Deep Bayesian Active Learning with Image Data

based on the original research article of the same name by Gal et al. (2017)

Munich, July 16th 2024
Seminar “Machine learning with limited labels”

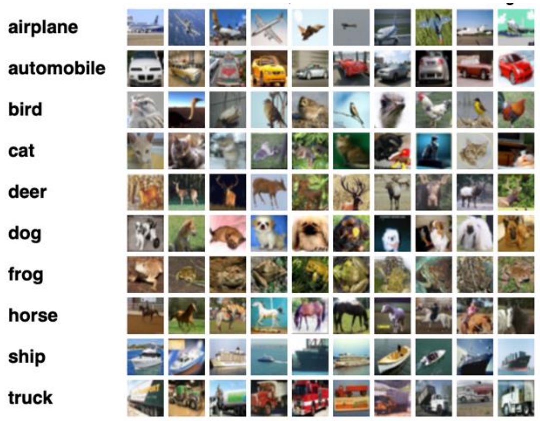
Nicolas Malz
Computer Science and Statistics
LMU Munich

Outline

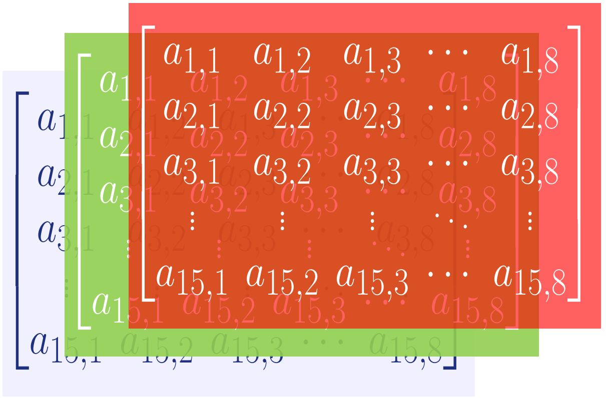
1. The active learning problem for high-dimensional tasks
2. Related work: The early days of active learning for images
3. Method: Uncertainty in neural networks; Monte Carlo dropout; Bayesian CNN
4. Experiments: Deterministic CNN vs. Bayesian CNN vs. Ensembles
5. Conclusion, Outlook, and Discussion

Outline

The Active Learning problem for Image Classification



Dataset



Features

- Given a set of images ("dataset"), we want to create a model that
  - 1) …accurately classifies images, i.e. assigns them to the correct class,
  - 2) …all while having to label as little data as possible.
- **Challenge: High dimensionality** of inputs. An ImageNet image has 544509 features
- Consequence: feature extraction is very challenging, labelling is expensive

Takeaway 1: We need a powerful model capable of handling all these high-dimensional features well

Sources: The author's own elaboration, CIFAR-10 dataset by Krizhevsky et al. (2009)

Outline

Related Work: What others had tried so far

MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Multi-Class Active Learning for Image Classification

CVPR 2009

ZHUXJ@CS.CMU.EDU
LAFFERTY@CS.CMU.EDU
ZOUBIN@GATSBY.UCL.AC.UK

This CVPR2013 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Adaptive Active Learning for Image Classification

Xin Li Yuhong Guo
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122
{xinli,yuhong}@temple.edu

Abstract

Recently active learning has attracted a lot of attention in computer vision field, as it is time and cost consuming to prepare a good set of labeled images for vision data analysis. Most existing active learning approaches employed in computer vision adopt most uncertainty measures as instance selection criteria. Although most uncertainty query selection strategies are very effective in many circumstances, they fail to take information in the large amount of unlabeled instances into account and are prone to querying outliers. In this paper, we present a novel adaptive active learning approach that combines an information density measure and a most uncertainty measure together to select critical instances to label for image classifications. Our experiments on two essential tasks of computer vision, object recognition and scene recognition, demonstrate the efficacy of the proposed approach.

1. Introduction

Image classification has a long history in computer vision research, and it remains a major challenge due to the broad intra-class diversity of images caused by shape, color, size, or environmental conditions. To build a robust image classifier, it typically requires a large number of labeled training instances. For example, 10,000 instances of handwriting digits are used for training classifiers in [33]. It is time and cost consuming to prepare such a large set of labeled training instances. On the other hand, one fascinating characteristic of human vision system is that we can categorize image objects with only few labeled training instances. Is it possible for a computer to achieve this with the solid support of machine learning techniques? This is the motivation of this research. We aim to develop an effective active learning method to build a competitive classifier with a limited amount of labeled training instances.

Training a good classifier with minimal labeling cost is a critical challenge posed in machine learning research. Randomly selecting unlabeled instances to label is inefficient in many situations, since non-informative or redundant instances might be selected. Aiming to reduce labeling effort, active learning methods have been adopted to control the labeling process. Recently, active learning has been studied in computer vision [3, 14, 13, 15, 16], focusing on pool-based setting. These works however merely evaluate the informativeness of instances with most uncertainty measures, which assume an instance with higher classification uncertainty is more critical to label. Although the most uncertainty measures are effective on selecting informative instances in many scenarios, they only capture the relationship of the candidate instance with the current classification model and fail to take the data distribution information contained in the unlabeled data into account. This may lead to selecting non-useful instances to label. For example, an outlier can be most uncertain to classify, but useless to label. This suggests representativeness of the candidate instance in addition to the classification uncertainty should be considered in developing an active learning strategy.

In this paper, we propose a novel adaptive active learning strategy that exploits information provided by both the labeled instances and the unlabeled instances for query selection. Our new query selection measure is an adaptive combination of two terms: an uncertainty term based on the current classifier trained on the labeled instances; and an information density term that measures the mutual information between the candidate instance and the remaining unlabeled instances. The combination of the two terms is given in a general weighted product form. We seek to obtain an adaptive combination of the two terms by selecting the weight parameter to minimize the expected classification error on unlabeled instances. We conduct experiments on a few benchmark image classification datasets and present promising results for the proposed active learning method.

2. Related Work

A large number of active learning techniques have been developed in the literature. Most of them have been focused

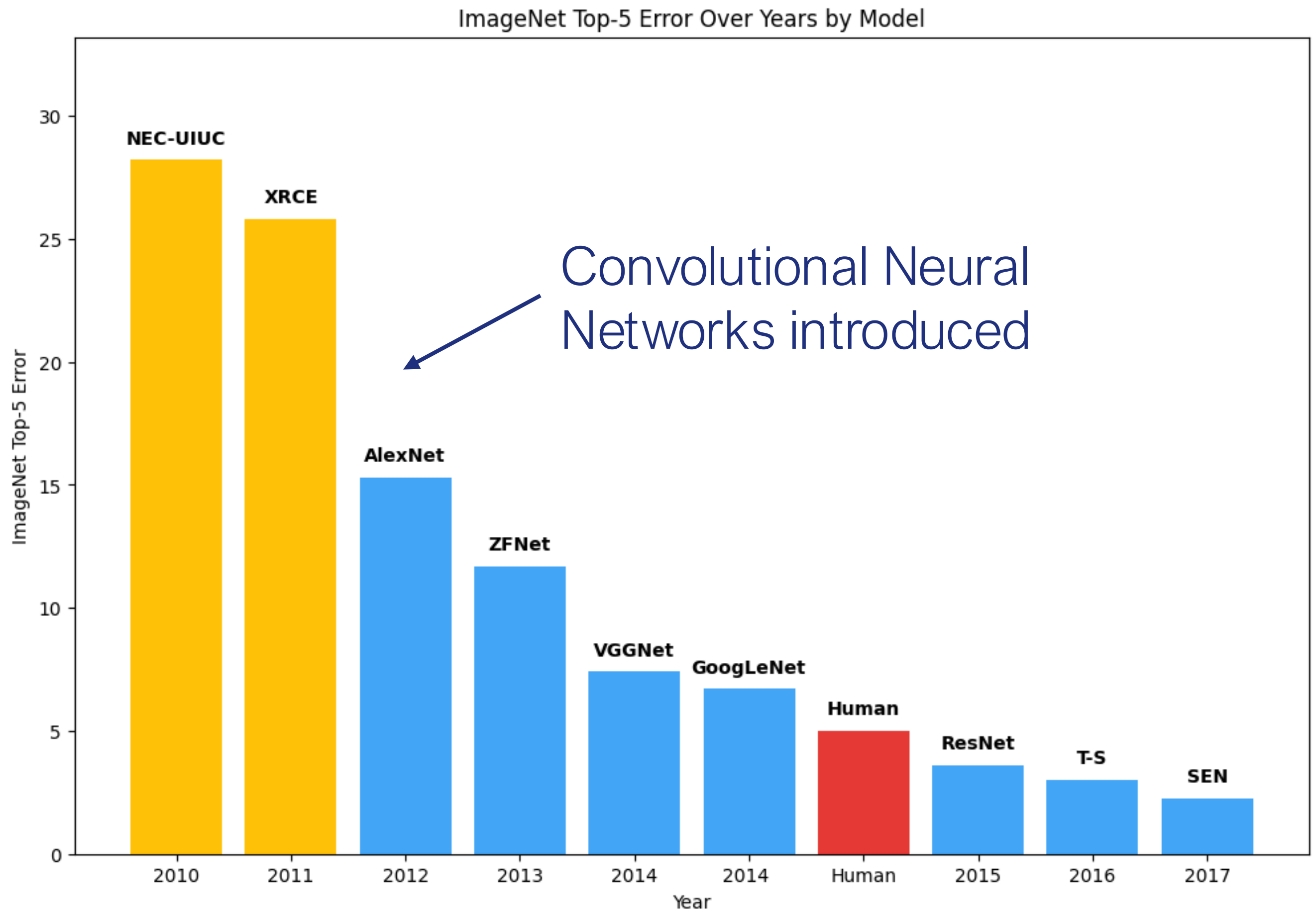
859

RBF Kernel downsides:

