Contributions by Feature Layers in Two-Class Deep Learning Image Classification Decisions

Debanjali Banerjee[1] and Henry Chu[2]

School of Computing and Informatics

University of Louisiana at Lafayette, U.S.A.

{[1]debanjali.banerjee1, [2]chu}@louisiana.edu

The Fourteenth International Conference on Pervasive Patterns and Applications
PATTERNS 2022

Debanjali Banerjee

Education

- Ph.D. candidate, Computer Science, University of Louisiana at Lafayette
- M.S., Computer Science, University of Louisiana at Lafayette
- B. Tech., Computer Science and Engineering, West Bengal University of Technology, India

Professional Experience

- Graduate Assistant, University of Louisiana at Lafayette
- Teaching Assistant Professor, Techno International Batanagar, India
- Research and Development Associate, Institute of Cybernetic Systems and Information Technology, India

Technical Interests

- Machine Learning, Deep Learning, Explainable AI
- Data Science

UNIVERSITY of LOUISIANA LAFAYETTE

Henry Chu

Professional Experience

- Professor, School of Computing and Informatics, University of Louisiana at Lafayette
- Executive Director, Informatics Research Institute, University of Louisiana at Lafayette

Education

- Ph.D., Electrical and Computer Engineering, Purdue University
- M.S.E., Computer, Information, and Control Engineering, University of Michigan
- B.S.E., Computer Engineering, University of Michigan

Technical Interests

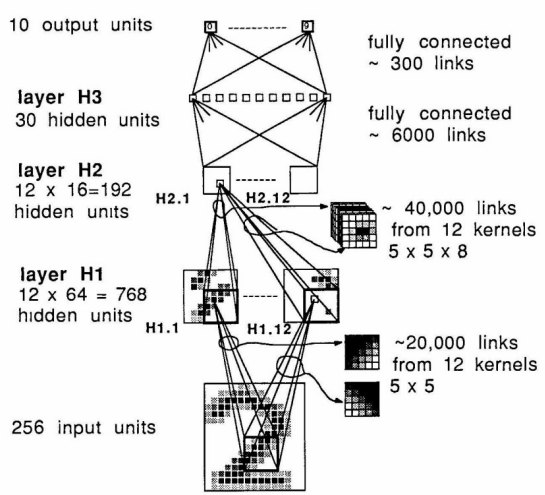
- Machine Vision
- Machine Learning
- Applied AI

UNIVERSITY of LOUISIANA LAFAYETTE

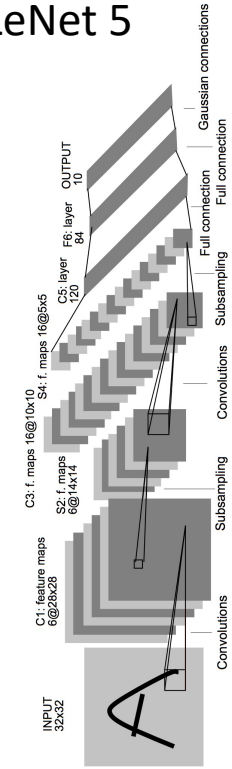
Deep Learning and Image Classification

1989 — 1998 — 2012 — 2015

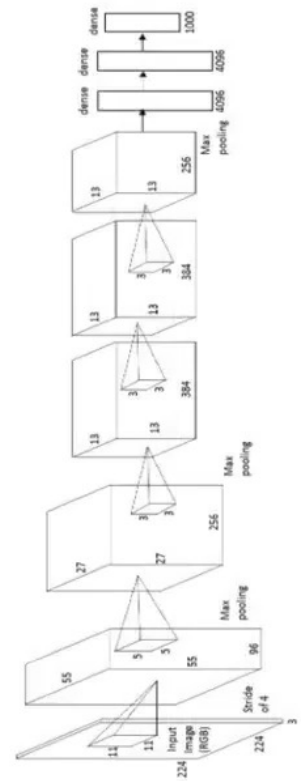
LeNet



LeNet 5



AlexNet



Convolutional neural networks outperform humans in the *ImageNet* image classification challenge

Image classifier forms the core of machine vision object detectors

UNIVERSITY of LOUISIANA LAFAYETTE

Deep Learning-based Image Classification

A deep learning-based two-class image classifier

Input is a color image

Feature extraction layers

Classification layers

Output is a class label

Layer Number	Type	Output Shape
1	InputLayer	[(None, 224, 224, 3)]
2	Conv2D	(None, 224, 224, 64)
3	Conv2D	(None, 224, 224, 64)
4	MaxPooling2D	(None, 112, 112, 64)
5	Conv2D	(None, 112, 112, 128)
6	Conv2D	(None, 112, 112, 128)
7	MaxPooling2D	(None, 56, 56, 128)
8	Conv2D	(None, 56, 56, 256)
9	Conv2D	(None, 56, 56, 256)
10	Conv2D	(None, 56, 56, 256)
11	Conv2D	(None, 56, 56, 256)
12	MaxPooling2D	(None, 28, 28, 256)
13	Conv2D	(None, 28, 28, 512)
14	Conv2D	(None, 28, 28, 512)
15	Conv2D	(None, 28, 28, 512)
16	Conv2D	(None, 28, 28, 512)
17	MaxPooling2D	(None, 14, 14, 512)
18	Conv2D	(None, 14, 14, 512)
19	Conv2D	(None, 14, 14, 512)
20	Conv2D	(None, 14, 14, 512)
21	Conv2D	(None, 14, 14, 512)
22	MaxPooling2D	(None, 7, 7, 512)
23	Flatten	(None, 25088)
24	Dense	(None, 64)
25	Dropout	(None, 64)
26	Dense	(None, 2)

Each layer is a block of feature maps
E.g., here it is a block of 512 maps, each of which is 28x28

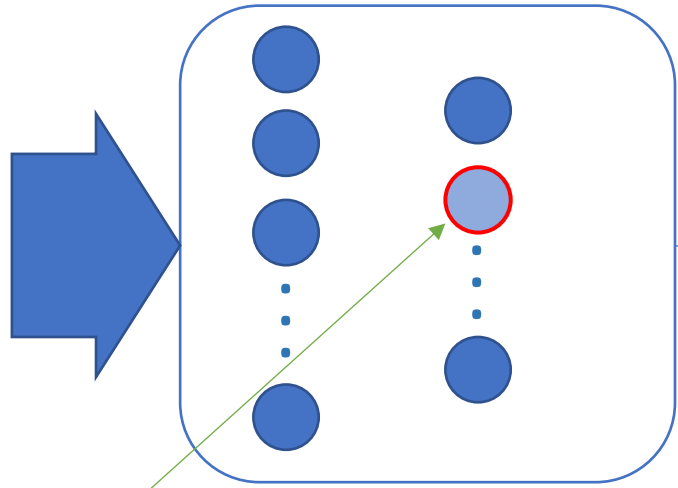
When a deep learning-based image classifier makes a decision, what is the basis for the decision?

One approach is to examine which region of an input image contributes most to a decision

Our approach in this work is to look at the contribution of each of the features in the feature extraction layers

UNIVERSITY of LOUISIANA LAFAYETTE

Shapley Values



Output

Computed as a function of the input variables through a network of intermediate features

The Shapley value is a reflection of the contribution by a feature by determining the change to the output when the feature is excluded from the output calculation

$$\sum_{S \subset F \setminus \{x_0\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} (f_{S \cup \{x_0\}} - f_S)$$

Set of all features excluding x_0, the feature under consideration

Weighting term so that all subsets of the same cardinality are weighted equally

Change in output by excluding x_0, the feature under consideration

UNIVERSITY of LOUISIANA LAFAYETTE

Experiments: 2-class classifier

Layer Number	Type	Output Shape
1	InputLayer	[(None, 224, 224, 3)]
2	Conv2D	(None, 224, 224, 64)
3	Conv2D	(None, 224, 224, 64)
4	MaxPooling2D	(None, 112, 112, 64)
5	Conv2D	(None, 112, 112, 128)
6	Conv2D	(None, 112, 112, 128)
7	MaxPooling2D	(None, 56, 56, 128)
8	Conv2D	(None, 56, 56, 256)
9	Conv2D	(None, 56, 56, 256)
10	Conv2D	(None, 56, 56, 256)
11	Conv2D	(None, 56, 56, 256)
12	MaxPooling2D	(None, 28, 28, 256)
13	Conv2D	(None, 28, 28, 512)
14	Conv2D	(None, 28, 28, 512)
15	Conv2D	(None, 28, 28, 512)
16	Conv2D	(None, 28, 28, 512)
17	MaxPooling2D	(None, 14, 14, 512)
18	Conv2D	(None, 14, 14, 512)
19	Conv2D	(None, 14, 14, 512)
20	Conv2D	(None, 14, 14, 512)
21	Conv2D	(None, 14, 14, 512)
22	MaxPooling2D	(None, 7, 7, 512)
23	Flatten	(None, 25088)
24	Dense	(None, 64)
25	Dropout	(None, 64)
26	Dense	(None, 2)

- VGG 19-based network with ReLU nonlinearity
- Trained to classify between cats and dogs
- Transfer learning
  - Layers 1-22 trained using the ImageNet data
  - Layers 23-26 trained using 2000 cats and 2000 dogs images
- Training accuracy about 95%
- After training:
  - Feed input image to network
  - Compute SHAP value for each feature

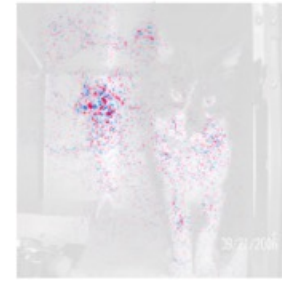
Visualize SHAP values of a feature map block by summing all maps in a block

Results

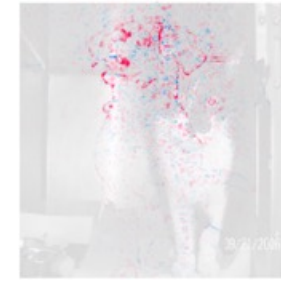
Layer Number	Type	Output Shape
1	InputLayer	[(None, 224, 224, 3)]
2	Conv2D	(None, 224, 224, 64)
3	Conv2D	(None, 224, 224, 64)
4	MaxPooling2D	(None, 112, 112, 64)
5	Conv2D	(None, 112, 112, 128)
6	Conv2D	(None, 112, 112, 128)
7	MaxPooling2D	(None, 56, 56, 128)
8	Conv2D	(None, 56, 56, 256)
9	Conv2D	(None, 56, 56, 256)
10	Conv2D	(None, 56, 56, 256)
11	Conv2D	(None, 56, 56, 256)
12	MaxPooling2D	(None, 28, 28, 256)
13	Conv2D	(None, 28, 28, 512)
14	Conv2D	(None, 28, 28, 512)
15	Conv2D	(None, 28, 28, 512)
16	Conv2D	(None, 28, 28, 512)
17	MaxPooling2D	(None, 14, 14, 512)
18	Conv2D	(None, 14, 14, 512)
19	Conv2D	(None, 14, 14, 512)
20	Conv2D	(None, 14, 14, 512)
21	Conv2D	(None, 14, 14, 512)
22	MaxPooling2D	(None, 7, 7, 512)
23	Flatten	(None, 25088)
24	Dense	(None, 64)
25	Dropout	(None, 64)
26	Dense	(None, 2)

Input

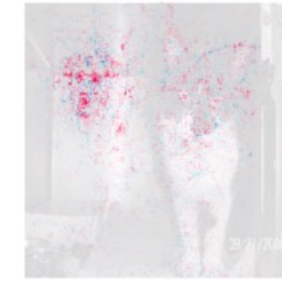
Layer 1



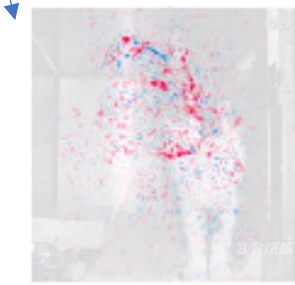
Layer 2



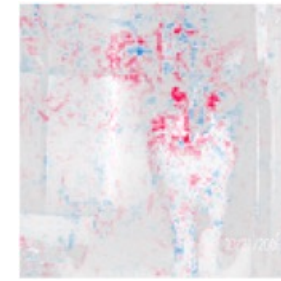
Layer 3



Layer 4



Layer 5



Layer 6

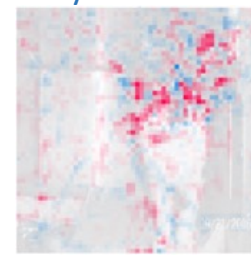
UNIVERSITY of LOUISIANA LAFAYETTE

Results

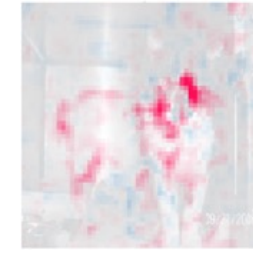
Layer Number	Type	Output Shape
1	InputLayer	[(None, 224, 224, 3)]
2	Conv2D	(None, 224, 224, 64)
3	Conv2D	(None, 224, 224, 64)
4	MaxPooling2D	(None, 112, 112, 64)
5	Conv2D	(None, 112, 112, 128)
6	Conv2D	(None, 112, 112, 128)
7	MaxPooling2D	(None, 56, 56, 128)
8	Conv2D	(None, 56, 56, 256)
9	Conv2D	(None, 56, 56, 256)
10	Conv2D	(None, 56, 56, 256)
11	Conv2D	(None, 56, 56, 256)
12	MaxPooling2D	(None, 28, 28, 256)
13	Conv2D	(None, 28, 28, 512)
14	Conv2D	(None, 28, 28, 512)
15	Conv2D	(None, 28, 28, 512)
16	Conv2D	(None, 28, 28, 512)
17	MaxPooling2D	(None, 14, 14, 512)
18	Conv2D	(None, 14, 14, 512)
19	Conv2D	(None, 14, 14, 512)
20	Conv2D	(None, 14, 14, 512)
21	Conv2D	(None, 14, 14, 512)
22	MaxPooling2D	(None, 7, 7, 512)
23	Flatten	(None, 25088)
24	Dense	(None, 64)
25	Dropout	(None, 64)
26	Dense	(None, 2)

Input

Layer 7



Layer 9



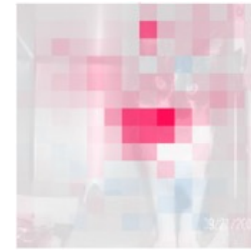
Layer 11

Layer 12

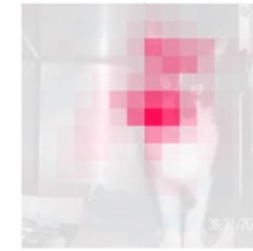
Layer 14

Layer 16

Layer 17



Layer 19



Layer 21

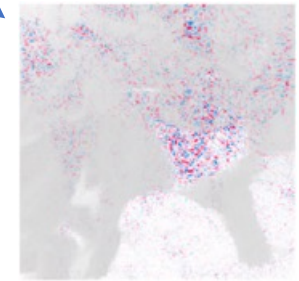
UNIVERSITY of LOUISIANA LAFAYETTE

Results

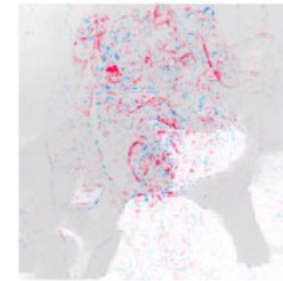
Layer Number	Type	Output Shape
1	InputLayer	[(None, 224, 224, 3)]
2	Conv2D	(None, 224, 224, 64)
3	Conv2D	(None, 224, 224, 64)
4	MaxPooling2D	(None, 112, 112, 64)
5	Conv2D	(None, 112, 112, 128)
6	Conv2D	(None, 112, 112, 128)
7	MaxPooling2D	(None, 56, 56, 128)
8	Conv2D	(None, 56, 56, 256)
9	Conv2D	(None, 56, 56, 256)
10	Conv2D	(None, 56, 56, 256)
11	Conv2D	(None, 56, 56, 256)
12	MaxPooling2D	(None, 28, 28, 256)
13	Conv2D	(None, 28, 28, 512)
14	Conv2D	(None, 28, 28, 512)
15	Conv2D	(None, 28, 28, 512)
16	Conv2D	(None, 28, 28, 512)
17	MaxPooling2D	(None, 14, 14, 512)
18	Conv2D	(None, 14, 14, 512)
19	Conv2D	(None, 14, 14, 512)
20	Conv2D	(None, 14, 14, 512)
21	Conv2D	(None, 14, 14, 512)
22	MaxPooling2D	(None, 7, 7, 512)
23	Flatten	(None, 25088)
24	Dense	(None, 64)
25	Dropout	(None, 64)
26	Dense	(None, 2)

Input

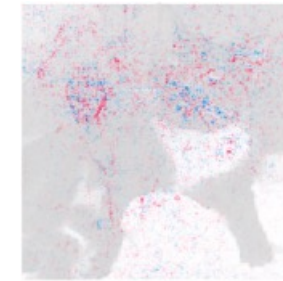
Layer 1



Layer 2

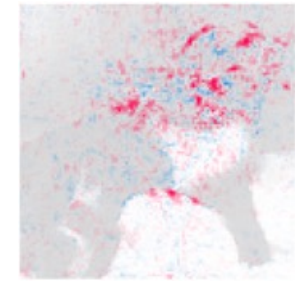


Layer 3

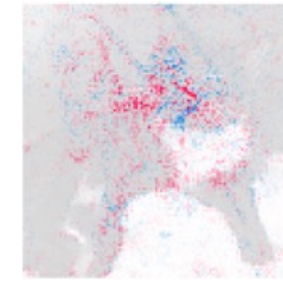


Layer 4

Layer 5



Layer 6



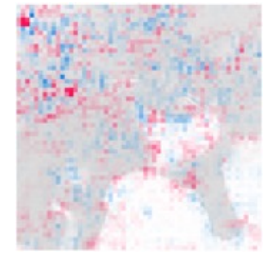
UNIVERSITY of LOUISIANA LAFAYETTE

Results

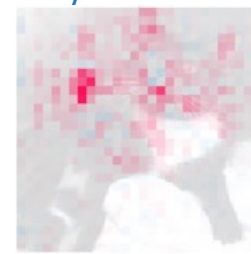
Layer Number	Type	Output Shape
1	InputLayer	[(None, 224, 224, 3)]
2	Conv2D	(None, 224, 224, 64)
3	Conv2D	(None, 224, 224, 64)
4	MaxPooling2D	(None, 112, 112, 64)
5	Conv2D	(None, 112, 112, 128)
6	Conv2D	(None, 112, 112, 128)
7	MaxPooling2D	(None, 56, 56, 128)
8	Conv2D	(None, 56, 56, 256)
9	Conv2D	(None, 56, 56, 256)
10	Conv2D	(None, 56, 56, 256)
11	Conv2D	(None, 56, 56, 256)
12	MaxPooling2D	(None, 28, 28, 256)
13	Conv2D	(None, 28, 28, 512)
14	Conv2D	(None, 28, 28, 512)
15	Conv2D	(None, 28, 28, 512)
16	Conv2D	(None, 28, 28, 512)
17	MaxPooling2D	(None, 14, 14, 512)
18	Conv2D	(None, 14, 14, 512)
19	Conv2D	(None, 14, 14, 512)
20	Conv2D	(None, 14, 14, 512)
21	Conv2D	(None, 14, 14, 512)
22	MaxPooling2D	(None, 7, 7, 512)
23	Flatten	(None, 25088)
24	Dense	(None, 64)
25	Dropout	(None, 64)
26	Dense	(None, 2)

Input

Layer 11



Layer 16



Layer 21

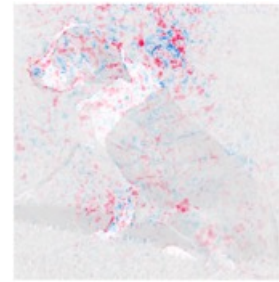
UNIVERSITY of LOUISIANA LAFAYETTE

Results

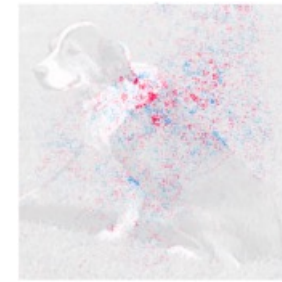
Layer Number	Type	Output Shape
1	InputLayer	[(None, 224, 224, 3)]
2	Conv2D	(None, 224, 224, 64)
3	Conv2D	(None, 224, 224, 64)
4	MaxPooling2D	(None, 112, 112, 64)
5	Conv2D	(None, 112, 112, 128)
6	Conv2D	(None, 112, 112, 128)
7	MaxPooling2D	(None, 56, 56, 128)
8	Conv2D	(None, 56, 56, 256)
9	Conv2D	(None, 56, 56, 256)
10	Conv2D	(None, 56, 56, 256)
11	Conv2D	(None, 56, 56, 256)
12	MaxPooling2D	(None, 28, 28, 256)
13	Conv2D	(None, 28, 28, 512)
14	Conv2D	(None, 28, 28, 512)
15	Conv2D	(None, 28, 28, 512)
16	Conv2D	(None, 28, 28, 512)
17	MaxPooling2D	(None, 14, 14, 512)
18	Conv2D	(None, 14, 14, 512)
19	Conv2D	(None, 14, 14, 512)
20	Conv2D	(None, 14, 14, 512)
21	Conv2D	(None, 14, 14, 512)
22	MaxPooling2D	(None, 7, 7, 512)
23	Flatten	(None, 25088)
24	Dense	(None, 64)
25	Dropout	(None, 64)
26	Dense	(None, 2)

Input

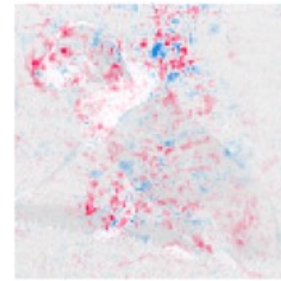
Layer 2



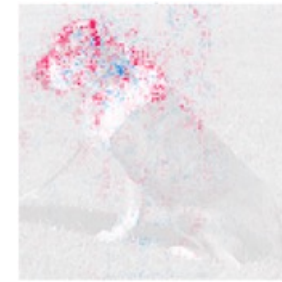
Layer 3



Layer 5



Layer 6



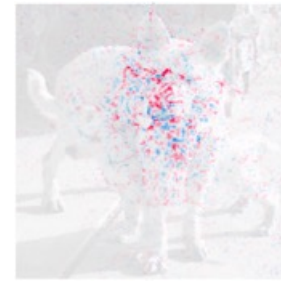
UNIVERSITY of LOUISIANA LAFAYETTE

Results

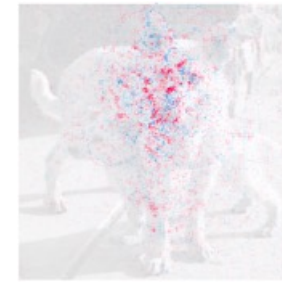
Layer Number	Type	Output Shape
1	InputLayer	[(None, 224, 224, 3)]
2	Conv2D	(None, 224, 224, 64)
3	Conv2D	(None, 224, 224, 64)
4	MaxPooling2D	(None, 112, 112, 64)
5	Conv2D	(None, 112, 112, 128)
6	Conv2D	(None, 112, 112, 128)
7	MaxPooling2D	(None, 56, 56, 128)
8	Conv2D	(None, 56, 56, 256)
9	Conv2D	(None, 56, 56, 256)
10	Conv2D	(None, 56, 56, 256)
11	Conv2D	(None, 56, 56, 256)
12	MaxPooling2D	(None, 28, 28, 256)
13	Conv2D	(None, 28, 28, 512)
14	Conv2D	(None, 28, 28, 512)
15	Conv2D	(None, 28, 28, 512)
16	Conv2D	(None, 28, 28, 512)
17	MaxPooling2D	(None, 14, 14, 512)
18	Conv2D	(None, 14, 14, 512)
19	Conv2D	(None, 14, 14, 512)
20	Conv2D	(None, 14, 14, 512)
21	Conv2D	(None, 14, 14, 512)
22	MaxPooling2D	(None, 7, 7, 512)
23	Flatten	(None, 25088)
24	Dense	(None, 64)
25	Dropout	(None, 64)
26	Dense	(None, 2)

Input

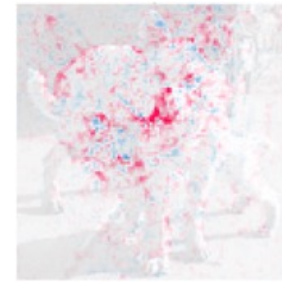
Layer 2



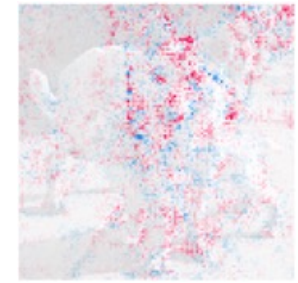
Layer 3



Layer 5



Layer 6



UNIVERSITY of LOUISIANA LAFAYETTE

Conclusion

- An empirical study of SHAP values of feature layers in a two-class deep learning-based image classifier
- The SHAP values are a special case of the Shapley value that explains the factors in a machine learning decision by measuring the output change due to change in each factor
- Results showed that the lower layers exhibited shapes that fit other faces, some of them not even from the same class. It appeared that the network worked on assembling the outlines of a shape much earlier in the layered architecture than expected, as early as Layer 2 which was immediately connected to the input layer

For more information

Debanjali Banerjee debanjali.banerjee1@louisiana.edu

Henry Chu chu@louisiana.edu

School of Computing and Informatics

University of Louisiana at Lafayette, U.S.A.

UNIVERSITY of LOUISIANA LAFAYETTE