Do representations know what they don’t know?

Denis Janiak

Katedra Inteligencji Obliczeniowej

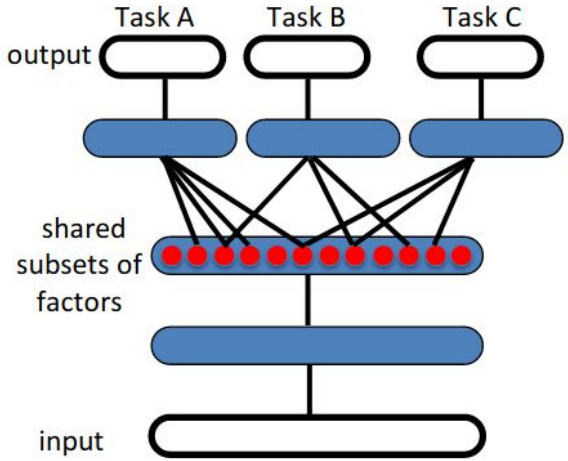
Presentation based on
http://www.gatsby.ucl.ac.uk/~balaji/DL4Sci-Uncertainty-in-Deep-Learning-overview.pdf

Politechnika Wrocławska

Introduction

Representation learning

- Representations matter…
  - The performance of machine learning algorithms depends directly on the data representation
  - We can leverage **generalization properties** to learn different downstream tasks
- The goal is to yield more **abstract** and ultimately more **useful** representations



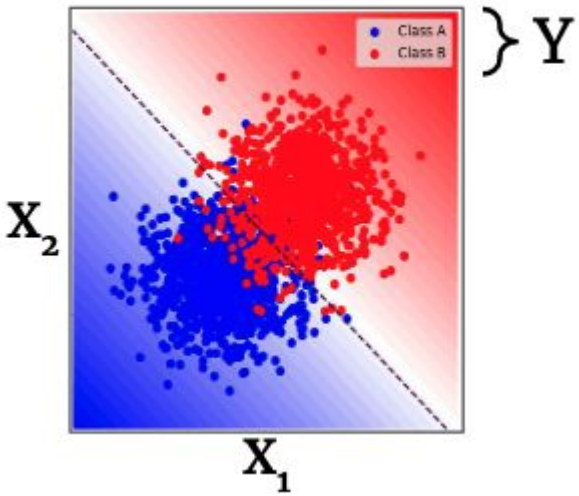
lllustration of representation learning discovering explanatory factors, some explaining the input, and some explaining target for each task (Bengio et al. 2012)

Representation learning

Supervised	Generative	Contrastive	Non-contrastive
• significant advancements over the past decade • reliance on manual annotations • susceptibility to attacks	• ability to capture the full data distribution • can generate new samples • do not need labels • do not produce discriminative representation	• do not need labels • robust and discriminative representations • necessity for hard negative samples	• do not need labels • robust and discriminative representations • do not need negative samples • more vulnerable to representation collapse

Predictive Uncertainty

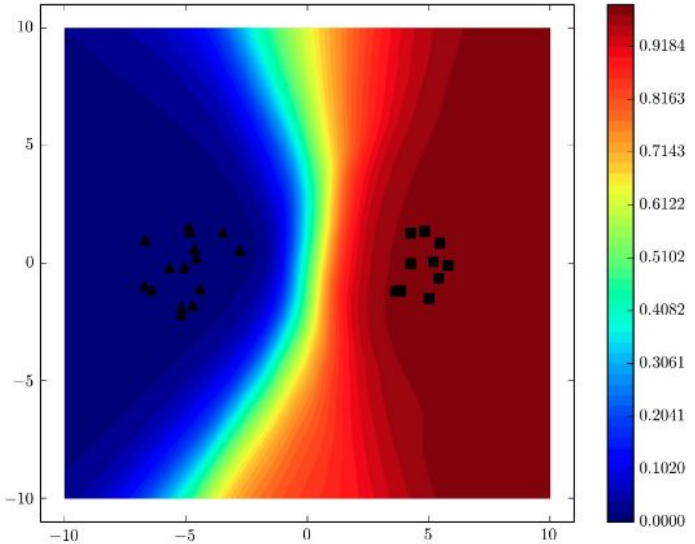
- Predict output distribution **p(y|x)** rather than point estimate, e.g.
  - **Classification**: output label along with confidence
  - **Regression**: output mean and variance



Credits:
http://www.gatsby.ucl.ac.uk/~balaji/DL4Sci-Uncertainty-in-Deep-Learning-overview.pdf