- No spatial awareness
- Loss of edge information
- Still (too) high-dimensional
- Uniform treatment of all parts of an image

Sources: The author's own elaboration, scikit-learn image dataset (2024)

The Rise of Deep Learning for Image Classification

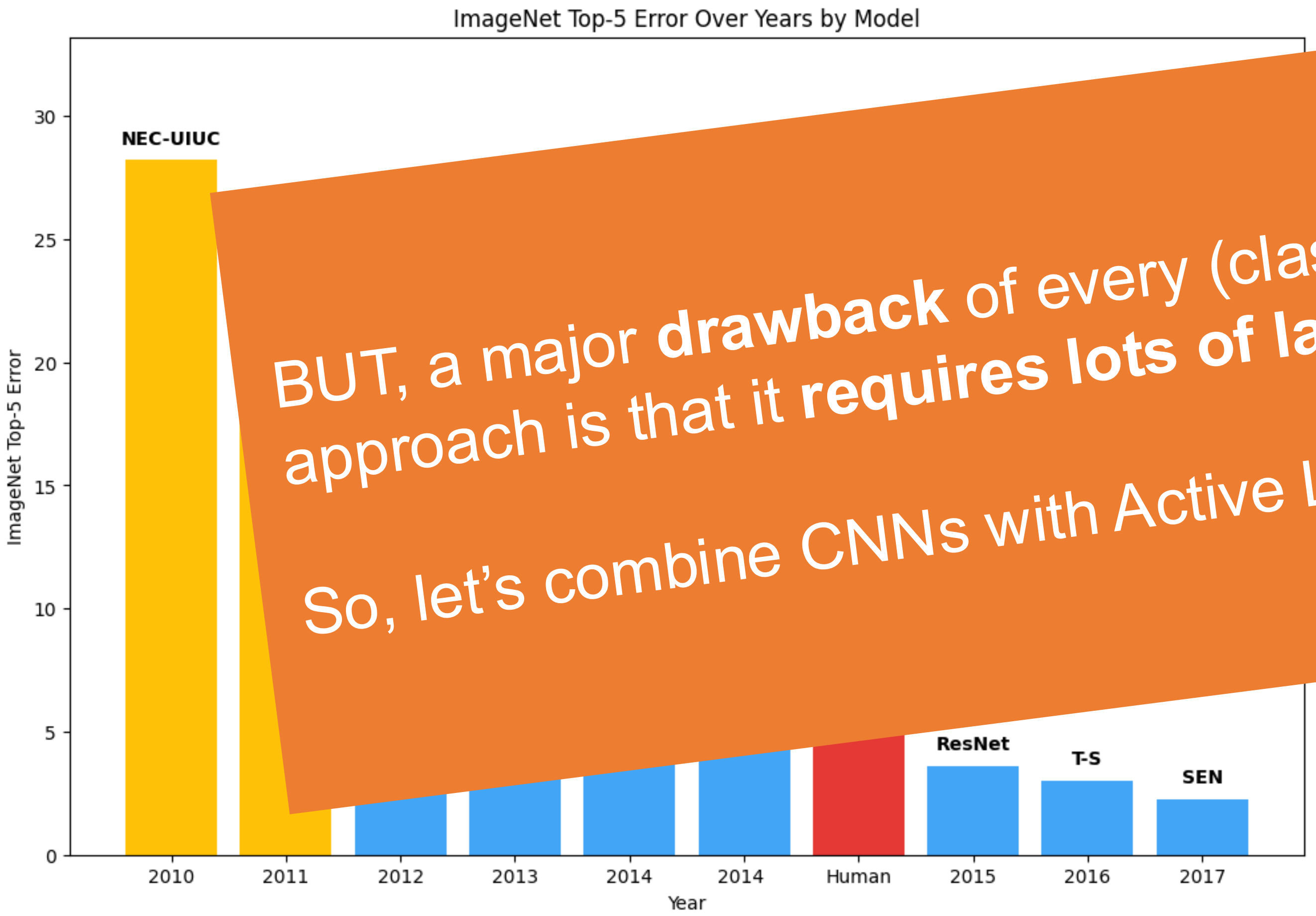


- **Neural Networks** are just powerful, complex compositions of functions, capable of capturing linear and non-linear relationships in the data
- **Convolutional Neural Networks** (CNNs) have proven very effective at image classification tasks
- CNNs **perform feature selection efficiently** as opposed to using a classic fully connected neural network (aka Multi-Layer Perceptron)

Takeaway 2: Using a Convolutional Neural Network motivates the active learning approach!

Sources: The author's own elaboration, ImageNet Competition Benchmarks (2022)

The Rise of Deep Learning for Image Classification



BUT, a major **drawback** of every (classic) deep learning approach is that it **requires lots of labelled data!**

So, let's combine CNNs with Active Learning…

owerful,
ctions,
d non-
rks
ctive at
on
as opposed to using a classic fully connected neural network (aka Multi-Layer Perceptron)

Takeaway 2: Using a Convolutional Neural Network motivates the active learning approach!

Sources: The author's own elaboration, ImageNet

Outline

The Challenge of Integrating Deep and Active Learning

Active Learning Checklist

- ✓ Model
- ✓ Dataset with a few labelled data points
- ✓ Rest of unlabelled data points in a pool
- ? **Acquisition mechanism** to add new labelled data points from pool to training set
  - **Find those datapoints that are likely to improve the model's performance**
  - **Assumption:** Model can learn most by looking at its most uncertain predictions

Acquisition Function

⟺ Find those unlabelled data points that maximize the acquisition function

Options for this function:

- Max-Entropy
- Mutual Information (Bayesian Active Learning by Disagreement, BALD)
- Variation Ratios
- …
- Random (baseline)

Sources: The author's own elaboration, Gal et al. (2017)

The Challenge of Integrating Deep and Active Learning

Acquisition Functions

Entropy

$$\mathbb{H}[y|\mathbf{x};\mathcal{D}_{\text{train}}]$$

$$:= -\sum_c p(y=c|\mathbf{x};\mathcal{D}_{\text{train}}) \log p(y=c|\mathbf{x};\mathcal{D}_{\text{train}})$$

BALD

$$\mathbb{I}[y;\boldsymbol{\omega}|\mathbf{x};\mathcal{D}_{train}]$$

$$= \mathbb{H}[y|x;\mathcal{D}_{train}] - \mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D}_{train})}[\mathbb{H}[y|\mathbf{x};\mathcal{D}_{train}]]$$

Var-R

$$\textit{Var-Ratio}[\mathbf{x}] = 1 - \max_y p(y|x,\mathcal{D}_{train})$$

Current CNN setup

- Input of our model: an image
- Output of our model: class that image belongs to
    - **Where's the uncertainty?**

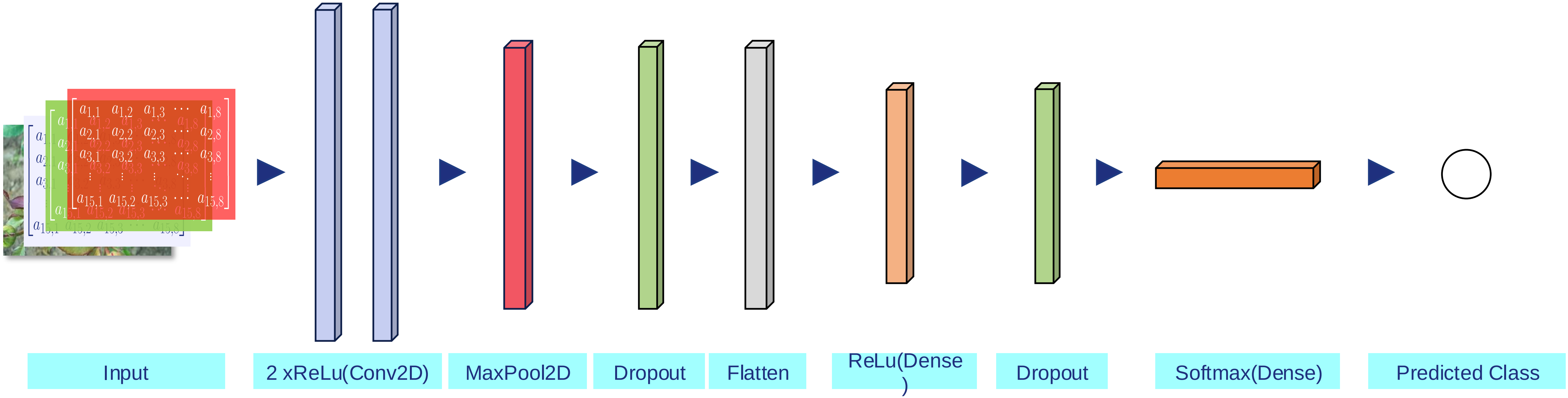
How do we get $p(y=c|\mathbf{x};\mathcal{D}_{train})$**?**

In other words: How do we quantify model prediction uncertainty if Neural Networks only provide us with a deterministic point estimate?

Takeaway 3: We need to quantify how confident the model is in its predictions to use it in an active learning setting.

