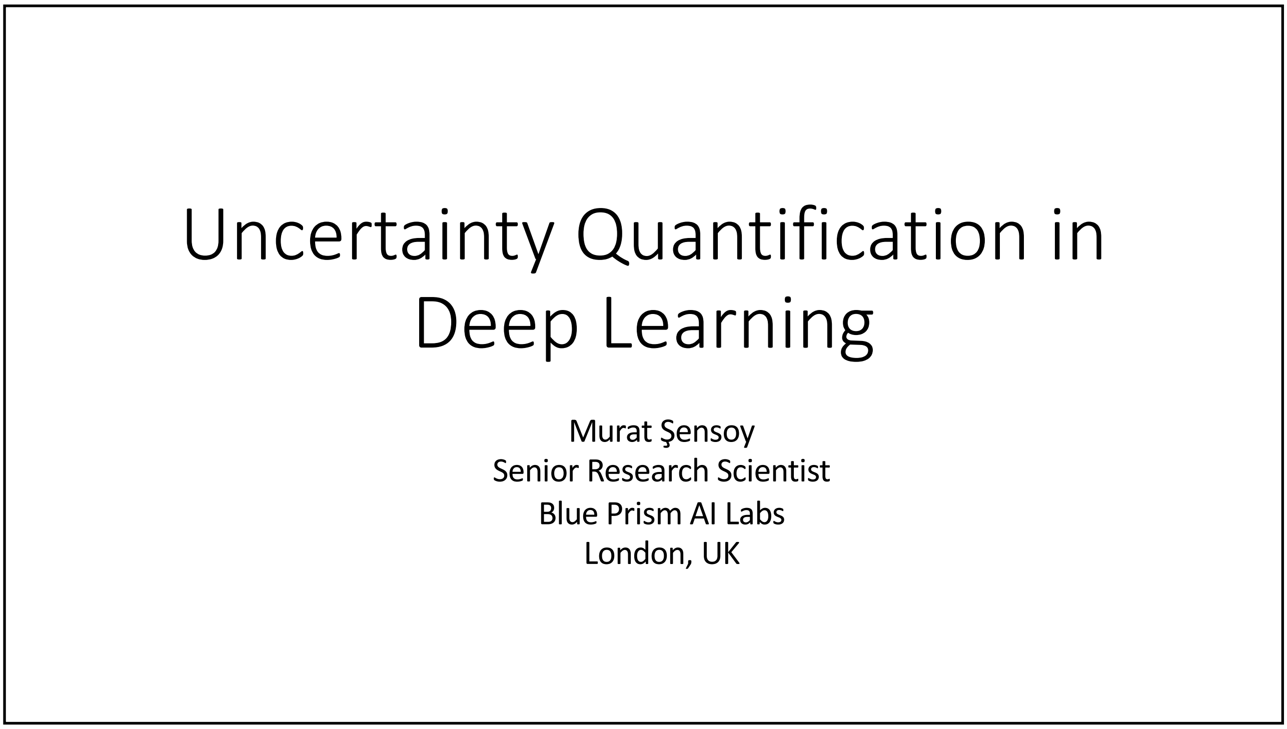
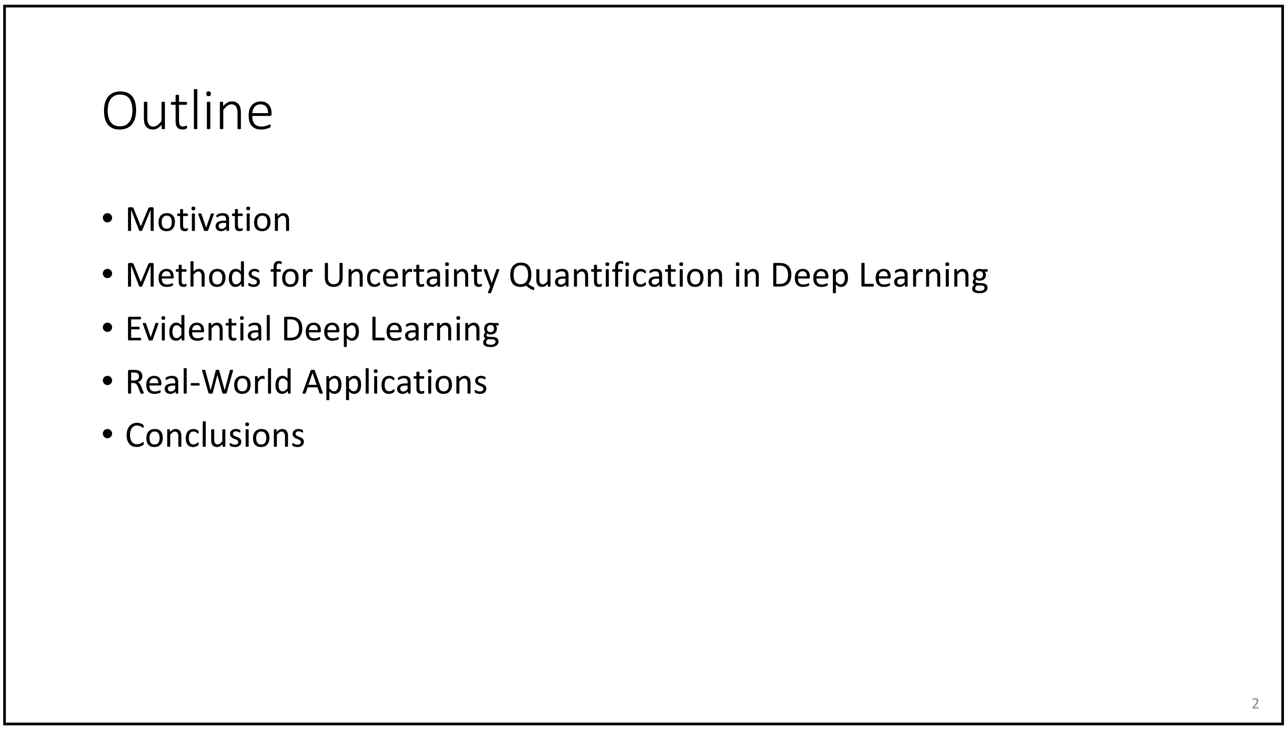


Uncertainty Quantification in Deep Learning

Murat Şensoy
Senior Research Scientist
Blue Prism AI Labs
London, UK

1

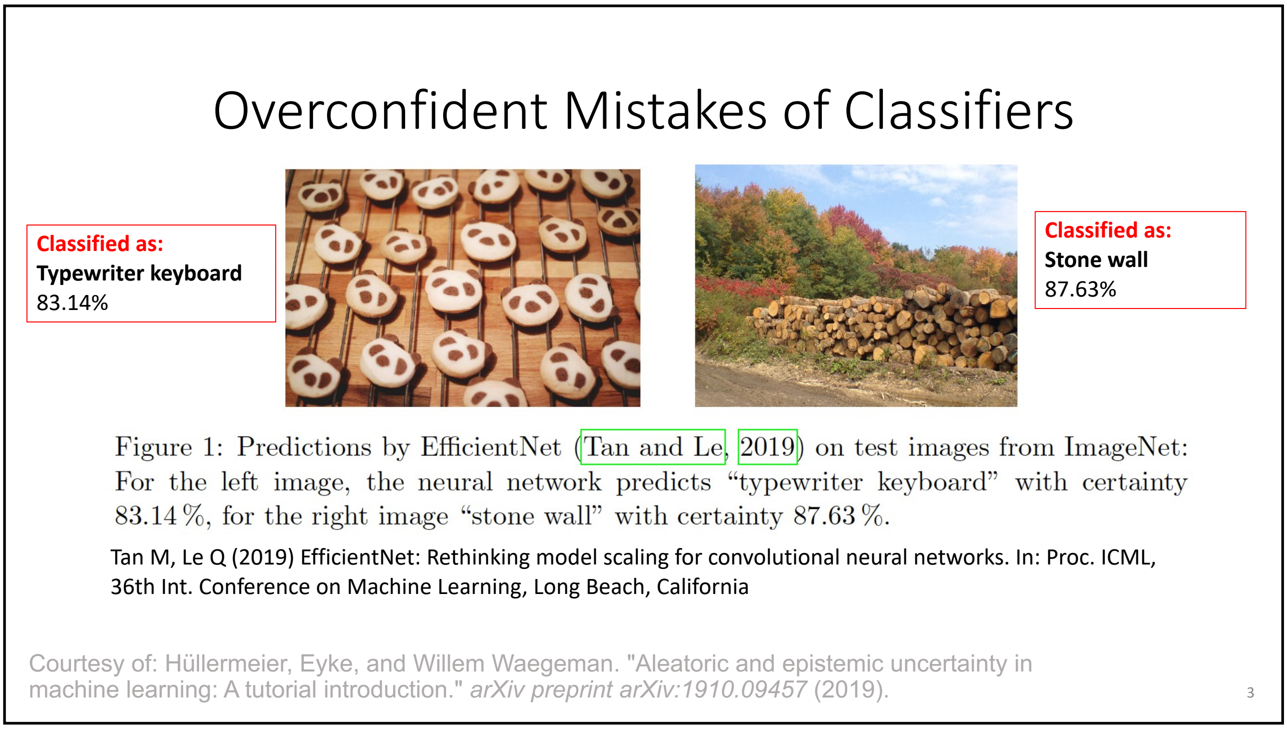


Outline

- Motivation
- Methods for Uncertainty Quantification in Deep Learning
- Evidential Deep Learning
- Real-World Applications
- Conclusions

2

2



Overconfident Mistakes of Classifiers

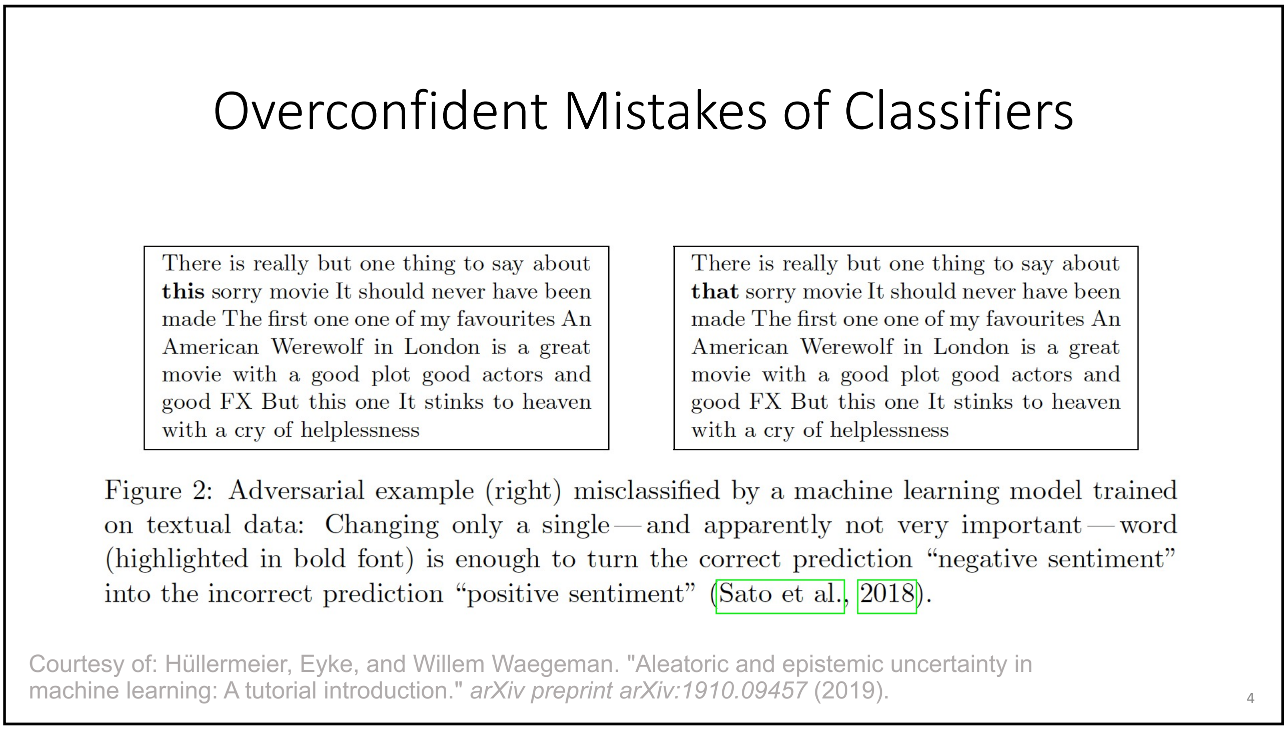
Classified as:
Typewriter keyboard
83.14%

Classified as:
Stone wall
87.63%

Figure 1: Predictions by EfficientNet (Tan and Le, 2019) on test images from ImageNet: For the left image, the neural network predicts "typewriter keyboard" with certainty 83.14 %, for the right image "stone wall" with certainty 87.63 %.

