DeeperHyperion

CS454 Project Final Presentation

Team 6

20180459 Subeom
20224734 Arogya
20228163 Xiangchi
20180650 Hyunjoon

“DeepHyperion: Exploring the Feature Space of Deep Learning-Based Systems through Illumination Search”

Problem

- Dependability of Deep Learning (DL) Systems is more crucial than ever.
  - DL systems are now being used in many safety-critical domains.

- How do we ensure DL systems can be trusted with diverse real-world inputs?
  - Traditional code coverage metrics are not effective.
  - White box approaches are not sufficient to understand misbehaving input features.

- What if we could see a detailed view of the system's behavior with diverse inputs?
  - What about a feature map to interpret system behavior based on input characteristics?

DeepHyperion

- An automated test input generator for DL systems.
  - Generate diverse set of high-performing test inputs.

- “Illuminates” the input space by returning the highest-performing solution.
  - User can define the search space by features of interest.

- Provide a feature map where inputs are positioned based on their characteristics.
  - User can understand which inputs expose which misbehaviours.

DeepHyperion

- An automated test
    - Generate divers
- “Illuminates” the in ming solution.
    - User can define
- Provide a feature r n their characteristics.
    - User can under rs.

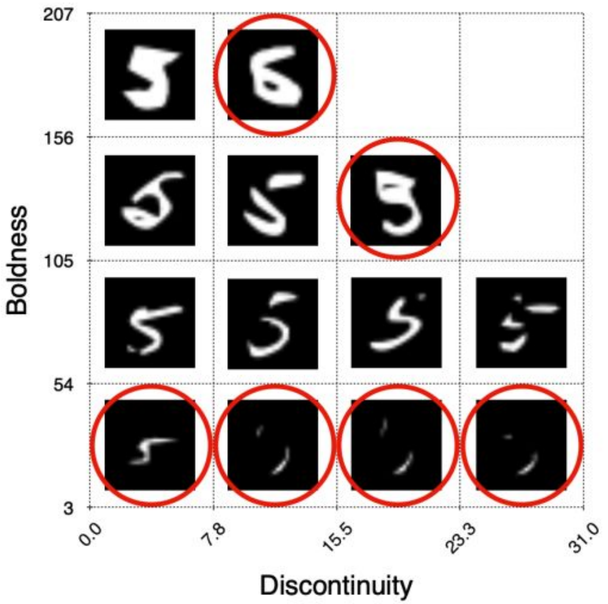
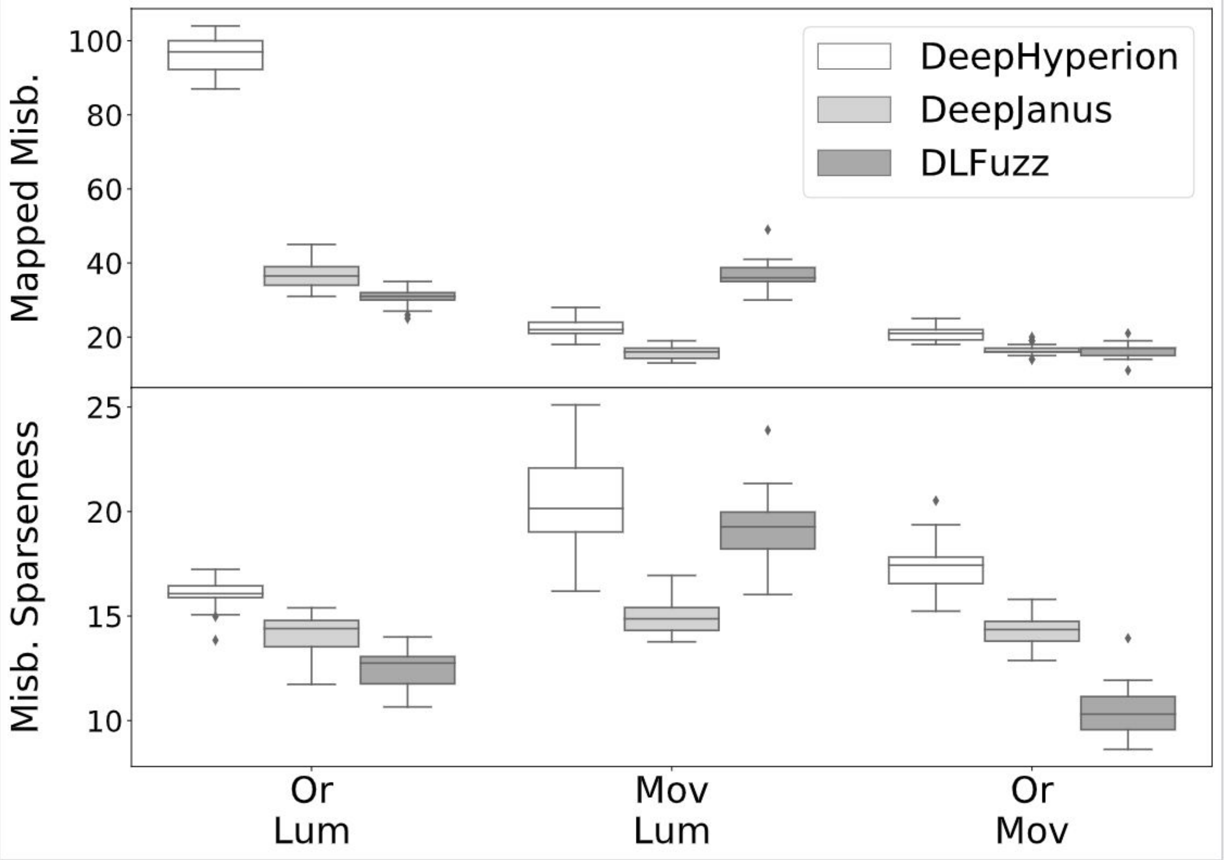


Figure 1: Feature map produced by DeepHyperion for a handwritten digit classifier. The two axes quantify two features: *discontinuity* and *boldness*. Cells show inputs that are either misclassified (marked with a circle) or close to being misclassified.