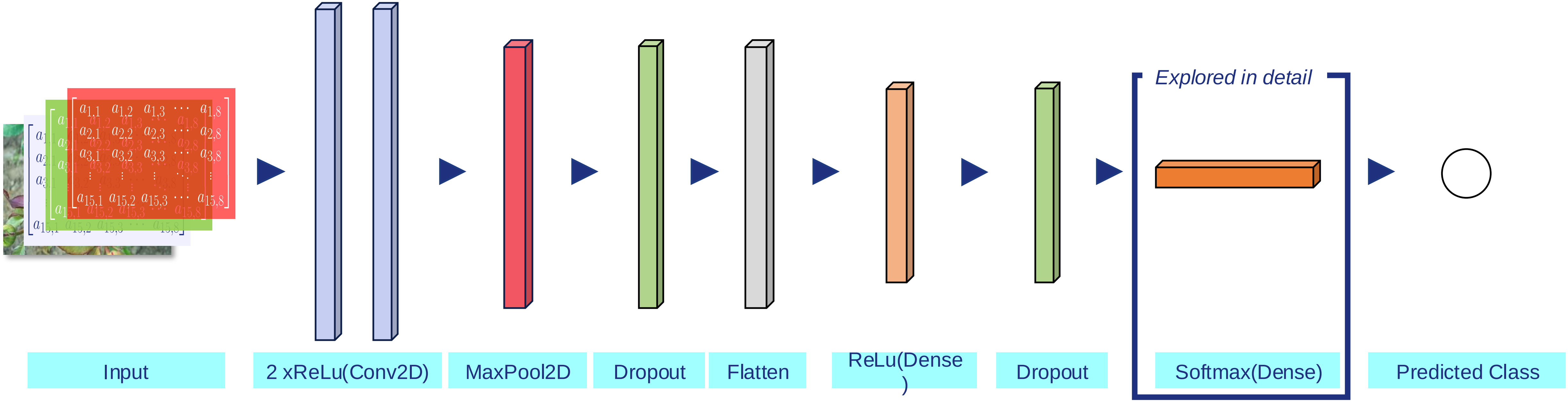
Sources: The author's own elaboration, Gal et al. (2017)

Where is the model prediction uncertainty in a CNN?



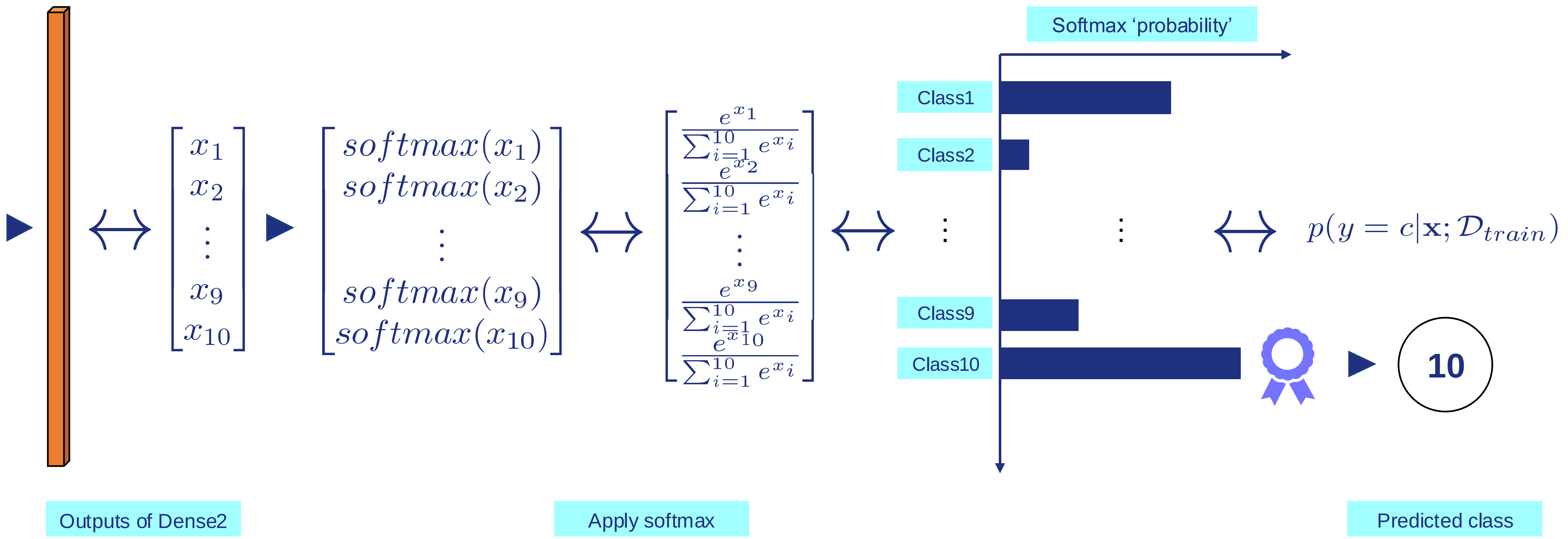
Sources: The author’s own elaboration based on the model used by Gal et al. (2017) and Keras standard CNN implementation example for MNIST (GitHub: @fchollet, 2015)

Where is the prediction uncertainty in a CNN?



Sources: The author's own elaboration based on the model used by Gal et al. (2017) and Keras standard CNN implementation example for MNIST (fchollet, 2015)

Where is the prediction uncertainty in a CNN?

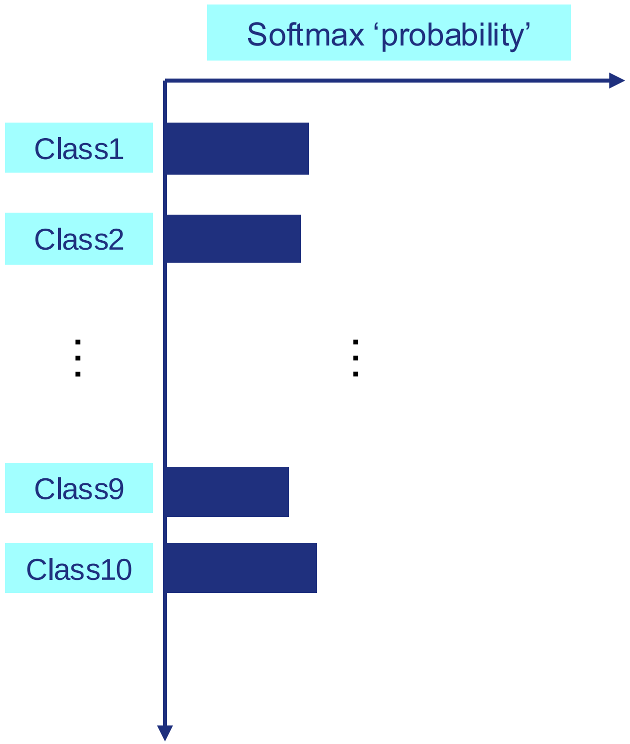


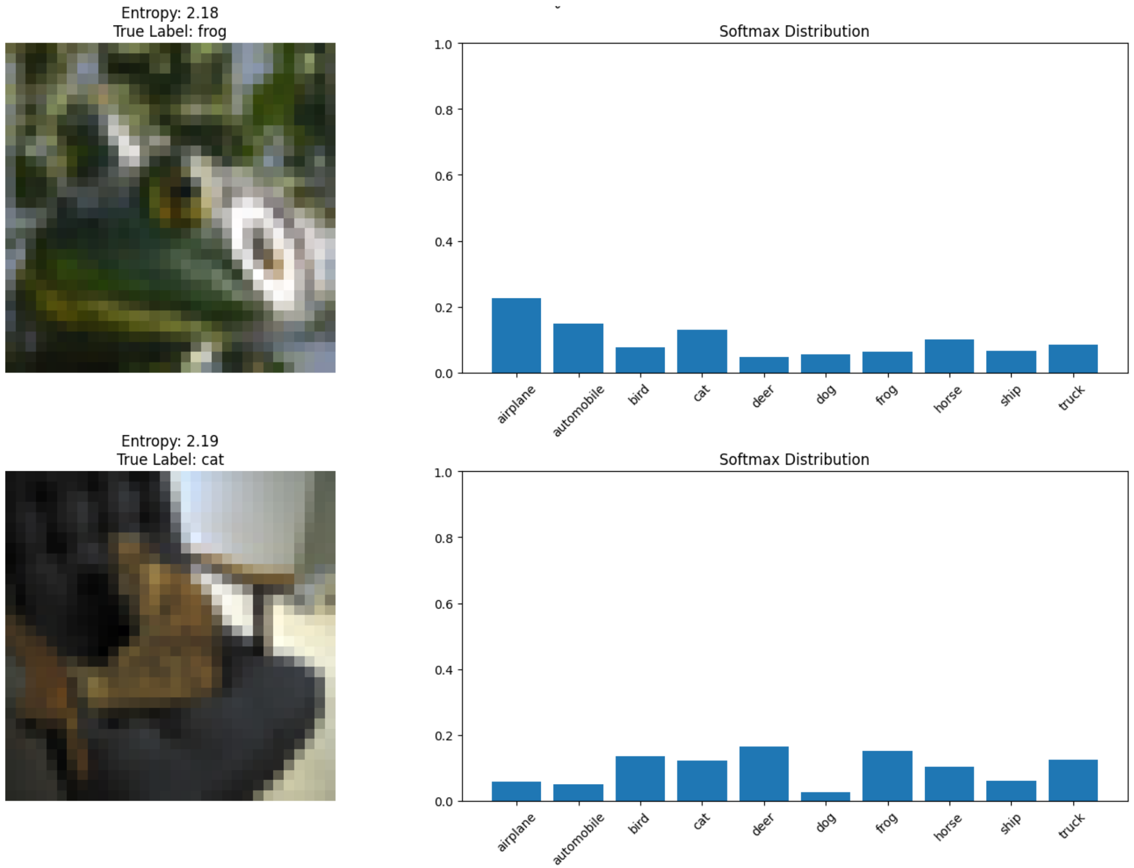
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_9 \\ x_{10} \end{bmatrix} \quad \begin{bmatrix} softmax(x_1) \\ softmax(x_2) \\ \vdots \\ softmax(x_9) \\ softmax(x_{10}) \end{bmatrix} \quad \begin{bmatrix} \frac{e^{x_1}}{\sum_{i=1}^{10} e^{x_i}} \\ \frac{e^{x_2}}{\sum_{i=1}^{10} e^{x_i}} \\ \vdots \\ \frac{e^{x_9}}{\sum_{i=1}^{10} e^{x_i}} \\ \frac{e^{x_{10}}}{\sum_{i=1}^{10} e^{x_i}} \end{bmatrix}$$

Takeaway 4: In a classic CNN for classification, uncertainty is displayed in the final Softmax layer.

