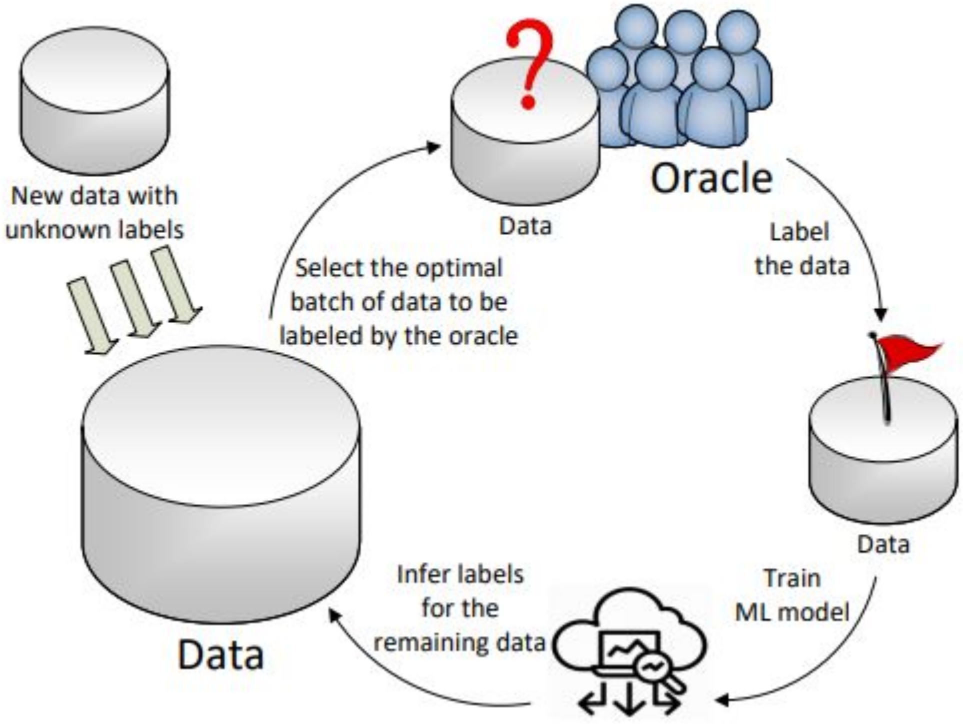
Theory of Evidence in Active Learning

Daniel Kałuża

Active Learning

Goal: Obtain the best possible model with limited labelling capabilities, assuming possibility of experts-model interaction.

Active Learning cycle



Source: A. Janusz, Ł. Grad, M. Grzegorowski, “Clash Royale Challenge: How to Select Training Decks for Win-rate Prediction”

Active Learning - approaches

Usually based on:

- Informativeness (E. g. Max Entropy, Prediction Margin)
- Representativeness (E. g. Clustering based, Distance based)
- Dissimilarity (E. g. Distance to the current batch)

Theory of Evidence - Basics

A different view on probability, distinguishing:

- subjective beliefs
  from
- objective chances

Focuses on sets of random events instead of single events.

Theory of Evidence - Rules

Let θ be a finite set of possible states. Then if function Bel: 2^θ -> [0, 1] satisfies conditions:

1. Bel(ø) = 0
2. Bel(θ) = 1
3. For every positive n and every collection of subsets A_1, A_2, ... , A_n of θ:

$$Bel(A_1 \cup A_2 \cup ... \cup A_n) >= \sum_{i=1}^{n} Bel(A_i) - \sum_{j=i+1}^{n} Bel(A_i \cap A_j) + ... + (-1)^{n+1} Bel(A_1 \cap A_2 \cap ... \cap A_n)$$

Then Bel is called a belief function over θ.

Theory of Evidence - Example

Let $\theta = \{\theta_1, \theta_2\}$

θ_1 - genuine

θ_2 - counterfeit

$Bel(\theta_1) = a$

$Bel(\theta_2) = b$

$Bel(\{\}) = 0$

$Bel(\theta) = 1$

By Mary and Jon Hirschfeld Workshop https://www.flickr.com/photos/24449799@N03/3316554665, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=55189743

Theory of Evidence - Uncertainty Intuition

Lets consider the following random events:

A - the dice number will be even

B - the dice number will be odd

C1 - the dice number will be 1

C3 - the dice number will be 3

C5 - the dice number will be 5

Bayesian uncertainty:

- $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$

What about:

- $P(A) = \frac{1}{2}$, $P(C1) = \frac{1}{6}$, $P(C3) = \frac{1}{6}$, $P(C5) = \frac{1}{6}$

In Theory of Evidence we can say:

$Bel(\{X\}) = 0$, for X in {A, C1, C3, C5}

$Bel(\{A, C1, C3, C5\}) = 1$

Example of application - Neural Networks

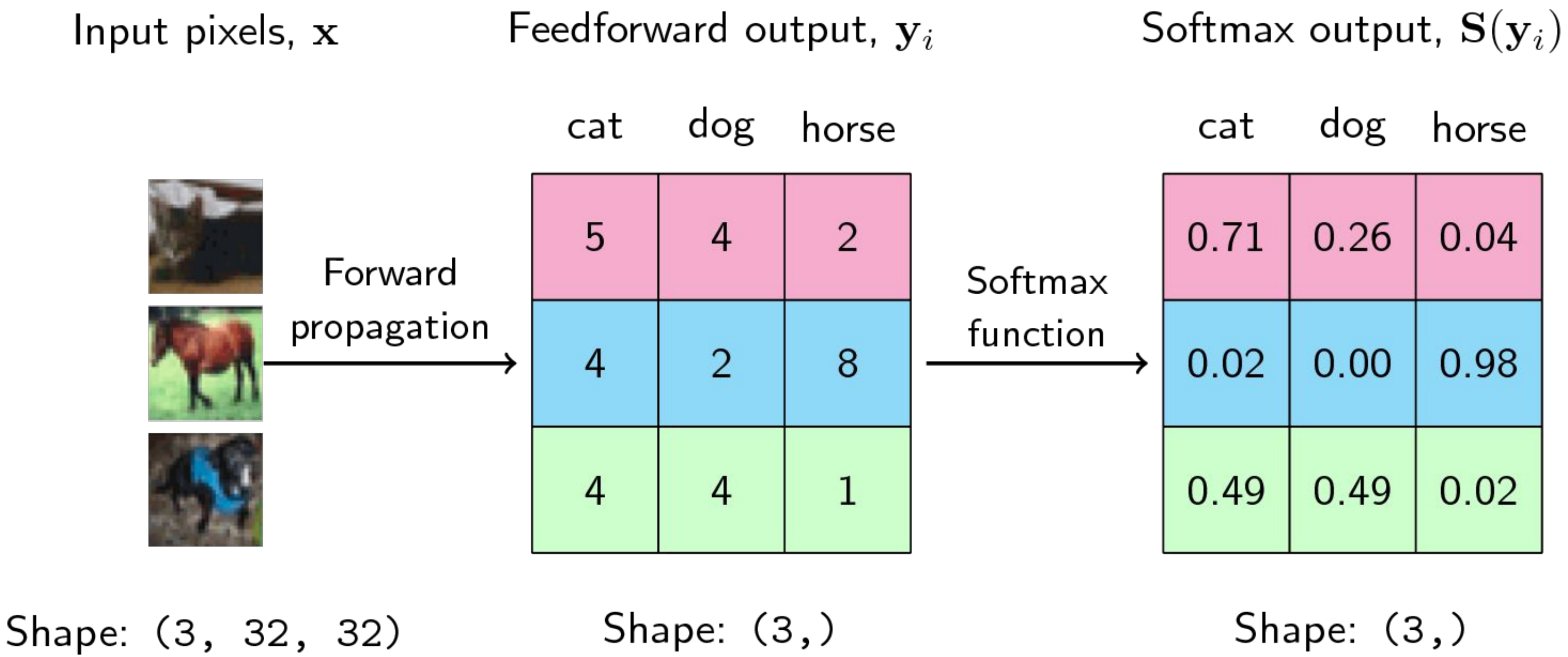
Regular neural network classifier:

- softmax as an output
- output interpreted as probability distribution
- uncertainty measured on output, e.g. entropy
- optimized with cross-entropy and gradient based methods

Softmax - inflating the probabilities

$$\sigma(x)_j = \frac{e^{x_j}}{\sum_k e^{x_k}}$$

Softmax - inflating the probabilities

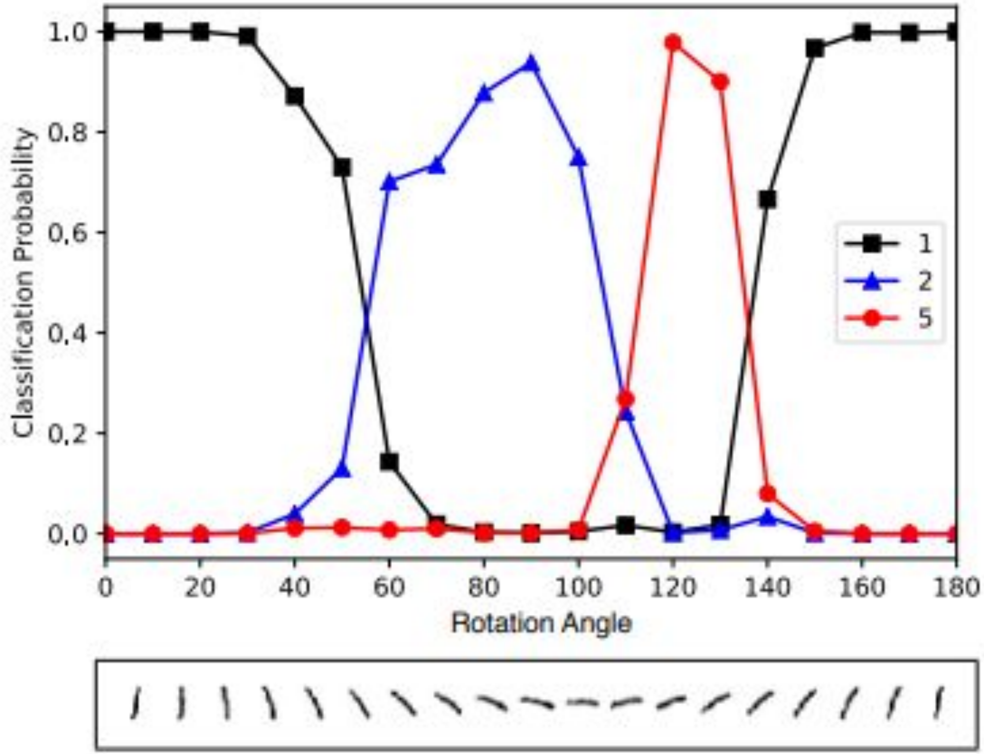


Source: https://ogunlao.github.io/images/softmax.png

Cross-entropy loss ~ Maximum Likelihood Estimation

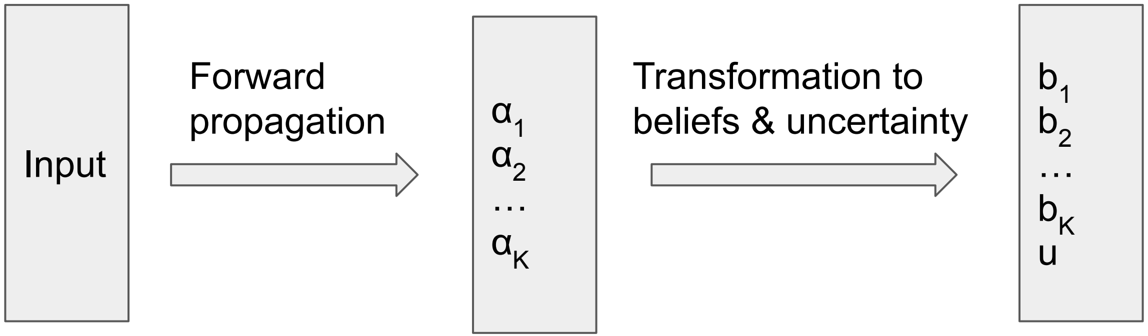
MLE as a frequentist method, therefore it isn't capable to describe the distribution variance!