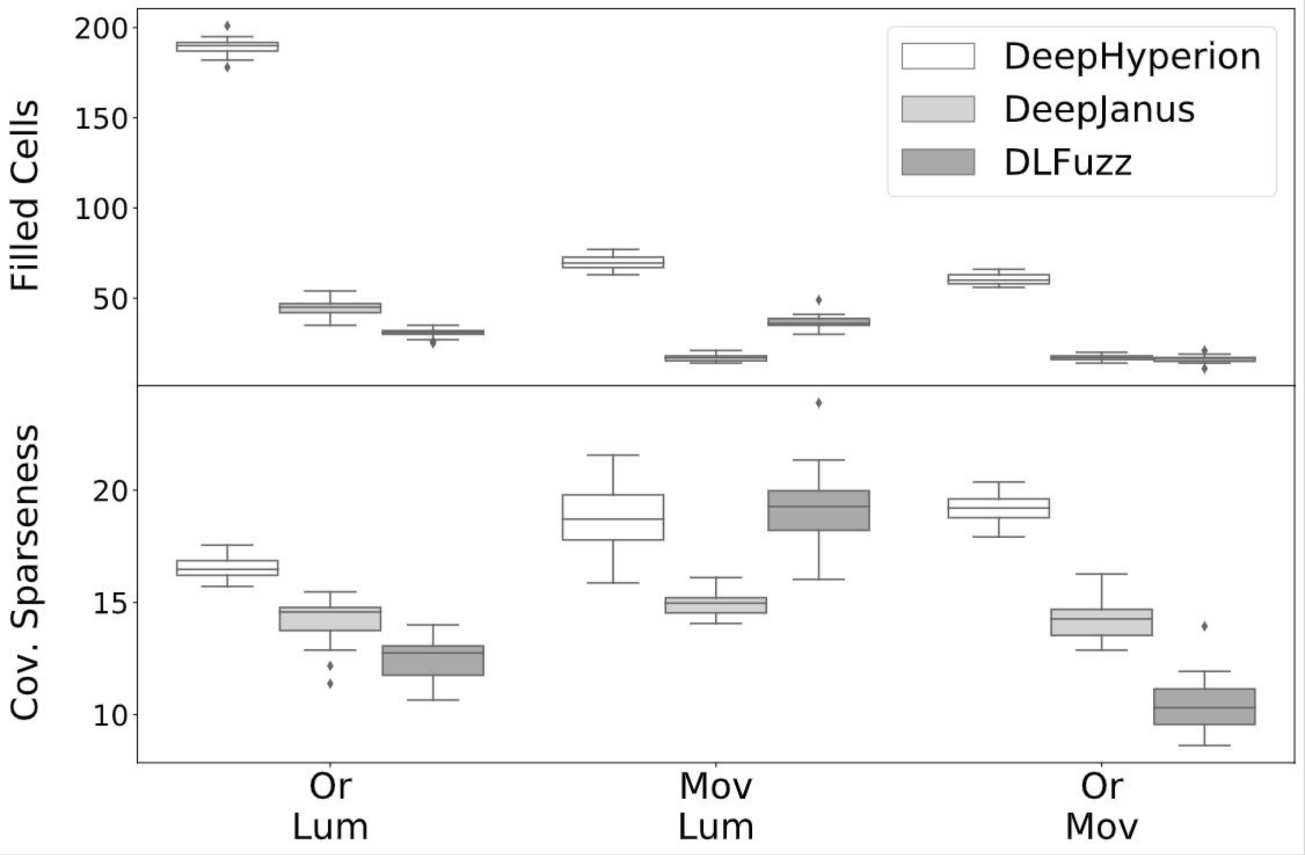
DeepHyperion: Replication

- Misbehaviours found by DeepHyperion, DeepJanus and DLFuzz on MNIST



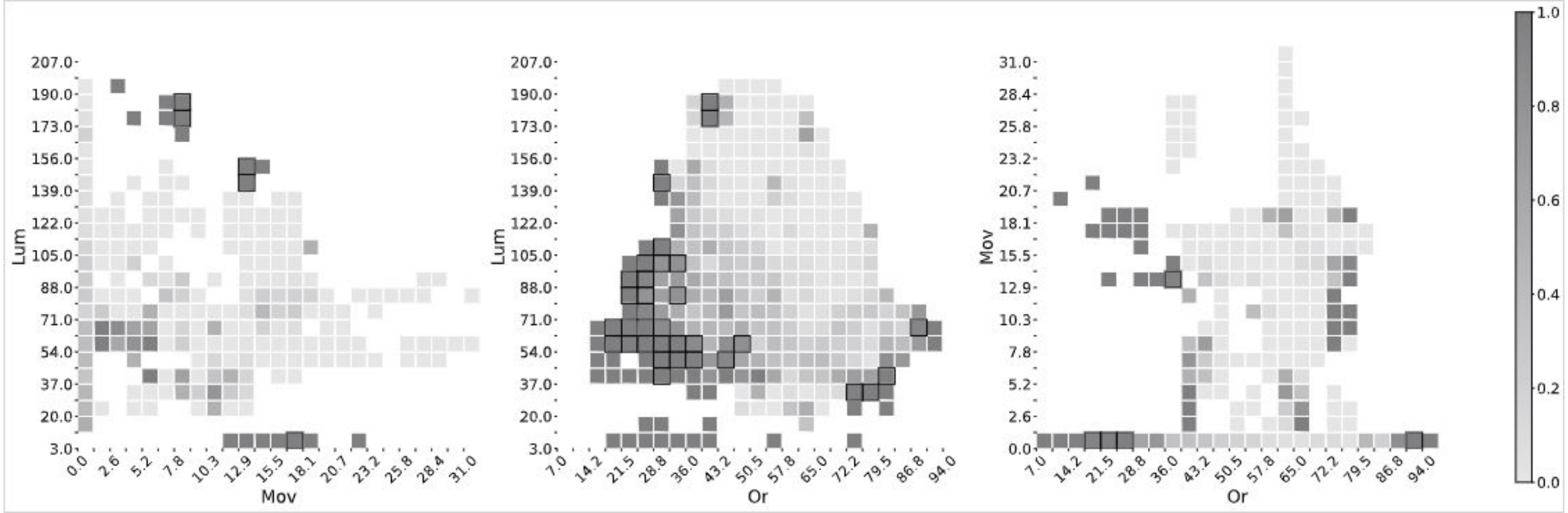
DeepHyperion: Replication

- Map cells filled by DeepHyperion, DeepJanus and DLFuzz on MNIST



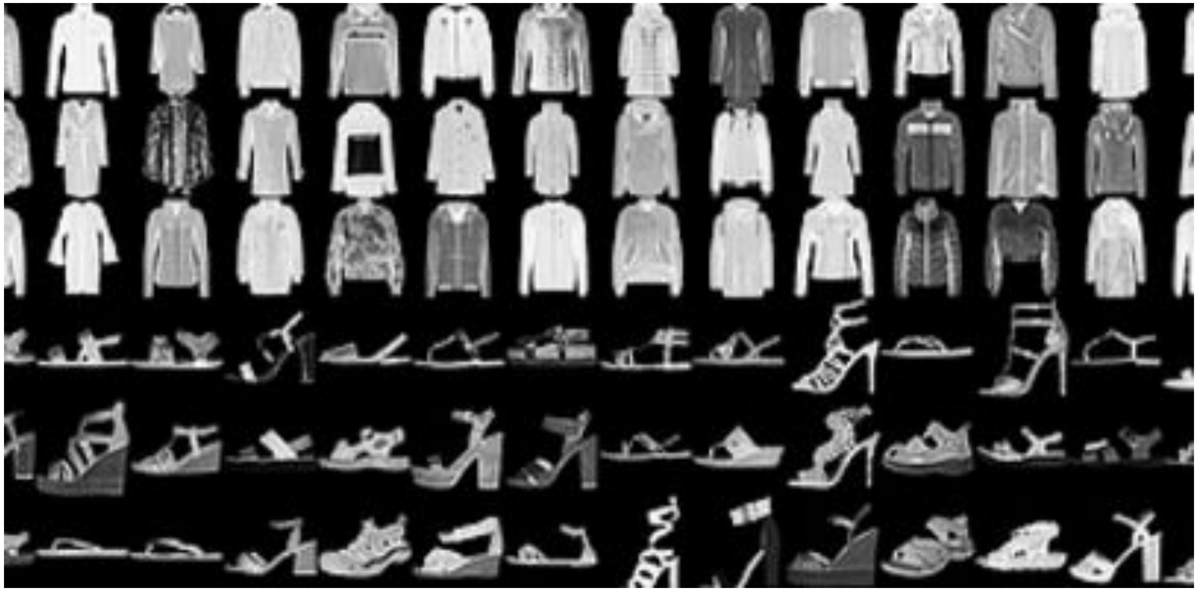
DeepHyperion: Replication

- Probability maps and feature discrimination for MNIST



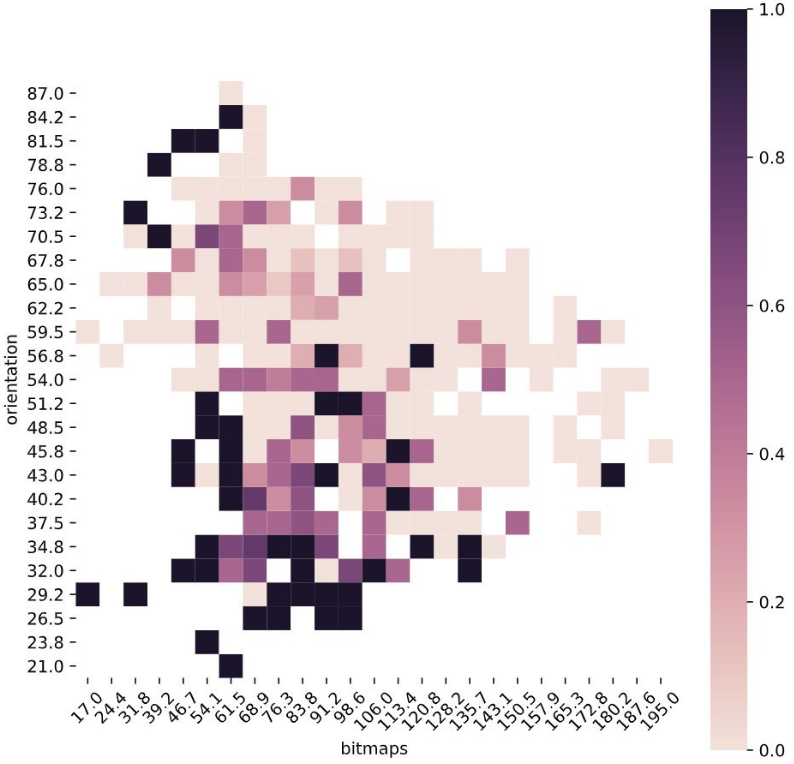
DeeperHyperion: Dataset Expansion

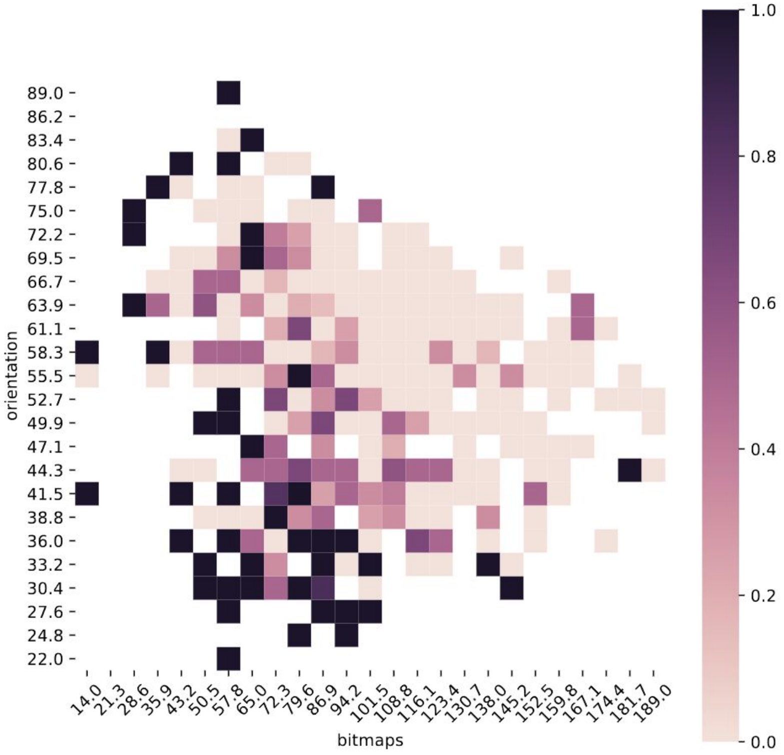
- F-MNIST