Tan M, Le Q (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proc. ICML, 36th Int. Conference on Machine Learning, Long Beach, California

Courtesy of: Hüllermeier, Eyke, and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction." *arXiv preprint arXiv:1910.09457* (2019).



Overconfident Mistakes of Classifiers

There is really but one thing to say about **this** sorry movie It should never have been made The first one one of my favourites An American Werewolf in London is a great movie with a good plot good actors and good FX But this one It stinks to heaven with a cry of helplessness

There is really but one thing to say about **that** sorry movie It should never have been made The first one one of my favourites An American Werewolf in London is a great movie with a good plot good actors and good FX But this one It stinks to heaven with a cry of helplessness

Figure 2: Adversarial example (right) misclassified by a machine learning model trained on textual data: Changing only a single — and apparently not very important — word (highlighted in bold font) is enough to turn the correct prediction "negative sentiment" into the incorrect prediction "positive sentiment" (Sato et al., 2018).

Courtesy of: Hüllermeier, Eyke, and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction." *arXiv preprint arXiv:1910.09457* (2019).

> “Being able to assess the reliability of a probability score for each instance is much more powerful than assigning an aggregate reliability score [...] independent of the instance to be classified.”

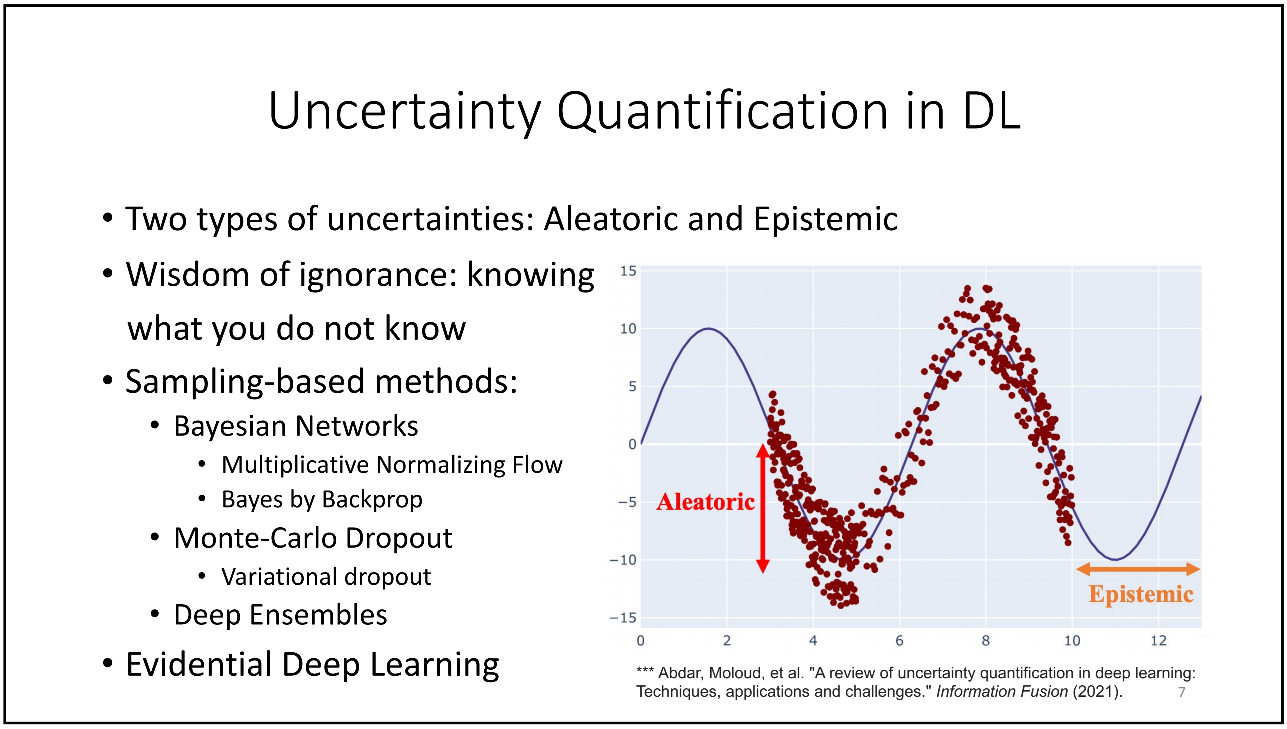
Kull and Flach (2014). Reliability maps: A tool to enhance probability estimates and improve classification accuracy. In: Proc. of ECML’14.

Do not worry!
The success rate of this operation is 95%
what if I am in 5%?
mariascrivan.com
Maria Scrivan

adapted from
10/3 © 2015 Maria Scrivan Dist. by Tribune Content Agency, LLC.

Not all mistakes are equals!

On 7th May 2016, a car operating with automated vehicle control systems crashed with a truck near Williston, Florida, USA. Unfortunately, the car driver died due to the severe injury. The car manufacturer reported that ***the car’s vision system classified the white side of the truck as the sky***.



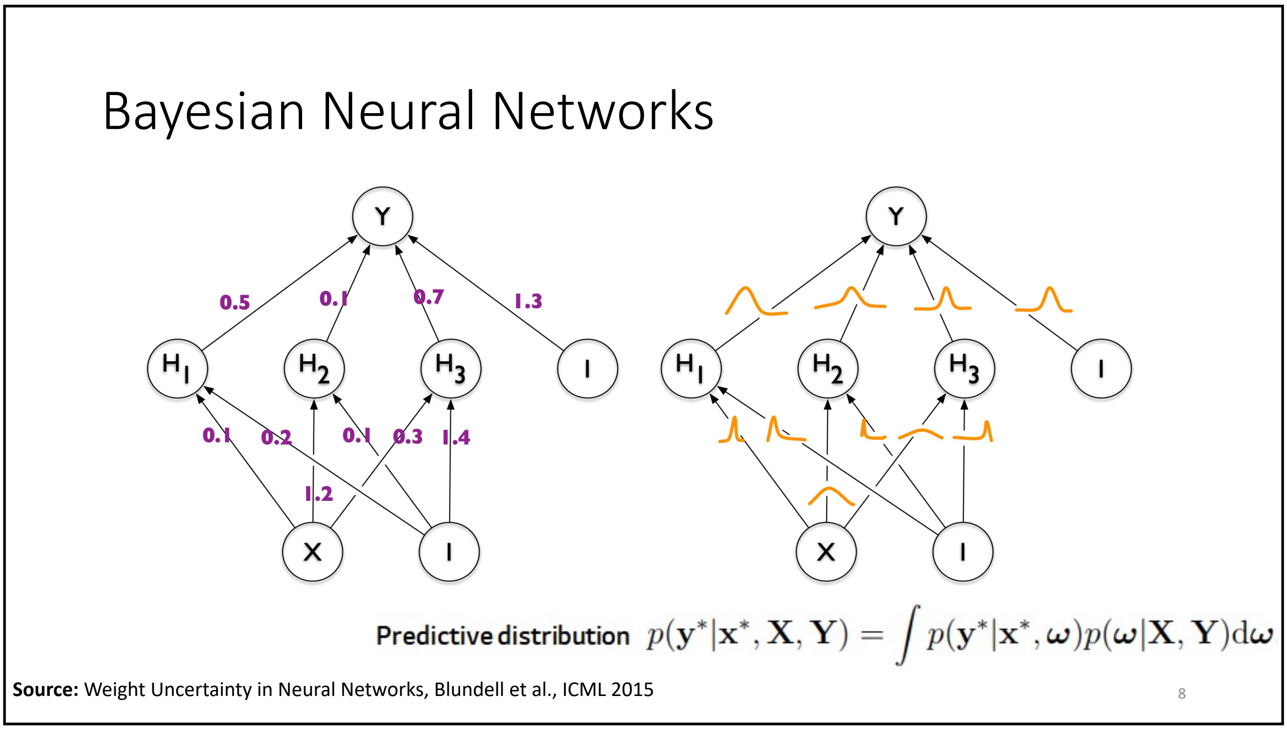
Uncertainty Quantification in DL

- Two types of uncertainties: Aleatoric and Epistemic
- Wisdom of ignorance: knowing what you do not know
- Sampling-based methods:
  - Bayesian Networks
    - Multiplicative Normalizing Flow
    - Bayes by Backprop
  - Monte-Carlo Dropout
    - Variational dropout
  - Deep Ensembles
- Evidential Deep Learning

*** Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." *Information Fusion* (2021).

7

7



Bayesian Neural Networks

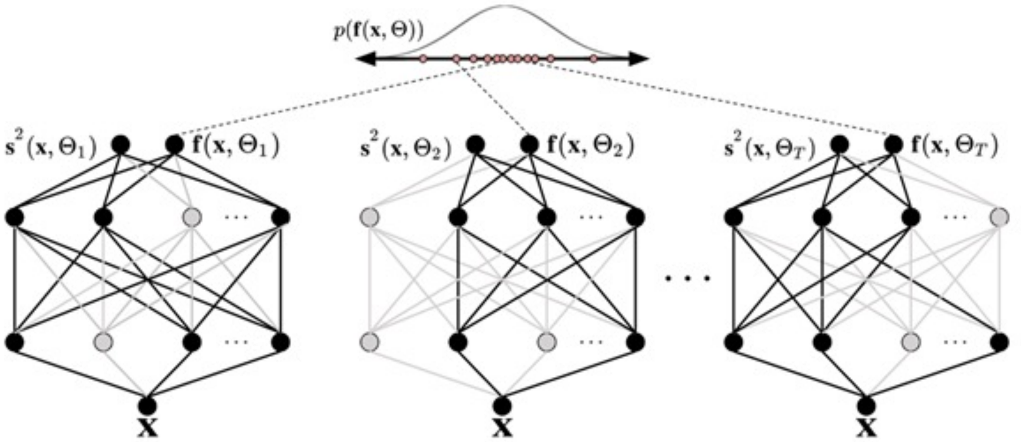
Predictive distribution $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})d\boldsymbol{\omega}$

Source: Weight Uncertainty in Neural Networks, Blundell et al., ICML 2015

8

8

Monte-Carlo Dropout (Bernoulli Approx. VI)


$p(\mathbf{f}(\mathbf{x},\Theta))$
$\mathbf{s}^2(\mathbf{x},\Theta_1)$
$\mathbf{f}(\mathbf{x},\Theta_1)$
$\mathbf{s}^2(\mathbf{x},\Theta_2)$
$\mathbf{f}(\mathbf{x},\Theta_2)$
$\mathbf{s}^2(\mathbf{x},\Theta_T)$
$\mathbf{f}(\mathbf{x},\Theta_T)$
$\mathbf{x}$
$\mathbf{x}$
$\mathbf{x}$

Gal, Y., and Z. Ghahramani. 2016. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." *33rd International Conference on Machine Learning (ICML 2016)*. https://arxiv.org/abs/1506.02142.

