**Book**: XAI: Interpreting, Explaining, and Visualizing Deep Learning – State of the Art Survey.

Reading the scoring System – Batarseh

Read the paper on XAI Method 10 vs 9 DL

Apply SHAP on simple methods – Pytorch

Apply: Understanding Neural Networks via Feature Visualization: A Survey – Chapter 2

-----------------------------------------------------------------------------------------------------------------------------------’

**References**

1. Alber, M.: **Software and application patterns for explanation methods**. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS, vol. 11700, pp. 399–433, Springer, Cham (2019)

2. Alber, M., et al**.: iNNvestigate neural networks!.** J. Mach. Learn. Res. 20, 1–8 (2019)

3. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: **Gradient-based attribution methods.** In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS, vol. 11700, pp. 169–191, Springer, Cham (2019)

4. Graves, A., Mohamed, A., Hinton, G.: **Speech recognition with deep recurrent neural networks**. In: IEEE ICASSP, pp. 6645–6649 (2013)

5. Krizhevsky, A., Sutskever, I., Hinton, G.E**.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems**, pp. 1106–1114 (2012)

6. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.-R**.: Layer-wise relevance propagation: An overview**. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS, vol. 11700, pp. 193–209, Springer, Cham (2019)

7. Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.-R., Tkatchenko, A**.: Quantumchemical insights from deep tensor neural networks**. Nat. Commun.

8, 13890 (2017) 8. Thomas, A.W., Heekeren, H.R., Müller, K.-R., Samek, W**.: Analyzing neuroimaging data through recurrent deep learning models (2018**). arXiv preprint arXiv:1810.09945

**1/ Preface:**

* Before deploying an AI system, we see a strong need to validate its behavior, and thus establish guarantees that it will continue to perform as expected when deployed in a real-world environment.
* More recent deep learning based neural networks provide far superior predictive power, but at the price of behaving as a ‘black-box’ where the underlying reasoning is much more difficult to extract

**References**

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: **Concrete Problems in AI Safety.** arXiv preprint arXiv:1606.06565 (2016)

2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: **On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation**. PLoS ONE 10(7), e0130140 (2015)

3. Fong, R.C., Vedaldi, A**.: Interpretable explanations of black boxes by meaningful perturbation**. In: IEEE International Conference on Computer Vision (CVPR). pp. 3429–3437 (2017)

4. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R.: **Unmasking clever hans predictors and assessing what machines really learn.** Nature Communications 10, 1096 (2019)

5. Montavon, G., Samek, W., Müller, K.-R.: **Methods for interpreting and understanding deep neural networks**. Digital Signal Processing 73, 1–15 (2018)

6. Rosenblatt, F.: **The perceptron, a perceiving and recognizing automaton (Project Para).** Report No. 85-460-1. Cornell Aeronautical Laboratory (1957)

7. Samek, W., Müller, K.-R**.: Towards explainable artificial intelligence**. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS, vol. 11700, pp. 5–22. Springer, Cham (2019)

8. Zeiler, M.D., Fergus, R**.: Visualizing and understanding convolutional networks**. In: European Conference Computer Vision (ECCV). pp. 818–833 (2014)

9. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.-R.: How to explain individual classification decisions. J. Mach. Learn. Res. 11, 1803–1831 (2010)

10. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: ICLR Workshop (2014)

**Chapter 1** : Towards AI Transparency:

* Towards Explainable AI
* Transparency: Motivations & Challenges
* Interpretability in Intelligent Systems – A New Concepts

**1/ Towards XAI:**

* the urge to understand their complex and non-linear models
* With the emergence of kernel machines and deep learning, the wish for AI transparency grew even stronger.
* With explainable AI it may be possible to also identify such novel patterns and strategies in domains like health, drug development or material sciences, moreover, the explanations will ideally let us comprehend the reasoning of the system and understand why the system has decided.

