Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

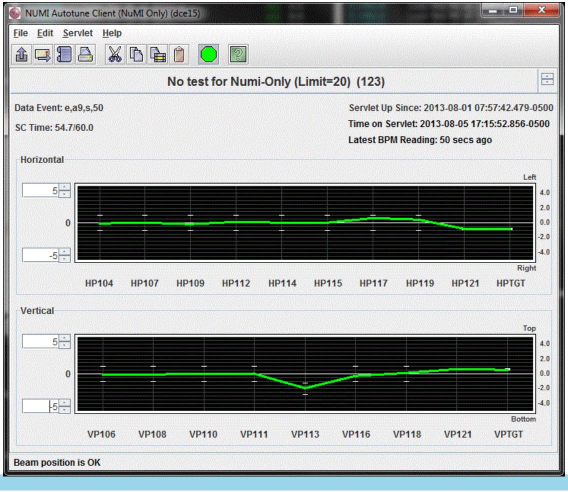
NOICE: Deep Ensemble Confidence Levels for Multi-hot Categorization

Giovani Leone

NOICE (Neural Optical Image Categorizer for the E-log)

Small collaboration tasked with categorizing the images in the Fermilab Accelerator Division electronic logbook by using Artificial Intelligence

- Manually categorized 7177 images (~300,000 Images in the E-log)
- Multi-hot-encoding
  - "Application", "Parameter Page", "Plot", "Document", "Drawing", "Photograph", "Diagram", "NOICE", and "Undefined”



No test for Numi-Only (Limit=20) (123)

Data Event: e,a9,s,50

SC Time: 54.7/60.0

Servlet Up Since: 2013-08-01 07:57:42.479-0500

Time on Servlet: 2013-08-05 17:15:52.856-0500

Latest BPM Reading: 50 secs ago

Beam position is OK

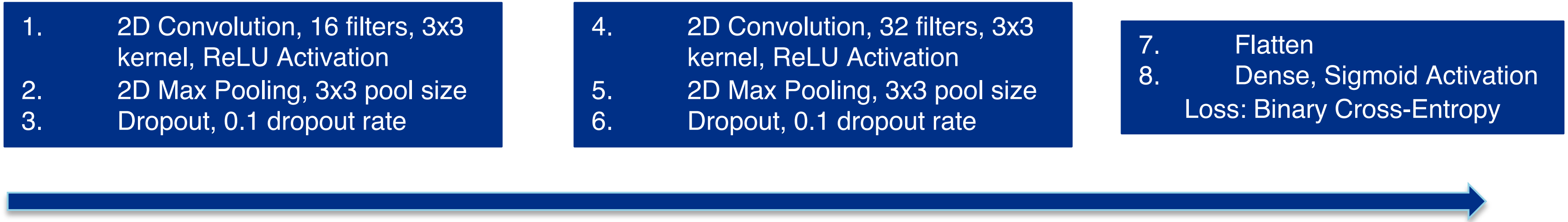
Ground truth is an “Application” and a “Plot”

- [1,0,1,0,0,0,0,0,0]

Deep Ensembles

Ensemble of deep neural networks[1]

- 100 random initializations of a deep neural network
  - built in TensorFlow2 (version 2.3)



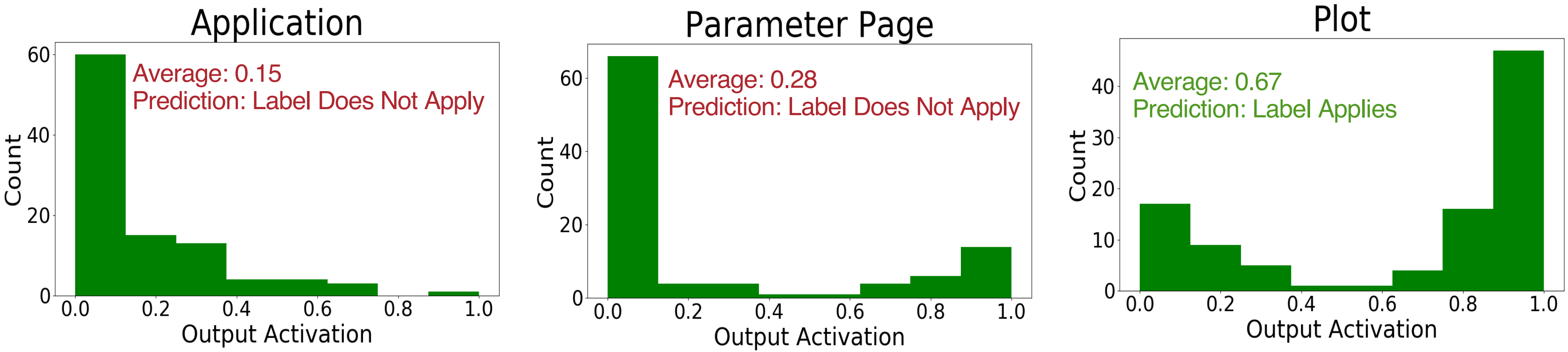
1. 2D Convolution, 16 filters, 3x3 kernel, ReLU Activation
2. 2D Max Pooling, 3x3 pool size
3. Dropout, 0.1 dropout rate
4. 2D Convolution, 32 filters, 3x3 kernel, ReLU Activation
5. 2D Max Pooling, 3x3 pool size
6. Dropout, 0.1 dropout rate
7. Flatten
8. Dense, Sigmoid Activation