MNIST example



Source: Sensoy et al. “Evidential Deep Learning to Quantify Classification Uncertainty”

Draft of idea - replace softmax with Dirichlet Distribution



Modeling DST with Subjective Logic & Dirichlet Distribution

$$u + \sum_{k=1}^{K} b_k = 1, \qquad u = \frac{K}{S}, \qquad b_k = \frac{e_k}{S} \qquad S = \sum_{i=1}^{K}(e_i + 1)$$

b_k - belief of mass corresponding to k-th singleton class

u - uncertainty

e_i - evidence for the i-th singleton class

K - number of classes

Dirichlet Distribution

$$D(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1} & \text{for } \mathbf{p} \in \mathcal{S}_K, \\ 0 & \text{otherwise,} \end{cases} \qquad \alpha_k = e_k + 1 \qquad b_k = \frac{e_k}{S}$$

b_k - belief of mass corresponding to k-th singleton class

u - uncertainty

e_i - evidence for the i-th singleton class

K - number of classes

α_k - parameter of Dirichlet distribution corresponding to k-th class

Loss & Training

$$\mathcal{L}_i(\Theta) = \sum_{j=1}^{K} (y_{ij} - \mathbb{E}[p_{ij}])^2 + \mathrm{Var}(p_{ij}) + \lambda_t \sum_{i=1}^{N} KL[D(\mathbf{p_i}|\tilde{\boldsymbol{\alpha}}_i) \;||\; D(\mathbf{p}_i|\langle 1, \ldots, 1 \rangle)],$$

$\lambda_t = \min(1.0, t/10) \in [0, 1]$ where t is an index of learning epoch

KL - Kullback-Leibler divergence

$\tilde{\boldsymbol{\alpha}}_i = \mathbf{y}_i + (1 - \mathbf{y}_i) \odot \boldsymbol{\alpha}_i$