  - a dataset consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example being a 28x28 grayscale image associated with one of 10 classes.



DeeperHyperion: Dataset Expansion

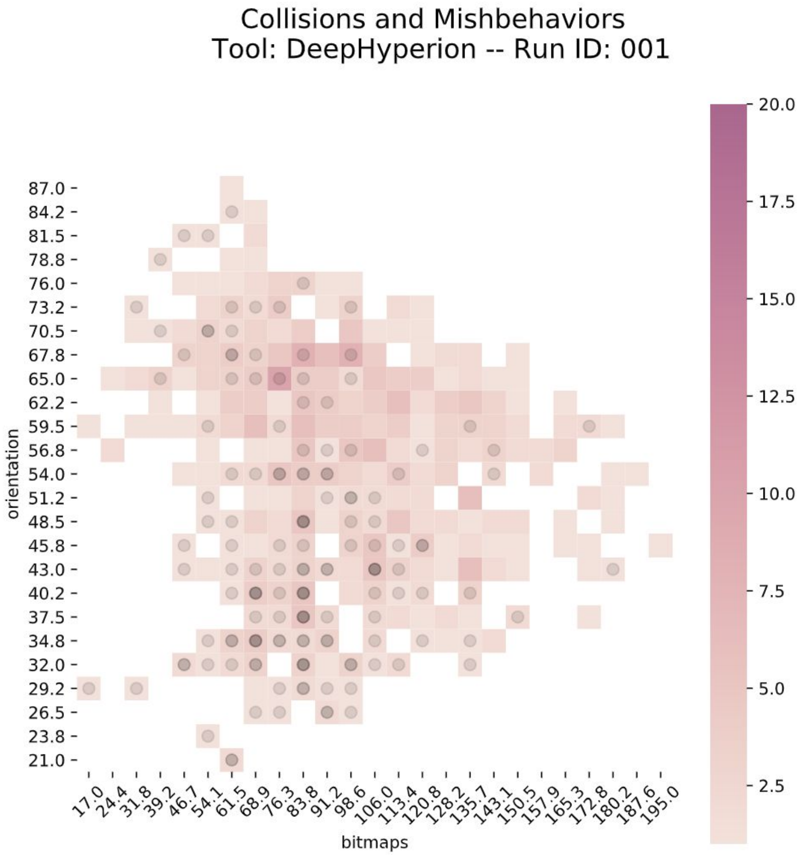
- MNIST[Left] vs F-MNIST[Right]
  - Misbehaviours Analysis