Source of uncertainty

1. **Aleatoric uncertainty**
   a. Noise in the labeling process (humans disagree on the label, e.g. CIFAR-10-H)
   b. Measurement noise in y
   c. Considered to be "**irreducible** uncertainty"
      i. Persists even in the limit of infinite data
2. **Epistemic uncertainty**
   a. Multiple parameters could be consistent with the observed training data
   b. Considered to be "**reducible** uncertainty"
      i. Vanishes in the limit of infinite data



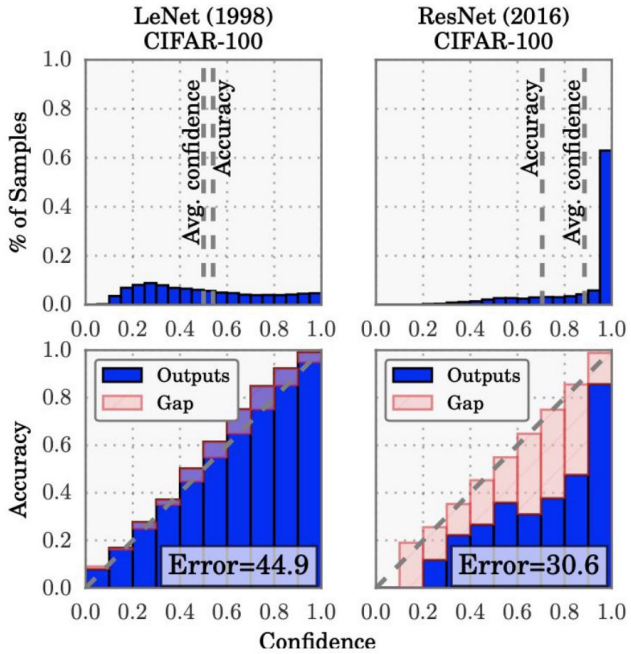
Credits:
http://www.gatsby.ucl.ac.uk/~balaji/DL4Sci-Uncertainty-in-Deep-Learning-overview.pdf

Uncertainty measures - Calibration

Calibration measures how well predicted confidence (probability of correctness) aligns with the observed accuracy.

- **Expected Calibration Error (ECE)**
- Computed as the average gap between within-bucket accuracy and within-bucket predicted probability for M buckets.

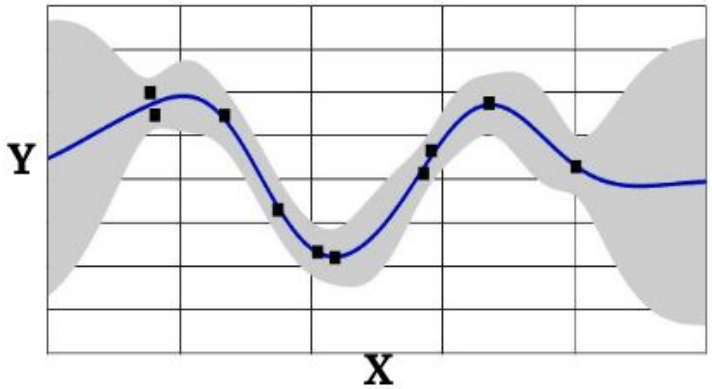
$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \Big|,$$



Credits: Guo et al. 2017 “On calibration of modern neural networks”

Uncertainty measures - other *scoring rules*

- Scoring rules measure the quality of predictive uncertainty
- Many common NN loss functions are proper scoring rule
  - Negative log likelihood (NLL)
  - Brier score (BS) - minimizing the squared error between the predictive probability of a label and one-hot encoding of the correct label