Loss: Binary Cross-Entropy

  - Prediction gives a distribution of 100 output sigmoid activation scores for each label of each image (range (0,1))
  - The scores are compiled on a given label to determine the verdict
  - Used the output sigmoid activation scores as a measure of a model's *self-confidence*
    - collection of the models' *self-confidences* used to generate an ensemble confidence level

[1] B. Lakshminarayanan, A. Pritzel, and C. Blundell, ArXiv:1612.01474 [Cs, Stat] (2017).

Ensemble Output Distributions



- Output sigmoid activation scores treat each label independently
- Ensemble decides that a label applies if and only if the average of the output sigmoid activation scores is greater than 0.5
- Distribution of output sigmoid activation scores can vary in spread and modality

Confidence Level Calculation

- For each sigmoid output activation score on a label
  - score of 0: 100% confidence that the label does not apply
  - score of 0.5: 0% confidence that the label applies and does not apply
  - score of 1: 100% confidence that the label does apply
- Define confidence level C on a choice of labeling made by an N model ensemble:

$$C = \left| \sum_{n=1}^{N} c(s(n)) \right|$$

  - $c(s(n))$ is the *self-confidence functional*
  - $s(n)$ is the output sigmoid activation score of the n^{th} model
  - Normalization condition of $\mathbf{1} = \left|\sum_{\boldsymbol{n}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{c}(\mathbf{1})\right|$

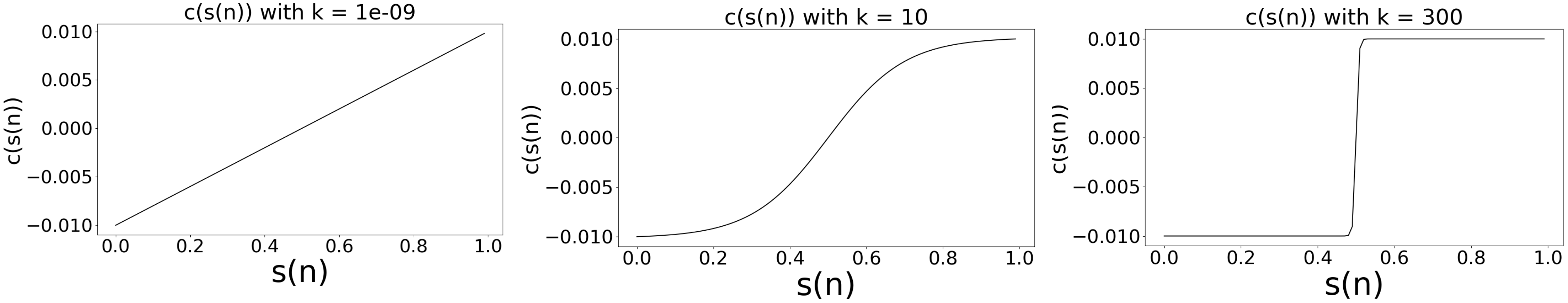
The Sigmoid-Shaped Self-Confidence Functional

- Used a sigmoid-shaped self-confidence functional, centered at $s(n) = 0.5$

$$c_{sigmoid}(s(n)) = A\left(\frac{e^{k(s(n)-0.5)}}{e^{k(s(n)-0.5)}+1} - 0.5\right)$$

