Detecting Data Drift in Image Classification Neural Networks

Dr. Spyros Tragoudas
Professor and School Director
Southern Illinois University Carbondale

Joint work with: Danushka Senarathna[1], Kiriti Nagesh Gowda[2], Mike Schmit[2]

1. School of ECBE Southern Illinois University Carbondale, IL
2. ML Computer Vision Group, Advanced Micro Devices, Inc. Santa Clara, CA

Overview

- Introduction
- Contribution
- Motivation
- Proposed Methodology
- Experimental Results
- Conclusions

© 2023 Southern Illinois University Carbondale

Introduction

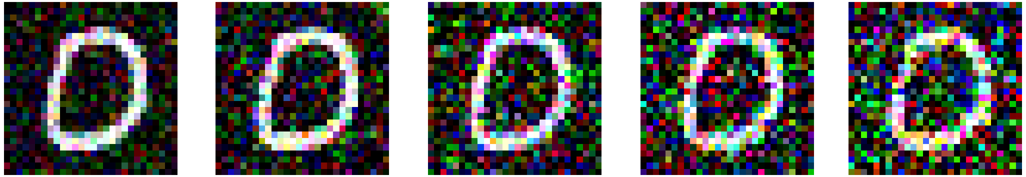
- Image classification applications
  - Self-driving cars
  - Mobile computing devices
  - Medical imaging, etc.
- Data drift: A change in the input data distribution that impacts the accuracy of machine learning model
- Reasons for data drifts:
  - Noise effects
  - Weather effects
  - Camera degradation
  - Change in the lighting, etc.

© 2023 Southern Illinois University Carbondale

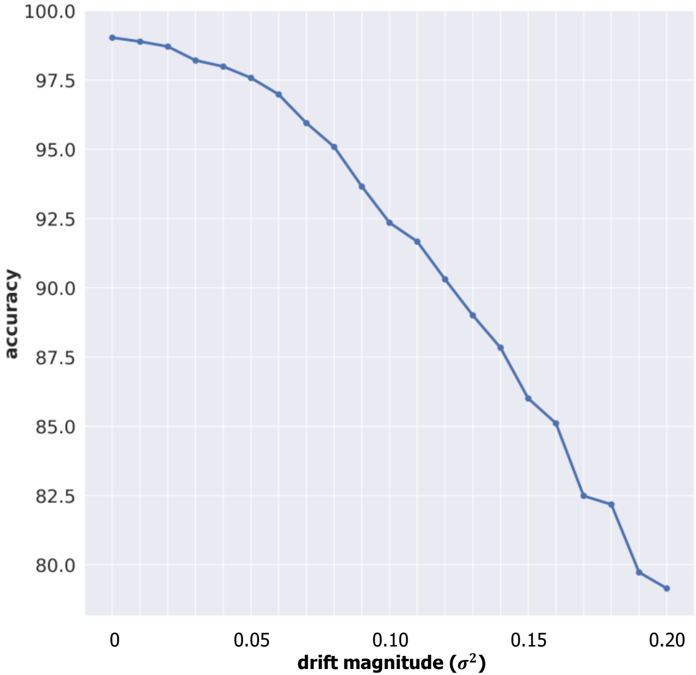
Introduction

- Accuracy of a neural network model degrades due to data drift
- Example: MNIST digits with Gaussian noise
  - Classify digits using a convolutional neural network
  - Validation accuracy = 99.10%
  - Test the model for images with different magnitudes of Gaussian noise
  - Accuracy drops below 80% when noise variance (σ^2) is 0.2

Sample images with different levels of Gaussian noise



$\sigma^2 = 0.04$ | $\sigma^2 = 0.08$ | $\sigma^2 = 0.12$ | $\sigma^2 = 0.16$ | $\sigma^2 = 0.20$



© 2023 Southern Illinois University Carbondale

Introduction

- Objective:
  - Develop an approach to cope with data drifts of any type
  - Detect and identify the drift type and estimate the drift magnitude during the network's operation
- Data drifts considered in this work,
  - Noise effects
    - Gaussian
    - Poisson
    - Salt & Pepper
  - Weather effects:
    - Snow
    - Fog
    - Shadow

© 2023 Southern Illinois University Carbondale

Contribution

- Previous work
  - [1] considers data drifts due to outliers (unseen classes)
  - Not aware of detection methods that identify the drift magnitude in images except for noise effects [2]

[1] *Dube, Parijat, and Eitan Farchi. "Automated detection of drift in deep learning based classifiers performance using network embeddings." In Engineering Dependable and Secure Machine Learning Systems: Third International Workshop, EDSMLS 2020, New York City, NY, USA, February 7, 2020.*

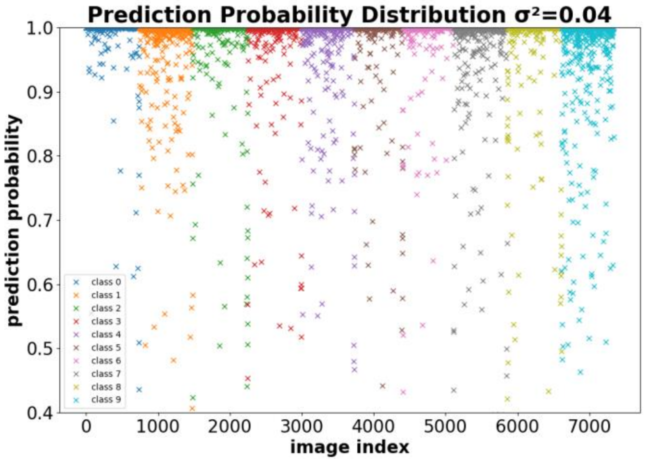
[2] *Kokil, Priyanka, and Turimerla Pratap. "Additive white gaussian noise level estimation for natural images using linear scale-space features." Circuits, Systems, and Signal Processing 40, no. 1 (2021): 353-374.*

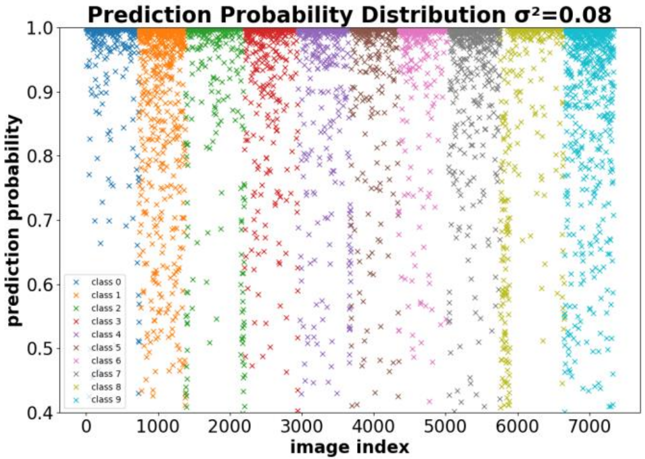
- Contribution
  - A novel statistical method to estimate the drift magnitude for any each possible drift type

© 2023 Southern Illinois University Carbondale

Motivation

- The distribution of the prediction probabilities is sensitive to the drift magnitude
- Prediction probabilities tend to decrease as the drift magnitude increases
- <u>Example:</u>
  - MNIST digit classification with Gaussian noise for two different noise levels.
  - For clean images, most of the prediction probabilities for all classes are >0.99.





© 2023 Southern Illinois University Carbondale

Thresholding

- A “threshold probability” value is introduced for each class
- The percentage of predictions above the “threshold probability” is computed for each class
- This percentage varies for each class at different drift magnitudes
- Drift magnitude is determined based on the percentage of predictions above class thresholds

© 2023 Southern Illinois University Carbondale

Thresholding

- **Example**
  - MNIST digit classification with Gaussian noise, threshold set to 0.98 for all classes
  - Three noise levels $\sigma^2 = 0, \sigma^2 = 0.04, \sigma^2 = 0.08$

Class	Percentage above threshold		
	$\sigma^2 = 0$	$\sigma^2 = 0.04$	$\sigma^2 = 0.08$
0	99.32%	95.41%	85.99%
1	98.93%	87.03%	43.67%
2	97.49%	92.05%	81.14%
3	98.01%	92.11%	75.95%
4	98.49%	88.18%	75.65%
5	96.89%	93.13%	80.42%
6	98.18%	94.37%	82.69%
7	97.46%	87.06%	68.85%
8	96.98%	91.82%	76.46%
9	95.70%	76.75%	43.70%

© 2023 Southern Illinois University Carbondale

Proposed Methodology: Thresholds

- Thresholds for each class, drift type, and drift magnitude:
  - For each drift type: Obtain prediction probabilities and prediction labels for different drift magnitudes
  - $for\ each\ class:$

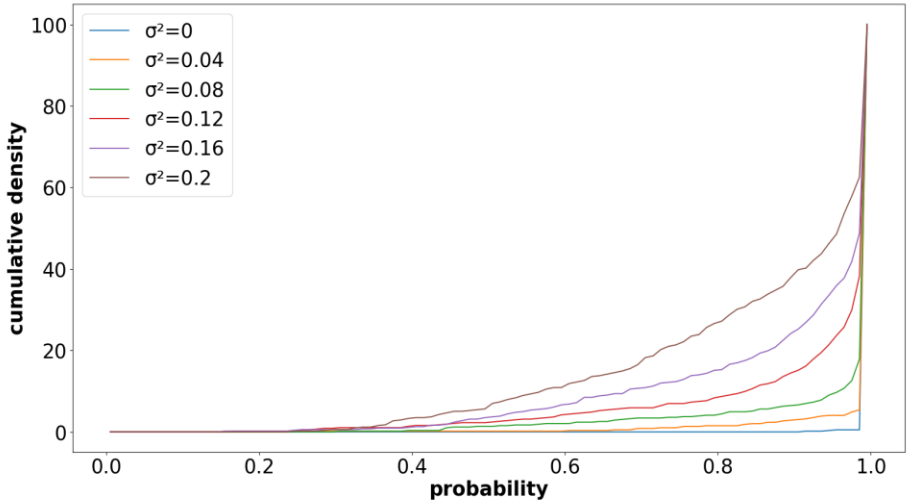
    $for\ each\ magnitude:$

    $compute\ cumulative\ distribution\ function\ (CDF)$

    $for\ each\ magnitude:$

$$threshold = \left\{ \begin{matrix} probability\ where\ difference \\ between\ \ CDFs\ \ of\ adjacent \\ magnitudes\ is\ maximum \end{matrix} \right\}$$

  - Compute the percentages above the threshold for each class and for each drift magnitude



© 2023 Southern Illinois University Carbondale

Proposed Methodology: Data Structures

- Example: MNIST with Gaussian noise
  - Data Structures: Threshold Dictionary and Expected Percentage Table

Magnitude \ Class	0	1	2	3	4	5	6	7	8	9
None	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
M1 ($\sigma^2 = 0.04$)	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
M2 ($\sigma^2 = 0.08$)	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
M3 ($\sigma^2 = 0.12$)	0.98	0.94	0.98	0.98	0.98	0.98	0.98	0.96	0.98	0.89
M4 ($\sigma^2 = 0.16$)	0.98	0.71	0.98	0.96	0.95	0.97	0.97	0.91	0.98	0.76
M5 ($\sigma^2 = 0.20$)	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98

Threshold Dictionary

Magnitude \ Class	0	1	2	3	4	5	6	7	8	9
None	99.3%	98.9%	97.5%	98.0%	98.5%	96.9%	98.2%	97.5%	97.0%	95.7%
M1 ($\sigma^2 = 0.04$)	95.4%	87.0%	92.1%	92.1%	88.2%	93.1%	94.4%	87.1%	91.8%	76.8%
M2 ($\sigma^2 = 0.08$)	86.0%	43.7%	81.1%	76.0%	75.7%	80.4%	82.7%	68.9%	76.5%	43.7%
M3 ($\sigma^2 = 0.12$)	71.1%	32.8%	66.1%	55.7%	51.7%	64.8%	63.9%	56.0%	59.7%	47.2%
M4 ($\sigma^2 = 0.16$)	56.1%	56.4%	50.0%	42.7%	47.7%	51.1%	51.2%	57.7%	45.6%	50.6%
M5 ($\sigma^2 = 0.20$)	43.2%	7.5%	39.1%	21.9%	26.3%	35.0%	31.4%	34.9%	36.6%	10.2%

Expected Percentage Table (percentages above thresholds)

© 2023 Southern Illinois University Carbondale

Proposed Methodology: Magnitude Estimation

- Magnitude Estimation:
  - Obtain prediction probabilities for a given set of images
  - Apply the threshold dictionary for the prediction probabilities
  - Obtain the percentages above the thresholds of each magnitude
  - Compute the difference between the observed percentages and the expected percentages for each magnitude
  - Estimated magnitude: Corresponds to the minimum difference between the expected and the observed percentages

  - Example: MNIST digit classification with $\sigma^2 = 0.04$ Gaussian noise

Magnitude \ Class	0	1	2	3	4	5	6	7	8	9
None	95.4%	87.0%	92.1%	92.1%	88.2%	93.1%	94.4%	87.1%	91.8%	76.8%
M1 ($\sigma^2 = 0.04$)	95.4%	87.0%	92.1%	92.1%	88.2%	93.1%	94.4%	87.1%	91.8%	76.8%
M2 ($\sigma^2 = 0.08$)	95.4%	87.0%	92.1%	92.1%	88.2%	93.1%	94.4%	87.1%	91.8%	76.8%
M3 ($\sigma^2 = 0.12$)	95.4%	91.1%	92.1%	92.1%	88.2%	93.1%	94.4%	90.0%	91.8%	89.7%
M4 ($\sigma^2 = 0.16$)	95.4%	98.5%	92.1%	94.3%	91.9%	93.9%	95.5%	93.5%	91.8%	94.5%
M5 ($\sigma^2 = 0.20$)	95.4%	87.0%	92.1%	92.1%	88.2%	93.1%	94.4%	87.1%	91.8%	76.8%

Observed percentages after applying the class thresholds of all magnitudes

Magnitude	Absolute sum of errors
None	5.4%
M1 ($\sigma^2 = 0.04$)	2.7%
M2 ($\sigma^2 = 0.08$)	20.9%
M3 ($\sigma^2 = 0.12$)	37.9%
M4 ($\sigma^2 = 0.16$)	45.6%
M5 ($\sigma^2 = 0.20$)	63.7%

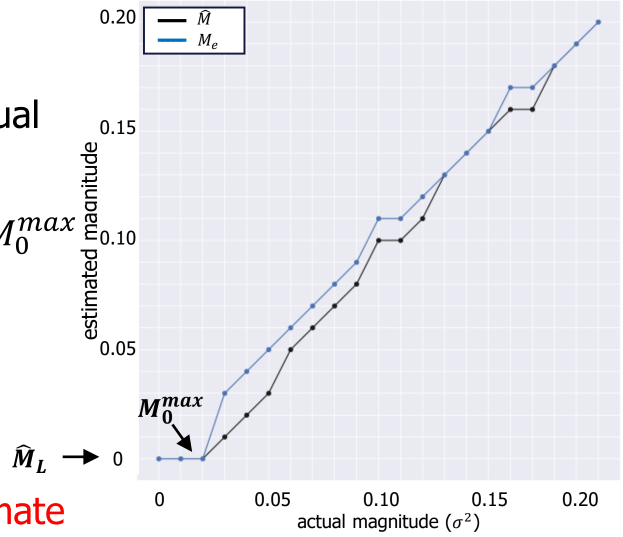
Estimated Magnitude

Error between the expected and the obtained percentage

© 2023 Southern Illinois University Carbondale

Alleviating Underestimation

- The estimated magnitude does not always exactly match the actual one
- Underestimation can be harmful and over-estimation can raise false alarms
- Adjust the initial estimation to alleviate underestimation
- For a given estimated magnitude $\widehat{M}$, let $M_{\widehat{M}}^{max}$ be the highest possible actual magnitude
- Let $\widehat{M}_L$ be the highest estimated magnitude among all magnitudes below M_0^{max}
- In the example, $M_0^{max} = 0.02$ and $\widehat{M}_L = 0$
- Adjust the initial estimation $\widehat{M}$ to final estimation M_e
  - For all $\widehat{M}$ such that $\widehat{M} \leq \widehat{M}_L$, $M_e = 0$ -> Region we underestimate
  - For all $\widehat{M}$ such that $\widehat{M} > \widehat{M}_L$, $M_e = M_{\widehat{M}}^{max}$ -> Region we may overestimate

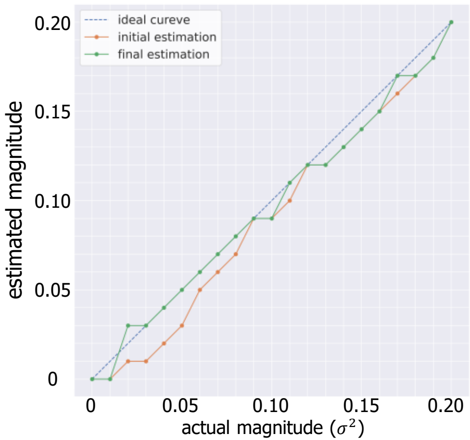


Mapping of $\widehat{M}$ to M_e for Gaussian noise in MNIST dataset

© 2023 Southern Illinois University Carbondale

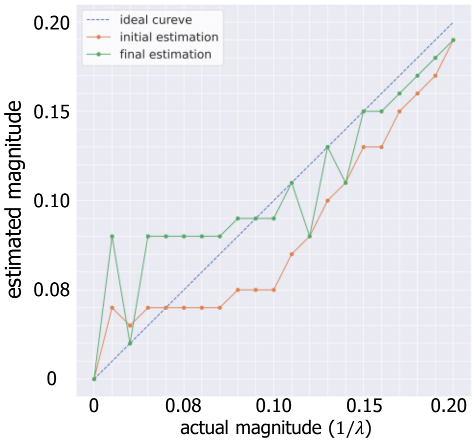
Experimental Results

- MNIST digit classification neural network
  - 7500 images used in the setup
  - 2500 images used in the testing
  - Gaussian, Poisson and Salt & Pepper are considered with 20 magnitudes per type