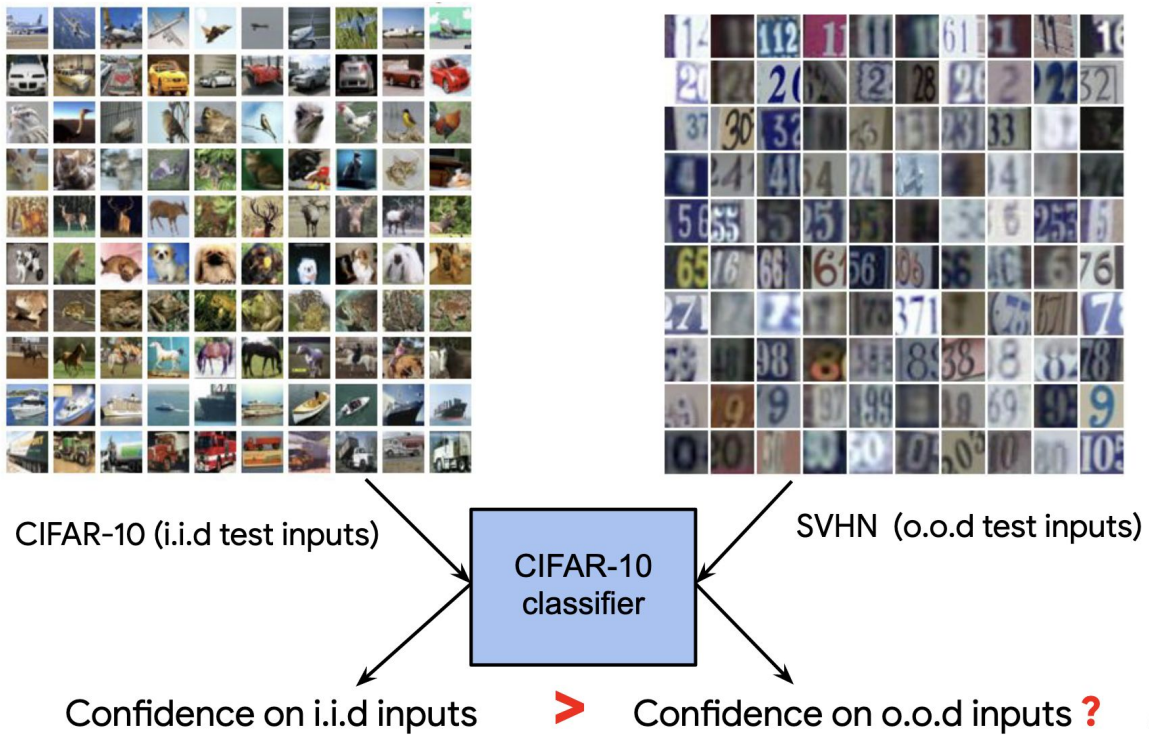


$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Credits:
http://www.gatsby.ucl.ac.uk/~balaji/DL4Sci-Uncertainty-in-Deep-Learning-overview.pdf

How to benchmark uncertainty estimation?

- evaluate model on out-of-distribution (OOD) inputs which do not belong to any of the existing classes
- scores based on max confidence



Credits:
http://www.gatsby.ucl.ac.uk/~balaji/DL4Sci-Uncertainty-in-Deep-Learning-overview.pdf

Uncertainty estimates in AI are important because they:

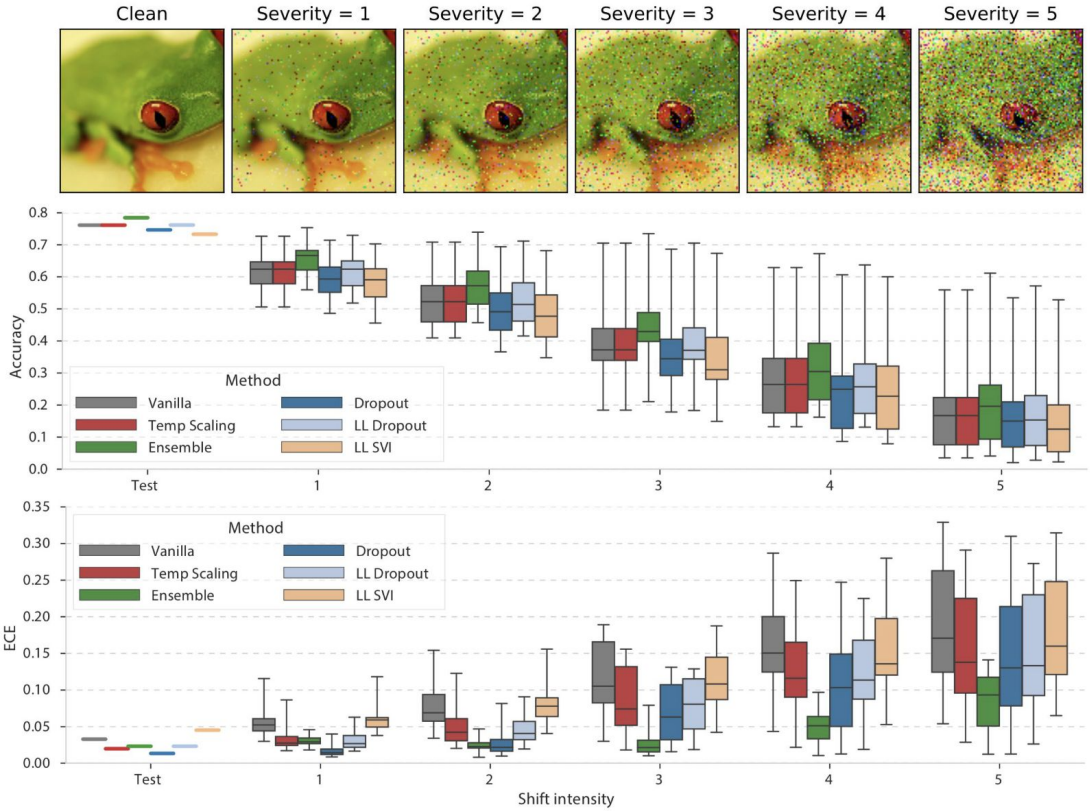
1. Aid decision making by providing confidence measures for predictions.
2. Assess the risk associated with AI predictions.
3. Help select and evaluate models.
4. Guide active learning and data collection.
5. Facilitate human-AI collaboration and trust.
6. Detect out-of-distribution inputs or anomalies.