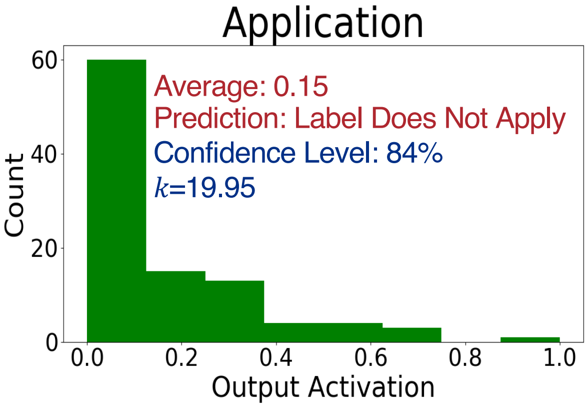
  - A determined by the normalization condition
    - $A = \left(N\left(\frac{e^{0.5k}}{e^{0.5k}+1} - 0.5\right)\right)^{-1}$
  - k was chosen for each label by a calibration condition
    - The accuracy of a label over all images equals the average of the confidence levels on that label over all images
- Range over [0,1]: [-N,N]

The Sigmoid-Shaped Self-Confidence Functional



- $c_{sigmoid}\left(s(n)\right)$ of a 100-model ensemble for $k = 10^{-9},\ 10,\ 300$ (from left to right)
  - As k approaches 0, $c_{sigmoid}\left(s(n)\right)$ becomes symmetric about $s(n) = 0.5$
  - As k becomes large, $c_{sigmoid}\left(s(n)\right)$ approaches a signum function, scaled by A

Preliminary Results

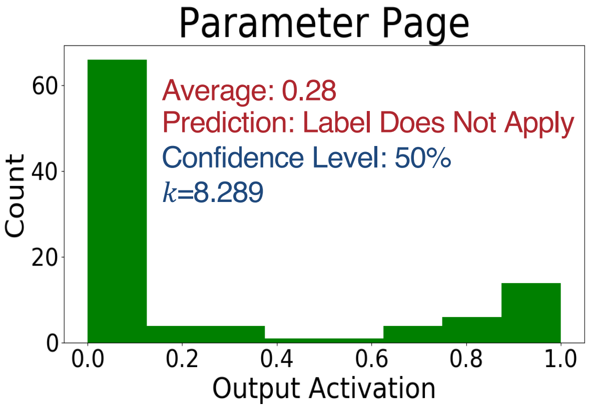

Application

Average: 0.15

Prediction: Label Does Not Apply

Confidence Level: 84%

k=19.95

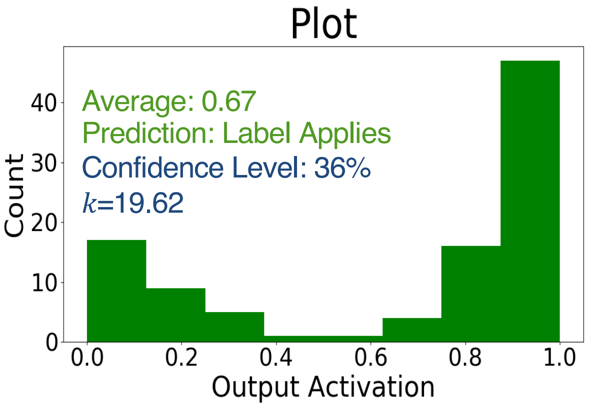

Parameter Page

Average: 0.28

Prediction: Label Does Not Apply

Confidence Level: 50%

k=8.289


Plot

Average: 0.67

Prediction: Label Applies

Confidence Level: 36%

k=19.62

- Average label accuracies
  - "Application": 0.836; "Parameter Page": 0.708; "Plot": 0.625
- Wider spread or bimodality yields a lower confidence level
  - Lowest calculable confidence level being 0 for a perfectly symmetric distribution
    - By the symmetry of the sigmoid-shaped *self-confidence functional*

Conclusions

- A *self-confidence functional* can be calibrated to a deep ensemble's accuracy and used to calculate the confidence levels on labels for a multi-hot-encoded Deep Ensemble
  - Also applicable to single-hot-encoded models utilizing sigmoid output activation functions.
- Future explorations of this technique
  - Evaluating the confidence levels of labels across a large data set
  - Comparing the average confidence level calculation of a label to the Deep Ensemble's accuracy on a large, unseen data set.
    - Test the predictive nature of the Confidence Level Calculation
  - Utilize the Confidence Level Calculation on other machine learning uncertainty estimation tools
    - Concrete Dropout

References and Acknowledgements

[1] B. Lakshminarayanan, A. Pritzel, and C. Blundell, ArXiv:1612.01474 [Cs, Stat] (2017).

– Jason St. John
– Kyle Hazelwood
– Justin Rower and Terence Njekeu
– Fermilab
– This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI)