Can we interpret the black box?

Eunsu Baek
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826
esbaek@hcil.snu.ac.kr

JiSoo Jang
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826
simonjisu@snu.ac.kr

Cho Gyung min
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826
km950501@snu.ac.kr

Abstract

With a lot of deep learning models being used in safety-critical scenarios, such as self-driving cars and financial industries, there is a growing sense that neural networks should be interpretable by humans. To shed the light on neural networks which are known as black-box, many studies have utilized feature visualizations that maximize patterns detected by neurons. However, we questioned the practicality of feature visualizations and conducted experiments to measure their quantitative/qualitative performance. In addition, we proposed a noble Interpretability metric, I-metric, to measure the interpretability of feature visualizations. As a result, we demonstrate our visualization support with feature visualization can help humans understand patterns detected by neurons based on I-metric

1. Introduction

Deep learning has recently seen rapid development and received significant attention due to its state-of-the-art performance on previously-thought hard problems [11]. However, because of the internal complexity and nonlinear structure of deep neural networks, the underlying decision-making processes for why these models are achieving such performance are challenging and sometimes are called black-box [11]. As a result, many threat cases and studies have been reported on AI vulnerabilities, not intended on model building step, such as fairness [19](‘Lee-Luda’ scandal [13]), data poisoning [16], and adversarial attacks [8, 17] (Tesla’s burger king issue [24]) and can be affect trust on AI technologies. Furthermore, the limitations of non-interpretable AI, which makes debugging issues difficult, are fatal to deployment in safety-critical scenarios such as autonomous vehicles, financial services, and healthcare.

Many works have strived to achieve the interpretability of deep learning. Gilpin, Leilani H., et al. [9] introduced fundamental concepts of explainability and its utility. F.Hohman., et al [11] surveyed visual analytics in deep learning. An approach to understanding how neural networks work internally is to study neurons’ activation patterns. To interpret what concept a neuron is detecting, feature visualization [3, 7, 20, 23] creates a visualization that maximizes such neurons. Circuits [21] and Summit [4] visually explain how higher-level concepts can be constructed by neural connections. Activation Atlas [3] visualizes neuron activations per layer and analyzes how models can be exploited when predicting on manipulated input.

In some studies, the limitations of feature visualizations(not interpretable and wide search space, etc.,) are reported [23]. Therefore, we were curious about the feasibility and application of feature visualization techniques. To reveal that, in this work, we designed and conducted experiments to evaluate quantitative/qualitative aspects of feature visualizations. In Particular, we proposed a noble interpretability metric, I-metric , to evaluate feature visualization for patterns detected by neurons. In this project, our contributions the contents below.

- User study of feature visualizations
- New prototype for visualization supports with requirements to provide explainability
- Experiment design and I-metric

2. Backgrounds and related works

2.1. Feature Visualization

Feature visualization answers questions about what a network (or parts of a network)are looking for by generating examples [23]. By optimizing the inputs to an optimization objective, we can generate an example of visualized features. Optimization objectives can be different by selecting different targets that we are interested in. Targets can be a single neuron, single channel of output, whole layer, class logits or class probabilities.

There is another method to achieve interpretability of explanation for neuron activation: attribution, which studies what part of an example is responsible for the network activating a particular way [23]. We can combine two methods like [22] to build more interpretable visualization to help human understanding

2.2. Interpretability Metrics

2.2.1 What are good metrics?

[1, 18, 15] introduced the requirements of good metrics

- **Comparative** : The probability of success is a better metric than the number of successes. Using a metric defined by ratio or rate is a good strategy.
- **Understandable** : It should be easy to understand when explaining metrics to third parties.
- **Behavior changing** : the metric we use should be tracking the target we want.

2.2.2 What is good interpretability?

There is no consensus or definition of what is good interpretability for now. In a rough sense, interpretability can be defined as the ability to explain or to provide the meaning in understandable terms to a human [7]. Then, the next question will be 'what is a good explanation to a human?'. [20, 23] shared the answers to the questions.

- **Contrastive** : The best explanation is the one that highlights the greatest difference between the object of interest and the reference object.
- **Selected** : Humans are used to selecting one or two causes from a variety of possible causes as the explanation.
- Social, focusing on the target scenario, Truthful,.. etc.

2.2.3 Interpretability metrics

Previous research on interpretability metrics has been conducted in many ways. These can be divided into two main categories: model interpretability and attribution interpretability. In Model Interpretability, some metrics were studied to measure the interpretability of latent representations of CNNs by introducing network dissection and interpretable units [2]. In Attribution Interpretability, a metric was studied to measure how much the model's prediction changed when we removed the input pixel with high attribution scores [12].

3. Pilot Study

3.1. Experiment Settings

CNN has the automatic feature extraction ability and can learn good internal representation from raw pixels. [14] So, we can extract a 'representation vector' of an input image from CNN architecture. We'll focus on this representation vector. Using the feature visualization technique to see if we humans can understand what a model learned.

3.1.1 Methods

We compared 2 methods: Madry [6] and Lucid [23]

Madry is related to the 'robustness' package which students in the MadryLab created to make training, evaluating, and exploring neural networks flexible and easy. [5]

They [6] propose using the robust optimization framework to enforce (user-specified) priors on features that models should learn. It can visualize recognizable features of the model easily, by directly maximizing the coordinates of robust representations suffices.

Lucid [23] is a collection of infrastructure and tools for research in neural network interpretability. They researched various aspects on how neurons work and tried different regularization and parameterization techniques to make the feature visualization more recognizable to humans(Lucent [23] is a PyTorch version of the Lucid method).

3.1.2 Models and details

We selected the ResNet50 model which is trained by the Restricted-ImageNet dataset from MadryLab GitHub [5]. Next, we took the output of the average pooling layer as the representation vector with a size of 2048. Each single neuron in the representation vector can be our target to get the feature visualization image.

In Experiment 1, five neurons were randomly selected from a total of 2048 representation vector neurons to create a dataset. Using six random noise images with 0.5 standard deviation as input, we did the optimization on input images which activated the selected neuron mostly for each method(Lucent, Madry) through 200 iterations. The output images are in Appendix A.

In Experiment 2, we used the same neurons and methods in Experiment 1. We build data from the same random noise inputs for every 30 iterations(30 300 iterations) to see how feature visualization is working in a certain step and whether this can help humans to understand more about the representation vector neurons. The output images are in Appendix B.

3.2. Latency Experiment

To check the feasibility of the feature visualization, we tested the latency. The method are described as follows:

We initialized six random noise images with 0.5 standard deviation to feed the network and randomly selected 1000 activation index from the length of representation size as our target optimization. At last, we ran 200 iterations for each method to do the optimization.

The experiment was runned on a computer which has 32GB memory and NVIDIA GTX 1080 GPU with 11GB memory.

3.3. User study

3.3.1 qualitative interview

In addition to quantitative evaluations, we performed semi-structured interviews for qualitative evaluations. We selected 3 interviewee who has pre-knowledge about CNN and machine learning backgrounds. First, we explained the definition of a representation vector and how the feature visualization process has been performed. And the specific question is as follows.

Experiment 1

- Q1.1: What patterns do you think it represents?
- Q1.2: How many patterns do you think it represents?
- Q1.3: Can you understand what this feature visualization means?

Experiment 2:

- Q2: How the 6 images(2 groups between starting from random noise images(3)and original images(3)) are different?

Experiment 3:

- Q3: In which number of iterations do you see the pattern the most clearly?

3.3.2 findings

Feature visualization provides a much richer representation than describing features dependent on the dataset. However, despite seeing the same feature visualizations for target neurons, the concepts people derived were inconsistent

- **Rich expression and flexibility** Feature visualizations represent a variety of visual concepts and a wide range of functional levels. (edge, pattern, texture, parts, objects, (Figure 1). It provides visual information about the concept more explicitly than dataset examples due to its operating principles, which generate visual features to maximize the activation of the target neuron.(Figure. 2 On the other hand, if we were limited to understand features on the fixed examples of dataset examples, lower-level concepts(like texture, pattern etc.) would be much more challenging to explore. Also, the probability of the presence of sample images with noticeable inclusion of features that neurons perceive would be sparse. [23]

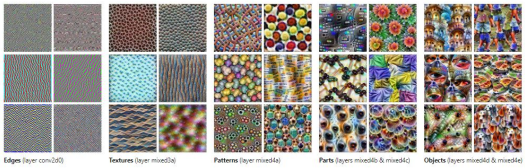
Edges (layer conv2d0)
Textures (layer mixed3a)
Patterns (layer mixed4a)
Parts (layers mixed4b & mixed4c)
Objects (layers mixed4d & mixed4e)

Figure 1: various concepts and levels of feature visualizations

- **Various Interpretations** Feature Visualization could be interpreted from person to person in various concepts. This implies that single feature visualization alone makes it difficult to convey consistent concepts of features perceived by neurons. The cause of these results is listed below based on the findings from the user study.[user study link]
  - Cause 1 : Feature visualizations vary depending on the parameters(seed img, iteration numbers,... etc)
  - Cause 2 : In some cases with single feature visualization were difficult to map the example in the real world and to find the regular patterns. Because neurons can learn non meaningful patterns, feature visualization does not always generate semantically meaningful patterns. [6]
  - Cause 3 : Since the part where neurons become highly active in feature visualization is unknown, we can recognize and derive various patterns like optical illusions. [10]

4. New visualization prototype

4.1. Design Requirements

By utilizing feature visualization rich expression and flexibility, as a result, we expected that humans could understand the features recognized by neurons. To achieve the goal, we tried to devise a visual support tool that can overcome the limitations in section . We have derived design requirements from the limitations of existing attributions

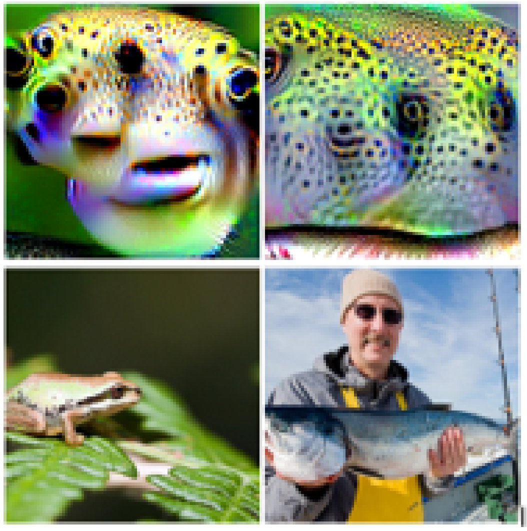

Figure 2: above: feature visualizations of nueron 356, below: dataset examples which make nueron 356 highly activate

- R1. cover the various concepts and wide levels of features that neurons perceive.
- R2. provide the information of activations of sample images by spatiality. (feature visualization limitation)
  - R1-1. explain the spatial features which cover semantic patterns, not pixel units. (Grad-cam limitation)
  - R2-2. reflect the various crop sizes of sample images and the sliding window concept. (Grid-cam limitation)
- R3. support exploration of multiple images to avoid obtaining concepts biased.
- R4. support to derive consistent concepts and map the concepts to real data.

4.2. New visualization support prototype

We designed the noble visualization tool with 4 main functions to satisfy the requirements in section to understand the features that highly activate neurons.(Fig. 3)

- A: Views with 4 example images that make neurons activate highly. [R1,R3] (2 from datasets and 2 from feature visualizations with random noise seed and a dataset example)
- B,C : Grid-cam views and sliders for each image example to compare and extract intersect features between samples that make neurons activate highly(R4). To explore various crop sizes, 4 levels grid cam views can be switched by sliders. (bigger value means bigger crop size). To reflect sliding window concepts, we averaged the activation values of windows which include each grid.(R2)

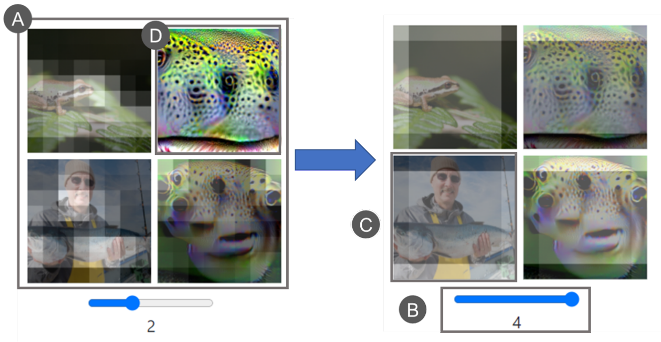

Figure 3: New prototype for visualization supports

- D : Hover interaction to visualize original images

5. Experiments Design

5.1. Interpretability metric for attributions

5.1.1 Interpretability metric for explanation for neuron activation

As far as we know, we were the first attempt to measure the interpretability of feature visualizations, and we devised a noble metric, I-metric , to evaluate how explanation for neuron activation, such as feature visualizations and Grad-CAM [25], affect a person's understanding of patterns perceived by the model. Our metric measures how much the level of recognition of features of target visualization matches the model and the human being after exploring attributions for feature visualizations. Our measurement of interpretability for explanations which visualize and explain the pattern detected by neurons proceeds in three steps Fig.5

1. Explore the user interface which contains feature visualizations and attributions for the pattern detected by target neurons and identify visual concepts of patterns.
2. Select the top 3 images(Human response) which most contain the patterns that are identified in step 1 among 9 options.
3. Obtain the top 3 images(Machine response) which activate the target neurons and measure the count of how many machine and human responses are matched.

In [Table.1], We summarized our metric coverage to requirements for good metrics and reflection of interpretability in sections 2.2.1 and 2.2.2. Therefore, our metric can be said to be appropriate for explaining the interpretability of attributions.

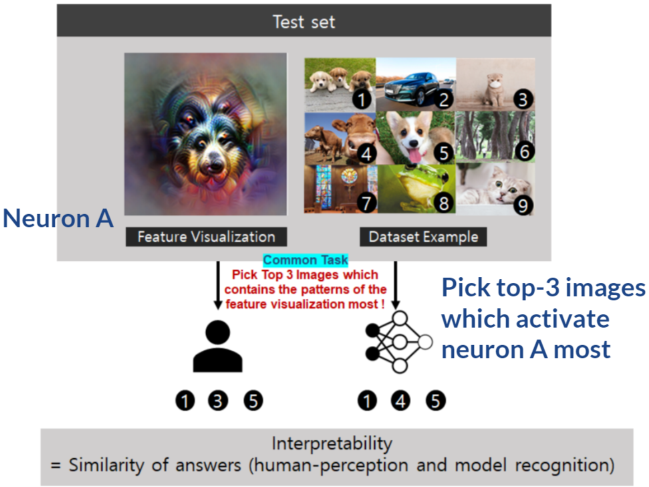
Test set
1
2
3
4
5
6
7
8
9
Neuron A
Feature Visualization
Dataset Example
Common Task
Pick Top 3 Images which contains the patterns of the feature visualization most !
Pick top-3 images which activate neuron A most
❶ ❸ ❺
❶ ❹ ❺
Interpretability
= Similarity of answers (human-perception and model recognition)

Figure 4: I-metric Measurement Scenario

Requirements	Descriptions
Comparative	The data type of our metric is the ratio type.
Understandable	Our interpretability metrics have been explained to several experts in the HCI(Human-Computer Interaction) field, and our metrics have received intuitive and simple evaluations.
Behavior changing, Contrastive, Selected	Behavior changing, Contrastive, Selected & Our indicators reflect how well people can detect the patterns that neurons activate, consequently, means that humans can choose images that maximize neurons.

Table 1: Good interpretability metric requirements satisfied by I-metric

5.1.2 Experiments scenarios for interpretability evaluation

We designed 3 experiment scenarios to measure the interpretability of explanations that visualize and support the exploration of the patterns detected by neurons. 3 scenarios cover different attribtutions each to check the scenergy of feature visualization and our techniques. Exeperiment Overview with test page examples are in Figure. 5. And Each experiments will provide 10 questions with target attributions and 9 options from dataset which are only 3 answers. (hihgly activated group)

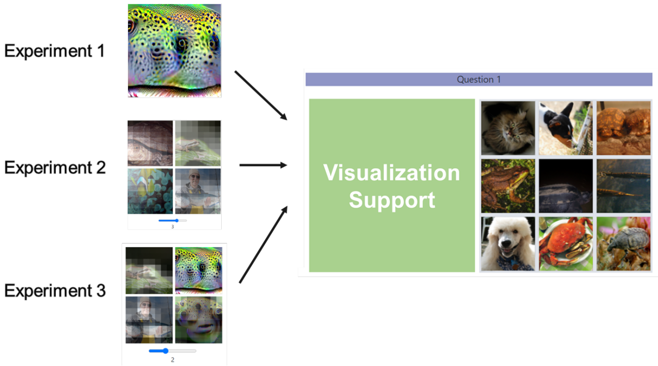

Figure 5: Test page for 3 experiments for I-metric
s

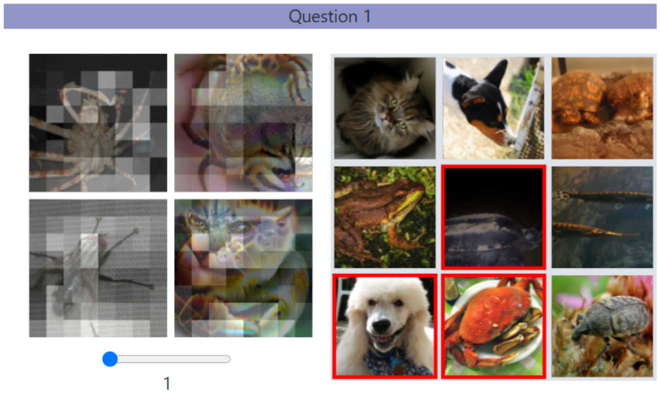

Figure 6: Test page example(exp3)
s

- feature visualization : 1 single feature visualization image from random noise will be provided as an attribution
- our visualization without feature visualization : 4 dataset examples which activates target nurons highly will be provided instead of 2 dataset example and 2 feature visualizations
- our visualization with feature visualization : 2 dataset examples and 2 feature visualizations will be provided.

Each experiment is given a total of 10 questions, with attributes describing the features of target neuron recognition. And participants have to find and 3 correct answers from 9 options from the dataset. At least seven participants were recruited per experiment to participate in only one experiment, and for fairness, the target neurons used in each experiment and the choices were the same.

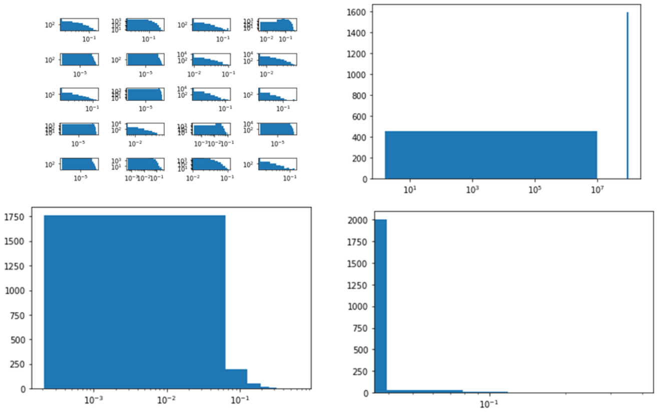

Figure 7: 1. various activation distributions 2. top-100 minimum/ bottom-100 maximum activation values 3.top-100 minimum activation values 4.bottom-100 minimum activation
s

5.1.3 Test set generations and challenges

To generate fair test sets, we had to control some factors that can affect the test set level and results.

- **Avoid too simple problems** The dataset was skewed to a particular class(dog(5000)/validation set(10500)). There would be many dogs in the options from the skewed dataset, which can encourage users to exclude them when picking the answers. So We adjust each class to be evenly distributed.(Debiased Dataset :we adjust each class to contain 150-250 samples)

- **Avoid too difficult problems** To distinguish between answers and wrong answers, 2 groups' activation values must be clearly different. Also, the answer group has to result in high activation values for target neurons sufficiently. If not, it means that samples contain less of the features that activate the target neuron. As a result, humans will not be able to perceive the features from the sample images. The activation value distribution of neurons has been shown in a wide variety of distributions and ranges.(Fig. 7) We selected neurons by developing a module that can be obtained by separating groups of answer options with sufficiently high activations and options of incorrect answers that are sufficiently low activations from them according to the examiner's criteria.

To control the problem quality, we selected 13 neurons which satisfying our criteria below in debiased dataset and remove the neurons which feature visualizations are similar.

- topEdgeNum=100, topMin =0.11 : The minimum activation value for the top 100 images activating the target neuron must be greater than 0.11.

- bottomEdgeNum = 100, minMaxRatioMin = 10000 : The ratio between the minimum value of the upper activation image for the target neuron and the maximum value of the lower activation image is at least 10000 times.

As a result, we piked 10 neurons of representation vectors(291, 356, 660, 906, 908, 1526, 1591, 1943, 1994, 2031) as target neurons of questions. And we picked 3 answers and dataset examples of attributions from top-100 images activating the target neurons and 6 incorrect answers were picked in bottom-100 images for each selected target neuron.

We generate 2 feature visualization images for attributions, with random noise and dataset example by Madry's method with iteration 200. It was not a noticeable difference in use study results and pilot experience(exp 1) between Madry's and Lucent's method. But Madry's method can represent more various concept because the method highlights the parts of the original image which activates target neuron highly. Meanwhile, Lucent's generate consistent images. And from the user study, we could derive the iteration number of feature visualization that needs to be larger than 150.

6. Results

6.1. Latency

As we mentioned at section 3.2, we tested the latency by randomly choosing 1000 neurons in the representation vector neurons and running 200 iterations of optimization. The result is different between two methods, Madry took 2:54:38 in total (average 10.48 second for each optimization) and Lucid took 5:05:41 in total (average 18.34 second for each optimization).

6.2. User Study

We conducted user study through a couple of questions. We could see some findings as below and with Table. 2.

- Most had consistent results in both Lucent and Madry's method, But in some cases, people recognized different patterns and number of patterns even the same feature visualization(Table. 3).

- Both Madry and Lucent methods had significant pattern expressions after 150 iterations.

- However, It is reported to be more vivid and colourful in Madry's way

	Question	Findings
Experiment1	What patterns do you think it represents?	There is a case in which each person recognizes different patterns in the same image.
	How many patterns do you think it represents?	In some cases, people recognize multiple patterns in one feature visualization.
	Can you understand what this feature visualization means?	People can recognize clearly in about half feature visualization images.
Experiment2	How the 6 images are different?(2 groups between starting from random noise images(3) and original images(3))	In Madry method, participants said there are differences between the two groups. Vice versa. Images in Madry method have more clear pattern than Lucent one.
Experiment3	Which iteration number of feature visualizations make the most sense?	The number of iterations starting to express a pattern is similar in both ways. In both ways, patterns are better recognized after more than 120-150 iterations.

Table 2: User study results and findings

6.3. Interpretability of Attributions

6.3.1 Participants

A total of 25 students(CSE, DS, Statistics, Physics) who has deep learning backgrounds participated in the experiment.

- Experiment 1 : 7 participants(4 female), average age 26.9 (23-31)

	Lucent	Madry
Object	6	7
Parts	6	5
Pattern	5	6
Texture	2	3
Edge	3	0

Table 3: Count of patterns that participants described in User study

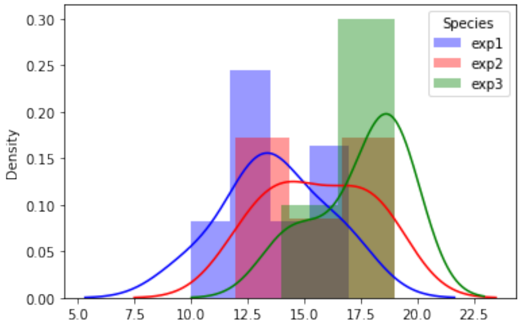

Figure 8: I-metric results for 3 scenarios

- Experiment 2 : 10 participants(7 male), average age 31.1 (25-42)
- Experiment 3 : 8 participants(7 male), average age 29.4 (23-35)

6.3.2 Results

Participants solved the problem at each experimental web interface we deployed and the results were stored in JSON file form and we recorded participants reactions during the test. We count the number of corrected answers of the 30 answers to be found in the test set, we counted how many answers the participants actually found. In experiment 1, participants got an average of 13.7(10-17). In experiment 2, participants got an average of 15.6(12-19). In experiment 3, participants got an average of 17.5(14-19). Also we plot a histogram for comparing the distributions of each experiments.(Figure. 8)

Participants in experiment 3 found more answers(2 4) than in experiment 1 and experiment 2. It can be interpreted to be synergistic when feature visualization and our new visualization techniques are combined than when used separately. Also it is consistent with what we expected.