Methods and research

Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift (Ovadia et al.)

Along with accuracy, the quality of uncertainty consistently degrades with increasing dataset shift regardless of method.

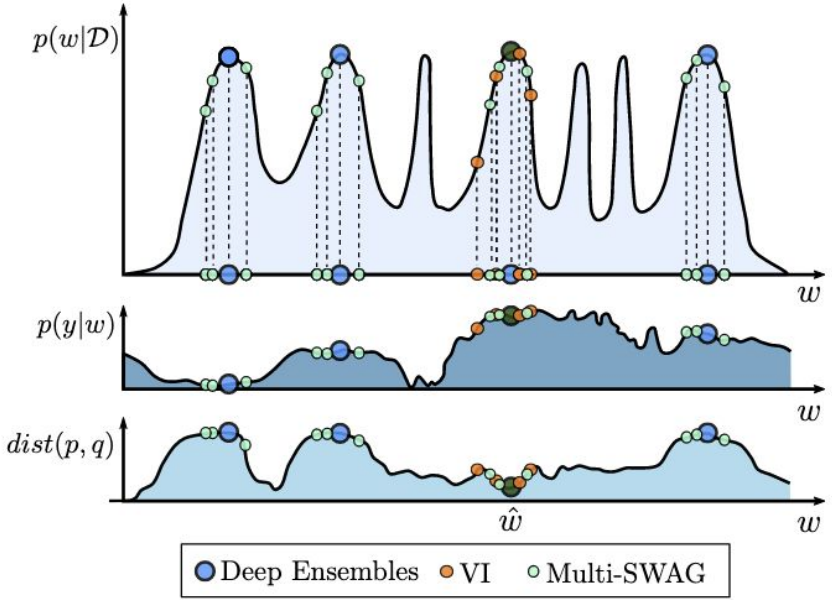
Deep ensembles seem to perform the best across most metrics and be more robust to dataset shift.



Deep ensemble

Variational Bayesian methods are effective at averaging uncertainty within a single mode, but fail to explore the diversity of multiple modes.

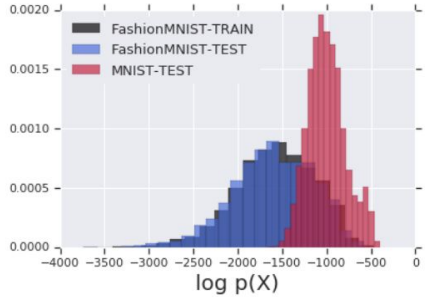
Deep ensembles can be interpreted as an approximate approach to Bayesian marginalization, which selects for functional diversity by representing multiple basins of attraction in the posterior.



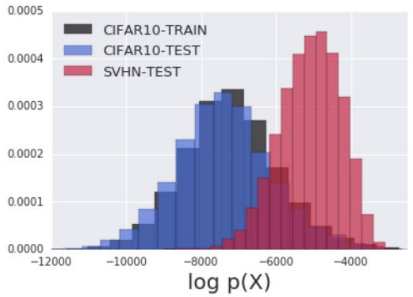
Credits: Bayesian Deep Learning and a Probabilistic Perspective of Generalization (Wilson and Izmailov)

Do Deep Generative Models Know What They Don't Know? (Nalisnick et al.)

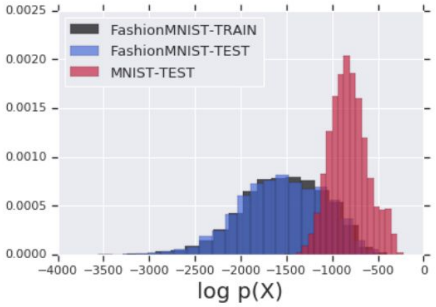
The density learned by flow-based models, VAEs, and PixelCNNs cannot distinguish images of common objects such as dogs, trucks, and horses (i.e. CIFAR-10) from those of house numbers (i.e. SVHN), assigning a higher likelihood to the latter when the model is trained on the former.