Sources: The author‘s own elaboration based on Bishop and Bishop (2024)

Example: Inherent ambiguity in the data



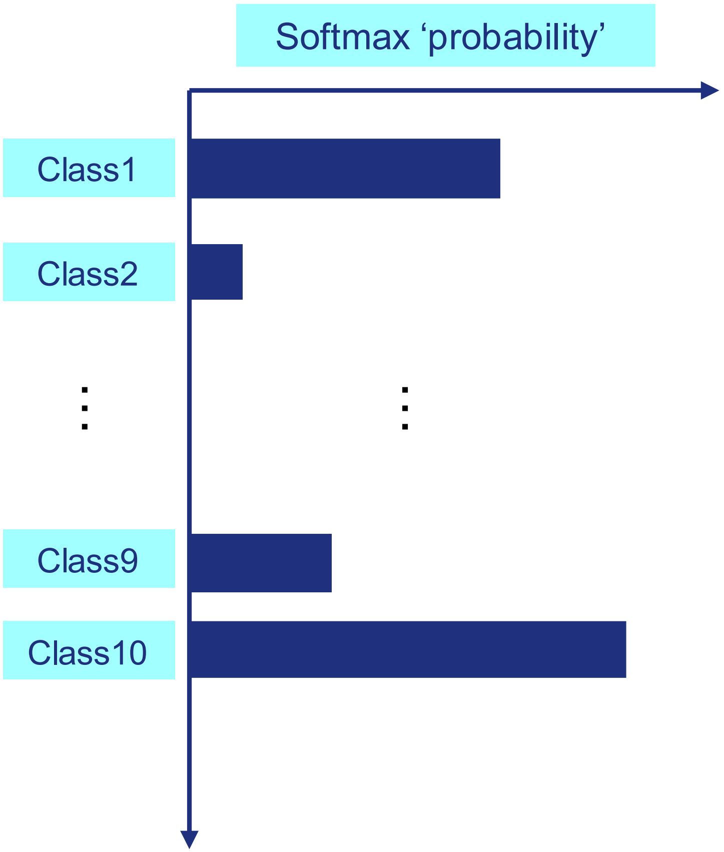


Max. uncertainty ⇔ Softmax yields a uniform distribution

What do the images show?
This is just a point estimate of model uncertainty!

Sources: The author‘s own elaboration based on Bishop and Bishop (2024), CIFAR-10 by Krizhevsky et al. (2009), and Gal et al. (2016, 2017)

The problem with output layer (Softmax) probabilities

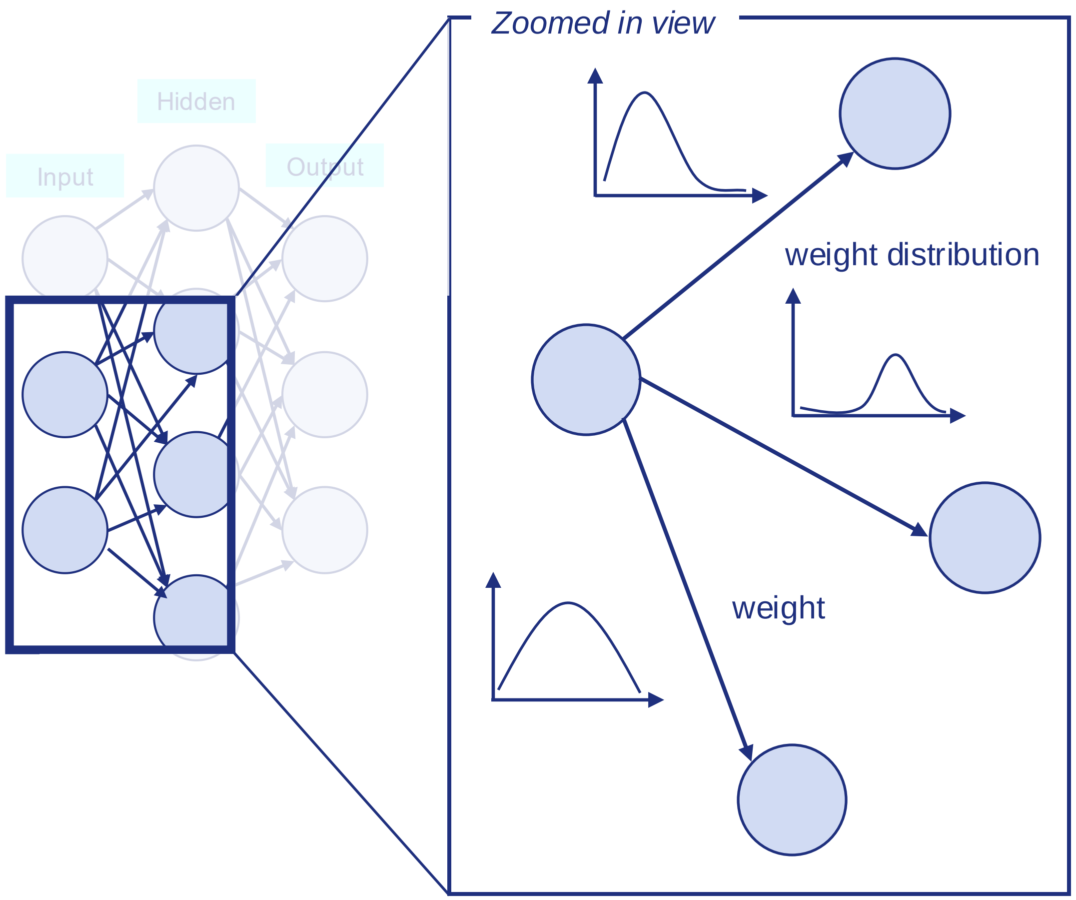


- In our model, the Softmax probabilities for a given input are a **reflection of inherent noise in the data**
  - **:= aleatoric uncertainty**
- **They do not capture the model’s uncertainty about its parameters**
  - We want to use active learning to improve our model, i.e. its parameters!
  - **:= epistemic uncertainty**
- **In essence: How certain are we about a single prediction vs. about our model’s parameters being correct**

Takeaway 5: A classic CNN first and foremost captures aleatoric uncertainty, but we want a measure of epistemic uncertainty, i.e. uncertainty w.r.t. our **model parameters**, for our active learning setting.

Sources: The author‘s own elaboration based on Bishop and Bishop (2024) and Gal et al. (2016, 2017)

New Perspective: Bayesian Neural Network



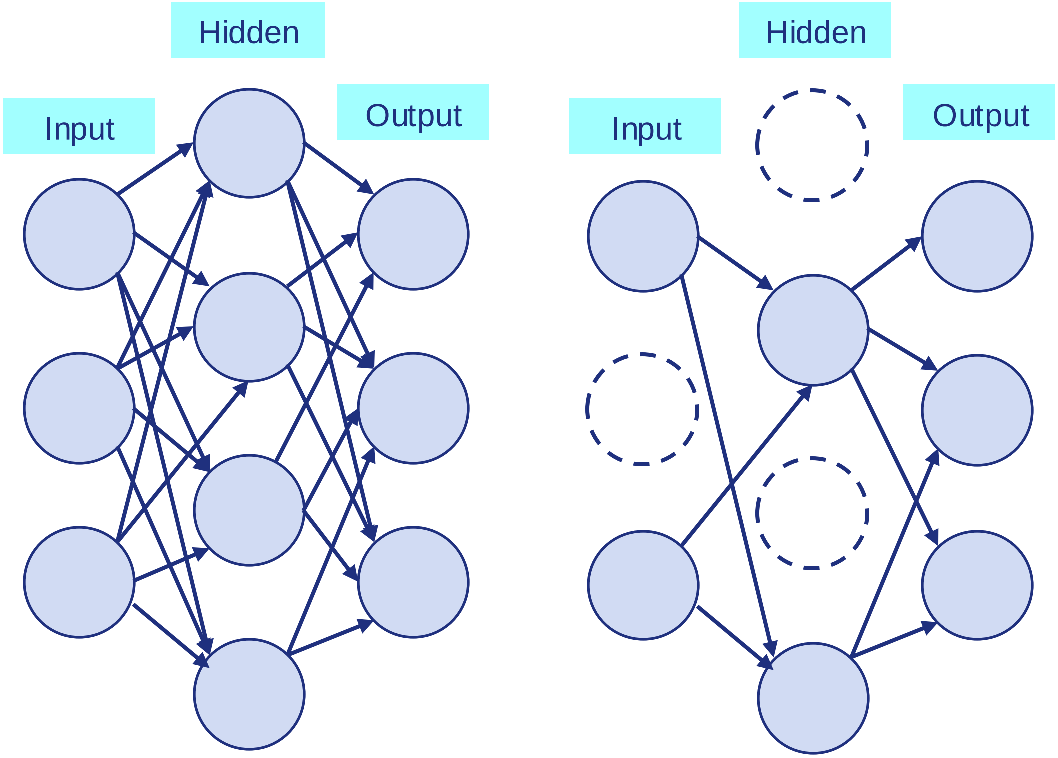
- In a Bayesian Neural Network, **weights are not fixed** (point estimates) but **have a distribution**
- We therefore include **uncertainty w.r.t. the model's weights** in training
- As data 'flows' through the model (training), these distributions are updated ("marginal" Bayes' Theorem)
- Computing the actual posterior distribution (marginals) of weights given the data and prior is computationally intractable:

$$p(y = c \mid \mathbf{x}, \mathcal{D}_{\text{train}}) = \int p(y = c \mid \mathbf{x}, \omega) p(\omega \mid \mathcal{D}_{\text{train}}) d\omega$$

Takeaway 6: We need a tractable approximation for the posterior.

Sources: The author's own elaboration based on Gal et al. (2016)

Solution: Use Dropout at Prediction Time for Inference

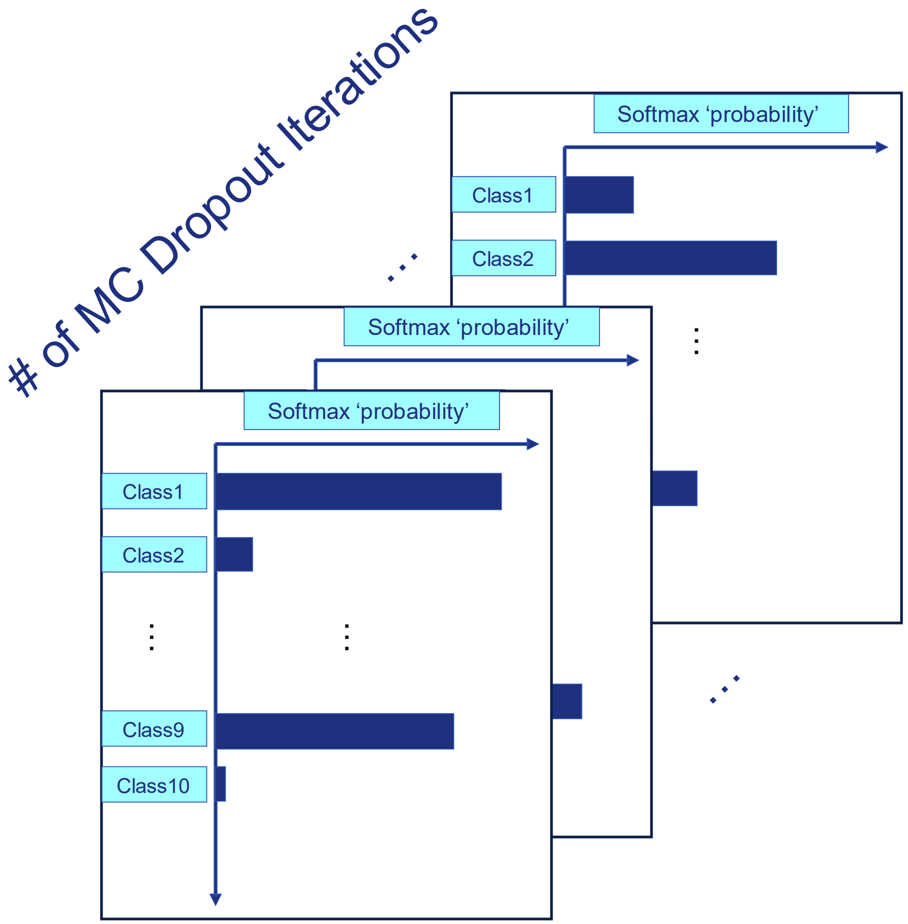


Dropout is a **regularization** technique **randomly sets nodes from the network to zero during training**; a way to simulate model averaging without training multiple models.

Sources: The author's own elaboration based on Srivastava et al. (2014)

What happens if we leave dropout on during prediction?

Solving the Uncertainty Problem in Deep Learning

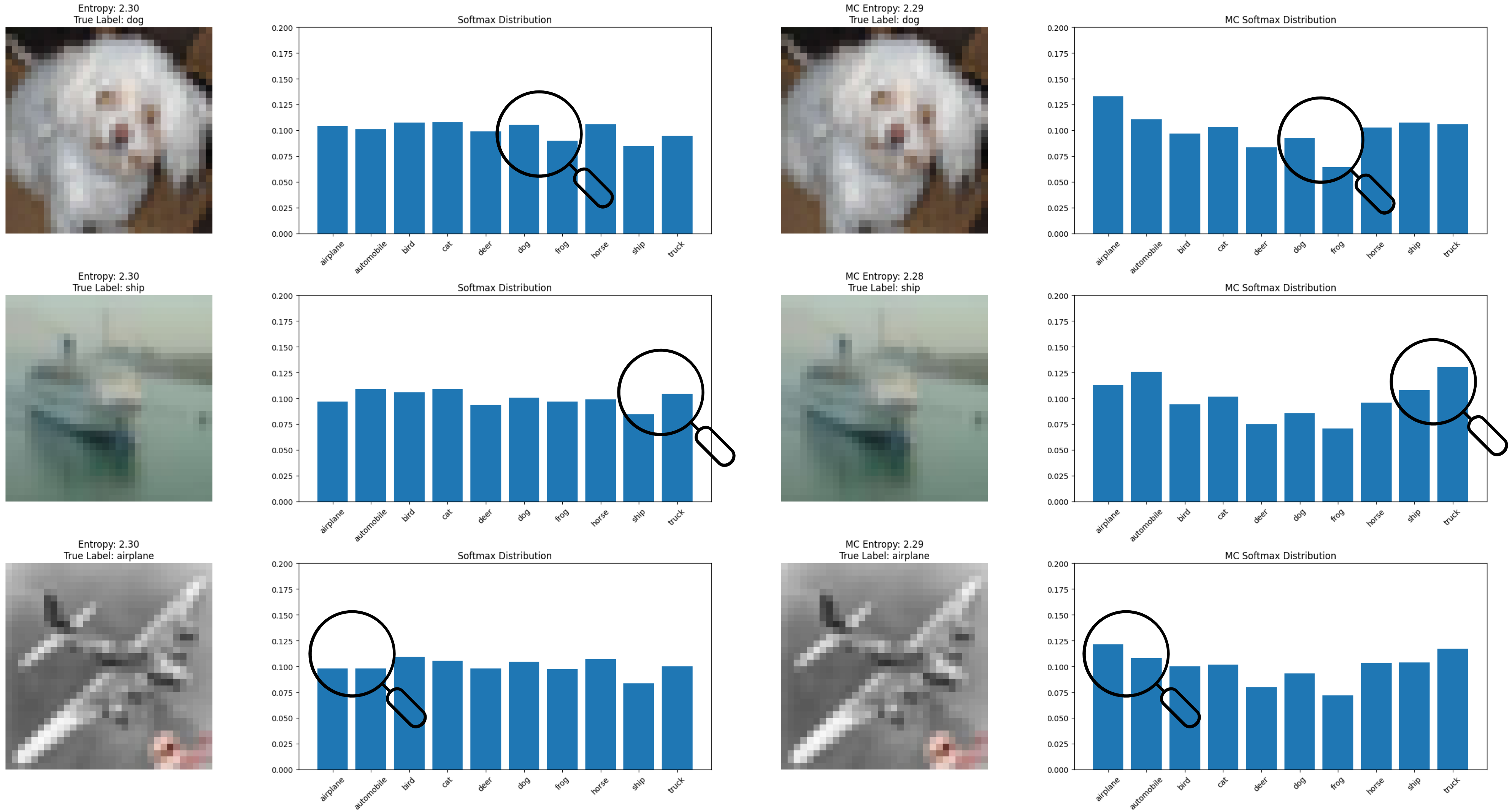


Sources: The author's own elaboration based on Gal et al. (2016)

- If we turn dropout off during prediction: *Deterministic*
  - **We always get the same Softmax distribution and therefore the same prediction!**
- If we turn dropout on during prediction: *Probabilistic*
  - We get different Softmax distributions every time we predict ("stochastic forward pass")
  - We get a distribution over the model's predictions conditional on its weights and input data
  - Much more informative measure of model uncertainty
- **Bayesian interpretation:** Posterior distribution for a given input given the training data and model parameters

Takeaway 7: Dropout at prediction time provides us with a measure of epistemic and aleatoric uncertainty.

Example: Softmax Layer Probabilities vs. MC Dropout



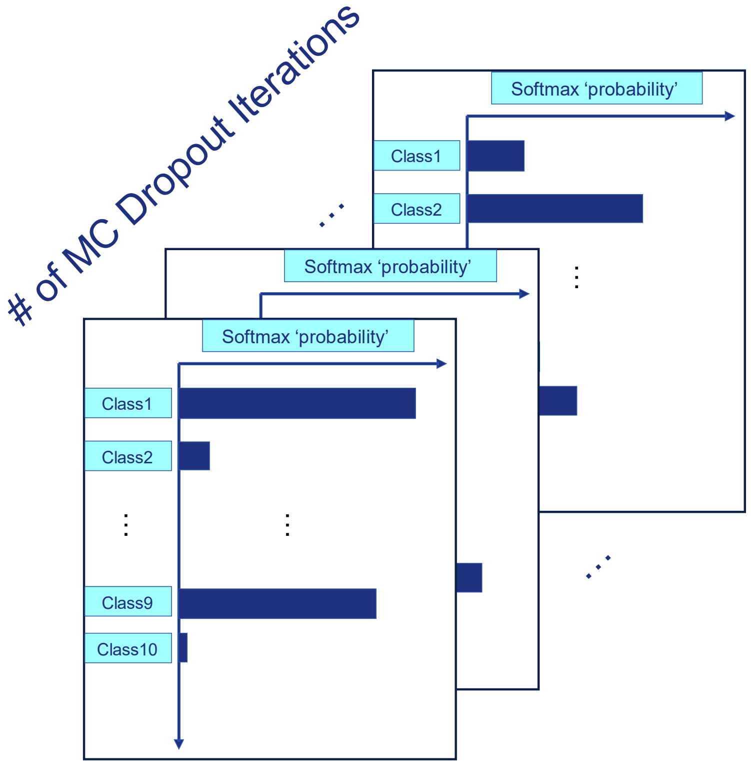
Model is actually more **off target** than it appeared

Model is actually more **off target** than it appeared

Model is actually more **on target** than it appeared

Sources: The author's own elaboration based on Bishop and Bishop (2024), CIFAR-10 by Krizhevsky et al. (2009), and Gal et al. (2016, 2017)
Remarks: 1000 stochastic forward passes were used in this example.