Deep Ensembles

1. Let each neural network parametrize a distribution over the outputs, i.e. $p_\theta(y|\mathbf{x})$. Use a **proper scoring rule** as training criterion
   - Classification: cross entropy loss
   - Heteroscedastic Regression: net outputs mean $\mu_\theta(\mathbf{x})$ and variance $\sigma^2_\theta(\mathbf{x})$

$$\ell(\boldsymbol{\theta}, \mathbf{x}_n, y_n) = \frac{1}{2}\log\sigma^2_\theta(\mathbf{x}) + \frac{\left(y - \mu_\theta(\mathbf{x})\right)^2}{2\sigma^2_\theta(\mathbf{x})} + \text{const.}$$

2. Augment with **adversarial training**
3. Train an **ensemble of *M* networks in parallel** with random initialization
4. Combine predictions at test time

$$p(y|\mathbf{x}) = \frac{1}{M}\sum^{M} p_{\theta_m}(y|\mathbf{x}, \theta_m)$$

Source: https://www.gatsby.ucl.ac.uk/~balaji/deep-ensembles-poster.pdf

Lakshminarayanan, B., A. Pritzel, and C. Blundell. 2017. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles." Advances in Neural Information Processing Systems. https://arxiv.org/abs/1612.01474.

Evidential Deep Learning

Murat Sensoy, Lance Kaplan, and Melih Kandemir, "Evidential deep learning to quantify classification uncertainty." *Advances in Neural Information Processing Systems (NuerIPS)*. pp. 3179-3189 , 2018.

Murat Sensoy, Lance Kaplan, Federico Cerutti, “Uncertainty-Aware Deep Classifiers using Generative Models”, The 34rd Conference on Artificial Intelligence (AAAI), 2020.

Murat Sensoy, Maryam Saleki, Simon Julier, John Reid. "Not all Mistakes are Equal." Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). 2020.

Murat Sensoy, Maryam Saleki, Simon Julier, John Reid. "Misclassification Risk and Uncertainty Quantification in Deep Classifiers." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2021.

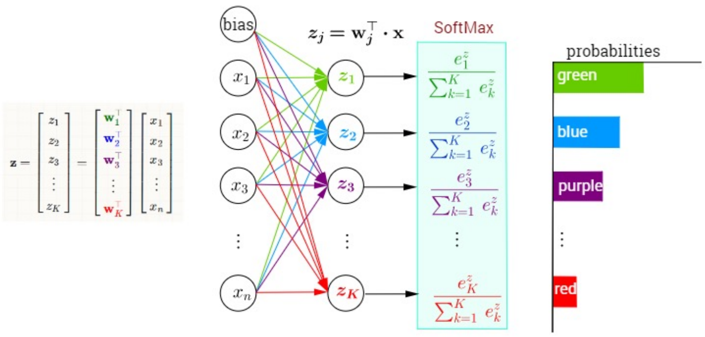
Standard Approach for Classification

Usually **SoftMax** function is used to estimate class probabilities based on the output of a Deep Neural Network.

SoftMax Function

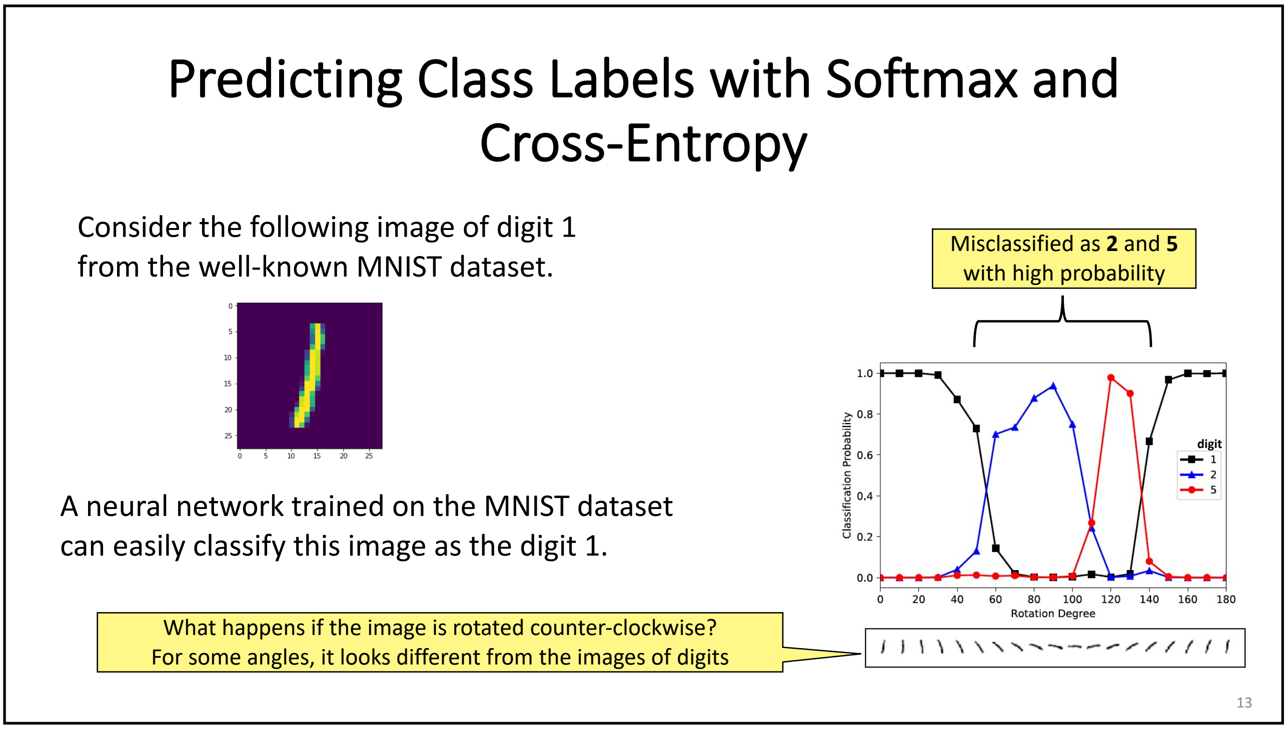
$$\boxed{\sigma(x_j) = \frac{e^{x_j}}{\sum_i e^{x_i}}}$$

Multi-Class Classification with NN and SoftMax Function


$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_K \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^\top \\ \mathbf{w}_2^\top \\ \mathbf{w}_3^\top \\ \vdots \\ \mathbf{w}_K^\top \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$
bias
x_1
x_2
x_3
x_n
$z_j = \mathbf{w}_j^\top \cdot \mathbf{x}$
z_1
z_2
z_3
z_K
SoftMax
$\frac{e_1^z}{\sum_{k=1}^K e_k^z}$
$\frac{e_2^z}{\sum_{k=1}^K e_k^z}$
$\frac{e_3^z}{\sum_{k=1}^K e_k^z}$
$\frac{e_K^z}{\sum_{k=1}^K e_k^z}$
probabilities
green
blue
purple
red

SoftMax leads to neural networks that are over-confident when they encounter samples which are highly different from examples in the training set.

Lets check the example in the following slide



Predicting Class Labels with Softmax and Cross-Entropy

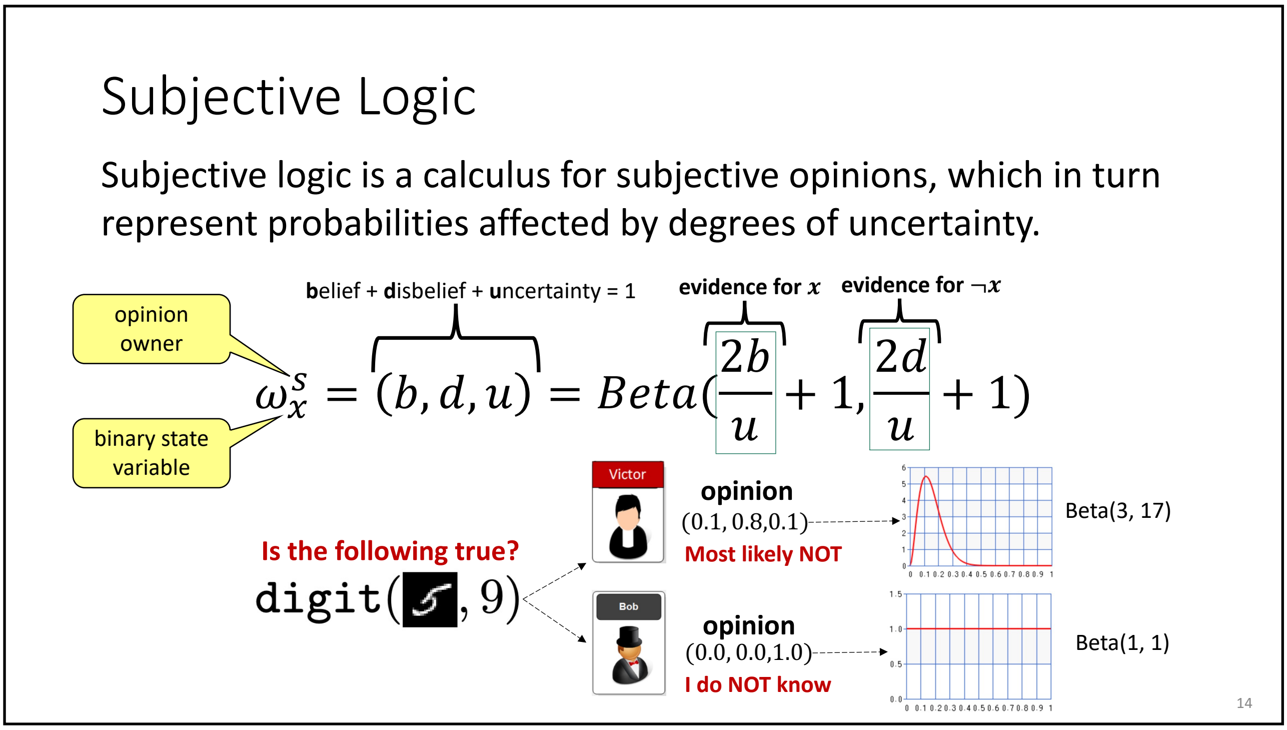
Consider the following image of digit 1 from the well-known MNIST dataset.

A neural network trained on the MNIST dataset can easily classify this image as the digit 1.

