Uncertainty Quantification in Deep Learning

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

Outline

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


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 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 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).

“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.”


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

Bayesian Neural Networks

**Source:** Weight Uncertainty in Neural Networks, Blundell et al., ICML 2015
Monte-Carlo Dropout (Bernoulli Approx. VI)


Deep Ensembles

1. Let each neural network parametrize a distribution over the outputs, i.e. pθ(y|x). Use a proper scoring rule as training criterion
   - Classification: cross entropy loss
   - Heteroscedastic Regression: net outputs mean μθ(x) and variance σθ(x)

   \[ \ell(\theta, x_n, y_n) = \frac{1}{2} \log \sigma_\theta(x) + \frac{(y - \mu_\theta(x))^2}{2\sigma_\theta(x)} + \text{const.} \]

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

\[ p(y|x) = \frac{1}{M} \sum_{m=1}^{M} p_{\theta_m}(y|x, \theta_m) \]

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

Evidential Deep Learning


Standard Approach for Classification

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

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

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

*Let's 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.

What happens if the image is rotated counter-clockwise? For some angles, it looks different from the images of digits 2 and 5 with high probability.

Subjective Logic

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

\[ \omega^s_x = (b, d, u) = \text{Beta} \left( \frac{2b}{u} + 1, \frac{2d}{u} + 1 \right) \]

Is the following true? digit(5, 9)

Opinion owner

Binary state variable

Belief + disbeliefs + uncertainty = 1

Evidence for \( x \)

Evidence for \( \neg x \)

Most likely NOT

I do NOT know

Beta(3, 17)

Beta(1, 1)
Dirichlet Distribution

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 simplex in $\mathbb{R}^K$ for various values of the $\alpha$ parameters and, at the bottom, 500 categorical distributions sampled from each of these Dirichlet distributions.

$$\text{Dirichlet}(\pi|\alpha) = \frac{1}{\mathcal{B}(|\alpha|)} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1} \text{ for } \pi \in S_K, \quad \text{otherwise}, \quad (1)$$

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

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

Subjective Logic Interpretation

Why is this useful?

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

$c = [9, 9, 9] \quad \alpha = [10, 10, 10]$

$\pi = \left( \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)$

Aleatoric Uncertainty

Epistemic Uncertainty

$c = [0, 0, 0] \quad \alpha = [1, 1, 1]$

$\pi = \left( \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)$

Uniform Categorical Distribution
Evidential Deep Classifiers

\[ q_\theta(\pi | x) = \text{Dirichlet}(\pi | f_\theta(x) + 1) \]

EDL loss functions

- For each input \( x_i \), we define a base loss function parametrized by the latent categorical distribution \( \pi_i \) and compute its expectation using the predicted Dirichlet distribution: \( \text{Dirichlet}(\pi_i | f_\theta(x_i) + 1) \).

- Expected cross-entropy loss:

\[
\mathcal{L}_i(\theta) = \mathcal{L}_{\text{cross-entropy}} + \mathcal{L}_{\text{Dirichlet}} = \sum_{j=1}^{K} -y_{ij} \log(\pi_{ij}) + \frac{1}{B(\alpha_i)} \prod_{j=1}^{K} \pi_{ij}^{\alpha_{ij}-1} \, d\pi_i = \sum_{j=1}^{K} y_{ij} \left( \psi(\sum_{k=1}^{K} \alpha_{ik}) - \psi(\alpha_{ij}) \right)
\]

\( \psi \) is the digamma function, i.e., the derivative of the log gamma function.
\( y_i \) is the one-hot label vector for the sample \( x_i \).
EDL loss functions

**Type II maximum likelihood:**

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

$$
\mathcal{L}_i(\theta) = -\log \left( \int \left[ \prod_{j=1}^{K} \pi_{ij}^{y_{ij}} \right] \frac{1}{B(\alpha_i)} \prod_{j=1}^{K} \pi_{ij}^{\alpha_{ij} - 1} d\pi_i \right) = \sum_{j=1}^{K} y_{ij} \left( \log \left( \sum_{k=1}^{K} \alpha_{ik} \right) - \log(\alpha_{ij}) \right)
$$

---

**EDL loss functions**

**The expected sum square error loss (Brier score):**

$$
\mathcal{L}_i(\theta) = \int \frac{1}{B(\alpha_i)} \prod_{j=1}^{K} \pi_{ij}^{\alpha_{ij} - 1} d\pi_i \\
= \sum_{j=1}^{K} \mathbb{E} \left[ y_{ij}^2 - 2y_{ij}\pi_{ij} + \pi_{ij}^2 \right] = \sum_{j=1}^{K} \left( y_{ij}^2 - 2y_{ij}\mathbb{E}[\pi_{ij}] + \mathbb{E}[\pi_{ij}^2] \right)
$$
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
\[ \mathcal{L}_i(\theta) = \sum_{j=1}^{K} (y_{ij} - \mathbb{E}[\pi_{ij}])^2 + \text{Var}(\pi_{ij}) = \sum_{j=1}^{K} \left( \frac{(y_{ij} - \hat{\pi}_{ij})^2}{\mathcal{L}_{ij}^{err}} + \frac{\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})}{1 + \sum_k \alpha_{ik}} \right). \]

- **Proposition 1.** For any \( \alpha_{ij} \geq 1 \), the inequality \( \mathcal{L}_{ij}^{err} < \mathcal{L}_{ij}^{err} \) is satisfied,
  i.e. The loss prioritizes data fit over variance estimation.

