Short Report on VISxXAI
MSBD5014: Human-eye on Deep Learning

Name: ZHOU Jiawei
Student ID: 20556794 & ITSC: jzhoubu
Supervisor: Huamin Qu
Hong Kong University of Science and Technology, HongKong
jzhoubu@connect.ust.hk

***Abstract*—Artificial Intelligence (AI) has grown rapidly and brought more impact to our daily lives these years. There has been a surge of complex black-box models with high performance. On the other hand, the application of these models especially in high-risk domains is more stagnant due to lack of explainability and interpretability, which lead to a disconnect between these models and domain experts. In this report, we will first discuss when we need explainable artificial intelligence (XAI). Then We introduce formulations to describe when and how we visualize the black-box model and categorize the visualization technology applied in the XAI domain. Last we use this formula to describe our main work in MSBD5014 with focus on what different layer of CNN focus on during a image classification task.**

The contents of this report are organized as:

***Index Terms*—XAI, visualization, CNN**

I. Introduction

"The role of visualization in artificial intelligence (AI) gained significant attention in recent years. With the growing complexity of AI models, the critical need for understanding their inner-workings has increased. Visualization is potentially a powerful technique to fill such a critical need"[19]. The purpose of this independent project is to visualize what is inside CNN model. Thus, we reproduced and organized three advanced methods: class activation mapping, saliency maps, and guided backpropagation. We have packed them into a python package and put it on Github[19]. Moreover, believing the clear explanation lead to better product, we go further by introducing a practical system to formulize *where* and *how* we do visualization on the black-box model. Last, we present what we have developed during MSBD5014 project using the formula we introduce.

II. Overview of XAI

A. Opinions upon XAI

In recent years, the rapidly growing application of AI has created new anxieties for the high-risk domain and has drawn attention to the need for model reasoning. Such a concern has made a contribution to putting XAI on the workshop of both IJCAI and VIS 2018 where experts spread different opinions on XAI: Believing the prospective future of interactive visualization technology. ShixiaLiu [2] present the concept of interactive model analysis with three categories: understanding, diagnosis, and refinement, and Josua Krause [3] brought out a practical, interactive visual analytics system for assessing the interpretability and actionable insights of trained predictive models in VIS x AI workshop. On the other hand, Cynthia Rudin [4] provide us with a new point of view by suggesting creating models that are interpretable in the first place rather than explain the black box models. All opinions are fantastic since theyre with detailed explanation, which makes XAI the most remarkable part of AI for some reasons.

B. When do we Need XAI

In this session, we will discuss the situations for them respectively. Moreover, we talk about the relationship between XAI and AI.

Based on the opinion of Jim [6], the human mind consists of two different systems the brain uses to form a thought. System 1 uses association and metaphor to produce a quick and dirty draft of reality, which System 2 draws on to arrive at explicit beliefs and reasoned choices. System 1 proposes, System 2 disposes. System 1 is a fast, intuitive, unconscious system based on experience to draw a conclusion, which is more like AI. System 2 is a slow, conscious, logical system with high-level reasoning, which is related to XAI. The point is that these two systems are not contrary but complementary: having an accurate AI model may meet the need of most tasks, but explainability would lead to better product and help to build more general AI.

C. Overview of XAI technology

The problem of explainability is as old as AI and maybe the result of AI itself, and recently has become a prominent problem with the dramatic growth of deep learning field.

In computer vision tasks, CNN (convolutional neural network) has acquired a monopoly for years with the advantage of learning the filters that in traditional algorithms without prior knowledge and human effort in feature design. However, the inability to explain the black box models not only raise potential risk but also become an obstacle to the breakthrough of the research. In early year, Matthew [7] proposed a novel visualization technique that gives insight into CNN by mapping intermediate activities back to the input pixel space with deconvolutional networks [8], while saliency maps[9] could be obtained from a convnet by projecting things back from the fully connected layers. By considering activation function on backward deconvnet based on saliency maps, guided backpropagation[12] achieve better in visualizing features learned by CNN. Unlike technique above, class activation mapping(CAM)[11] succeed in localizing the discriminative image regions without projecting things back with the practical design of the global average pooling [10]. Later, Grad-CAM[13] uses the gradients to produce a coarse localization map which gets rid of the requirement of global average pooling in CAM. For tasks in natural language processing, Ming [14] proposed RNNvis, a glyph-based sequence visualization tool to analyze the behavior of an RNN's hidden state at the sentence-level. Besides visualization on deep learning, LIME [15] increase the explainability of black-box models by approximating a black-box model by a simple model locally. Inspired by a set of methods like LIME, Lundberg S M and Lee S I proposed SHAP[16], a unified approach to interpreting model predictions with theoretical guarantees about consistency and local accuracy from game theory.

These days, explainable AI appears in several top workshops (eg.VIS, IJCAI), which indicates people are growing more concern on model reasoning over the performance. Before we give different categorizations of XAI via visualization, we first describe the notations used in this report.

III. VIS x XAI

In my opinion, visualization provides us front-end explanation while advanced research in algorithm could give us back-end mathematical proof. Combining them together can make things black more clear. Thus XAI via visualization is as practical as powerful. To make this work, we need to figure out two simple problems: for a black-box model, *where to visualize and how to visualize?* In order to give people an accurate answer, we are going to introduce two formula systems: algorithm formula and visualization formula.

A. Notations and Definitions

In this section, we introduce some notations and definitions that are used in this report. First of all, we give the definitions of complex algorithm with little explainability, noted as *F* and visualization technology, noted as *V*, respectively.

In this survey, an algorithm system consists two components: an input space *X* and an algorithm *F* which can be formulated from a set of simple functions or say lower level function *f*. For example, if our algorithm is a random forest model. Then *F* can be written as $F = f(f_1, f_2, ..., f_n)$ where f_i is a decision tree model and *f* is an overall function applied on the results of these n decision trees. To be more specific, we call *F* a parallel algorithm and denote it as F_p if *F* can be formulated by a parallel set of f_i as like:

$$F = f(f_1, f_2, ..., f_n) \tag{1}$$

Otherwise, we call *F* a sequential algorithm noted as F_s if *F* can be formulated by a sequential set of f_i as like:

$$F = f_1 \circ f_2 \circ ... \circ f_n \tag{2}$$

Noted: Here we use symbol $\circ$ to denote the nested functions so that $(f \circ g)(x) := f(g(x))$. For a sequential model F_s, we use F_i to denote $f_1 \circ f_2 ... \circ f_i$.

B. A Categorization of XAI via Visualization

Given a specific algorithm system (*X*,*F*), a visualization system involves two parts: a feed-in data and a visualization technology noted as *V*. The feed-in data is the input for any visualization technology *V*. It could be the same as the input of algorithm *F* or any internal product of *F*, such as $f_i(X)$ and $F_i(X)$. If the visualization technology is applied on *X*, we can note the visualization system as *V(X)*.

Definition 1 *(End-to-end Visualization)* Given a algorithm system (*X*,*F*) and visualization technology $V(\cdot)$, the visualization system is an end-to-end visualization if $V(\cdot)$ is directly applied on *G(X,F(X))* where *G* is an outside function that not related to *F*.

Definition 2 *(Internal Visualization)* Given a algorithm system (*X*,*F*) and visualization technology $V(\cdot)$, the visualization system is an internal visualization if $V(\cdot)$ is involved with internal product f_i rather than only the input *X* and output *F(X)*.

Definition 3 *(Multiple Graph)* If a visualization system is consist of different lower level visualization systems, we will use symbol $\oplus$ to note it such that:

$$V(X_1, X_2, ..., X_n) = V_1(X_1) \oplus V_2(X_2) ... \oplus V_n(X_n) \tag{3}$$

IV. Case Study: MSBD5014

In this session, we will use the formula above to introduce my MSBD5014 independent project *Human-eye on Deep Learning*[17] with focus on what different layer of CNN see during a image classification task. The main contribution of my MSBD5014 independent project is to reproduce and organize advanced work to visualize what's inside CNN model. During the project, we developed three methods, respectively they are class activation mapping[11], saliency maps[9], and guided backpropagation[12].

First of all, we declare that most CNN architecture can be regarded as a sequential model. Let's take a example on the simplest one, *AlexNet*[18]. AlexNet is consist of 5 convolution layers(including pooling function) and 3 fully connected layers sequentially, which could be written as:

$$F_{Alex} = \prod_{i=1}^{5} f_{C_i} \circ \prod_{j=1}^{3} f_{FC_j} \tag{4}$$

where f_{C_i} means the i^{th} convolution layer with pooling function, and f_{FC_j} means the j^{th} fully connected layer. Based on this point of view, next we present the three methods mentioned above.

A. Class Activation Mapping

Class activation mapping requires a specific CNN architecture with global average pooling designed before fully connected layer:

$$F = F_{n-k-1} \circ f_{gap} \circ \prod_{i=1}^{k} f_{FC_i} \tag{5}$$

where F_{n-k-1} is the general convolution architecture in the first n-k-1 layer, f_{gap} is the gobal average pooling(GAP) function followed up and $\prod_{i=1}^{k} f_{FC_i}$ is the last k fully connected layers. What class activation mapping do is to remove the last GAP layer, letting the fully connected layer applied on the feature maps rather than the numerical value through GAP:

$$F_{CAM_Y} = F_{n-k-1} \circ \prod_{i=1}^{k} f_{FC_{Y\,i}} \tag{6}$$

where Y is the target index, so that $f_{FC_{Y_i}}$ is the linear combination to target Y rather than to all the targets. Hence, what we get from F_{CAM_Y} is a 2-dimensional filter which is able to localize the image regions from a trained CNN. To conclude, this visualization is $V(X, F_{CAM_Y}(X))$ where $V(\cdot)$ is to resize the 2-dimension matrix $F_{CAM_Y}(X)$ and apply a color map onto original image as Figure 1.

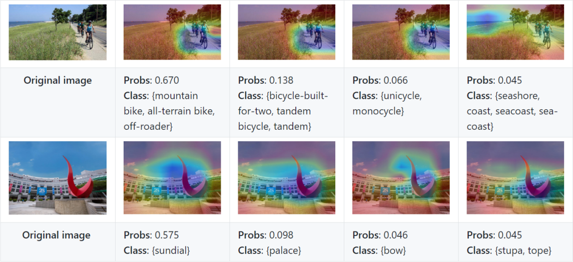

Fig. 1. Class Activation Mapping.

B. Saliency Maps

Unlike class activation mapping, saliency maps can be used in any kind of CNN by taking the derivative of any internal product. Theoretically, we could visualize the relationship between the input image and any pixel inside CNN. In practice, in order to achieve better explanation, we would visualize the relationship between input image X and the target Y like Figure 2 shown. In this case, saliency maps can be written as:

$$F_{SM_Y} = \max_{channels}\left(\frac{\partial F_Y(X)}{\partial X}\right) \tag{7}$$

and what we visualize is $V(F_{SM_Y})$ in which $V(\cdot)$ is to apply a color map.

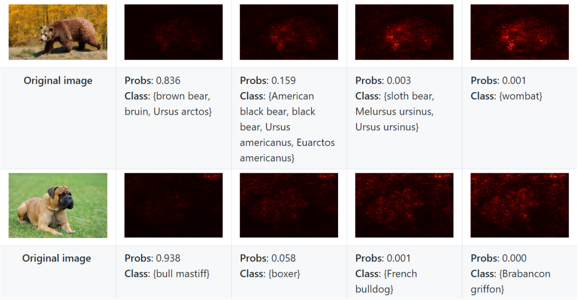
Original image
Probs: 0.836
Class: {brown bear, bruin, Ursus arctos}
Probs: 0.159
Class: {American black bear, black bear, Ursus americanus, Euarctos americanus}
Probs: 0.003
Class: {sloth bear, Melursus ursinus, Ursus ursinus}
Probs: 0.001
Class: {wombat}
Original image
Probs: 0.938
Class: {bull mastiff}
Probs: 0.058
Class: {boxer}
Probs: 0.001
Class: {French bulldog}
Probs: 0.000
Class: {Brabancon griffon}

Fig. 2. Saliency Maps.

C. Guided BackPropagation

The design of guided backpropagation(GBP) is very similar to saliency maps, both computing gradient of neuron value with respect to image pixels. The only diffence is that guided backpropagation only backpropagate positive gradients through each ReLU which makes the result nicer. Indeed both saliency maps and guided backpropagation Visualize image that correspond to maximal activations, which is somehow like a Gobor filter. We have experimented GBP with different CNN networks (eg. AlexNet,VGG,ResNet,DenseNet) and different types of gradient (eg. positive,negative,absolute maximal) which is shown in Figure 3.

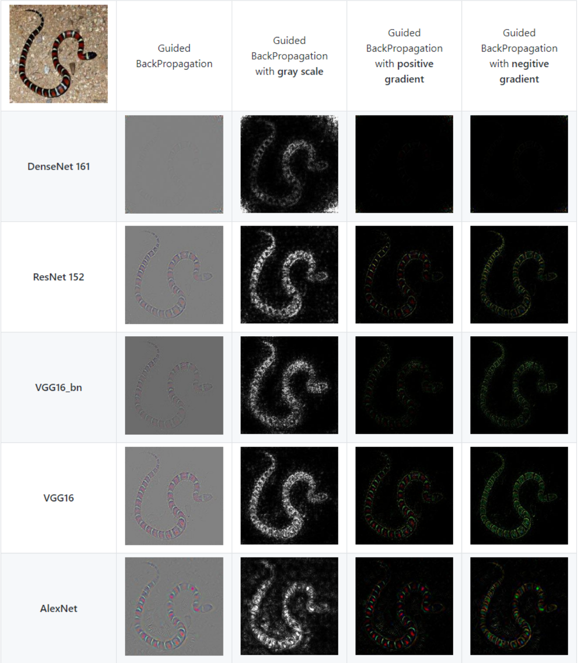
Guided BackPropagation
Guided BackPropagation with gray scale
Guided BackPropagation with positive gradient
Guided BackPropagation with negitive gradient
DenseNet 161
ResNet 152
VGG16_bn
VGG16
AlexNet

Fig. 3. Guided BackPropagation.

The algorithm we visualize here is:

$$F_{GBP_Y} = \prod_i (ReLu \circ \frac{\partial F_{Y_{n-k_i}}(X)}{\partial F_{Y_{n-k_{i-1}}}(X)}) \tag{8}$$

k_i is the number of layers that between the i^{th} ReLu function and $(i-1)^{th}$ ReLu function. Specifically, k_0 is X. The visualization function $V(\cdot)$ is the same as that in saliency maps, which is to apply a color map.

References

[1] Holzinger A, Biemann C, Pattichis C S, et al. What do we need to build explainable AI systems for the medical domain?[J]. arXiv preprint arXiv:1712.09923, 2017.

[2] Liu S, Wang X, Liu M, et al. Towards better analysis of machine learning models: A visual analytics perspective[J]. Visual Informatics, 2017, 1(1): 48-56.

[3] Krause J, Perer A, Ng K. Interacting with predictions: Visual inspection of black-box machine learning models[C]//Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 2016: 5686-5697.

[4] Rudin C. Please Stop Explaining Black Box Models for High Stakes Decisions[J]. arXiv preprint arXiv:1811.10154, 2018.

[5] Explainable AI: The data scientists new challenge

[6] Two Brains Running

[7] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.

[8] Zeiler M D, Taylor G W, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning[C]//Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011: 2018-2025.

[9] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. arXiv preprint arXiv:1312.6034, 2013.

[10] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.

[11] Zhou, Bolei, et al. ”Learning deep features for discriminative localization.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[12] Springenberg, Jost Tobias, et al. ”Striving for simplicity: The all convolutional net.” arXiv preprint arXiv:1412.6806 (2014).

[13] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization[C]//ICCV. 2017: 618-626.

[14] Ming Y, Cao S, Zhang R, et al. Understanding hidden memories of recurrent neural networks[J]. arXiv preprint arXiv:1710.10777, 2017.

[15] Ribeiro M T, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016: 1135-1144.

[16] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]//Advances in Neural Information Processing Systems. 2017: 4765-4774.

[17] Zhou J, Human-eye on Deep Learning, Github Link

[18]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[19]”Visxai Workshop At Ieee Vis 2018.” Insert Name of Site in Italics. N.p., n.d. Web. 10 Jan. 2019