Misclassified as **2** and **5** with high probability

What happens if the image is rotated counter-clockwise?
For some angles, it looks different from the images of digits

13



Subjective Logic

Subjective logic is a calculus for subjective opinions, which in turn represent probabilities affected by degrees of uncertainty.

opinion owner

binary state variable

$$\omega_x^s = \overbrace{(b, d, u)}^{\textbf{b}\text{elief} + \textbf{d}\text{isbelief} + \textbf{u}\text{ncertainty} = 1} = Beta(\overbrace{\boxed{\frac{2b}{u}}}^{\textbf{evidence for } x} + 1, \overbrace{\boxed{\frac{2d}{u}}}^{\textbf{evidence for } \neg x} + 1)$$

Is the following true?

digit(, 9)

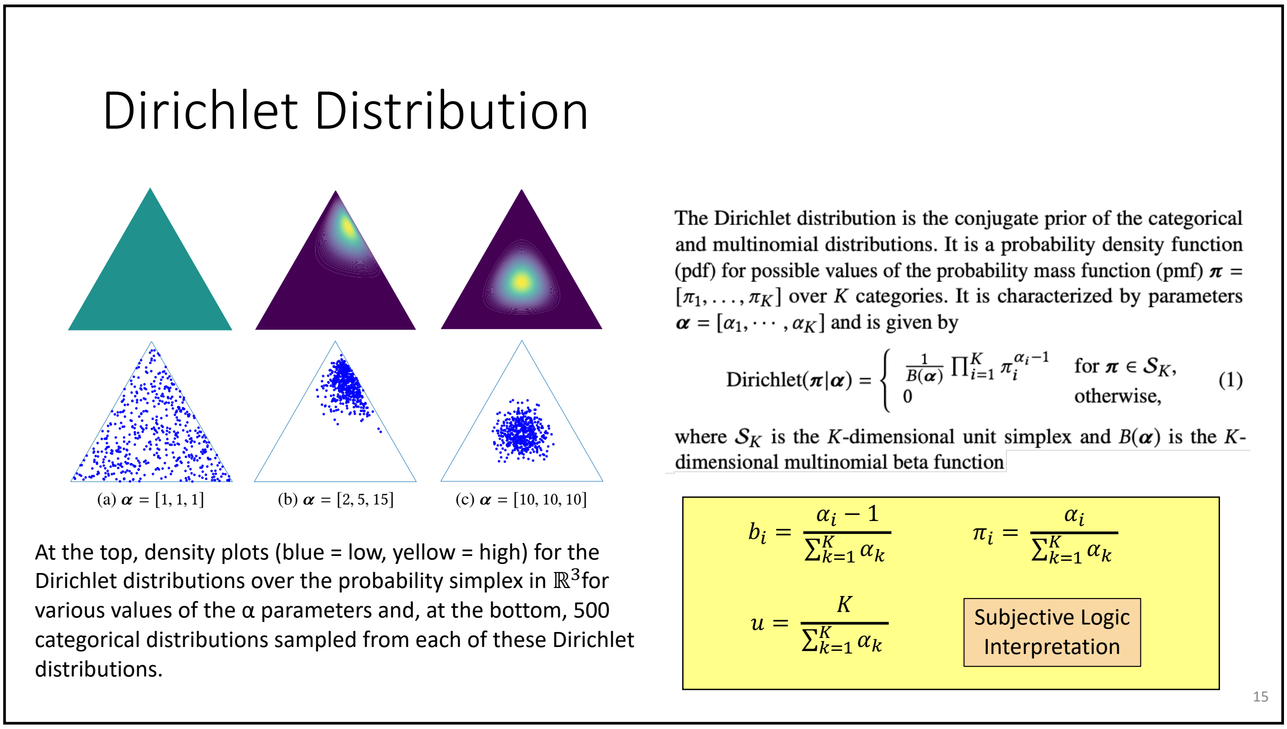
opinion
$(0.1, 0.8, 0.1)$
Most likely NOT

Beta(3, 17)

opinion
$(0.0, 0.0, 1.0)$
I do NOT know

Beta(1, 1)

14



Dirichlet Distribution

(a) $\boldsymbol{\alpha} = [1, 1, 1]$ (b) $\boldsymbol{\alpha} = [2, 5, 15]$ (c) $\boldsymbol{\alpha} = [10, 10, 10]$

At the top, density plots (blue = low, yellow = high) for the Dirichlet distributions over the probability simplex in $\mathbb{R}^3$for various values of the α parameters and, at the bottom, 500 categorical distributions sampled from each of these Dirichlet distributions.

The Dirichlet distribution is the conjugate prior of the categorical and multinomial distributions. It is a probability density function (pdf) for possible values of the probability mass function (pmf) $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ over K categories. It is characterized by parameters $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K]$ and is given by

$$\text{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1} & \text{for } \boldsymbol{\pi} \in \mathcal{S}_K, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\mathcal{S}_K$ is the K-dimensional unit simplex and $B(\boldsymbol{\alpha})$ is the K-dimensional multinomial beta function

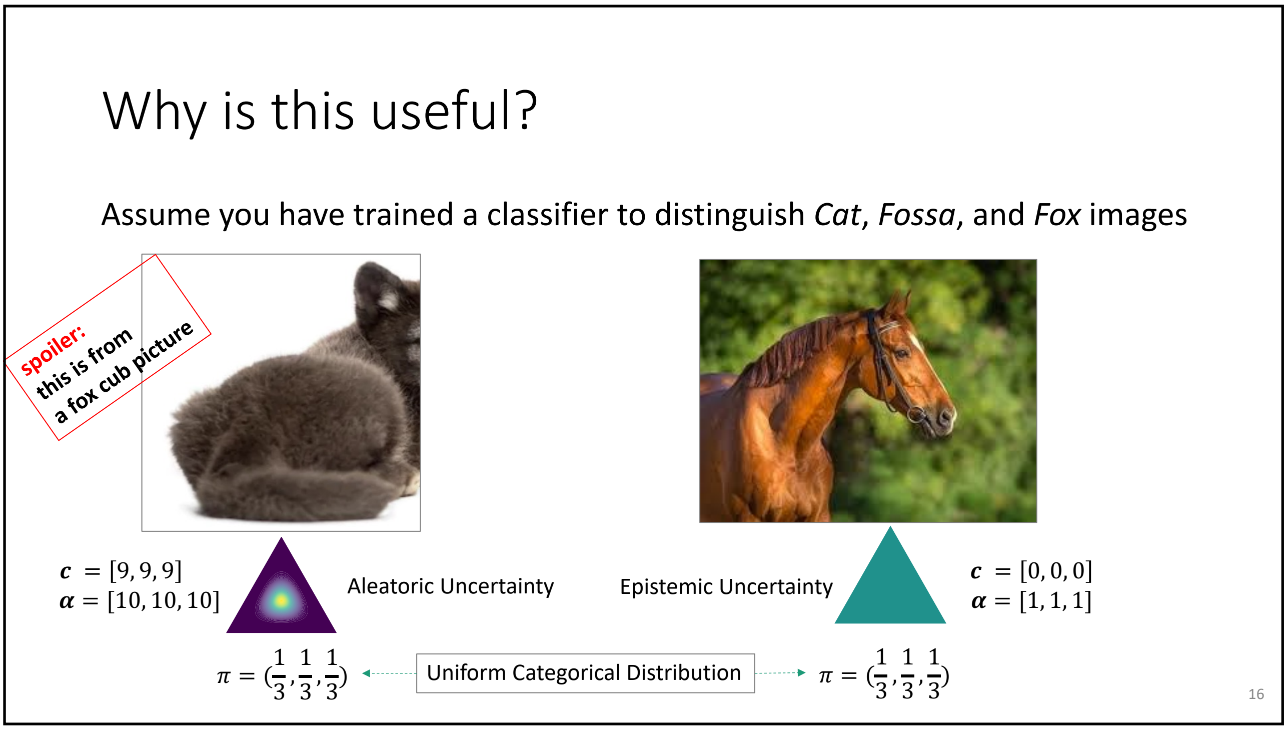
$$b_i = \frac{\alpha_i - 1}{\sum_{k=1}^{K} \alpha_k} \qquad \pi_i = \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}$$

$$u = \frac{K}{\sum_{k=1}^{K} \alpha_k}$$

Subjective Logic Interpretation

15

15



Why is this useful?

Assume you have trained a classifier to distinguish *Cat*, *Fossa*, and *Fox* images

spoiler: this is from a fox cub picture

$\boldsymbol{c} = [9, 9, 9]$
$\boldsymbol{\alpha} = [10, 10, 10]$

Aleatoric Uncertainty

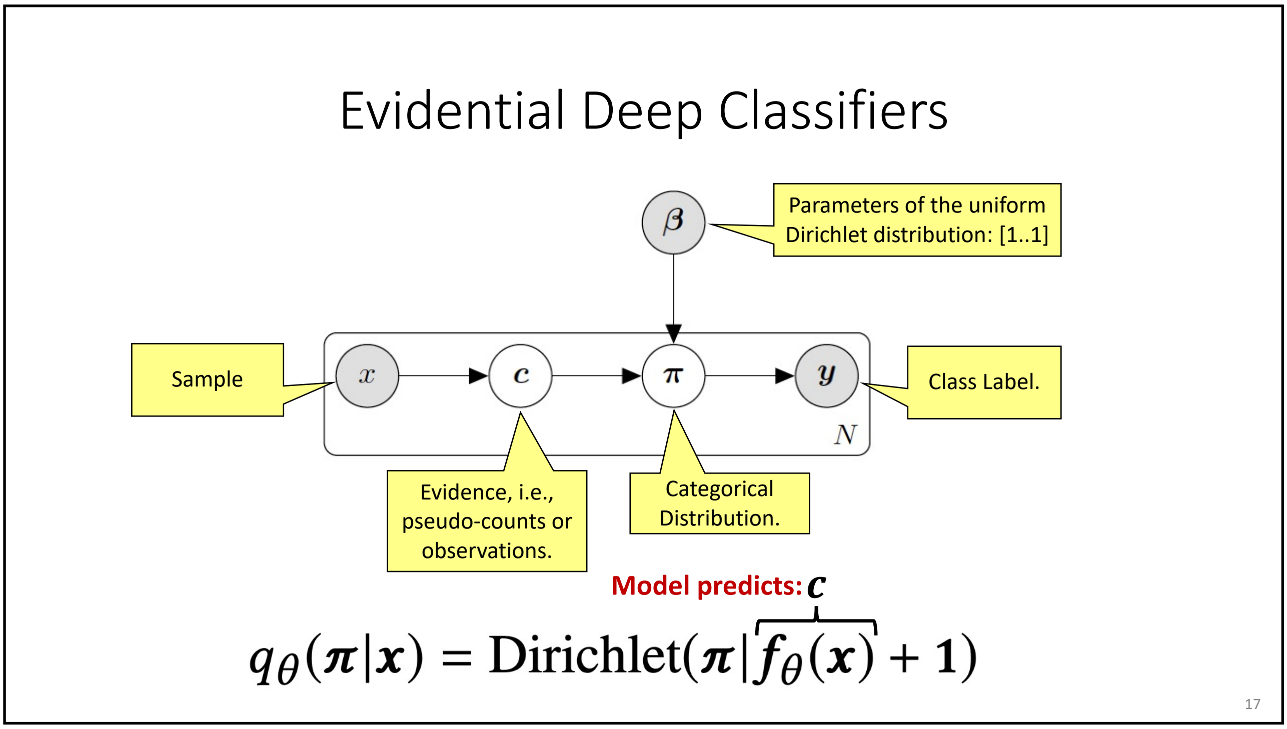
Epistemic Uncertainty

$\boldsymbol{c} = [0, 0, 0]$
$\boldsymbol{\alpha} = [1, 1, 1]$

$\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ ← Uniform Categorical Distribution → $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

16

16



Evidential Deep Classifiers

Parameters of the uniform Dirichlet distribution: [1..1]

Sample

Class Label.

Evidence, i.e., pseudo-counts or observations.