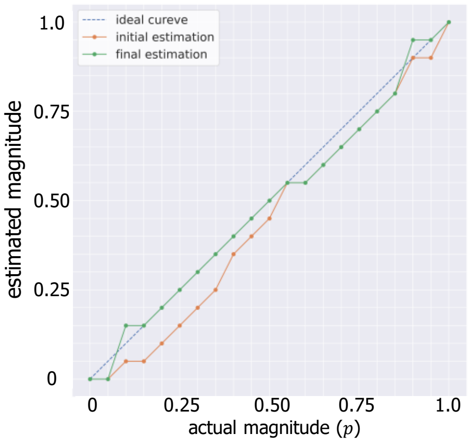
Gaussian Noise

$\sigma^2 = variance$



Poisson Noise

$\lambda = average\ number\ of\ events$



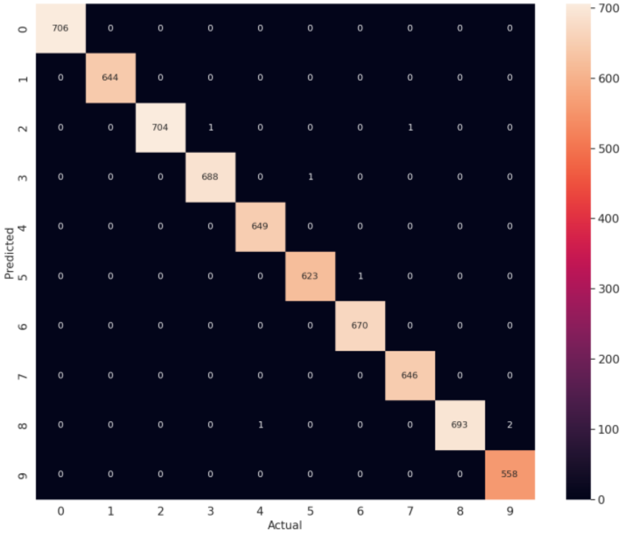
Salt & Pepper Noise

$p = percentage\ of\ impacted\ pixels$

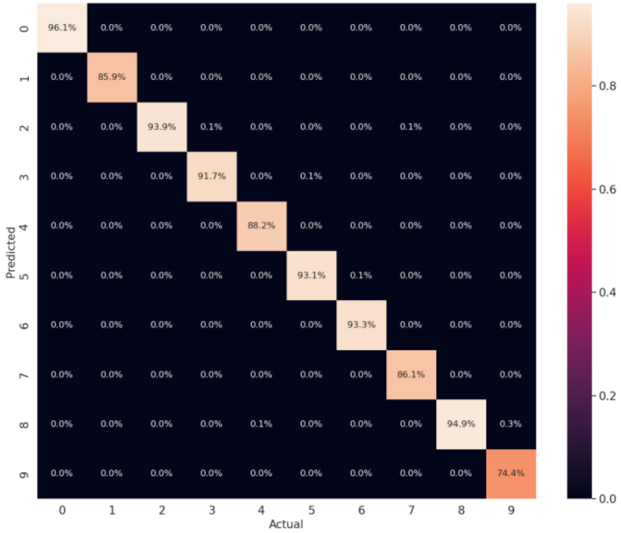
© 2023 Southern Illinois University Carbondale

Generalization: Uneven Distribution in Classes

- Expected percentage table: Sensitive to the percentage of images from each class
- The method is extended to cope with different distributions of images among the classes
- Uses the confusion matrix (and a normalization) after applying the thresholds for each type and magnitude



Confusion Matrix after thresholding for MNIST digits with Gaussian noise of $\sigma^2 = 0.04$



Coefficient Matrix: The normalized Confusion matrix

Coefficient Matrix $A_{n\times n}$ for n is classes

$$A_{n\times n} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix}$$

$a_{i,j}$ = percentage of images from class j predicted as class i above the threshold for a given magnitude of the drift type

© 2023 Southern Illinois University Carbondale

Generalization

- Obtain a system of linear equations

$$y_1 = a_{1,1}.x_1 + a_{1,2}.x_2 + \ \dots + a_{1,n}.x_n$$
$$y_2 = a_{2,1}.x_1 + a_{2,2}.x_2 + \ \dots + a_{2,n}.x_n$$
$$\vdots$$
$$y_n = a_{n,1}.x_1 + a_{n,2}.x_2 + \ \dots + a_{n,n}.x_n$$

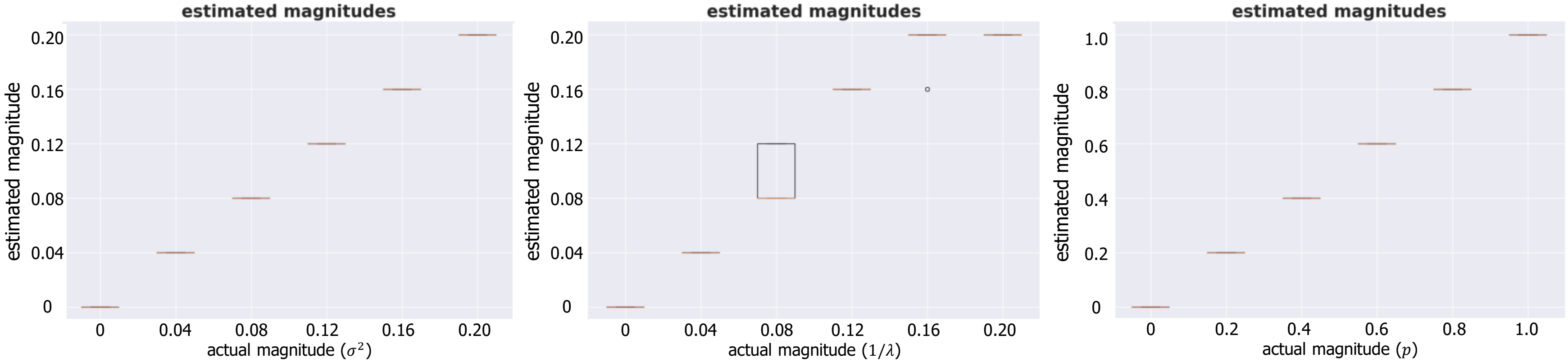
x_i = number of images from class i in the dataset
y_i = number of images predicted as class i above the respective threshold