Also, We can find features ,for each same questions,of similar concepts mentioned by participants in experiments 3. For examples, in Figure. 9 (exp3), many of the participants mentioned fur after looking at the left attribute and the right attribute and the leg. It means that attribution in exp3

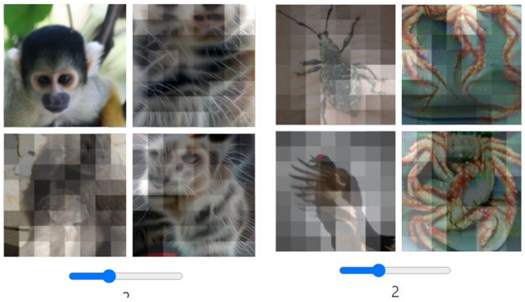

Figure 9: consistent mentioned features: 'fur'(left), 'leg'(right)

provided some consistent concepts to explain the features of neuron perceived and we satisfied R4 in section 4.1.

Moreover, some participation mentioned that levels 1 to 4 of the grid-cam actually help them analyze. Also, some participation required on/off function to look at once 4 original images instead of hover interaction for one. And some participation mentioned learning curve of our visualization support. "I think I have a sense of what to do when I get used to it a little bit."

7. Discussion

There seem to be a number of factors in the part where there were no people who got more than 20 correct answers. There is a reason why our tool's learning curve requires time to get used to it. Also, existing threats like adversarial attacks can be the reason for a lower score of results. It reveals the model itself learns non-semantic concepts.

Like the visualization tool proposed this time, feature visualization is likely to help us understand the model further. In addition, there are many places to look into the model, so a system that effectively navigates such a search process must be supported to be deployed in real-life situations.

8. Conclusion

In this project, we analyzed the limitation of existing attribution techniques from user studies. And we developed the visualization support for understanding the features which make neurons highly activate by extracting requirements that can better understand the model. Also, we devised interpretability metrics and experiments to demonstrate the effectiveness of our prototype. As a result, we proved our tool to be able to overcome the limitations of existing attributions. If we reflect feedbacks from experiments in future work, we can expect to provide more useful interpretability used in other visualization tools.

9. Deployment

You can experience the testsets provided in the 3 experiments on the site below.

- Experiment 1: https://edw2n.github.io/MLVU-Exp1/.
- Experiment 2: https://edw2n.github.io/MLVU-exp2/.
- Experiment 3: https://edw2n.github.io/MLVU-exp3/.

References

[1] B. Y. Alistair Croll. *Lean Analytics*. O'Reilly Media, Inc., 2013. 2

[2] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 2

[3] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah. Activation atlas. *Distill*, 4(3):e15, 2019. 1

[4] B. Class. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. 1

[5] L. Engstrom, A. Ilyas, H. Salman, S. Santurkar, and D. Tsipras. Robustness (python library), 2019. 2

[6] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019. 2, 3

[7] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. 1, 2

[8] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. 1

[9] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018. 1

[10] U. Hasson, T. Hendler, D. B. Bashat, and R. Malach. Vase or face? a neural correlate of shape-selective grouping processes in the human brain. *Journal of cognitive neuroscience*, 13(6):744–753, 2001. 3

[11] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics*, 25(8):2674–2693, 2018. 1

[12] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. *arXiv preprint arXiv:1806.10758*, 2018. 2

[13] hyo@yna.co.kr. ”ai hates homosexuality and the disabled?”...controversy over ’ai ethics’. *(C) Yonhapnews*. 1

[14] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, 2020. 2

[15] L. Klein. *UX for Lean Startups: Faster, Smarter User Experience Research and Design*. O’Reilly Media, Inc., 2018. 2

[16] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. *arXiv preprint arXiv:1608.08182*, 2016. 1

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[18] A. Maurya. *Running Lean: Iterate from Plan A to a Plan That Works*. O’Reilly Media, Inc., 2013. 2

[19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019. 1

[20] A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks. 2015. 1, 2

[21] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020. 1

[22] C. Olah et al. The building blocks of interpretability. distill 3, e10 (2018). 2

[23] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 1, 2, 3

[24] G. Rapier. Tesla’s autopilot confused a burger king sign for a stop sign. the fast-food chain turned it into an ad. *Insider*. 1

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 4