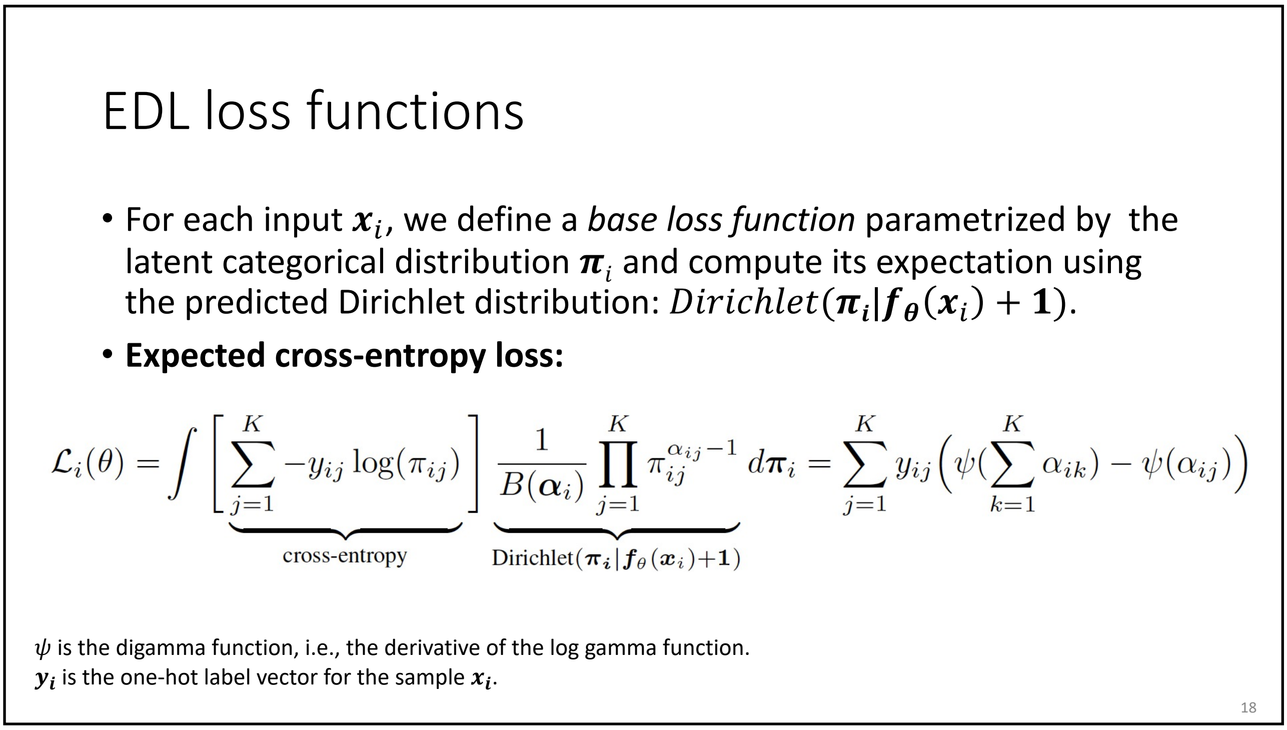
Categorical Distribution.

Model predicts: c

$$q_\theta(\boldsymbol{\pi}|\boldsymbol{x}) = \text{Dirichlet}(\boldsymbol{\pi}|\overbrace{f_\theta(\boldsymbol{x})}^{c} + \mathbf{1})$$

17

17



EDL loss functions

- For each input $\boldsymbol{x}_i$, we define a *base loss function* parametrized by the latent categorical distribution $\boldsymbol{\pi}_i$ and compute its expectation using the predicted Dirichlet distribution: $Dirichlet(\boldsymbol{\pi}_i|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + \mathbf{1})$.
- **Expected cross-entropy loss:**

$$\mathcal{L}_i(\theta) = \int \Bigg[\underbrace{\sum_{j=1}^{K} -y_{ij} \log(\pi_{ij})}_{\text{cross-entropy}} \Bigg] \underbrace{\frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} \pi_{ij}^{\alpha_{ij}-1}}_{\text{Dirichlet}(\boldsymbol{\pi}_i|\boldsymbol{f}_\theta(\boldsymbol{x}_i)+\mathbf{1})} d\boldsymbol{\pi}_i = \sum_{j=1}^{K} y_{ij} \Big(\psi\big(\sum_{k=1}^{K} \alpha_{ik}\big) - \psi(\alpha_{ij}) \Big)$$

ψ is the digamma function, i.e., the derivative of the log gamma function.
$\boldsymbol{y}_i$ is the one-hot label vector for the sample $\boldsymbol{x}_i$.

18

18

EDL loss functions

Type II maximum likelihood:

We can treat $Dirichlet(\boldsymbol{\pi_i}|\boldsymbol{f_\theta}(\boldsymbol{x}_i) + \mathbf{1})$ as a prior on the likelihood $Mult\ (\boldsymbol{y_i}|\boldsymbol{\pi_i})$ and obtain the negated logarithm of the marginal likelihood by integrating out the class probabilities.

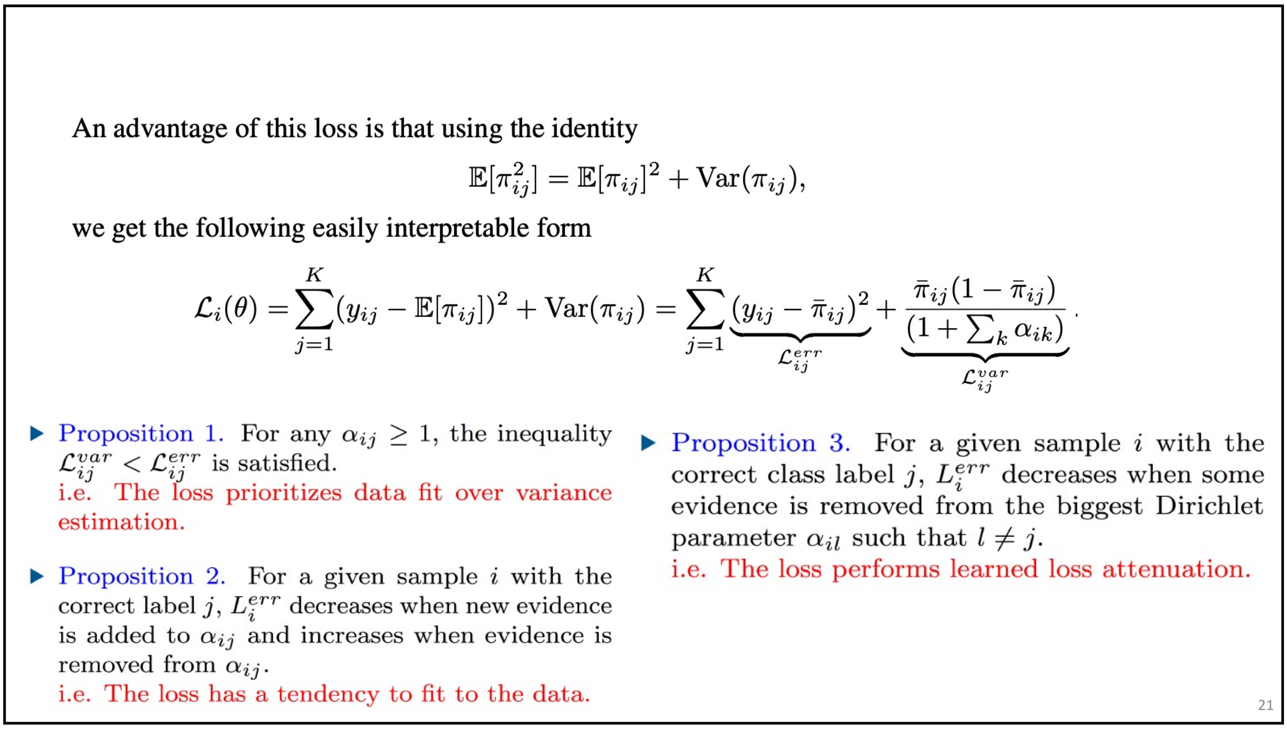
$$\mathcal{L}_i(\theta) = -\log\left(\int \underbrace{\left[\prod_{j=1}^{K} \pi_{ij}^{y_{ij}}\right]}_{Mult(\boldsymbol{y}_i|\boldsymbol{\pi}_i)} \underbrace{\frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} \pi_{ij}^{\alpha_{ij}-1}}_{\text{Dirichlet}(\boldsymbol{\pi_i}|\boldsymbol{f}_\theta(\boldsymbol{x}_i)+\mathbf{1})} d\boldsymbol{\pi}_i\right) = \sum_{j=1}^{K} y_{ij}\Big(\log(\sum_{k=1}^{K}\alpha_{ik}) - \log(\alpha_{ij})\Big)$$

EDL loss functions

The expected sum square error loss (Brier score):

$$\mathcal{L}_i(\theta) = \int \underbrace{||\boldsymbol{y}_i - \boldsymbol{\pi}_i||_2^2}_{\text{SSE loss}} \underbrace{\frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} \pi_{ij}^{\alpha_{ij}-1}}_{\text{Dirichlet}(\boldsymbol{\pi_i}|\boldsymbol{f}_\theta(\boldsymbol{x}_i)+\mathbf{1})} d\boldsymbol{\pi}_i$$

$$= \sum_{j=1}^{K} \mathbb{E}\Big[y_{ij}^2 - 2y_{ij}\pi_{ij} + \pi_{ij}^2\Big] = \sum_{j=1}^{K}\Big(y_{ij}^2 - 2y_{ij}\mathbb{E}[\pi_{ij}] + \mathbb{E}[\pi_{ij}^2]\Big)$$



An advantage of this loss is that using the identity

$$\mathbb{E}[\pi_{ij}^2] = \mathbb{E}[\pi_{ij}]^2 + \text{Var}(\pi_{ij}),$$

we get the following easily interpretable form

$$\mathcal{L}_i(\theta) = \sum_{j=1}^{K} (y_{ij} - \mathbb{E}[\pi_{ij}])^2 + \text{Var}(\pi_{ij}) = \sum_{j=1}^{K} \underbrace{(y_{ij} - \bar{\pi}_{ij})^2}_{\mathcal{L}_{ij}^{err}} + \underbrace{\frac{\bar{\pi}_{ij}(1 - \bar{\pi}_{ij})}{(1 + \sum_k \alpha_{ik})}}_{\mathcal{L}_{ij}^{var}}.$$

- Proposition 1. For any $\alpha_{ij} \geq 1$, the inequality $\mathcal{L}_{ij}^{var} < \mathcal{L}_{ij}^{err}$ is satisfied.
  i.e. The loss prioritizes data fit over variance estimation.
- Proposition 2. For a given sample i with the correct label j, L_i^{err} decreases when new evidence is added to α_{ij} and increases when evidence is removed from α_{ij}.
  i.e. The loss has a tendency to fit to the data.
- Proposition 3. For a given sample i with the correct class label j, L_i^{err} decreases when some evidence is removed from the biggest Dirichlet parameter α_{il} such that $l \neq j$.
  i.e. The loss performs learned loss attenuation.

Overall loss function

Evidence for the correct class is removed.