- $X_{n\times 1}$ = number of images from each class
- $Y_{n\times 1}$ = number of predicted images for each class
- T = the total number of images
- Find a unique solution for $X_{n\times 1}$ by solving $Y_{n\times 1} = A_{n\times n} \times X_{n\times 1}$
- For each drift magnitude, compute estimation error $E = T - (x_1 + x_2 + \ \dots + x_n)$
- Estimated magnitude: The drift magnitude with the minimum estimation error

© 2023 Southern Illinois University Carbondale

Experimental Results: MNIST with Uneven Class Distributions

- 7500 images in the setup
- Three noise types are considered with 5 drift magnitudes in each
- 5 different distributions of the input data are considered for each magnitude

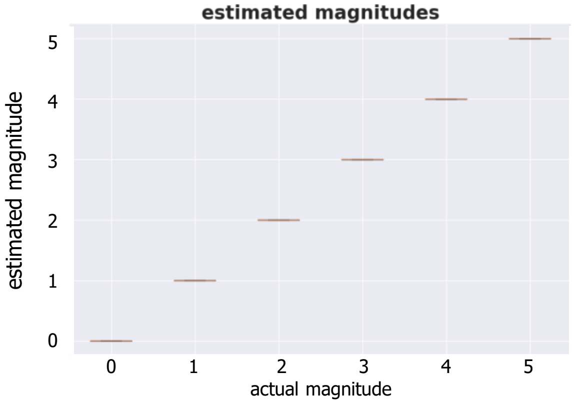


Gaussian Noise **Poisson Noise** **Salt & Pepper Noise**

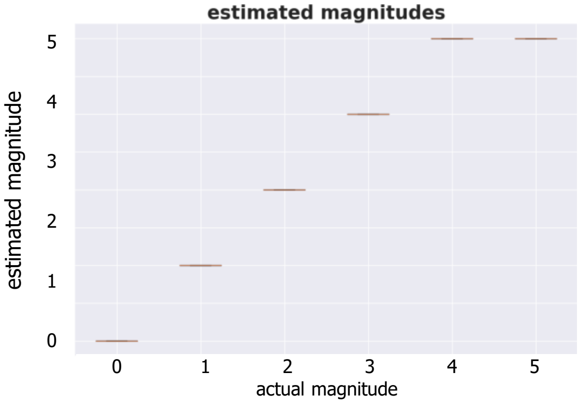
© 2023 Southern Illinois University Carbondale

Experimental Results: CIFAR10 with Uneven Class Distributions

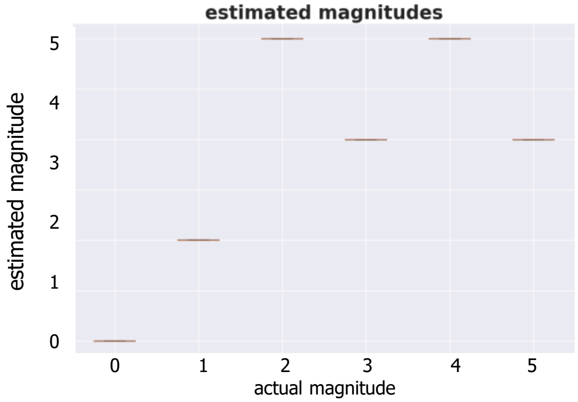
- CIFAR10 classification using ResNet18
  - 7500 images used in the setup phase
  - Three weather effects are considered with 5 drift levels in each
  - 5 different distributions of the input data are considered for each drift level



Snow



Fog



Shadow

© 2023 Southern Illinois University Carbondale

Conclusions

- Proposed method can be used to detect and estimate the magnitude of data drifts in image classification neural networks due to various effects on images
- Experimental results show that the method has a 100% detection rate
- Drift magnitude can be estimated with high accuracy
- The method is applicable in classification applications where different types of data drifts may occur
- This methodology can be used to detect any type of data drifts

© 2023 Southern Illinois University Carbondale

Resources

A preliminary version of this work will appear in:

- The proceedings of the IEEE International Conference on High-Performance Switching and Routing (HPSR 2023),
- The International Workshop on Resource-Constraint Machine Learning (RCML 2023) that is co-hosted with HPSR 2023.

© 2023 Southern Illinois University Carbondale