- **Proposition 2.** For a given sample \( i \) with the correct label \( j \), \( \mathcal{L}_{ij}^{err} \) decreases when new evidence is added to \( \alpha_{ij} \) and increases when evidence is removed from \( \alpha_{ij} \),
  i.e. The loss has a tendency to fit to the data.

- **Proposition 3.** For a given sample \( i \) with the correct class label \( j \), \( \mathcal{L}_{ij}^{err} \) decreases when some evidence is removed from the biggest Dirichlet parameter \( \alpha_{il} \) such that \( l \neq j \),
  i.e. The loss performs learned loss attenuation.

---

**Overall loss function**

\[
\mathcal{L}(\theta) = \sum_{i=1}^{N} \mathcal{L}_i(\theta) + \lambda_t \sum_{i=1}^{N} \text{KL} [\text{Dirichlet}(\pi_i | \hat{\alpha}_i) \| \text{Dirichlet}(\pi_i | 1)]
\]

Maximize model fit

Minimize evidence on errors.

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

\( \lambda_t \) is the annealing coefficient; initially 0 and increased gradually to 1 during training.
\( \hat{\alpha} \) refers to the predicted Dirichlet parameters after removing the evidence for the true category.
Revisiting MNIST

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

Evaluations

We use the following approaches in our evaluations:

- Vanilla Neural Nets with weight decay (L2)
- Variational dropout and the local reparameterization trick (FFLU). Kingma et al. NIPS, 2015
Accuracy vs Uncertainty

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST</th>
<th>CIFAR 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td>99.4</td>
<td>76</td>
</tr>
<tr>
<td>Dropout</td>
<td>99.5</td>
<td>84</td>
</tr>
<tr>
<td>Deep Ensemble</td>
<td>99.3</td>
<td>79</td>
</tr>
<tr>
<td>FFG</td>
<td>99.1</td>
<td>78</td>
</tr>
<tr>
<td>FFLU</td>
<td>99.1</td>
<td>77</td>
</tr>
<tr>
<td>MNFG</td>
<td>99.3</td>
<td>84</td>
</tr>
<tr>
<td><strong>EDL</strong></td>
<td><strong>99.3</strong></td>
<td><strong>83</strong></td>
</tr>
</tbody>
</table>

Out-of-Distribution Samples

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 $\epsilon$ on the MNIST dataset.

Adversarial Robustness (against FGSM)

Figure 5: Accuracy and entropy as a function of the adversarial perturbation $\epsilon$ on CIFAR5 dataset.
Some Real-World Examples using Evidential Deep Learning

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

1Daniel Bauer and Lars Kuhnert are with the Ford Werke GmbH, Cologne, dbauer31@ford.de, lkuhnert@ford.de
2Lutz Eckstein is with the Institute for Automotive Engineering, RWTH Aachen University, office@iata.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.
Quantifying and Leveraging Classification Uncertainty for Chest Radiograph Assessment

Florin C. Ghese1, Bogdan Georgescu1, Eli Gibson1, Sebastian Gaende1, Manmohan K. Khair2,3, Ramanan Singh2,3, Subba R. Dugumarthi2,3, Sasa Grbic1, and Dorin Comaniciu1

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.ghese@siemens-healthineers.com

DenseNet121 is used in this study.

Table 1: Comparison between the reference method [4] and several versions of our method calibrated at sample rejection rates of 6%, 10%, 25% and 50% (based on the PLCO dataset [2]). Lesion refers to lesions of the bones or soft tissue.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Granuloma</td>
<td>0.83</td>
<td>0.85</td>
<td>0.87</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>Fibrosis</td>
<td>0.87</td>
<td>0.88</td>
<td>0.90</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>Scarring</td>
<td>0.82</td>
<td>0.81</td>
<td>0.84</td>
<td>0.89</td>
<td>0.93</td>
</tr>
<tr>
<td>Lesion</td>
<td>0.82</td>
<td>0.83</td>
<td>0.86</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td>Cardiac Ab.</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>Average</td>
<td>0.85</td>
<td>0.86</td>
<td>0.89</td>
<td>0.91</td>
<td>0.93</td>
</tr>
</tbody>
</table>

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 line) is determined using the method from [6] (working point at max. average of per-class F1 scores). Decision threshold for our method is fixed at 0.5.


Quantifying and Leveraging Classification Uncertainty for Chest Radiograph Assessment

Florin C. Ghese1, Bogdan Georgescu1, Eli Gibson1, Sebastian Gaende1, Manmohan K. Khair2,3, Ramanan Singh2,3, Subba R. Dugumarthi2,3, Sasa Grbic1, and Dorin Comaniciu1

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.ghese@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.

Fig. 4: ChestX-Ray8 test images assessed for pleural effusion (ie. est. uncertainty, 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.

Quantifying and Leveraging Classification Uncertainty for Chest Radiograph Assessment
Florin C. Gheneş1, Bogdan Georgescu1, Eli Gibson1, Sebastian Gaedel1, Mannosleep K. Koha3,4, Ramaanlodh Sughp3,4, Subba R. Digumarthy2,4, Saen Grbic1, and Dorin Comaniciu1
1 Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ, USA
2 Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
3,4 Harvard Medical School, Boston, MA, USA
florin.ghesen@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 \(\text{AUC; } F1\text{-score (pos. class); } F1\text{-score (neg. class)}\). 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 \([0.93; 0.69; 0.94]_{15\%}\) 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.


MedSpecSearch: Medical Specialty Search

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%.

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.


**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.

EDL is extended by researchers in MIT for regression tasks.
Deep Evidential Regression

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

**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, $(\mu, \sigma^2)$.

**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!