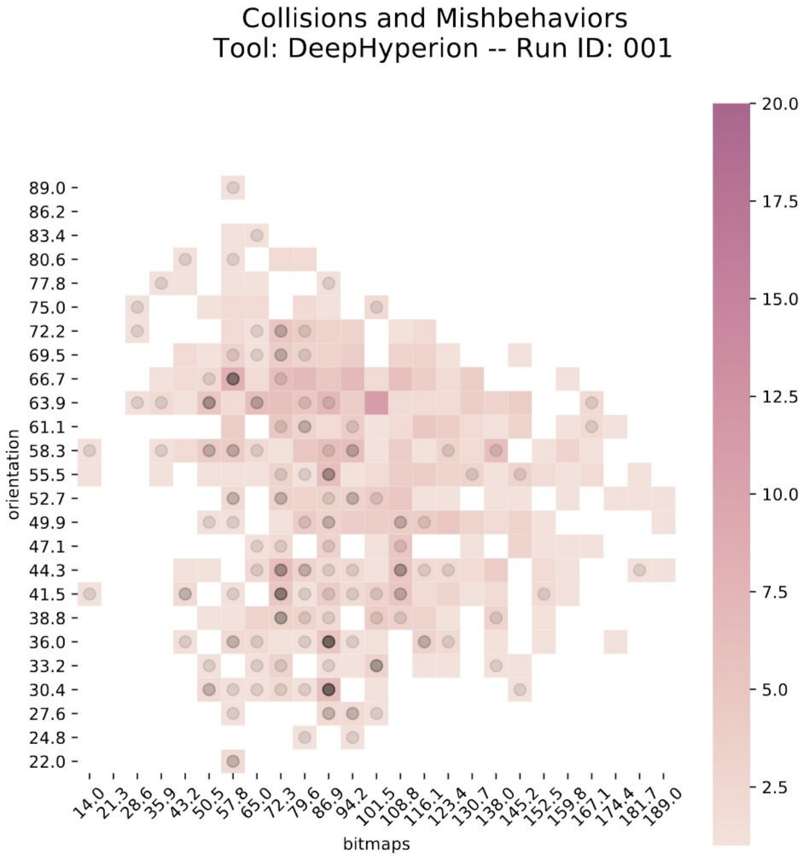

Mishbehavior Probability
Tool: DeepHyperion -- Run ID: 001


Mishbehavior Probability
Tool: DeepHyperion -- Run ID: 001

DeeperHyperion: Dataset Expansion

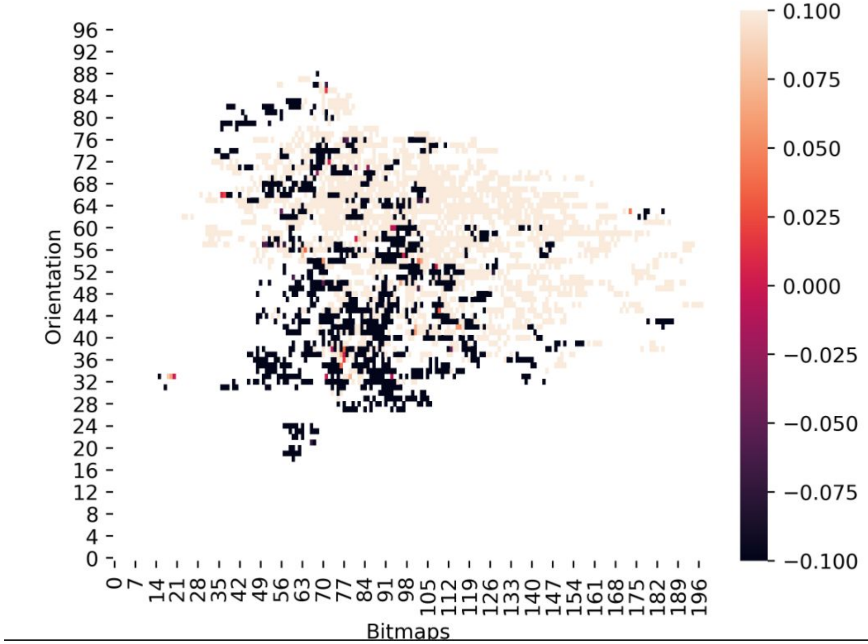
- MNIST[Left] vs F-MNIST[Right]
  - Collisions and Misbehaviors Analysis

