Title: ECE1512 – Homework 2 Spring 2019

Course Number: ECE1512H S

Student’s Name: Dae Yon Hwang

Student ID: 1004897150

Date Due: 2019-02-28

Date Handed In: 2019-02-26

Section 1. What are the main technical technological drivers behind the XAI solution reviewed?

Nowadays, machine learning technologies introduces high outputs in academic and industry areas. However, the problem is that they usually act as black-boxes, meaning designer cannot explain the reasons for particular decision and recommendation. Explainable machine learning (XAI) tries to open the black-boxes to explain the internal mechanics of a machine learning system in human terms. XAI is vey important when we consider applying in regulated domains like healthcare, finance and medicine because they require the high accountability and transparency.

For this assignment, [1] is chosen for literature survey and it mainly explains about one way of feature visualization. Briefly, feature visualization is to observe a specific trained feature (also can be layers, class logits and class probabilities) to understand what network is looking for. Convolutional neural network (CNN) gradually builds up abstractions: first it detects edges and orientations, then it uses this information to detect textures, the textures to detect patterns, and the patterns to detect parts of objects.

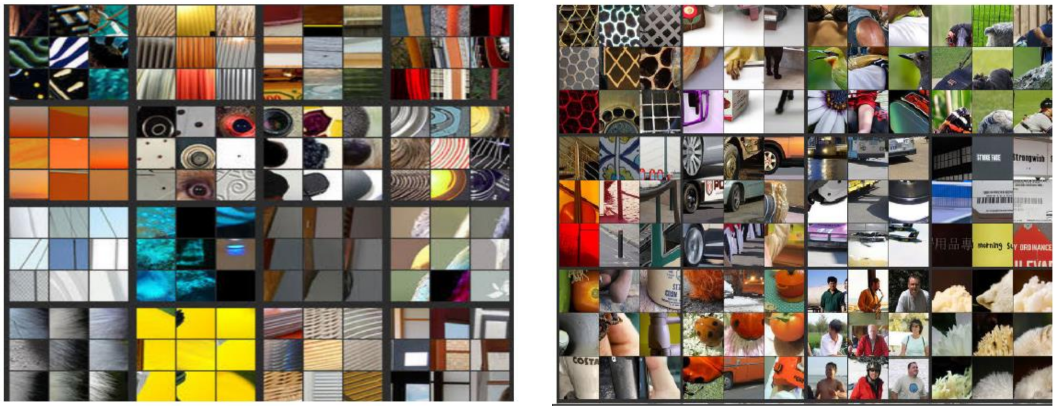

Figure 1.1 Figure 1.2

Figure 1. Building up abstractions in CNN [2]. Figure 1.1: Feature from second layer. Figure 1.2: Features from third layer. We can find deeper layer has more complexity and interpretability than shallow one

From [1], they use the derivate operation on feature to find out what kind of input would yield a certain behavior. This optimization-based approach is an effective way to understand what a model is really interested in, because it distinguishes the things causing behavior from the things that merely correlate with the causes. In addition, it has the advantage of flexibility which is helpful in visualizing how features evolve as the network trains.

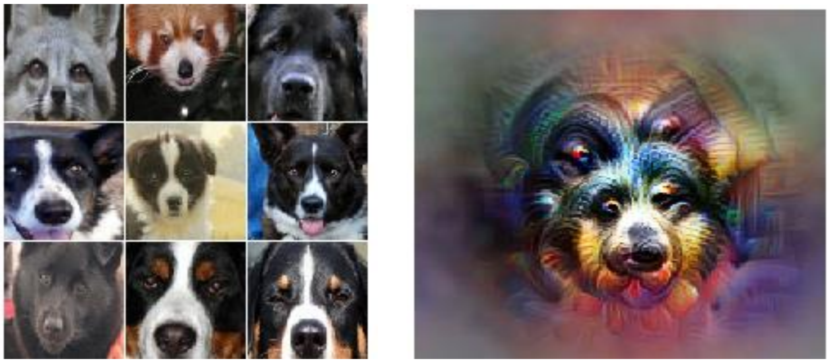

Figure 2.1 Figure 2.2

Figure 2. Feature visualization through [1]. Figure 2.1: Datasets used for training. Figure 2.2: Trained feature. Compared figure 1, trained feature is highly related with the causes of behaviors without considering the background

Diversity is important to broaden our expectation of what the neuron activates; in other words, we can understand the translations in dataset. For diversity, additional term is applied with the optimization objective to achieve this property. The problem from diversity in individual neuron is that it can cause unrelated artifacts, make examples in unnatural way and represent strange mixture of ideas. Because of these problems, the author said it is better to focus on interaction between neurons.

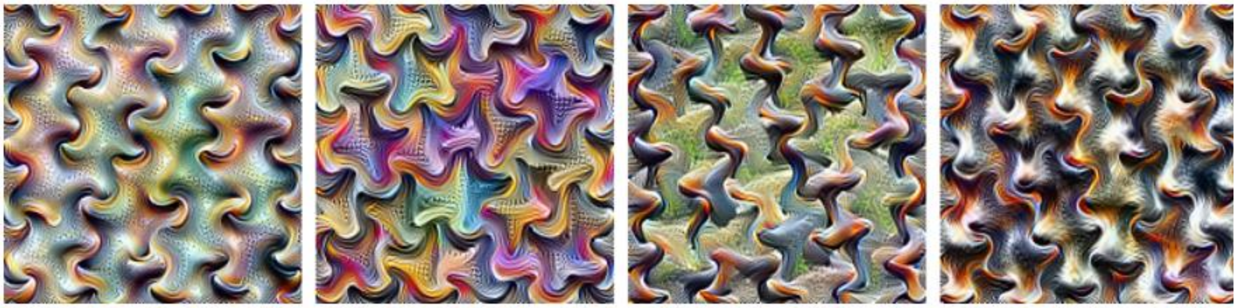

Figure 3. Feature visualization with diversity in optimization-based approach [1]. We can see different curvy facets which cannot be found with simple feature visualization without diversity

Combinations of neurons work together to represent images in neural networks. For understanding at interaction between neurons, arithmetic operations (ex. add, subtract) and linear interpolation can be applied to find out interesting patterns. However, this is not enough to obtain the meaningful directions for interpretation in gradients. Furthermore, they don't know how hundreds of directions in real case are interacting even if interpolation gives an insight how a few directions are interacting.

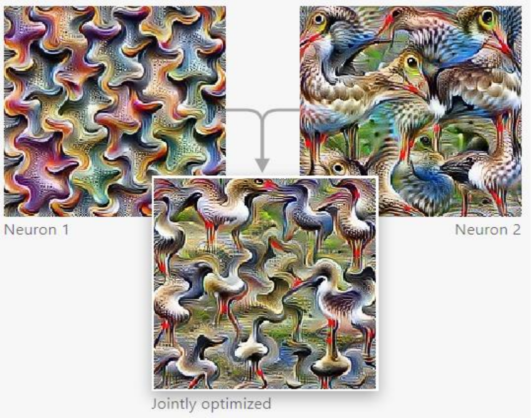

Figure 4. Example of combinations between neurons [1]. Jointly optimized image comes from interaction between neuron 1 and neuron 2

The author also explains about the enemy of feature visualization. Optimizing an image to make neurons to fire can yield high-frequency noise even if we optimize long enough. From [1], this noise comes from strided convolutions and pooling operations because these operations create high-frequency patterns during backpropagation. Thus, we can think as optimization-based feature visualization is a double-edged sword. For solving this problem, they impose the regularization on the model. Briefly, weak regularization avoids misleading correlation while it can yield adversarial examples (vice versa for strong regularization).

Three methods for regularization are mentioned in [1]: frequency penalization, transformation robustness and learned priors. All these methods try to reduce high-frequency noise in the gradient. Frequency penalization explicitly penalize the variance between neighboring pixels and/or implicitly penalize high-frequency noise by blurring. The problem for this method is it can reduce important high-frequency features like edges along with noise. Bilateral filter can be helpful to suppress this problem. Transformation robustness is to find examples that still activate the optimization target highly even though examples are slightly jittered, rotated or scaled. This stochastically transforming is also useful to decrease high-frequency noise. Lastly, learned prior is to learn a generator that maps points in a latent space to examples of real data and optimize within that latent space. It is similar with GAN and VAE. This method produces the most photorealistic visualization but there are still two questions to be solved before application: what came from the model being visualized and what came from the prior.

The author mentions that transforming the gradient (called preconditioner) is a powerful tool to make an optimization problem radically easier. Compared to three regularizations, it reduces high-frequency noise in the visualization itself. It is similar with gradient descent but doing steepest descent in another parameterization of the space or with different notation of distance. This preconditioner can make data decorrelated, where optimizing in the decorrelated space helps to reduce high-frequency noise. Thus, combining the preconditioner and transformation robustness improves the quality of feature visualization. However, there are some questions to be asked before applying preconditioner (ex. is the preconditioner merely accelerating descent and brining to same minimal as gradient descent?).

Section 2. Explain how a user perceives and assess the chosen XAI solution in varying contexts?

There are diverse visualization options in optimization-based approach that reveal different parts of networks. In other words, we can interpret from single neuron to class probability, depending on optimization objectives. When we want to understand individual features or channels, we need to find examples that yield high values of gradient in single neuron or entire channel. If we want to see a layer as a whole, we should search for images that layer shows the highest interesting in terms of gradient. Furthermore, we can also understand the feature visualization from class logits before the softmax and class probabilities after the softmax. Interesting point is that feature visualizations from optimization-based approach are not always show the highest interpretability in highest layers [3].

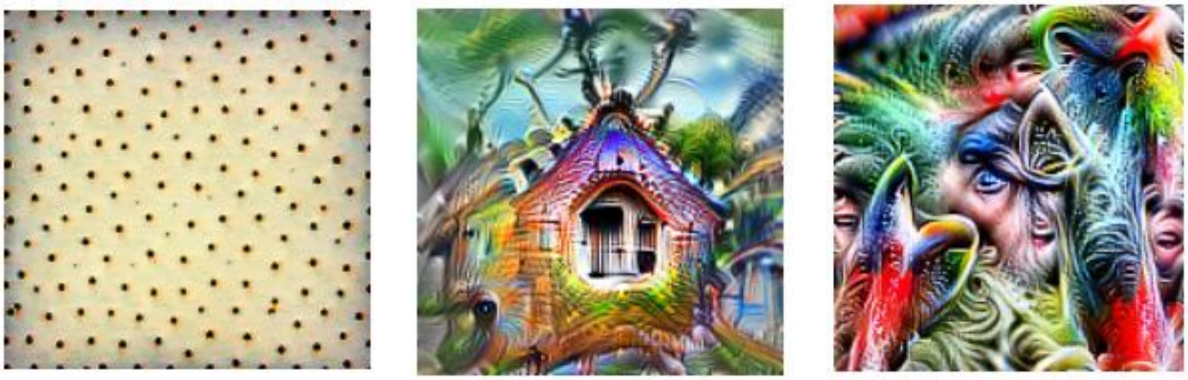

Figure 5.1 Figure 5.2 Figure 5.3

Figure 5. Feature visualizations in [3]. Figure 5.1: Visualization of feature from 3rd layer. Figure 5.2: Visualization of feature from 4th layer. Figure 5.3: Visualization of feature from 5th layer

From figure 5.1, neuron from the 3rd layer (the lowest layer) only focuses on simple textures and it does not possess complex interpretation. When we see figure 5.2, neuron get enough complex to understand what they are really looking for. As input datasets, it represents house. When more layer is stacked (figure 5.3), neuron is too complex to interpret and it does not seem to correspond to particularly meaningful semantic ideas. Thus, if user wants to understand individual neurons, they need to search the neurons between middle of layers and the highest layers to find most understandable ones.

As figure 3, diversity of individual neurons is important to broaden our expectation of what neurons are looking for. It means we can assess different view, angle, zoom and variety of same targets. However, there are problems when we apply diversity in individual neurons: it causes unrelated artifacts, makes examples in unnatural way and represents strange mixture of ideas. User should perceive this problem and consider feature visualization from interaction between neurons, instead of single neuron. Neural network represents images by working through combinations of neurons. User can find interesting interaction between neurons by applying arithmetic operations and linear interpolation on them (see figure 4). Even if we know interactions between a small number of neurons, there is no straightforward method to find interesting gradient directions in each neuron and interactions between huge number of neurons.

CNN requires the convolution and pooling operations which can yield high-frequency noise. Even if user control the learning rate carefully, he/she cannot perfectly remove noise in image. User should apply regularizations to solve this situation. There are three methods for regularization considered in [1] where each has pros and cons. User should decide which methods will be used, depending on structure of dataset and computational resources. In addition, preconditioner is explained in [1], which is similar with gradient descent but works in different parameterization of space or with different notation of distance. The better feature visualization can be achieved using combination of decorrelated space (preconditioner) and transformation robustness (regularization).

Section 3. Discuss implementation implications and the use of existing regulations in certain application areas (medicine, finance, etc.).

In high risk environments like medicine and finance, model should be transparent and accountable that user can fully understand why the model makes a certain prediction. This is very important since, without interpretability, a correct prediction only partially solves the original problem. There are several things to be considered in high risk environments [4], [5]: fairness, privacy, translucency, reliability, safety and mismatched objectives.

Table I. Regulations in high risk environment

Regulations	Explanations
Fairness	Model ensures that the predictions are unbiased and do not discriminate against protected groups. Also, it helps a human to judge whether the decision is based on a learned bias (ex. racism)
Privacy	Sensitive information should be protected properly. Without privacy, we cannot trust our personal information to model in high risk environment
Translucency	The degree to which an explanation method looks inside the model. In neural network, it can be individual feature visualization
Reliability	Model ensures that small changes in the input do not lead to large changes in the prediction
Safety	Model needs to consider a complete list of scenarios in which the system may fail. However, for complex tasks, the end-to-end system is almost never completely testable because of its size and computational property
Mismatched objectives	Model may optimize an incomplete objective. In other words, objective of model does not include all related cases or constraints that can influence our objective adversely

For this part, two certain areas will be considered: hospital and bank. First of all, machine learning model for detecting cancer should possess high reliability and safety. For example, if model is used to detect a high-risk cancer from normal person, it should never fail to detect a true patient. In other words, 99% accuracy is not enough to measure system performance; thus, we need high interpretability for model to make reliable decision. Feature visualization with diversity can be helpful to see the diverse examples of same targets. Also, jointly interaction between neurons is useful to expand our view of dataset. All these methods consider various scenarios of target (ex. detecting cancer) in which model may fail and thus, by including these scenarios, it will make the reliable and safe system. Optimization-based feature visualization [1] has high translucency because it can visualize features from individual neuron to class probability after the softmax. Thus, we can analyze all details in how our network is really looking for and this full understanding of model will give persuadable explanations to doctors and patients. Next, patient's information should be highly secured. In hospital, we should offer our personal information to doctors and machine learning model can be useful to protect and manage private information in huge database. Obviously, model needs to be highly secured and never permit the hacker's attack. Feature visualization can be applied to distinguish the access methods in terms of permittance through comparing the different visualization of those methods. Lastly, machine learning model in hospital should be optimized in complete objective (no mismatched objective). For example, a clinical system may be optimized for cholesterol control, without considering the likelihood of adherence for measurement. XAI should include all possible constraints in its objective and feature visualization may help to broaden our view of purpose and constraints.

In bank, fairness is the most important regulation to be considered. For example, machine learning model can reject a loan application even if it may be completely unexpected for the applicants. Bank should give persuadable reasons to applicants why they are rejected. We can see the specific feature visualization that reacts to permittance or rejection on loan application and then, we can extract some reasons for model's prediction. Furthermore, we need to find the bias in our datasets which is an inadequate factor for prediction. For instance, gender cannot be the cause for rejection of loan but, without fairness, model can be affected from this bias. For analyzing bias in model, we can put biased and unbiased data into model, then compare feature visualization between them to see which bias really affects model's prediction. Optimization-based feature visualization with diversity can consider various scenarios when our model is failed. This will expand our view of datasets, which is helpful to build reliability and safety in our model. In bank, it is obvious that personal information of client should be highly secured. Using feature visualization, we can analyze the specific features (or interaction of neurons) to see which features are related with access control, and it will help to understand which situation our model works properly and which situation our model fails to defend attack from hackers. Also, we can achieve high translucency from [1] that can visualize all the details in model, meaning we can understand model perfectly. Bank requires the high reliability from full understanding of model (meaning high translucency), thus feature visualization through optimization-based method can be suitable.

Section 4. Discuss in detailed implementation parameters like computational complexity, robustness.

In [6], they introduce implementation parameters used in XAI. Here, some of those parameters will be covered which seem valuable to be implemented in this assignment.

Table II. Implementation Parameters

Parameters	Explanations
Algorithmic complexity	Computational complexity of algorithms to produce explanations
Accuracy	The degree of explanations of a given prediction generalizes to other yet unseen data. Similar to generalization error
Fidelity	The degree of explanations reflects the behavior of the prediction model
Stability	The degree of similarity of explanations which come from same model with same task
Comprehensibility	Readability of explanations. It depends on the users
Representativeness	The degree of explanations represents of the model. For example, a model explanation can explain behavior of the whole model or one of layers
Debugging	The degree of interpretation that helps to understand the reason for the error, which also delivers a direction for how to fix the system

When we consider using [1], three factors should be considered in terms of algorithmic complexity. First, complexity depends on the scale of feature visualization (ex. one neuron, interaction between neuron, or whole layers). Second, high diversity included in optimization objective can show high computational complexity because it will yield more diverse examples which need additional computation. Third, the degree of methods for reducing high-frequency noise is also related with complexity. Obviously, we cannot eliminate all high-frequency noise without regularization and preconditioner, but they also needs some computational resource to do that. Depending on these factors, algorithmic complexity will be determined if we considered [1]. Furthermore, optimization-based approach considers using gradient for feature visualization, thus we should perceive the computational cost from gradient update (ex. backpropagation). Accuracy can be the basic measurement of system performance. From survey, high interpretable model is not always related with high accuracy in model because we need to tune some parameters (ex. optimizing the gradient [1]) for getting better feature visualization, but it does not guarantee to give good predictions. Optimization-based approach with diversity may not give high fidelity as database approach because [1] shows the mixed feature visualization which is hard to separate the specific target. However, from figure 2, feature visualization from optimization-based is good for eliminating the merely correlate with the causes, and thus it has good fidelity to focus on highly related with causes. Also, feature visualization is good for analyzing the stability. We can compare the feature visualization from same models with same datasets to see the stability in terms of explanations. Stability is important because model should give same trustworthy interpretation to users when they works on similar tasks. However, feature visualization has low comprehensibility. Obviously, all explanations can only be interpreted by experts who have experiences in machine learning areas. Most of users do not have knowledge in machine learning, thus they need interpreters to make them understand what feature visualizations are really meaning. Optimization-based feature visualization has flexible property in terms of representativeness. It can cover features from individual neurons to classification probabilities. In addition, we can specify which optimized features are related with our targets and which features are not useful to detect targets by feature visualization. Thus, we can debug our model properly and it will give persuadable explanations to users why they need to accept model's predictions. All these parameters are highly dependent on implementation environments; above all, we need to include all constraints related with tasks and then, consider these implementation parameters (table II) with regulations (table I) in environment to decide the structure of XAI.

Section 5. Given your research background / level of expertise, please share your opinions on the utility of such an approach.

During pursuing master’s degree, I was researched about fluorescence dataset in human oral to detect a cancer data from healthy one. Logistic regression with several classifiers was used for classifying dataset and feature extraction / selection was the most important and time-consumption part in that project. Without knowing XAI, I was tried to use principal component analysis (PCA) to visualize features in low-dimension because it was hard to understand data with high-dimension. The downside of PCA was less performance than data with high-dimension. For better system performance with interpretation, I can apply simple feature visualization through optimizing gradient and it will help to understand the trained high-dimension features.

I have worked in Hyundai which is the automobile company and mainly performed research about camera algorithms used for driver assistance system. The camera detects the lanes, signs, pedestrians and many other subjects, and gives specific signals, depending on what it recognizes. It was working with similar structure as CNN and there were many deterioration examples to decrease the performance of recognition. For example, if camera is applied in the country that never tested before, it cannot detect signs well and sometimes shows different labels. This is a big problem because confused detection can give wrong signals to vehicle and this will result in wrong occurrence of algorithm. From my experience, camera detected pedestrian signs as real person and vehicle suddenly stopped because of emergency algorithm used to protect pedestrian. The way to solve this problem was driving all new countries to collect new data from our camera or receiving the dataset from government or company in that country. At that time, size of model was huge, and thus we only considered a few trained feature visualizations for interpretation, but it was not enough to understand why they are working in certain way. Also, most of feature visualizations were hard to be interpreted. From [1], we may need optimization-based approach to build the useful feature visualizations to understand the model. Furthermore, it will give insights to understand the interaction between neurons. Above all, optimization-based feature visualization with diversity may be helpful to broaden our expectation of diverse view, angle and zoom of targets (ex. signs, lanes); thus, it will be useful to reduce time-consumption for driving test cars in new country. Clearly, we should analyze the optimization-based approach in deeply before applying in this camera because it is really high-risk environment. Regulations and implement parameters which are introduced in table I and II are good measurements for deciding permittance.

