction, and this can be measured by calibration.

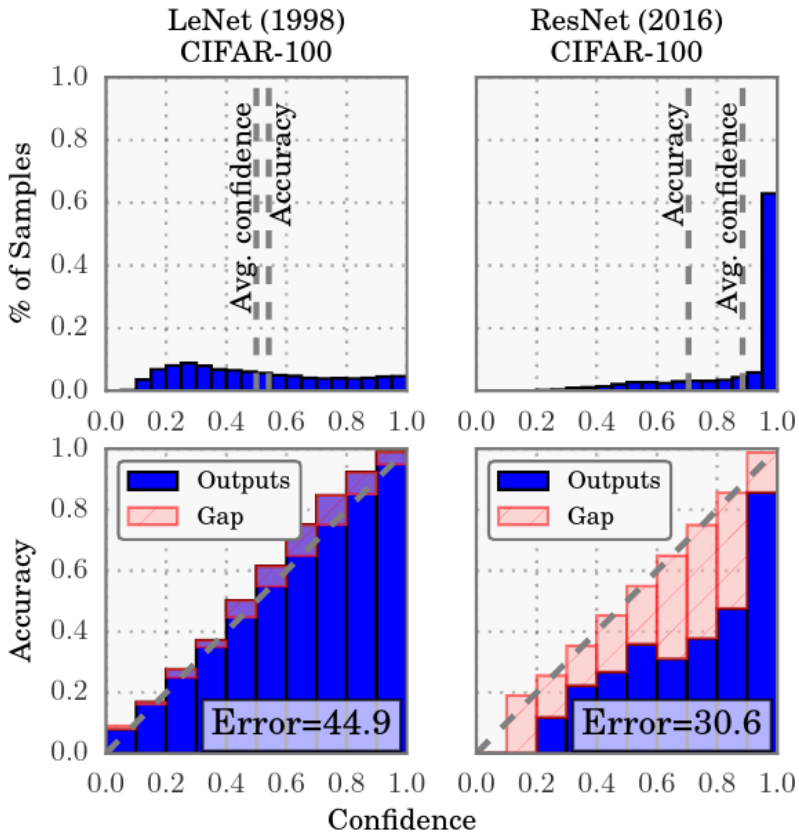


Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

Entropy

Entropy is a measurement of the "information content" in a probability distribution. It is defined as:

$$H = -\sum_{x} P(x) \log P(x) \tag{3}$$

The entropy is important as it is directly related to uncertainty. The units of entropy are called bits if one uses the base-two logarithm.
The uniform distribution is the one that maximizes entropy, as its results are harder to predict. For a fixed mean and variance, the Gaussian distribution is the one that maximizes the entropy.

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]