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \mathcal{L}_i(\theta) + \lambda_t \sum_{i=1}^{N} \text{KL}[\text{Dirichlet}(\boldsymbol{\pi}_i | \tilde{\boldsymbol{\alpha}}_i) \| \text{Dirichlet}(\boldsymbol{\pi}_i | \mathbf{1})]$$

Maximize model fit

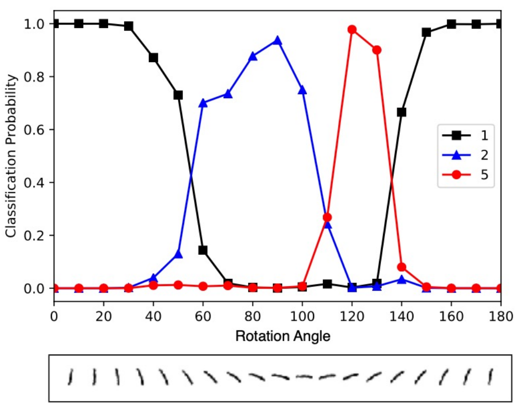
Minimize evidence on errors.

$$\text{KL}[\text{Dirichlet}(\boldsymbol{\pi}_i \tilde{\boldsymbol{\alpha}}_i) \| \text{Dirichlet}(\boldsymbol{\pi}_i | \mathbf{1})]$$
$$= \log\left(\frac{\Gamma(\sum_{k=1}^{K} \tilde{\alpha}_{ik})}{\Gamma(K) \prod_{k=1}^{K} \Gamma(\tilde{\alpha}_{ik})} \right) + \sum_{k=1}^{K} (\tilde{\alpha}_{ik} - 1) \left[\psi(\tilde{\alpha}_{ik}) - \psi\left(\sum_{j=1}^{K} \tilde{\alpha}_{ij} \right) \right]$$

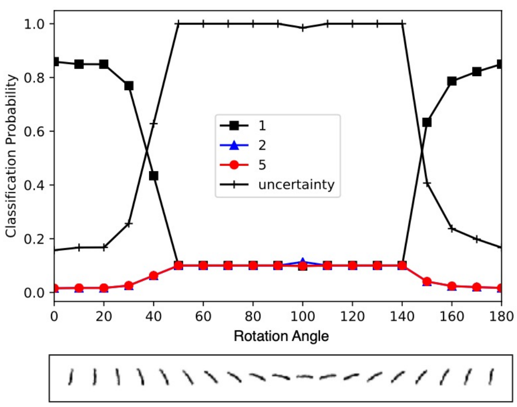
$\lambda_t$is the annealing coefficient; initially 0 and increased gradually to and 1 during training.
$\tilde{\boldsymbol{\alpha}}$ refers to the predicted Dirichlet parameters after removing the evidence for the true category.

Revisiting MNIST

LeNet5 with cross-entropy loss



LeNet5 with EDL loss



The demo is available at https://muratsensoy.github.io/uncertainty.html

Evaluations

Recent research on uncertainty quantification is centred around Bayesian Neural Networks using

- LeNet5* architecture with ReLU nonlinearities
- MNIST and CIFAR10 datasets

and evaluated on tasks

- Out of distribution detection
- Adversarial Robustness

*LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

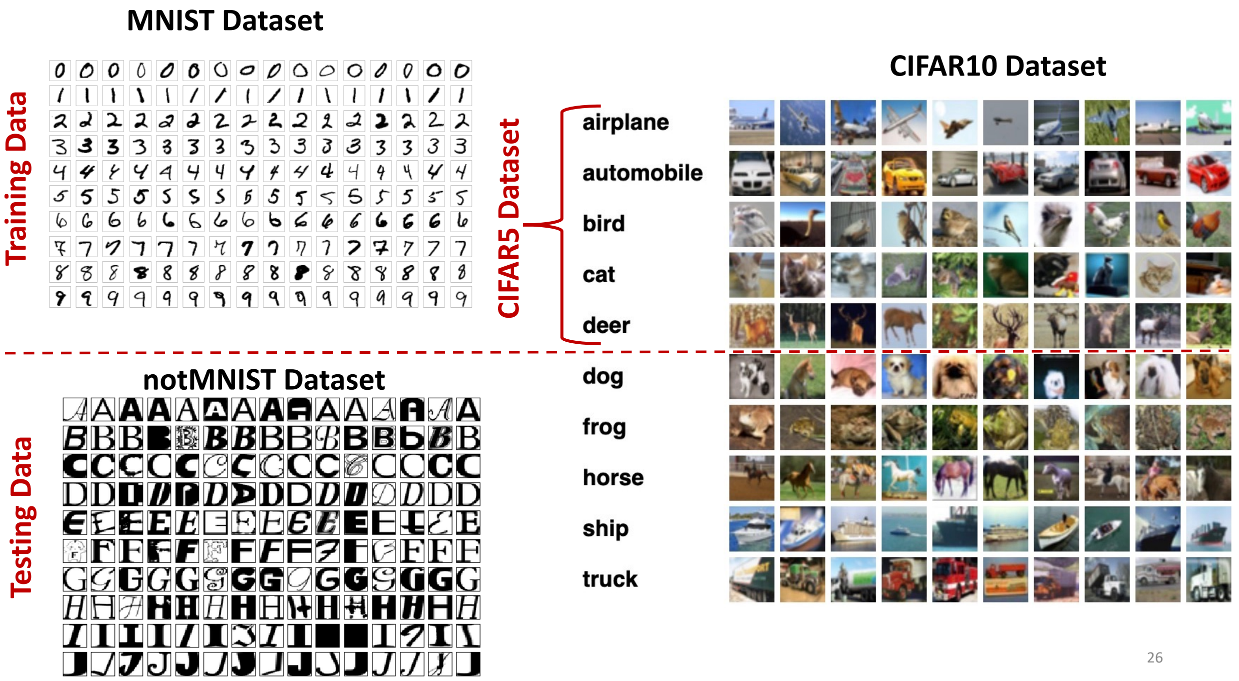
Evaluations

We use the following approaches in our evaluations:

- Vanilla Neural Nets with weight decay (L2)
- Multiplicative Normalizing Flow (MNFG). Louizos and Welling, ICML, 2017.
- Deep Ensembles. Lakshminarayanan et al., NIPS, 2017.
- Monte Carlo Dropout. Gal and Ghahramani, ICML, 2016.
- Variational dropout and the local reparameterization trick (FFLU). Kingma et al. NIPS, 2015
- Bayes by Backprop (FFG). Blundell, ICML, 2015.
- Evidential Deep Learning (EDL). Sensoy et al., NIPS, 2018.

25

25



26

26

Accuracy vs Uncertainty

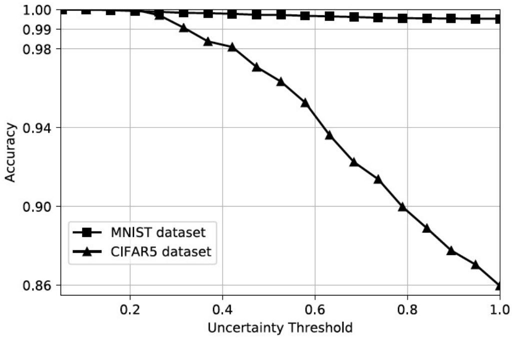


Figure 2: The change of accuracy with respect to the uncertainty threshold for *EDL*.

First 5 categories of CIFAR10

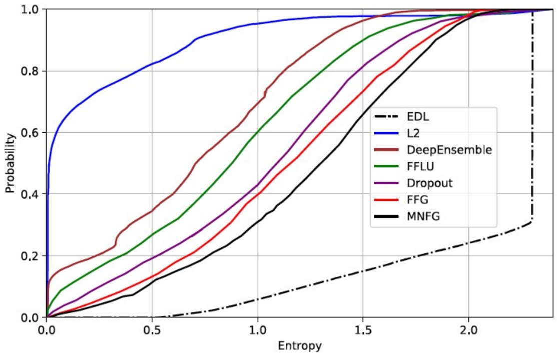
Method	MNIST	CIFAR 5
L2	99.4	76
Dropout	99.5	84
Deep Ensemble	99.3	79
FFG	99.1	78
FFLU	99.1	77
MNFG	99.3	84
EDL	99.3	83

Table 1: Test accuracies (%) for MNIST and CIFAR5 datasets.

Out-of-Distribution Samples

notMNIST dataset

Last 5 categories from CIFAR10



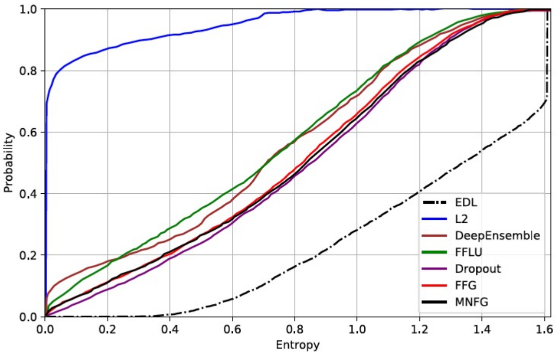
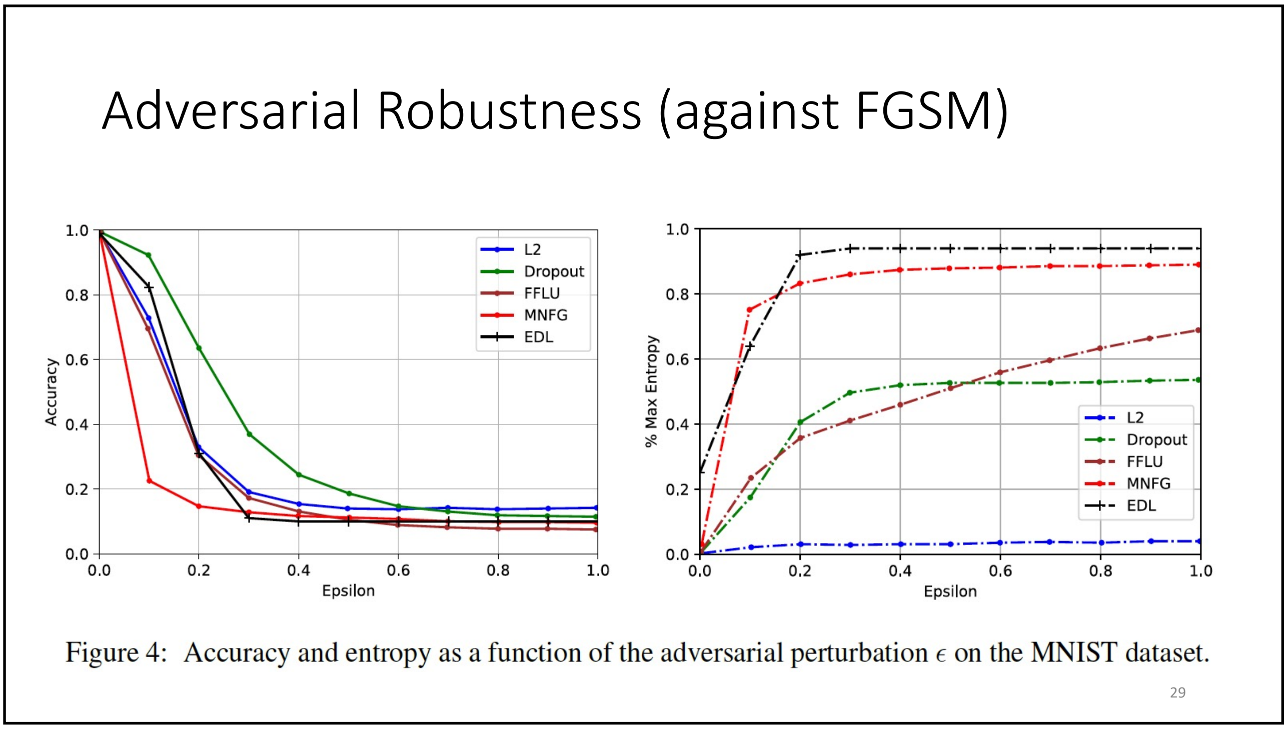
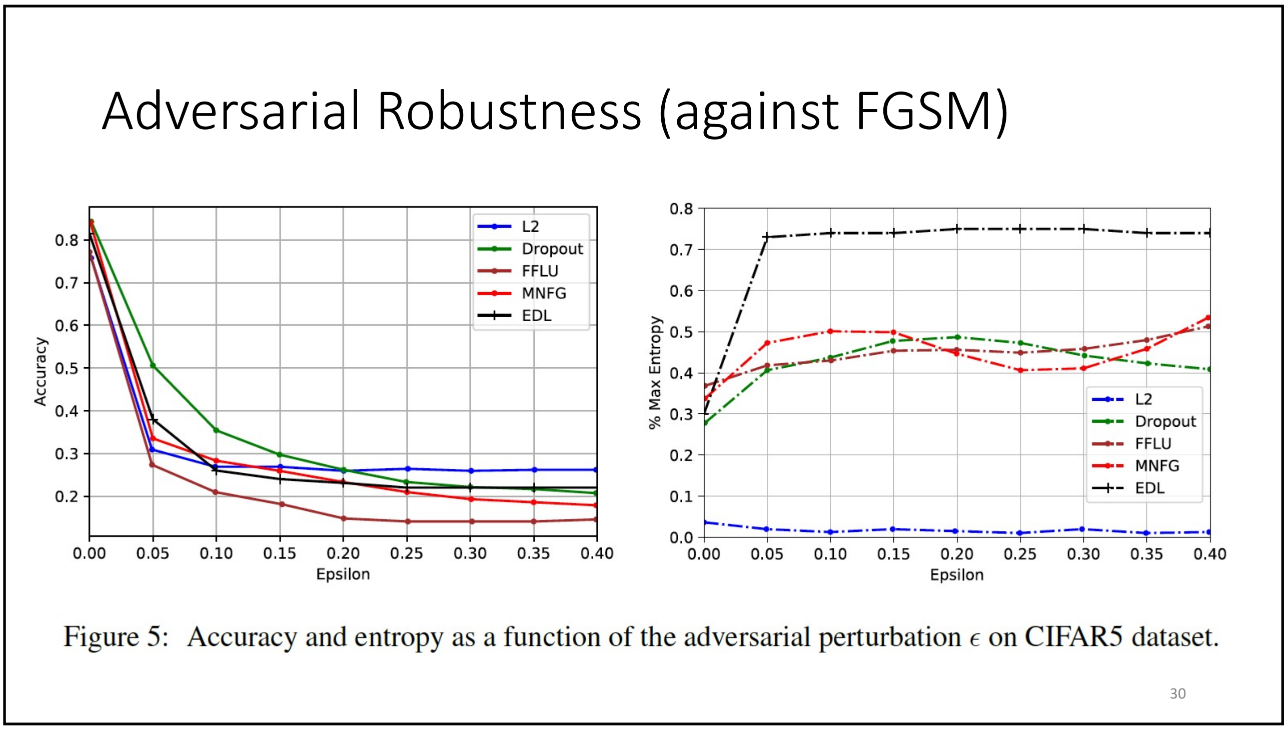