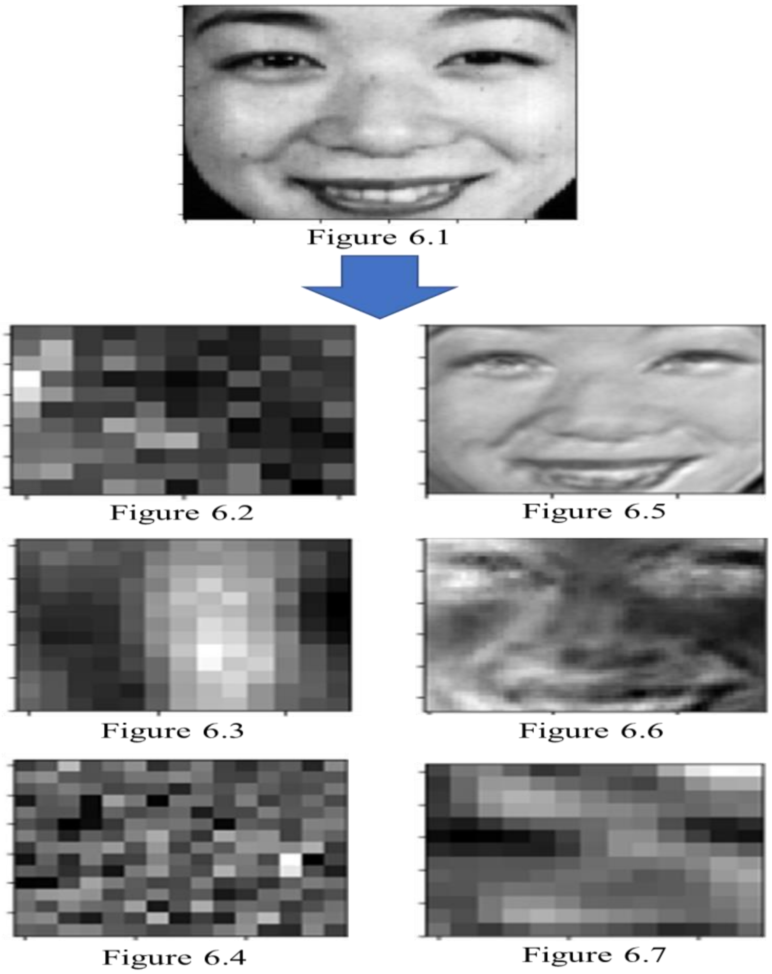
Figure 6.1

Figure 6.2

Figure 6.5

Figure 6.3

Figure 6.6

Figure 6.4

Figure 6.7

Figure 6. Feature visualizations in CNN for ECE1512 project. Figure 6.1: Input data into CNN. Figure 6.2, 6.3, 6.4: Visualization of filters in 1st, 2nd and 3rd CNN layers, respectively. Figure 6.5, 6.6, 6.7: Visualization of convolution layers in 1st, 2nd and 3rd CNN layers, respectively. Here, figures come from one of channel in each case.

For ECE 1512 project, facial expressions identification system will be built using CNN. Briefly, the network is constructed with 3 layers of CNN which are composed of convolution, rectified linear unit (RELU), maxpooling, dropout and 1 layer of fully-connection. As figure 6.1, cropped images from eyebrow to mouth are applied into the network to classify seven facial expressions. For this assignment, feature visualization of filters for convolution and layers in convolution are introduced in figure 6.

From figure 6.2, 6.3 and 6.4, filters try to find the distribution of intensity, edges and orientation in input data, but it is hard to understand. Also, it doesn't seem to gradually increase the interpretations as increment in layers since the size of filter is quite small (the biggest one is 15x15) compared to input (144x144) and the depth of layers in CNN is shallow (3 layers). Figure 6.5, 6.6 and 6.7 come from the visualization of convolution layers. Figure 6.5 and figure 6.6 show good interpretations that we can find what the expressions are, but figure 6.7 is hard to understand what it is looking for. This is because 3rd layer (figure 6.7) has small size (it is 13x13) to see interesting patterns as other layers. When we see feature visualization in 2nd layer (figure 6.6), it only seems to focus on the highly related with the causes of expressions like eyes, eyebrows and mouth. With this respect, the feature visualization of layers in convolution is still useful for understanding facial expressions even if layer is increased, because edges from facial parts (ex. eye, mouth) are remained well which are important to interpret the facial expressions.

References

[1] C. Olah, A. Mordvintsev, and L. Schubert, “Feature Visualization,” Distill, 07-Feb-2019. [Online]. Available: https://distill.pub/2017/feature-visualization/. [Accessed: 25-Feb-2019].

[2] M. D. Zeiler, and R. Fergus, “Visualizing and understanding convolutional networks,” In European conference on computer vision, Springer, Cham, 2014, pp. 818-833

[3] C. Olah, A. Mordvintsev, and L. Schubert, ″Feature Visualization - Appendix,” Distill. [Online]. Available: https://distill.pub/2017/feature-visualization/appendix/. [Accessed: 25-Feb-2019].

[4] C. Molnar, “Interpretable Machine Learning,” Christoph Molnar. [Online]. Available: https://christophm.github.io/interpretable-ml-book/interpretability-importance.html. [Accessed: 25-Feb-2019].

[5] F. Doshi-Velez, and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608 (2017).

[6] M. Robnik-Šikonja, and M. Bohanec, “Perturbation-Based Explanations of Prediction Models,” Human and Machine Learning. Human–Computer Interaction Series, Springer, Cham, 2018, pp. 159-175