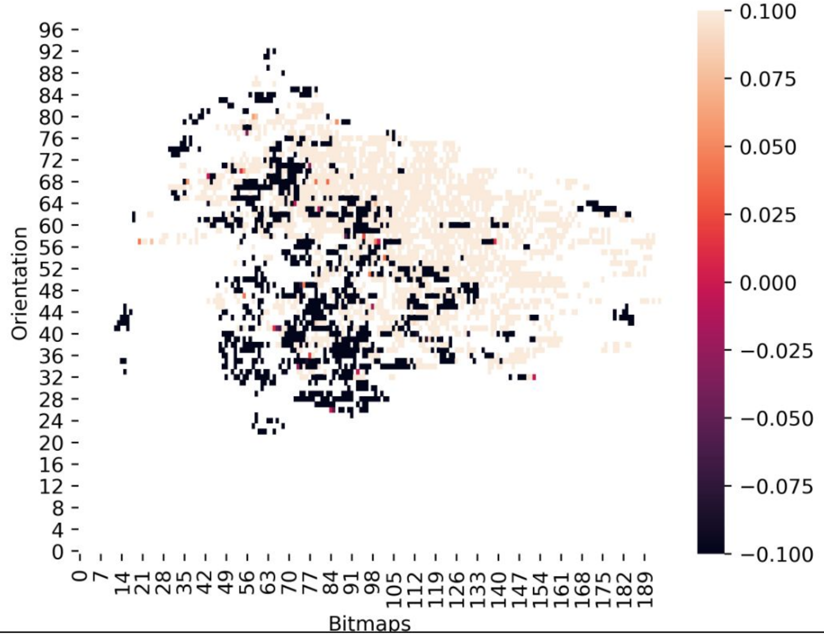




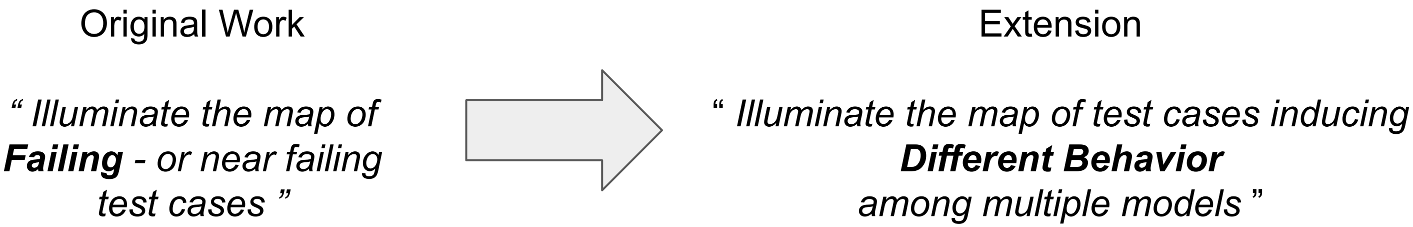
DeeperHyperion: Dataset Expansion

- MNIST[Left] vs F-MNIST[Right]
  - Heatmap Bitmap Orientation Analysis





DeeperHyperion: Different Behavior



Original Work

“ Illuminate the map of ***Failing*** *- or near failing test cases ”*

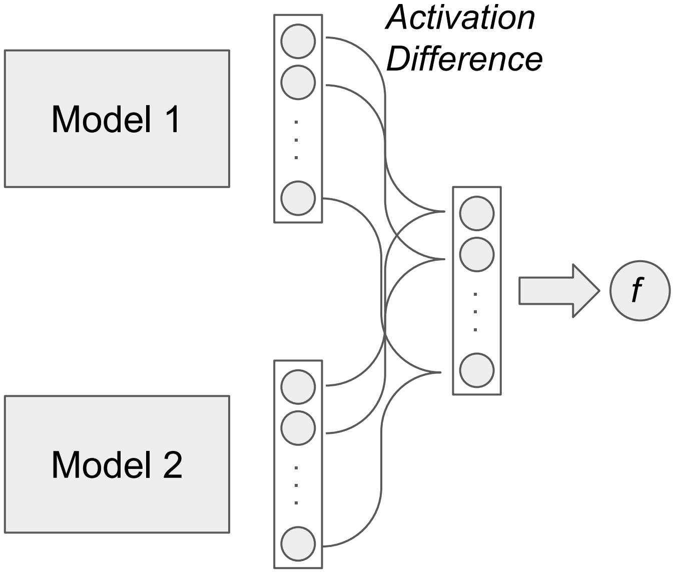
Extension

“ Illuminate the map of test cases inducing ***Different Behavior*** *among multiple models ”*

- *Baselines*
  - *a model from the original paper*
  - *a simple CNN model for comparison*

DeeperHyperion: Different Behavior

- Fitness



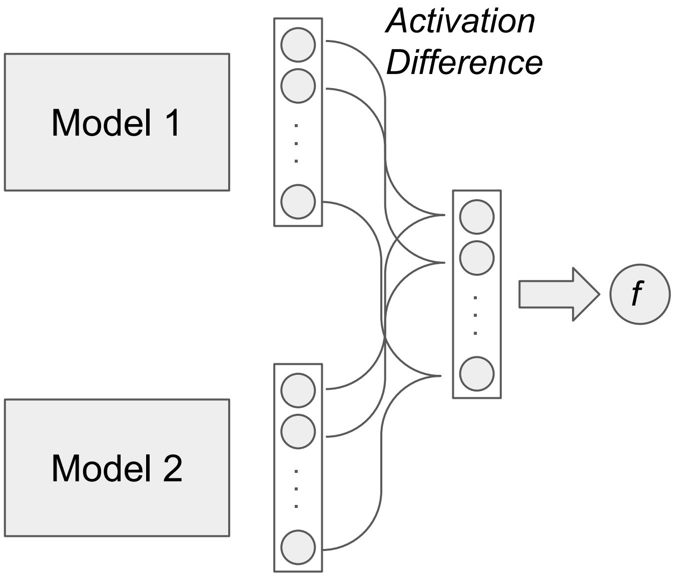
< Fitness under same prediction >

f = -0.1

< Fitness under different prediction >

DeeperHyperion: Different Behavior

- Fitness



< Fitness under same prediction >

f = -0.1

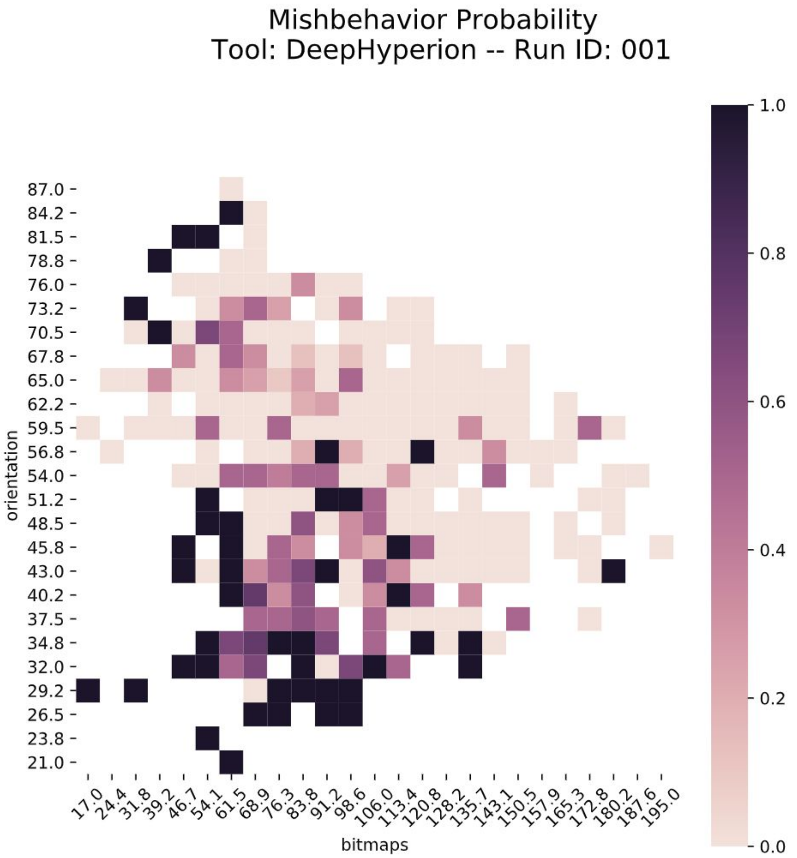
< Fitness under different prediction >

$$\boldsymbol{f} = \begin{cases} \textit{sum}(\text{activation difference}), & \\ \quad \textit{if model 1 \& 2 predict the same class} & \\ -\,0.1, \quad \textit{otherwise} & \end{cases}$$