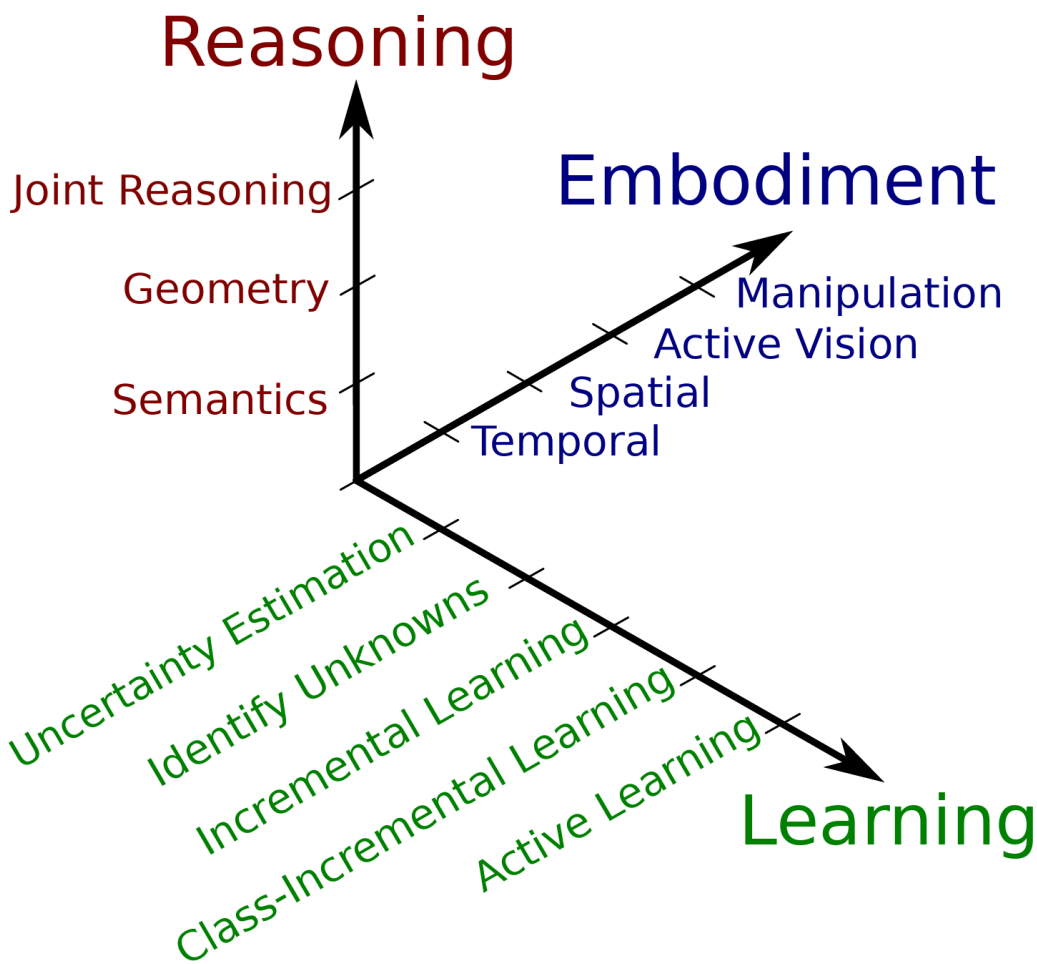


Figure from [Sünderhauf et al., 2018].

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

Level	Name	Description
0	Closed-Set Assumptions	The system can detect and classify objects of classes known during training. It provides uncalibrated confidence scores.
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.
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.

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

Level	Name	Description
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.
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.
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.

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.

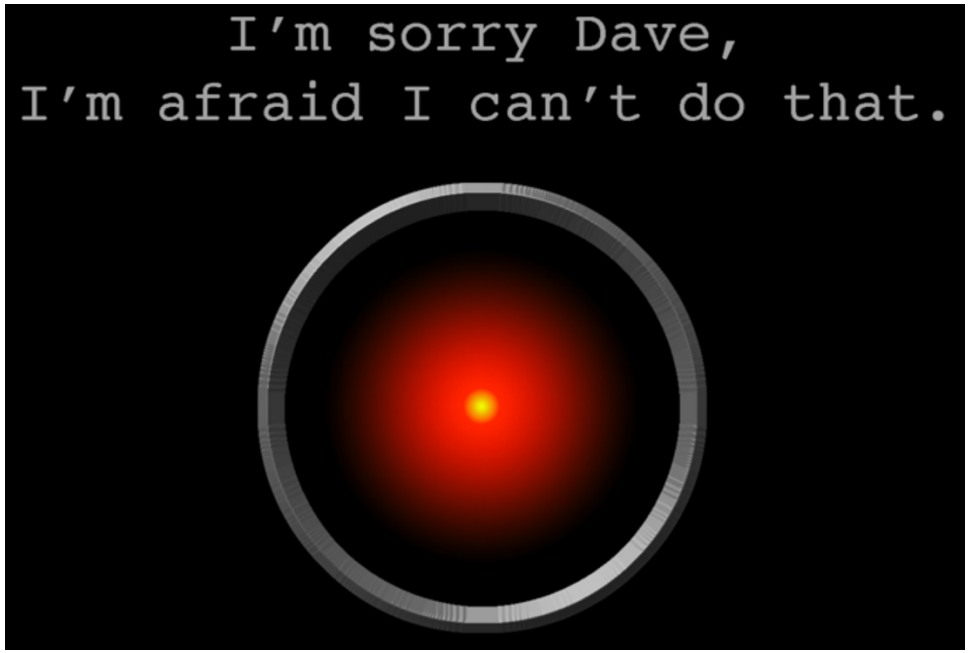
Challenges and Applications

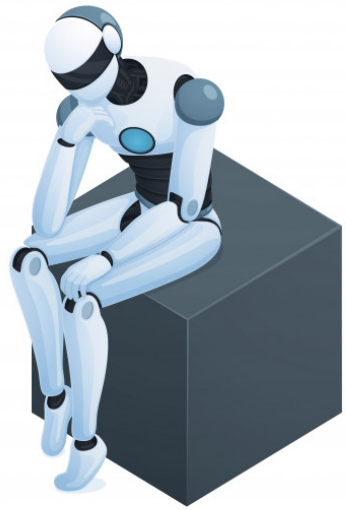
Autonomous Driving

Autonomous Driving is a very important application field, as the whole reason for AD to be desirable, is that it can provide safer driving than humans. But this is not automatic, safety has to be engineered and safe methods have to be developed and used.

- AD usually fails with variations of the test environment, like different weather, physical location of roads, and environmental conditions like snow and rain.
- A car using AD has to detect unusual and strange situations and alert the driver, and has to do this with nearly 100% precision.
- In many AD systems, usual situations are labeled by humans, which does not scale (And it is strange to claim super-human driving from human samples).

Objective - Safe and Trustable Learning Systems





Objective - Safe and Trustable Learning Systems

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.

Questions?

Questions to Think About

- What are Aleatoric and Epistemic Uncertainty? (Your own definitions).
- What is Disentangling of Uncertainty?
- Give some real-world examples of Uncertainty that are different to the ones mentioned in the lecture.

Bibliography I

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.

Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018.