**2/ Need for Transparency and Trust in AI:**

* if the AI system is a black box, it is very difficult to unmask such predictors. Explainability helps to detect these types of biases in the model or the data, moreover, it helps to understand the weaknesses of the AI system
* Explanations Foster Trust and Verifiability:
  + In cases where the AI system itself is deciding, it is even more critical to be able to comprehend the reasoning of the system in order to verify that it is not behaving like Clever Hans, but solves the problem in a robust and safe manner.
  + An AI system which interacts with humans should therefore be explainable
* Explanations are a prerequisite for new insights
* Explanations are part of the legislation:
  + The EU’s General Data Protection Regulation (GDPR) has even added the right to explanation to the policy in Articles 13, 14 and 22, highlighting the importance of human-understandable interpretations derived from machine decisions.
* **Different Facets of an Explanation:**
  + Recipient: Researcher may need different info from FDA.
  + **Info Content – 4 types of explanations:** Depending on the recipient of the explanations and his or her intent, it may be advantageous to focus on one particular type of explanation.
    - **1. Explaining learned representations:** 
      * This type of explanation aims to foster **the understanding of the learned representations**, e.g., neurons of a deep neural network. investigates the role of single neurons or group of neurons in encoding certain concepts.
        + Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6541–6549 (2017
        + Kim, B., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning (ICML), pp. 2673–2682 (2018)
      * Other methods aim to interpret what the model has learned **by building prototypes that are representative of the abstract learned concept.**
        + explain what the model has learned about the category “car” by generating a prototypical image of a car.
        + Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 3387–3395 (2016)
        + Nguyen, A., Yosinski, J., Clune, J.: Understanding neural networks via feature visualization: a survey. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., M¨uller, K.-R. (eds.) Explainable AI. LNCS, vol. 11700, pp. 55–76. Springer, Cham (2019)
        + Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: ICLR Workshop (2014)
        + Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
    - **2. Explaining individual predictions:**
      * Other types of explanations provide information about individual predictions, e.g., **heatmaps visualizing which pixels have been most relevant for the model to arrive at its decision**
        + Montavon, G., Samek, W., M¨uller, K.R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Process. 73, 1–15 (2018)
      * pixels have been most relevant for the model to arrive at its decision
        + Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: ICLR Workshop (2014)
      * Such explanations help to verify the predictions and establish trust in the **correct functioning on the systems and establish trust in the correct functioning on the system**.
      * Layer-wise Relevance Propagation (LRP) [9,58] provides a general framework for explaining individual predictions, i.e., it is applicable to various ML models, including neural network
    - **3. Explaining model behavior:**
      * This type of explanations go beyond the analysis of individual predictions towards a **more general understanding of model behaviour,** e.g., identification of distinct prediction strategies
        + The spectral relevance analysis (SpRAy) computes such metha Each cluster then represents a particular prediction strategy learned by the model

Lapuschkin, S., W¨aldchen, S., Binder, A., Montavon, G., Samek, W., M¨uller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nat. Commun. 10, 1096 (2019)

* + - * + Such explanations are useful for obtaining a global overview over the learned strategies and detecting “Clever Hans” predictors
    - **4. Explaining with representative examples:**
      * This type of explanations can be useful for obtaining a better **understanding of the training dataset** and **how it influences the model.** Furthermore, these representative examples can potentially help to identify biases in the data and make the model more robust to variations of the training dataset
        + Khanna, R., Kim, B., Ghosh, J., Koyejo, O.: Interpreting black box predictions using fisher kernels. arXiv preprint arXiv:1810.10118 (2018)
        + Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International Conference on Machine Learning (ICML), pp. 1885–1894 (2017)

**3/ Methods of XAI:**

* **model-agnostic** and rely on a simple surrogate function to **explain the predictions**
* methods which **compute explanations** by **testing the model’s response to local perturbations** (e.g., by utilizing gradient information or by optimization
* we present very **efficient propagation-based explanation** techniques which leverage the model’s internal structure.
* we consider methods which go beyond individual explanations towards a **meta-explanation of model behaviour**

**a/ Explaining with surrogates – model agnostic and explain predictions:**

* One approach to explain the predictions of complex models is to locally approximate them with a simple surrogate function, which is interpretable.
* Local Interpretable Model-agnostic Explanations (LIME)
  + Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: explaining the predictions of any classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
  + This method samples in the neighborhood of the input of interest, evaluates the neural network at these points, and tries to fit the surrogate function such that it approximates the function of interest
  + it requires several minutes for computing the explanation of a single prediction
* SmoothGrad does not leverage the internals of the model, however, it needs access to the gradients
  + Smilkov, D., Thorat, N., Kim, B., Vi´egas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)

**b/ Explaining with Local Perturbations – Explain model’s response to local change to the prediction - model agnostic – computational demanding:**