K - number of classes

α_k - parameter of Dirichlet distribution corresponding to k-th class

Results

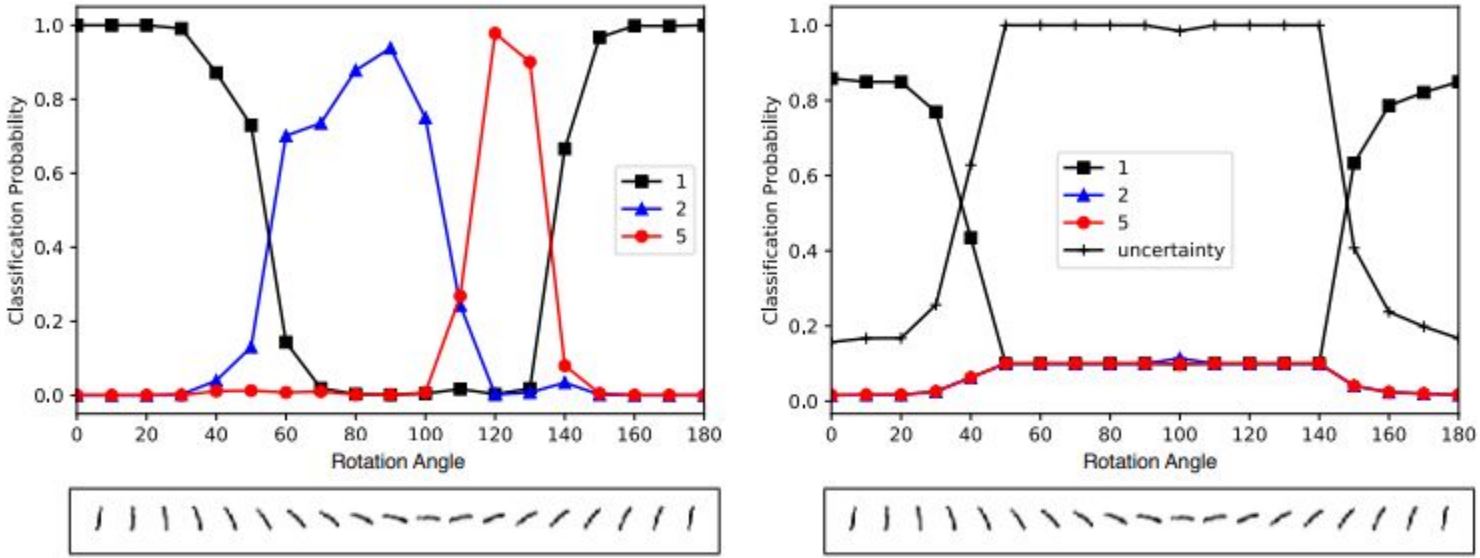


Figure 1: Classification of the rotated digit 1 (at bottom) at different angles between 0 and 180 degrees. **Left:** The classification probability is calculated using the *softmax* function. **Right:** The classification probability and uncertainty are calculated using the proposed method.

Results

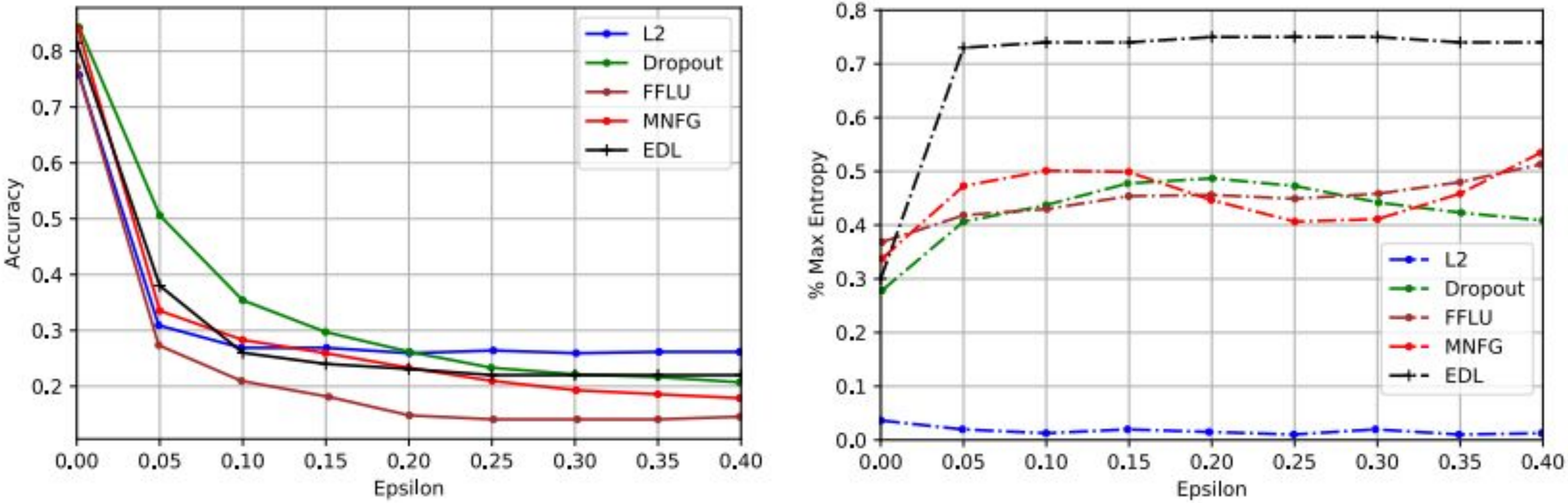
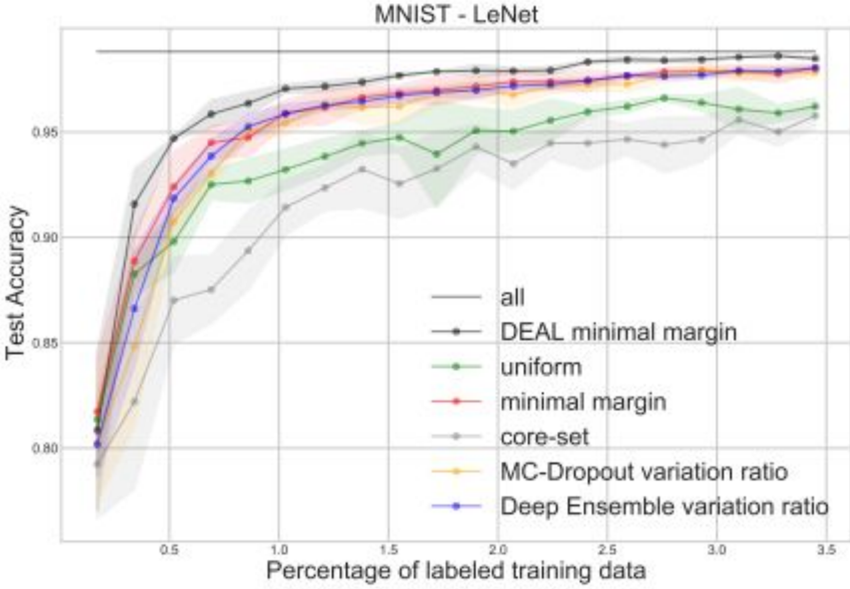
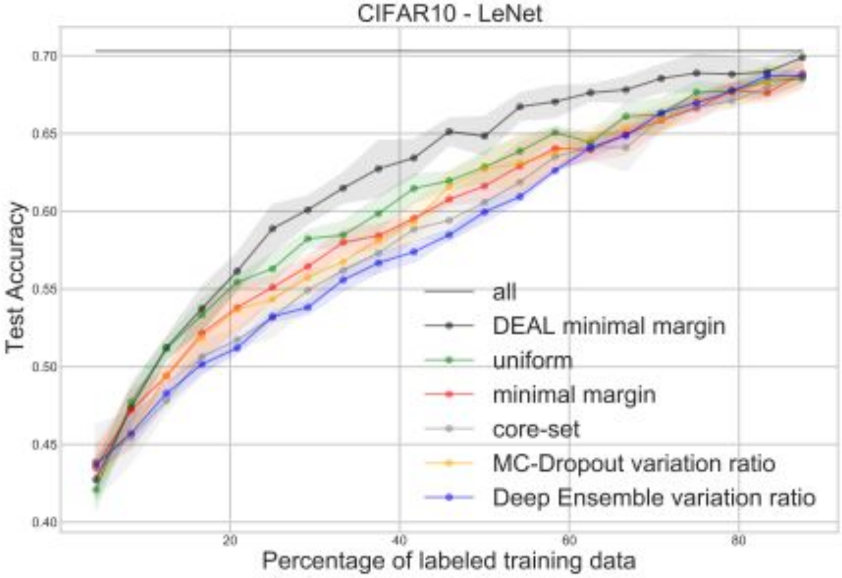


Figure 5: Accuracy and entropy as a function of the adversarial perturbation ϵ on CIFAR5 dataset.

Results Active Learning CNN





Conclusions and thoughts

- interesting usage of Dirichlet distribution
- why authors are not using uncertainty for AL?
- can the same be done for other softmaxed methods? e.g. xgboost
- maybe there is a better way to incorporate DS theory to machine learning models?

Bibliography

1. Glenn Shafer, “A Mathematical Theory of Evidence”, 1976
2. Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)
3. P. Hemmer, N. Kühl and J. Schöffer, "DEAL: Deep Evidential Active Learning for Image Classification," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)

Thank you for attention