Figure 3: Empirical CDF for the entropy of the predictive distributions on the notMNIST dataset (left) and samples from the last five categories of CIFAR10 dataset (right).



Adversarial Robustness (against FGSM)

Figure 4: Accuracy and entropy as a function of the adversarial perturbation ϵ on the MNIST dataset.



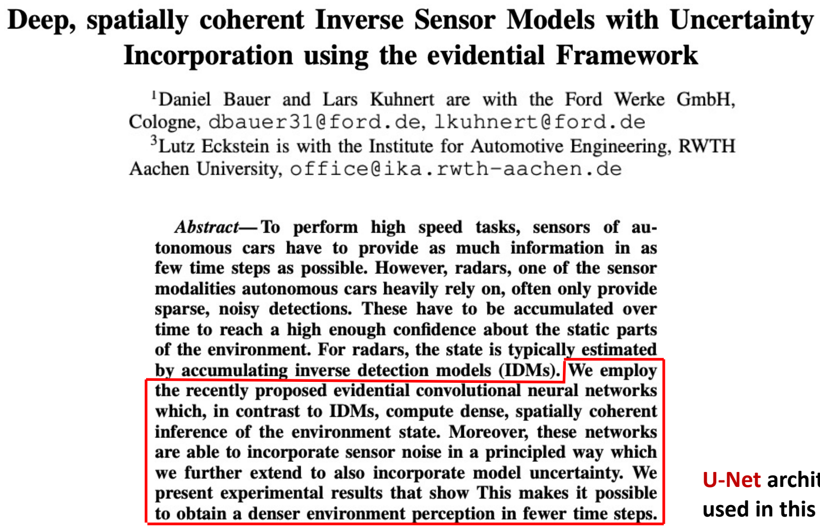
Adversarial Robustness (against FGSM)

Figure 5: Accuracy and entropy as a function of the adversarial perturbation ϵ on CIFAR5 dataset.

Some Real-World Examples using Evidential Deep Learning

31

31



Deep, spatially coherent Inverse Sensor Models with Uncertainty Incorporation using the evidential Framework

[1]Daniel Bauer and Lars Kuhnert are with the Ford Werke GmbH, Cologne, dbauer31@ford.de, lkuhnert@ford.de
[3]Lutz Eckstein is with the Institute for Automotive Engineering, RWTH Aachen University, office@ika.rwth-aachen.de

***Abstract*— To perform high speed tasks, sensors of autonomous cars have to provide as much information in as few time steps as possible. However, radars, one of the sensor modalities autonomous cars heavily rely on, often only provide sparse, noisy detections. These have to be accumulated over time to reach a high enough confidence about the static parts of the environment. For radars, the state is typically estimated by accumulating inverse detection models (IDMs). We employ the recently proposed evidential convolutional neural networks which, in contrast to IDMs, compute dense, spatially coherent inference of the environment state. Moreover, these networks are able to incorporate sensor noise in a principled way which we further extend to also incorporate model uncertainty. We present experimental results that show This makes it possible to obtain a denser environment perception in fewer time steps.**

U-Net architecture is used in this study.

D. Bauer, L. Kuhnert and L. Eckstein, "Deep, spatially coherent Inverse Sensor Models with Uncertainty Incorporation using the evidential Framework," *2019 IEEE Intelligent Vehicles Symposium (IV)*, Paris, France, 2019, pp. 2490-2495, doi: 10.1109/IVS.2019.8813826.

32

32

Quantifying and Leveraging Classification Uncertainty for Chest Radiograph Assessment

Florin C. Ghesu[1], Bogdan Georgescu[1], Eli Gibson[1], Sebastian Guendel[1],
Mannudeep K. Kalra[2,3], Ramandeep Singh[2,3], Subba R. Digumarthy[2,3],
Sasa Grbic[1], and Dorin Comaniciu[1]

[1] Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ, USA
[2] Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
[3] Harvard Medical School, Boston, MA, USA
florin.ghesu@siemens-healthineers.com

Table 1: Comparison between the reference method [4] and several versions of our method calibrated at sample rejection rates of 0%, 10%, 25% and 50% (based on the PLCO dataset [2]). Lesion refers to lesions of the bones or soft tissue.

	ROC-AUC				
Finding	Guendel et al. [4]	**Ours** [0%]	**Ours** [10%]	**Ours** [25%]	**Ours** [50%]
Granuloma	0.83	0.85	0.87	**0.90**	**0.92**
Fibrosis	0.87	0.88	0.90	**0.92**	**0.94**
Scaring	0.82	0.81	0.84	**0.89**	**0.93**
Lesion	0.82	0.83	0.86	**0.88**	**0.90**
Cardiac Ab.	0.93	0.94	0.95	**0.96**	**0.97**
Average	0.85	0.86	0.89	**0.91**	**0.93**

DenseNet121 is used in this study.

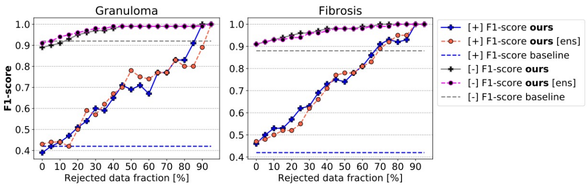


Fig. 2: Evolution of the F1-scores for the positive (+) and negative (−) classes relative to the sample rejection threshold - determined using the estimated uncertainty. We show the performance for granuloma and fibrosis based on the PLCO dataset [2]. The baseline (horizontal dashed lines) is determined using the method from [4] (working point at max. average of per-class F1 scores). Decision threshold for our method is fixed at 0.5.

In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. Springer, Cham

Quantifying and Leveraging Classification Uncertainty for Chest Radiograph Assessment

Florin C. Ghesu[1], Bogdan Georgescu[1], Eli Gibson[1], Sebastian Guendel[1],
Mannudeep K. Kalra[2,3], Ramandeep Singh[2,3], Subba R. Digumarthy[2,3],
Sasa Grbic[1], and Dorin Comaniciu[1]

[1] Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ, USA
[2] Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
[3] Harvard Medical School, Boston, MA, USA
florin.ghesu@siemens-healthineers.com

Labels in ChestX-Ray8 dataset are automatically generated by an NLP system from radiographic reports. A committee of 4 board certified experts analysed randomly selected test set of **689** cases; relabelled **120** samples are called *critical set*.

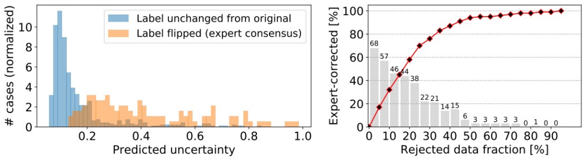
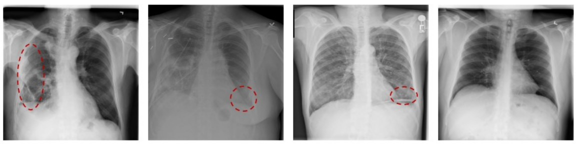


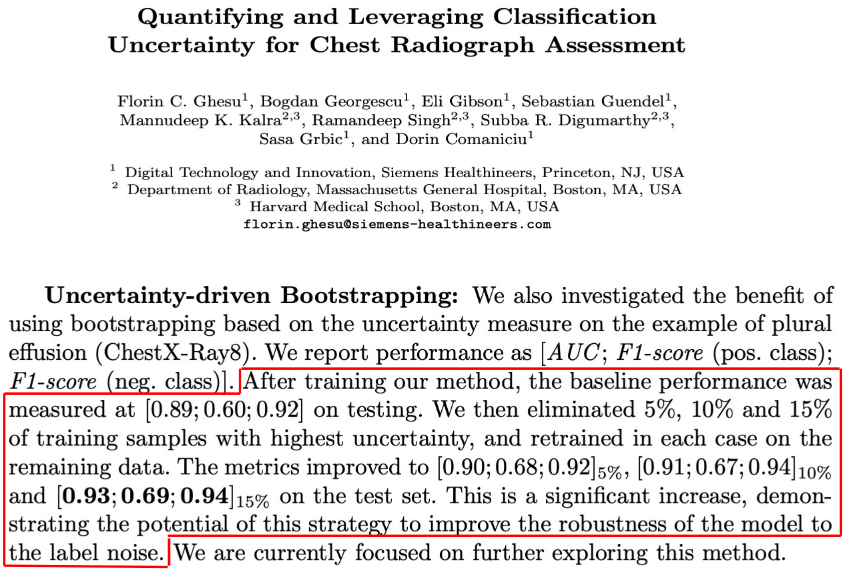
Fig. 3: **Left**: Predictive uncertainty distribution on 689 ChestX-Ray test images; a higher uncertainty is associated with cases of the critical set, which required a label correction according to expert committee. **Right**: Plot showing the capacity of the algorithm to eliminate cases from the critical set via sample rejection. Bars indicate the percentage of critical cases for each batch of 5% rejected cases.