* Another class of methods construct explanations by analyzing the model’s response to local changes. This includes methods which utilize the gradient information as well as perturbation- and optimization-based approaches.
* One example is the so-called Sensitivity Analysis (SA) – **Not computational feasible**
  + Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., M¨uller, K.R.: How to explain individual classification decisions. J. Mach. Learn. Res. 11, 1803–1831 (2010)
  + Morch, N., et al.: Visualization of neural networks using saliency maps. In: International Conference on Neural Networks (ICNN)
  + Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: ICLR Workshop (2014)
* Variants of Sensitivity Analysis exist which tackle some of these problems by locally averaging the gradients
  + Smilkov, D., Thorat, N., Kim, B., Vi´egas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2
* While the occlusion method of [94] measures the importance of input dimensions by masking parts of the input, the Prediction Difference Analysis (PDA) approach of [97] uses conditional sampling within the pixel neighborhood of an analyzed feature to effectively remove information
  + Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: prediction difference analysis. In: International Conference on Learning Representations (ICLR) (2017)

**c/ Propagation-based Approach (Leveraging Structure) – Analyze DATA:**

* This methods does not explain the model directly but integrate the internal structure of the model into the explanation process
* **LRP:**
  + is a propagation-based explanation framework, which is applicable to general neural network structures
  + Bach, S., Binder, A., Montavon, G., Klauschen, F., M¨uller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7), e0130140 (2015)’
  + Binder, A., Bach, S., Montavon, G., M¨uller, K.-R., Samek, W.: Layer-wise relevance propagation for deep neural network architectures. Information Science and Applications (ICISA) 2016. LNEE, vol. 376, pp. 913–922. Springer, Singapore (2016). [https://doi.org/10.1007/978-981-10-0557-2 87](https://doi.org/10.1007/978-981-10-0557-2%2087)
  + . Montavon, G., Binder, A., Lapuschkin, S., Samek, W., M¨uller, K.-R.: Layer-wise relevance propagation: an overview. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., M¨uller, K.-R. (eds.) Explainable AI. LNCS, vol. 11700, pp. 193– 209. Springer, Cham (2019)
  + LRP explains individual decisions of a model by propagating the prediction from the output to the input using local redistribution rules.
  + **“how much did the input feature contribute to the prediction”**
* The **iNNvestigate** toolbox [1] provides an efficient implementation for many of these propagation-based explanation methods
  + Alber, M., et al.: iNNvestigate neural networks!. J. Mach. Learn. Res. 20(93), 1–8 (2019)

**d/ Meta-explanations – find prediction strategy – DATA + MODEL:**

* A recently proposed method, spectral relevance analysis (**SpRAy**) [46], computes such meta explanations by clustering individual heatmaps.
* This approach allows to **investigate the predictions strategies of the classifier on the whole dataset** in a (semi-)automated manner and to systematically find weak points in models or training datasets.
* Another type of meta-explanation aims to better understand the learned representations and to provide
  + . Lapuschkin, S., W¨aldchen, S., Binder, A., Montavon, G., Samek, W., M¨uller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nat. Commun. 10, 1096 (2019)
  + Cire¸san, D., Meier, U., Masci, J., Schmidhuber, J.: A committee of neural networks for traffic sign classification. In: International Joint Conference on Neural Networks (IJCNN), pp. 1918–1921 (2011)
  + Zhou, B., Bau, D., Oliva, A., Torralba, A.: Comparing the interpretability of deep networks via network dissection. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., M¨uller, K.-R. (eds.) Explainable AI. LNCS, vol. 11700, pp. 243–252. Springer, Cham (2019)
  + Kim, B., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning (ICML), pp. 2673–2682 (2018)

**4/ Evaluating Quality of Explanation:**

* An alternative and indirect way to evaluate the quality of explanations is to use them for solving other tasks.
* Lastly, another promising approach to evaluate explanations is based on the fulfillment of a certain axioms. Axioms are properties of an explanation that are considered to be necessary and should therefore be fulfilled.
  + relevance conservation
  + explanation continuity:
    - Montavon, G., Samek, W., M¨uller, K.R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Process. 73, 1–15 (2018)
  + sensitivity
  + implementation invariance
    - . Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (ICML), pp. 3319–3328 (2017)

**5/ Challenges and Open Question:**

* Heat Map:
  + First, heatmaps computed with today’s explanation methods visualize “first-order” information, i.e., they show which input features have been identified as being relevant for the prediction.
  + However, the relation between these features, e.g., whether they are important on their own or only whether they occur together, remains unclear.
* Another limitation is the low abstraction level of explanations.
  + Heatmaps show that particular pixels are important without relating these relevance values to more abstract concepts such as the objects or the scene displayed in the image
  + Humans need to interpret the explanations to make sense them and to understand the model’s behaviour. This interpretation step can be difficult and erroneous
  + Some works (e.g., [43]) have already started to investigate human factors in explainable AI.0
    - **. Lage, I., et al.: An evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1902.00006 (2019)**

0