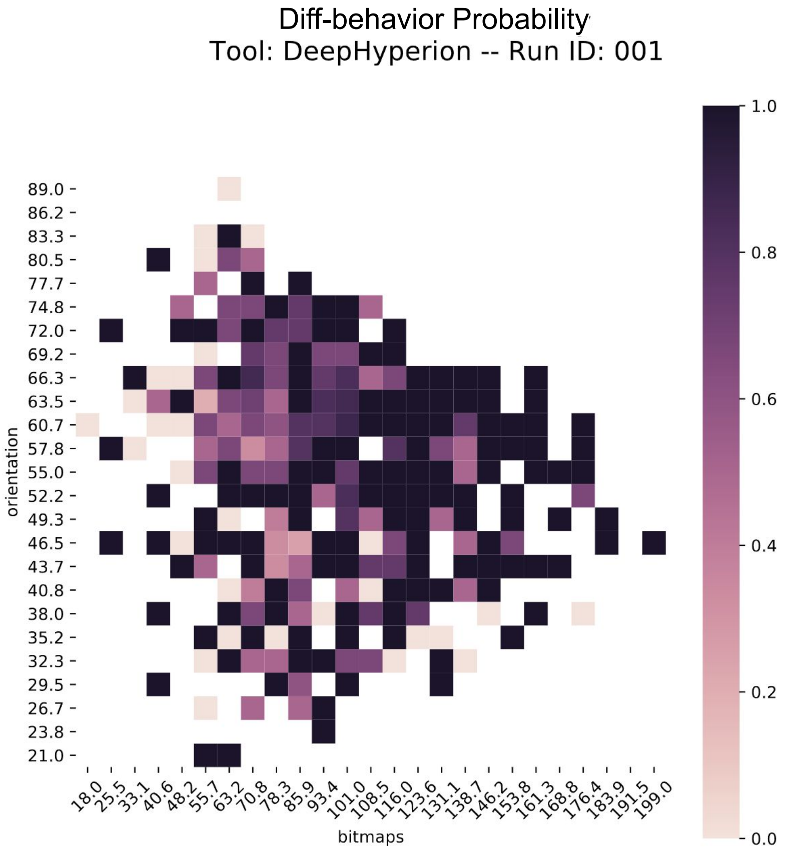
Objective: <u>minimize ***f***</u>

DeeperHyperion: Different Behavior

- Misbehavior[Left] vs Different Behavior[Right]
    - ○ Misbehavior / Diff-behavior Analysis



Mishbehavior Probability
Tool: DeepHyperion -- Run ID: 001

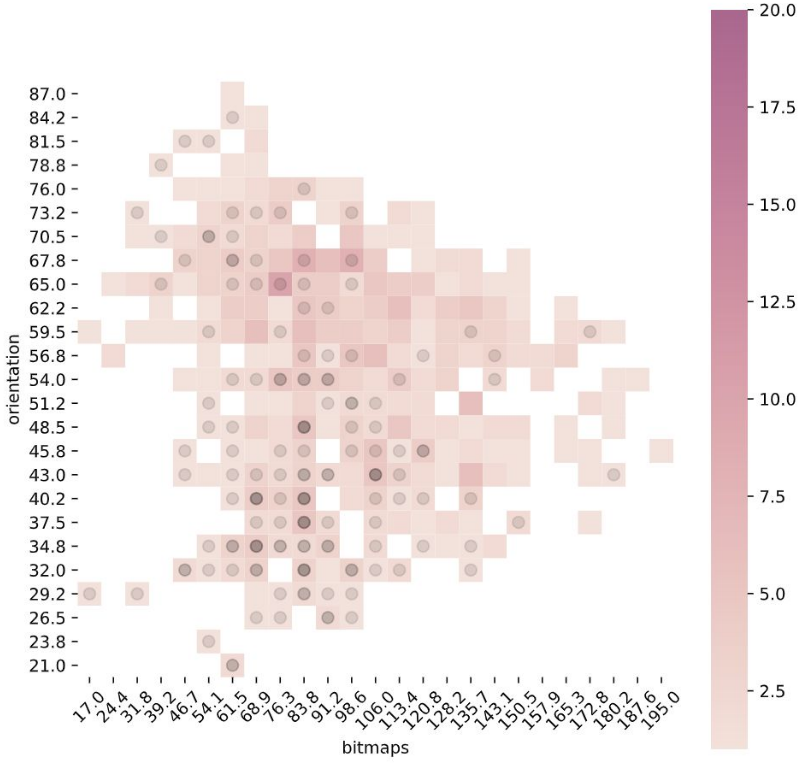


Diff-behavior Probability
Tool: DeepHyperion -- Run ID: 001

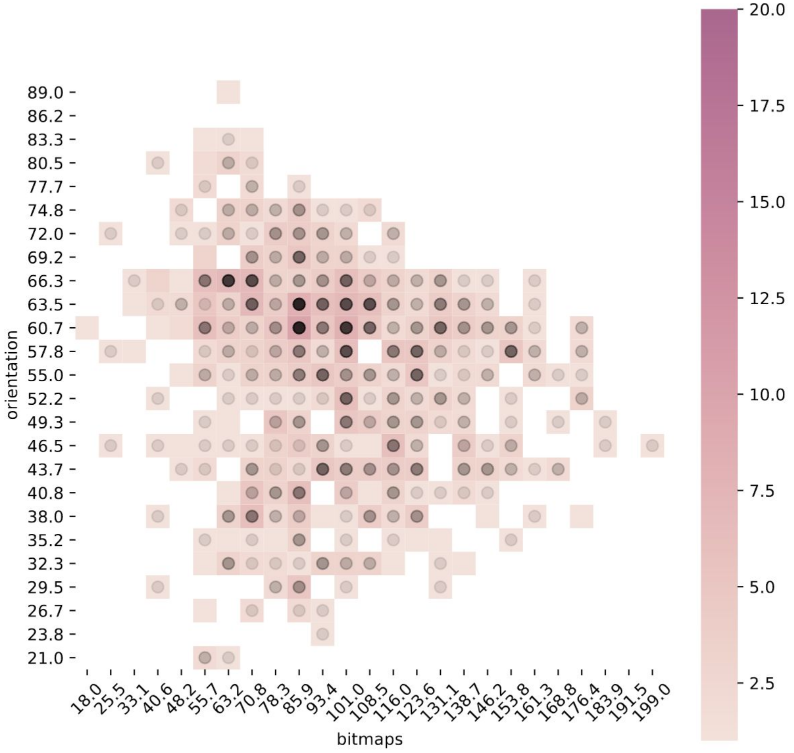
DeeperHyperion: Different Behavior

- Misbehavior[Left] vs Different Behavior[Right]
    - ○ Collisions and Misbehaviors / Diff-behaviors Analysis