(a) $\hat{u}, \hat{p} = 0.90, 0.45$ (b) $\hat{u}, \hat{p} = 0.93, 0.48$ (c) $\hat{u}, \hat{p} = 0.54, 0.65$ (d) $\hat{u}, \hat{p} = 0.11, 0.05$

Fig. 4: ChestX-Ray8 test images assessed for pleural effusion ($\hat{u}$: est. uncertainty, $\hat{p}$: output probability; with affected regions circled in red). Figures 4a, 4b and 4c show positive cases of the critical set with high predictive uncertainty – possibly explained by the atypical appearance of accumulated fluid in 4a, and poor quality of image 4b. Figure 4d shows a high confidence case with no pleural effusion.

In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. Springer, Cham



Quantifying and Leveraging Classification Uncertainty for Chest Radiograph Assessment

Florin C. Ghesu[1], Bogdan Georgescu[1], Eli Gibson[1], Sebastian Guendel[1], Mannudeep K. Kalra[2,3], Ramandeep Singh[2,3], Subba R. Digumarthy[2,3], Sasa Grbic[1], and Dorin Comaniciu[1]

[1] Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ, USA
[2] Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
[3] Harvard Medical School, Boston, MA, USA
florin.ghesu@siemens-healthineers.com

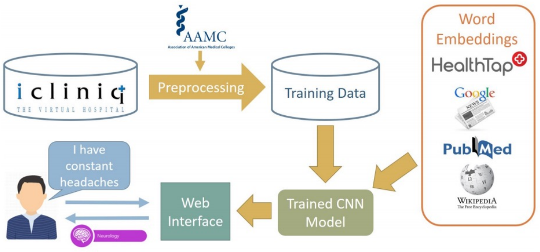
Uncertainty-driven Bootstrapping: We also investigated the benefit of using bootstrapping based on the uncertainty measure on the example of plural effusion (ChestX-Ray8). We report performance as $[AUC; \textit{F1-score}$ (pos. class); *F1-score* (neg. class)]. After training our method, the baseline performance was measured at $[0.89; 0.60; 0.92]$ on testing. We then eliminated 5%, 10% and 15% of training samples with highest uncertainty, and retrained in each case on the remaining data. The metrics improved to $[0.90; 0.68; 0.92]_{5\%}$, $[0.91; 0.67; 0.94]_{10\%}$ and $[\mathbf{0.93}; \mathbf{0.69}; \mathbf{0.94}]_{15\%}$ on the test set. This is a significant increase, demonstrating the potential of this strategy to improve the robustness of the model to the label noise. We are currently focused on further exploring this method.

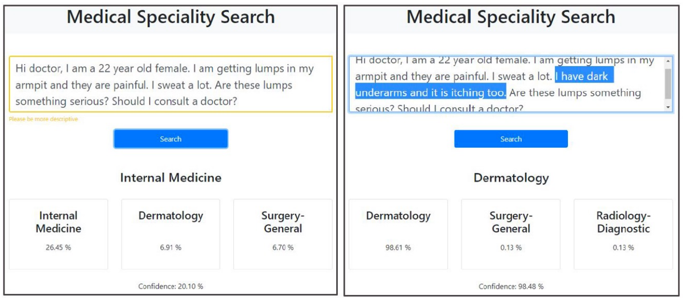
In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. Springer, Cham

35

35

MedSpecSearch: Medical Specialty Search





Hi doctor, I am a 22 year old female. I am getting lumps in my armpit and they are painful. I sweat a lot. Are these lumps something serious? Should I consult a doctor?

Please be more descriptive

Internal Medicine

Internal Medicine	Dermatology	Surgery-General
26.45 %	6.91 %	6.70 %

Confidence: 20.10 %

Hi doctor, I am a 22 year old female. I am getting lumps in my armpit and they are painful. I sweat a lot. I have dark underarms and it is itching too. Are these lumps something serious? Should I consult a doctor?

Dermatology

Dermatology	Surgery-General	Radiology-Diagnostic
98.61 %	0.13 %	0.13 %

Confidence: 98.48 %

Fig. 2. MedSpecSearch Front End with Two Example Queries

Using a confidence threshold like 90%, significantly reduces the amount of misclassifications of the system and increases the general prediction accuracy from 74% to 90.4%.

Şahin, M. U., Balatkan, E., Eran, C., Zeydan, E., & Yeniterzi, R. (2019, April). MedSpecSearch: Medical Specialty Search. In *European Conference on Information Retrieval* (pp. 225-229). Springer, Cham.

36

36

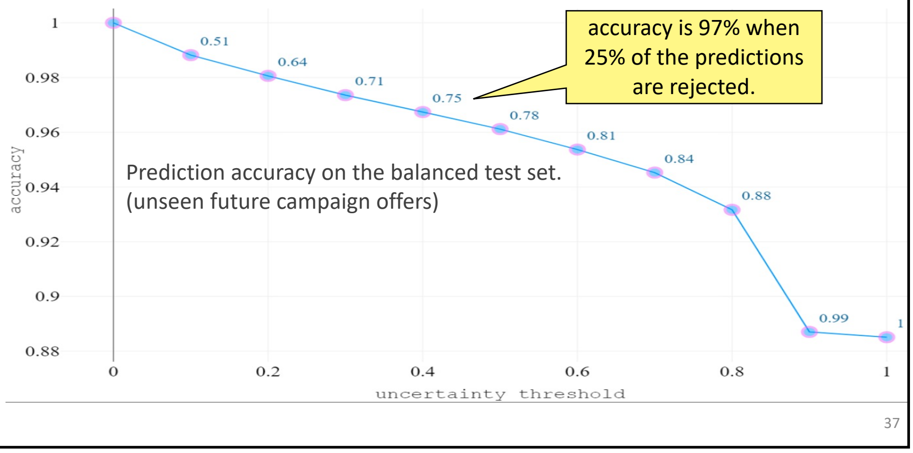
Campaign Participation Prediction for GSM Users

- Campaign offers are sent to users as short text messages.
- Binary classification is used to predict user participation based on Google’s *Wide&Deep* model.

Ayvaz, D., Aydoğan, R., Akçura, M. T., & Şensoy, M. (2021). Campaign participation prediction with deep learning. *Electronic Commerce Research and Applications*, *48*, 101058.

An example offer sent by the GSM operator:

“Get 500 minutes and 500MB only for 21$/month. You can join this campaign by texting ‘MONTHLY500MIN’ to 1111.”.



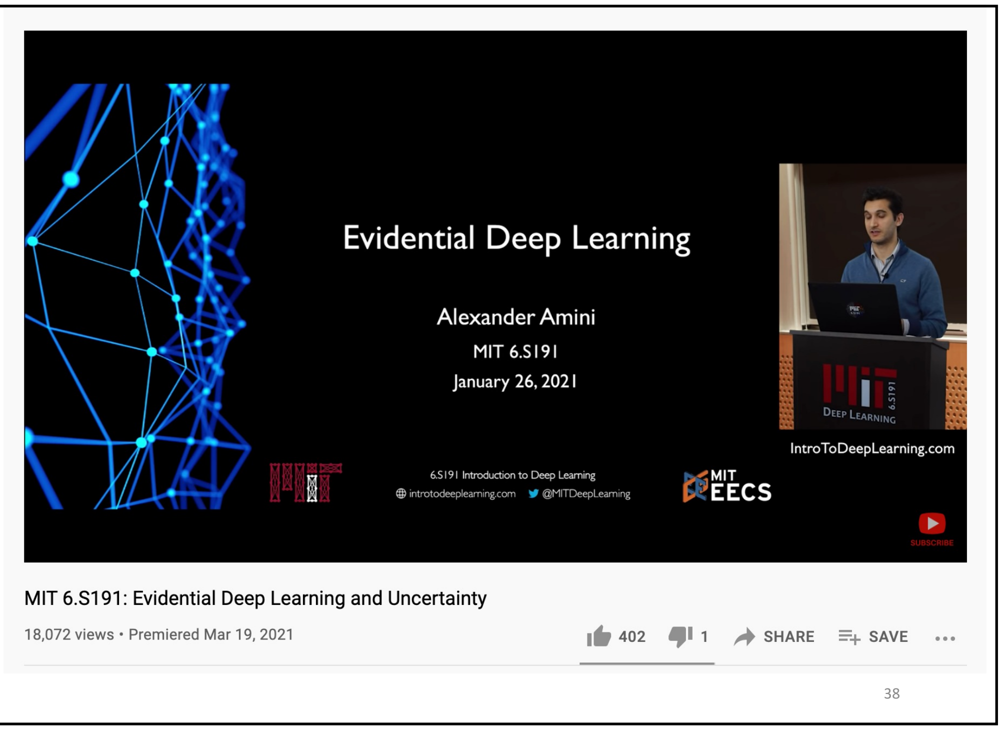
accuracy is 97% when 25% of the predictions are rejected.

Prediction accuracy on the balanced test set. (unseen future campaign offers)

37

37

EDL is extended by researchers in MIT for regression tasks.



Evidential Deep Learning

Alexander Amini

MIT 6.S191

January 26, 2021

6.S191 Introduction to Deep Learning

introtodeeplearning.com @MITDeepLearning

IntroToDeepLearning.com

MIT 6.S191: Evidential Deep Learning and Uncertainty

18,072 views • Premiered Mar 19, 2021

38

38

Deep Evidential Regression

Source: Alexander Amini, Wilko Schwarting, Ava Soleimany, Daniela Rus: Deep Evidential Regression. NeurIPS 2020

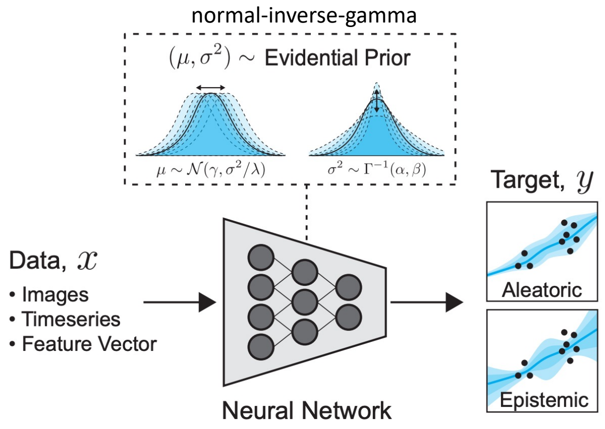

normal-inverse-gamma
$(\mu, \sigma^2) \sim$ Evidential Prior
$\mu \sim \mathcal{N}(\gamma, \sigma^2/\lambda)$
$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$
Target, y
Data, x
• Images
• Timeseries
• Feature Vector
Neural Network
Aleatoric
Epistemic

Figure 1: **Evidential regression** simultaneously learns a continuous target along with aleatoric (data) and epistemic (model) uncertainty. Given an input, the network is trained to predict the parameters of an evidential distribution, which models a higher-order probability distribution over the individual likelihood parameters, (μ, σ^2).

Conclusions

- Overconfidence of deep learning models stands as an important problem and require intensive research for the applicability of these models to real-world problems.
- Existing research focus on Bayesian approaches and sampling-based methods, while some promising work, such as EDL, can achieve calibrated uncertainties without using costly sampling methods.

Thank you!

41

41