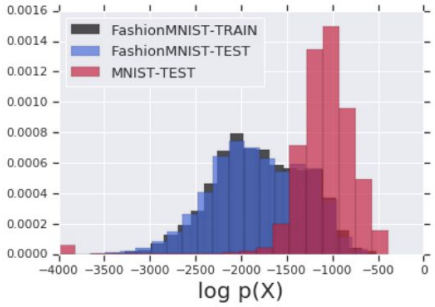
(a) Train on FashionMNIST, Test on MNIST



(b) Train on CIFAR-10, Test on SVHN



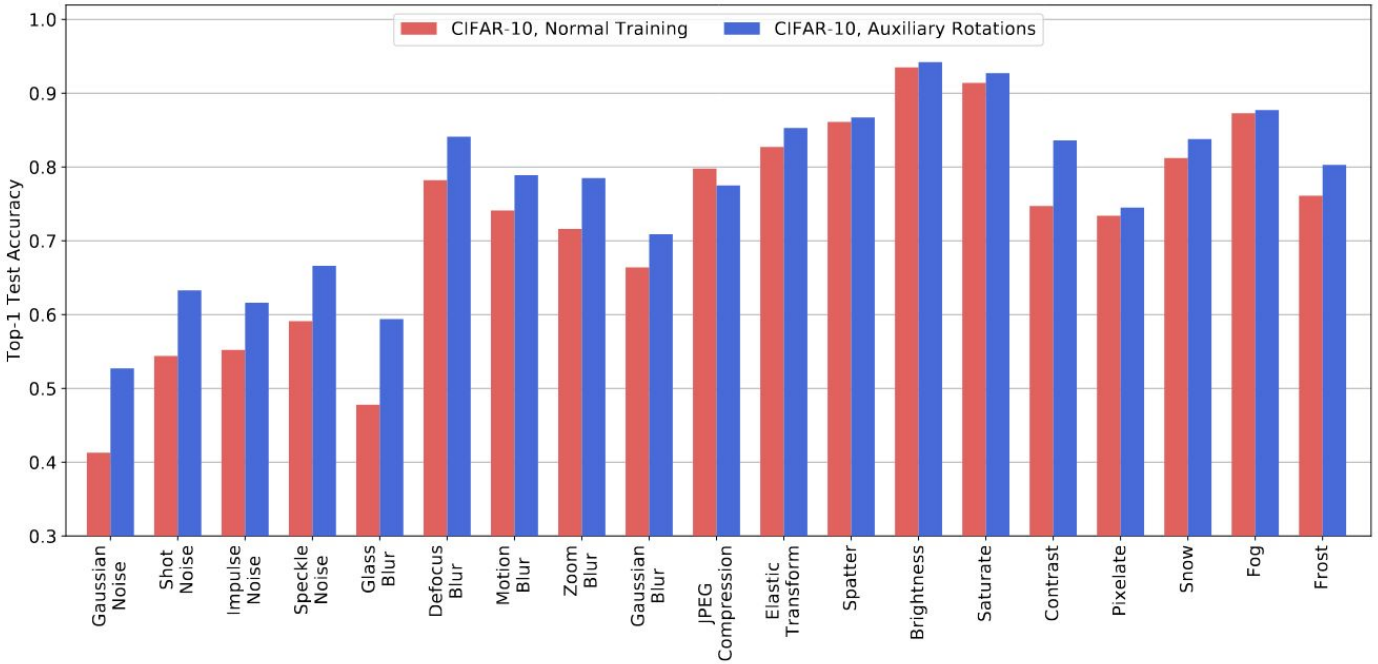
(a) **PixelCNN**: FashionMNIST vs MNIST



(b) **VAE**: FashionMNIST vs MNIST

Credits: Nalisnick et al.

Self-Supervised Learning Can Improve Model Robustness and Uncertainty (Hendrycks et al.)



Types of methods

1. Bayesian methods - posterior and predictive distributions
    a. BNN (VI/MCMC)
    b. Deep ensembles
    c. MC Dropout
    d. …
2. Hybrid models (Zhang et al.) - joint distribution
    a. flow-based
3. Generative models - data distribution
4. Features/logits-based models
    a. Maximum softmax probability
    b. ODIN
    c. Mahalanobis
    d. …

Implementation part

https://colab.research.google.com/github/djaniak/ai-tech-workshop-repr-ood/blob/main/workshops.ipynb