Collisions and Mishbehaviors
Tool: DeepHyperion -- Run ID: 001

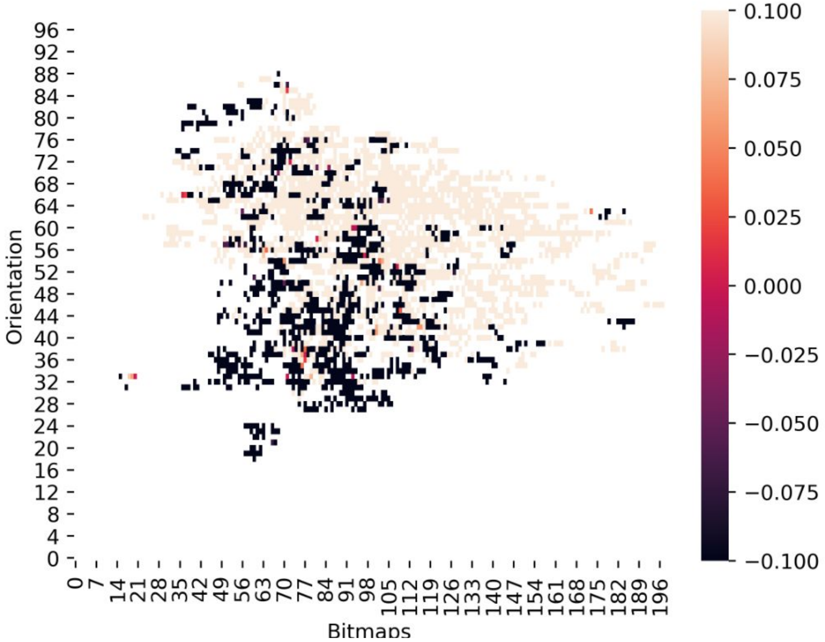


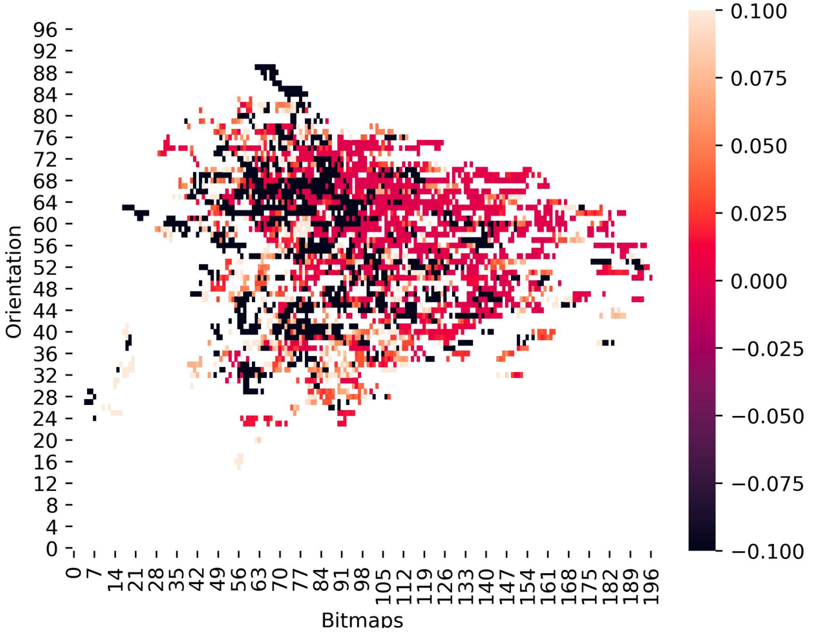
Collisions and Diff-behaviors
Tool: DeepHyperion -- Run ID: 001



DeeperHyperion: Different Behavior

- Misbehavior[Left] vs Different Behavior[Right]
  - ○ Heatmap Bitmap Orientation Analysis





Conclusion

- We replicated the DeepHyperion-MNIST experiments.

- We expand upon DeepHyperion-MNIST to FMNIST and presented our results.
  - FMNIST contains more complex and higher detailed examples.

- We expand DeepHyperion framework to consider differential behavior.
  - Useful for pinpointing precise weaknesses in the subject model.

Future Work

- Modifying fitness function
    - Current function is counter-intuitive, different function may provide insights.
- Investigate and compare with quantized models
    - Deephyperion may provide insights on where a quantized model breaks!
- Newer deep learning frameworks
    - Current implementation is based on TF 1.3, newer versions might improve efficiency.

Thank You

DeeperHyperion

CS454 Project Final Presentation

Team 6

“DeepHyperion: Exploring the Feature Space of Deep Learning-Based Systems through Illumination Search”

Reference

[1] Zohdinasab, Tahereh, et al. "Deephyperion: exploring the feature space of deep learning-based systems through illumination search." Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis. 2021.

[2] Mouret, Jean-Baptiste, and Jeff Clune. "Illuminating search spaces by mapping elites." arXiv preprint arXiv:1504.04909 (2015).

[3] Xiao H, Rasul K, Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms[J]. arXiv preprint arXiv:1708.07747, 2017.