Bringing it all together: Deep Bayesian Active Learning



Assuming **T** stochastic forward passes per sample[1] from unlabeled pool

$$T \to \infty$$

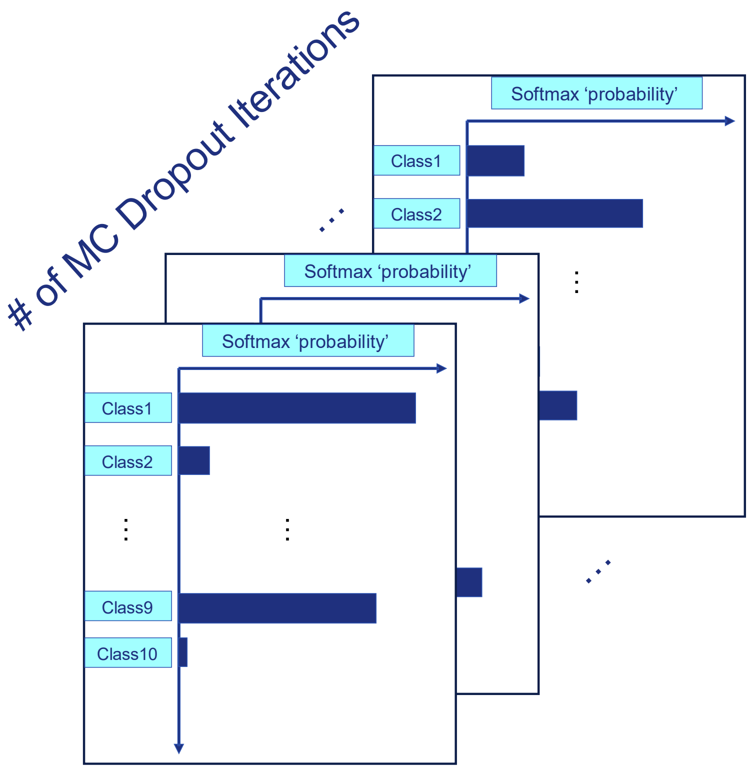
$$\approx \quad p(y = c|\mathbf{x}; \mathcal{D}_{train})$$

Takeaway 8: As the number of dropout iterations approaches infinity, the approximate class probability converges to the real probability. We can now compute our acquisition functions such as entropy from earlier.

1: Samples are processed in batches, mostly out of computational efficiency considerations. The problems this carries with it are addressed through BatchBALD in Kirsch et al. (2019)

Sources: The author's own elaboration based on Gal et al. (2016)

Bringing it all together: Deep Bayesian Active Learning



Assuming **T** stochastic forward passes per sample[1] from unlabeled pool

For model weights ω, input data **x,** and train set D_{train} , t stochastic forward passes it holds that:

$$p(y=c|\boldsymbol{x},\mathcal{D}_{\text{train}}) = \int p(y=c|\boldsymbol{x},\boldsymbol{\omega})p(\boldsymbol{\omega}|\mathcal{D}_{\text{train}})d\boldsymbol{\omega}$$

$$\approx \int p(y=c|\boldsymbol{x},\boldsymbol{\omega})q^*(\boldsymbol{\omega})d\boldsymbol{\omega}$$

$$\approx \frac{1}{T}\sum_{t=1}^{T} p(y=c|\boldsymbol{x},\boldsymbol{\omega}_t)$$

Takeaway 8: As the number of dropout iterations approaches infinity, the approximate class probability converges to the real probability. We can now compute our acquisition functions such as entropy from earlier.

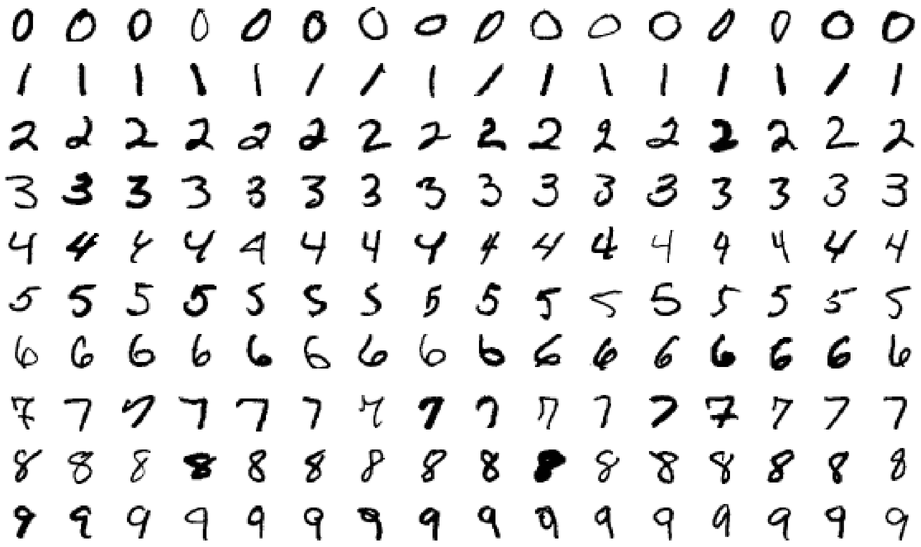
1: Samples are processed in batches, mostly out of computational efficiency considerations. The problems this carries with it are addressed through BatchBALD in Kirsch et al. (2019)

Sources: The author‘s own elaboration based on Gal et al. (2016)

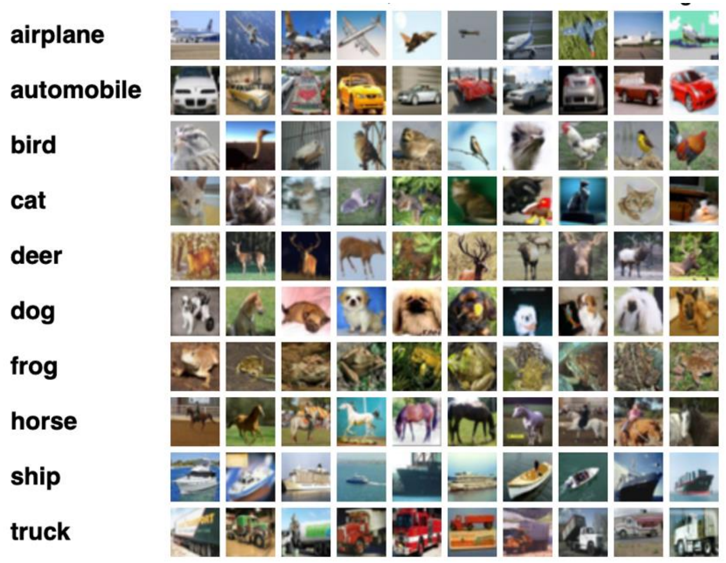
Outline

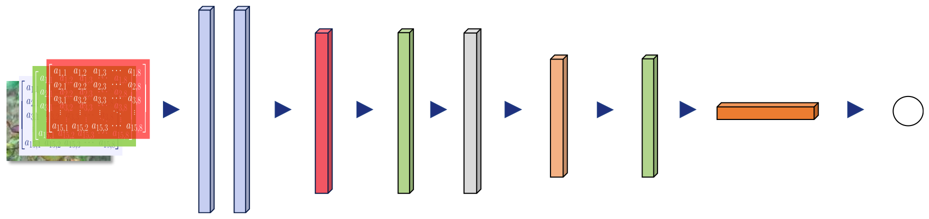
Experimental setup

1

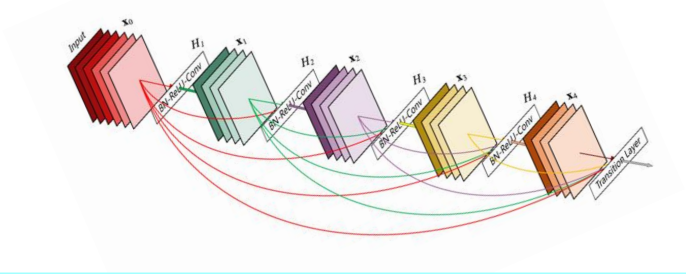


2





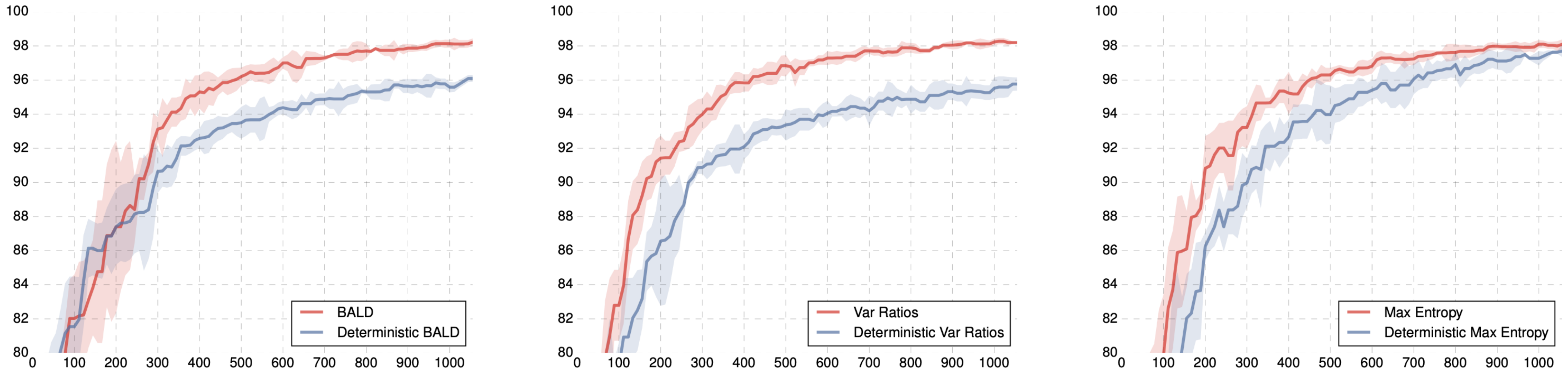
Standard Keras MNIST implementation



DenseNet architecture

Sources: MNIST by LeCun (1998), Keras implementation by fchollet (2015); CIFAR-10 by Krizhevsky et al. (2009) and DenseNet architecture by

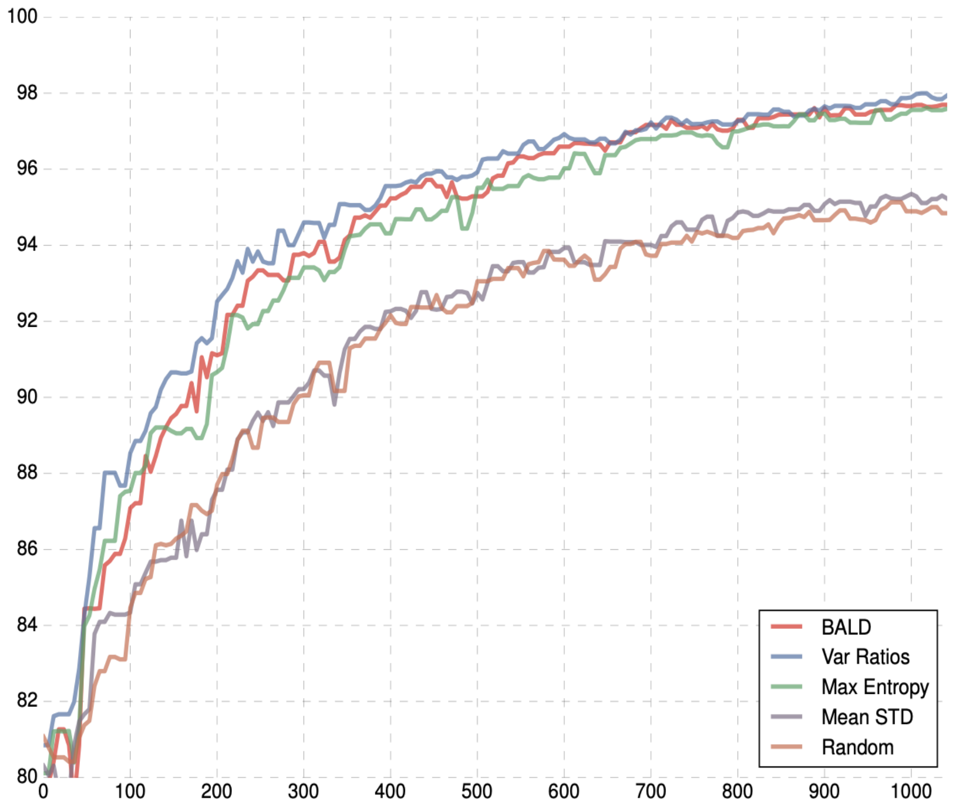
Experiment 1: Bayesian outperforms deterministic approach

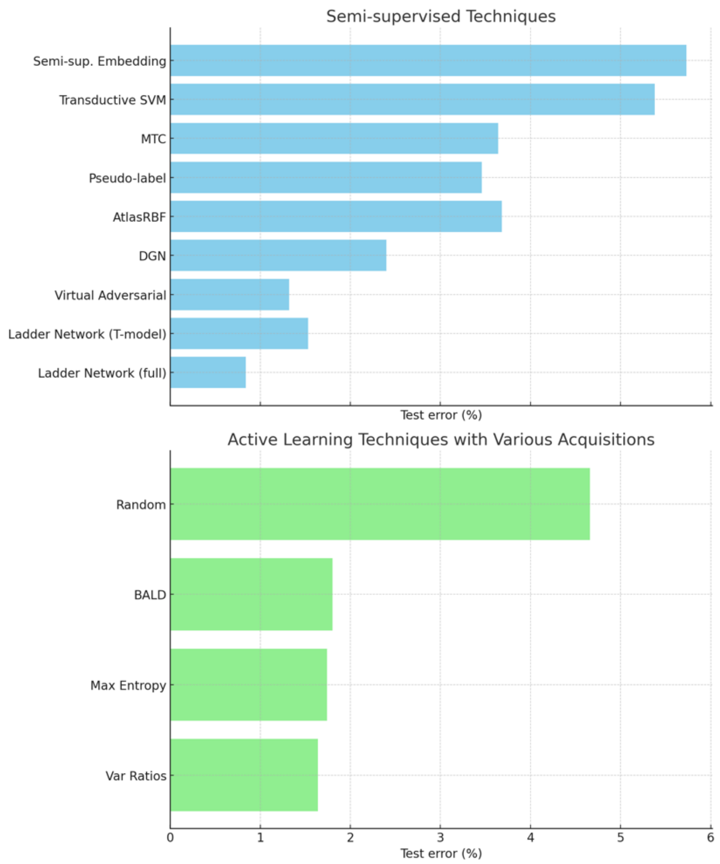


Key finding 1: Incorporating considerations about epistemic model uncertainty improves the active learning speed and converges to higher accuracy

Sources: Gal et al. (2017)

Experiment 2: Variation Ratios and BALD are the best-performing acquisition functions in this setting

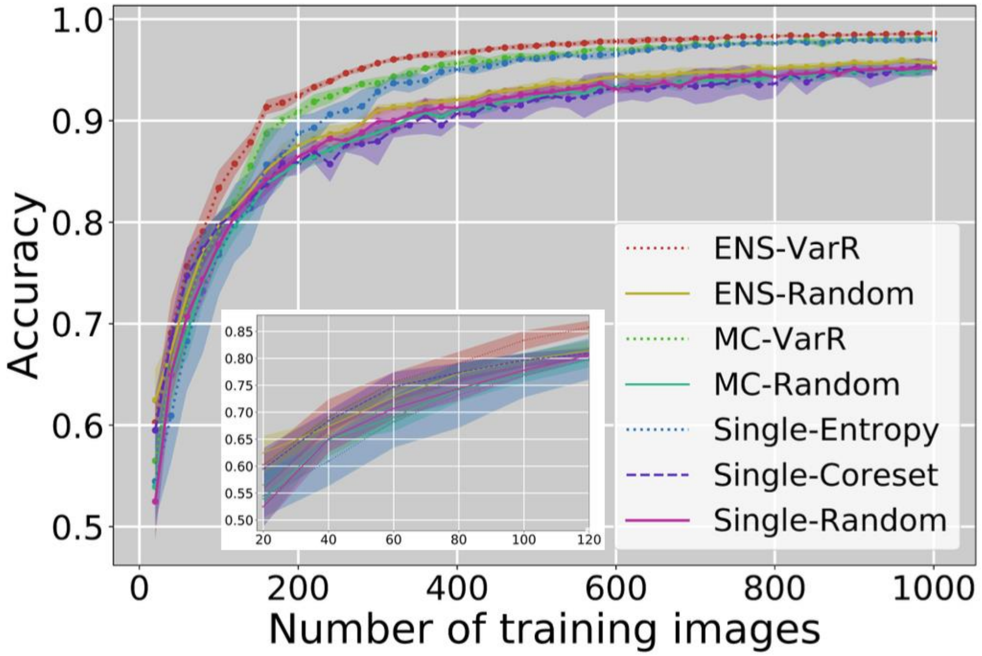




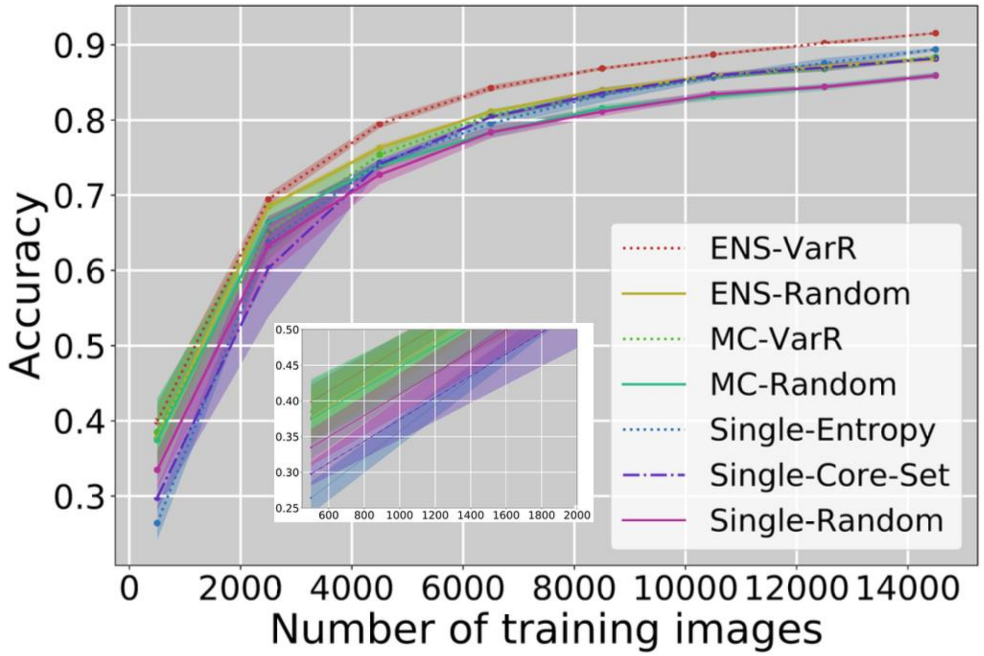
Key finding 2: Variation Ratios/Max Entropy/BALD perform significantly better than random and outperform numerous semi-supervised techniques available at the time.

Sources: Gal et al. (2017), own adaptation based on Gal et al. (2017)

Experiment 3: Why everything I told you could be considered outdated



Standard Keras CNN on MNIST

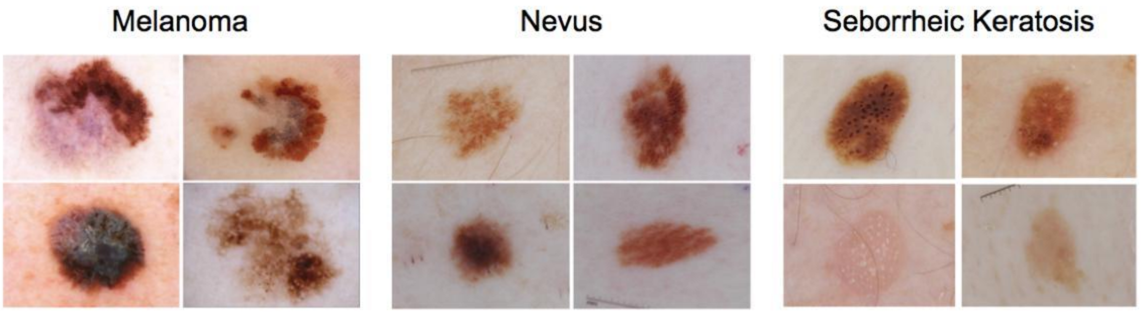


CIFAR-10 on DenseNet architecture

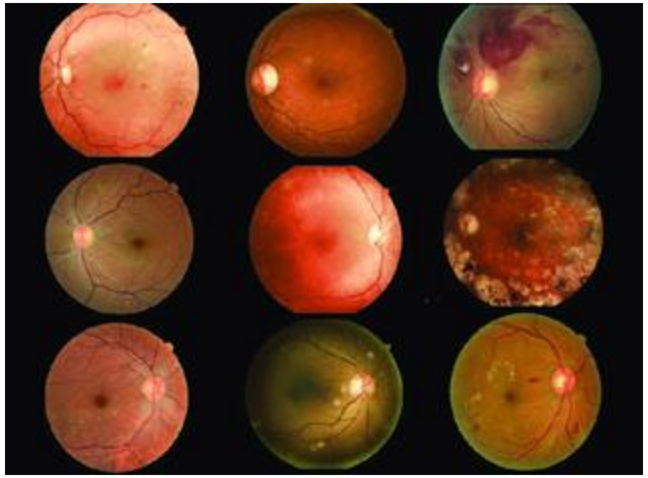
Key finding 3: Ensembles outperform Monte Carlo Dropout approach. Differences are more pronounced on complex tasks (CIFAR-10, -100 vs. MNIST). **Why?**

Sources: Beluch et al. (2018)

Experiment 4: Real-world applications



Melanoma detection



Diabetic Retinopathy detection

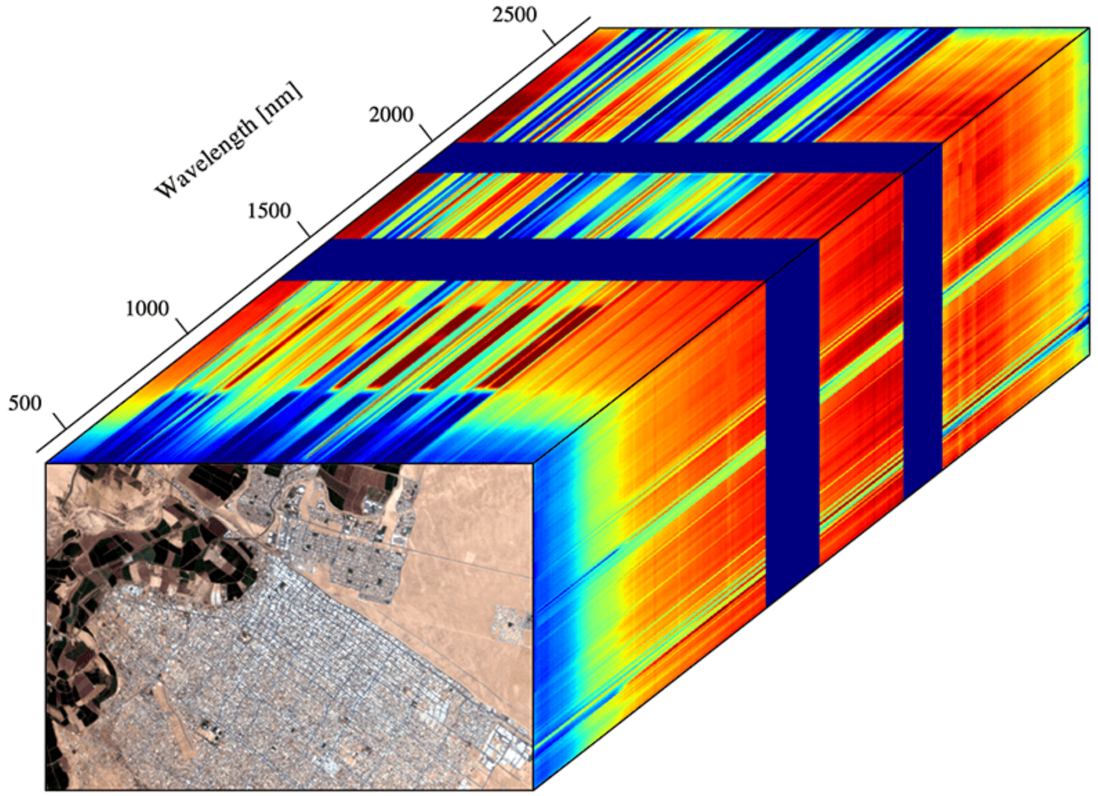
Key finding 4: Both MC dropout based approaches and ensembles have provided a significant performance boost in tricky domains with expensive data labelling such as medical imaging and diagnosis

Sources: ISIC 2016, Beluch et al. 2018

Outline

What does the future hold for high-dimensional active learning?

- Methods will become even more powerful and accurate as computational power and optimization capabilities increase
- Active learning paradigm can be applied to even higher-dimensional features, such as natural language or hyperspectral images (pictured)
- Even better uncertainty estimation for acquisition functions
- Even more powerful models, e.g. Transformer architectures in vision
- Quantum approaches?



Hyperspectral remote sensing images

Image Source: https://eos.org/opinions/realizing-machine-learnings-promise-in-geoscience-remote-sensing

THANK YOU!

Nicolas Malz (n.malz@campus.lmu-munich.de)

Discussion suggestion: Why did the ensemble models eventually outperform the Bayesian approach?

Works cited

Beluch, W. H., et al. “The Power of Ensembles for Active Learning in Image Classification”. In: Proceed- ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018, pp. 9368–9377.

Bishop, C. (2006). “Pattern Recognition and Machine Learning”. New York, USA: Springer Nature.

Bishop, C. and Bishop, H. (2024). “Deep Learning. Concepts and Foundations”. New York, USA: Springer Nature.

Gal, Y. and Ghahramani, Z. (2016). “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: Proceedings of The 33rd International Conference on Machine Learning. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050– 1059. URL: https://proceedings.mlr.press/v48/gal16. html.

Gal, Y., Islam, R. and Ghahramani, Z. (2017). “Deep Bayesian Active Learning with Image Data”. In: Proceedings of the 34th International Conference on Machine Learning. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Re- search. PMLR, June 2017, pp. 1183–1192. URL: https: //proceedings.mlr.press/v70/gal17a.html.

Kingma, D. P. et al. “Semi-supervised learning with deep generative models”. (2014). In: Advances in neural information processing sys tems 27 (2014).

Kirsch, A., van Amersfoort, J. and Gal, Y. (2018). BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. 2019. DOI: 10.48550/ ARXIV.1906.08158. URL: https://arxiv.org/abs/1906. 08158.

Krizhevsky, A., Nair, V. and Hinton, G. (2009). “CIFAR-10 (Canadian Institute for Advanced Research)”. In: (2009). URL: http://www.cs.toronto.edu/~kriz/cifar.html.

LeCun, Y. and Cortes, C. (1998). The MNIST database of handwritten digits. http : / / yann . lecun . com / exdb / mnist/. 1998.

Srivastava, N. et al. (2014). “Dropout: a simple way to prevent neural networks from overfitting”. In: The journal of machine learning research 15.1 (2014), pp. 1929– 1958.

Zhu, XJ, Lafferty, J. and Ghahramani, Z. (2003). “Combining active learning and semi-supervised learning using gaussian fields and harmonic functions”. In: ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining. Vol. 3. 2003, pp. 58–65.