Explainable and Interpretable ML (REVIEW)

Tejas Gaikwad(MT19AI021)
Dept. of Computer Science and Engineering
Indian Institute of Technology, Jodhpur

Interpretable Machine Learning for Computer Vision

- **Introduction to interpretable Machine Learning** by Been Kim, Google Brain

  This discussion arises as to the answer to the question, Is your machine learning system well known to you, Do we know on what basis the model is giving the decision. Well, the decision tree and linear models can be a solution but when a lot of examples are there, then it would be difficult to know the most important feature and now a day, the capabilities of machines are so huge that they can process huge amount of data, and to know what feature has resulted to what prediction, it is nearly impossible to do it with the help of tools like a decision tree. No one method fits all method. There was a paper published in the year 2008 stated that our machines will now have the knowledge, the world can never be able to understand. There is also a misconception that we need to know every detail of the model mathematically to know the learnings of the model which is not true. In fact, interpretability is about knowing enough for your goals. The goal is to have our values aligned and our knowledge should be reflected. Another misunderstanding is, more is the data for training more cleverly the algorithm will solve interpretability. Also, every model should be interpretable is not the compulsory condition as there might be some operation whose output or result is not critically impactful. Interpretability includes fairness, accountability, trust, causality but the inverse is untrue.
  There are basically 3 types of interpretable models. The first one is before building a model like data visualization. We can have 2D plot, clustering, etc. type of methods to visualize the categories of the data we are having. The second one is while the model is being built. These can be the rules, examples, Sparsity, and Monotonicity, like we can set some thresholds, rules to visualize the effect over the data, sparsity, etc. The last is the one when the model is built. We can visualize the last layer of the trained model to know the parameter the model has learned, or we can have saliency maps, heat maps, etc. to visualize the trained parameter. Now we know where to find the interpretability, now the question arises, how to evaluate the interpretation of the model, i.e. whether they are interpreting what they actually need to. For this there are plenty of methods like we can have Intersection over union(IoU) of the selected unclassified object over the one present in the dataset will labeling which means we can formulate an experiment where we can have ground truth.

- **Using t-SNE to understand Vision Models** by Laurens van der Maaten, (Facebook AI Research)

  t-SNE can be utilized to visualize high dimensional data by bringing it to a 2-dimensional or 3-dimensional space, understandable by humans. But, this involves a lot of complexities several Do's and Dont's to use high dimensional data. Tough PCA is also used for the same purpose but PCA has its own limitations and drawbacks. PCA is very restrictive, it learns linear mapping. It focuses on preserving large pairwise distances. Whereas t-SNE compute pairwise similarities between data with a normalized Gaussian kernel, then measure normalized Student-t similarities in the t-SNE map and lastly minimize the divergence between both distributions. In this way, we can have dimensionality reduction using t-SNE. There is some loss associate with this, which as the Kullback-Leibler divergence preserves local data structure and the heavy-tailed distribution corrects volume differences between both space, because of which the loss occurs. t-SNE should be used to

get some qualitative hypotheses on what your features capture. The creativity to visualize the outputs of t-SNE like if we plot the MNIST number dataset into a 2-D scatter plot with the same color vs. the one with unique colors to the numbers, we can easily visualize which color represents what. t-SNE should be used to present proof of the decision made, instead of just showing different outputs which have no meaning to the prediction provided. If there are some irregularities, t-SNE can be used to give a reason for that irregularities, it should be used to present a proof of concept instead of just giving some unmeaningful plots and hypothesizing something out of it without any base for it. t-SNE can help us to find outliers, or assign meaning to point densities in clusters, etc. To conclude, t-SNE is a powerful tool to generate hypotheses and understand it but it does not produce conclusive evidence.

- **Importance of Individual Units in CNNs:** Bolei Zhou(MIT)

  This talk discusses the visualization in the deep learning models. It basically answers the way to visualize the learned parameters of Deep Neural Network. The debate on if interpretability is really necessary has caught the eyes of many researchers who started their research in Interpretability and Visualizing the models. Interpretability of a deep neural network is necessary for the safety of AI models, to trust the decisions made by the model in some critical cases like medical diagnosis where false prediction may cause deaths, it becomes utmost importance. It is also required for the Policy and Regulation, i.e. right to the explanation for algorithmic decisions, thus interpretability is a need in today's world of AI.
  Interpretability can be of various types. One would relate to a network as a whole and observe the accuracies on the test, train, and validation set to check if overfitting or underfitting is taking place or not. Observing the performance of the model at different combinations of hyperparameters is also a part of interpreting the model. Also, observing the feature space by dimension reducibility as discussed above in the previous section can be one of the important methods to interpret what model is learning or trying to learn. Then comes, understanding the networks at different granularity using individual units. Units refer to a particular set of the object in an image like trees, streets doors, windows, etc. This visualization can be done by several methods like deconvolutions(its just opposite to convolutional, flipped horizontally, it helps to regenerate the trained or learned parameters. Backpropagation, image synthesis, gradient-based visualization are also some methods for visualization of the deep models. The units based visualization can be obtained by taking ratio over the heat maps of the neurons from the final dense layer which results in showing the object which has been activated in that particular neuron and the other one with the segmented image, with the help of this ration which is called as intersection over union, the activated objects wrt each neuron can be identified.More the value of this intersection better confidence the model can give over the object. The model gives a ranking of several objects probabilities and selects one with the highest probability. Also, one can visualize the effect on these predictions by removing the objects in an image and see the confidence over that image, if the confidence over the prediction of a particular image is large then it means that the object was a critical element of an image.

- **Understanding models via visualizations, attribution, and semantic identification:** Dr. Andrea Vedaldi

Dr. Vedaldi explained the representation and understanding of the model via visualization, attribution, and semantic identification.

The talk started with a question, How much information about Image X does the output side layer Y contains. This can be done by reconstruction of the image with the help of the output layer itself. But the reconstruction of the same image is not a good idea to visualize as it won't be interpretable. But by using a reference image and starting with a noise sample, minimizing the distance between them may result in getting an output that is somewhat similar to class to which it actually belongs. Several possible implementations for the above can be by Regularised energy (Understanding deep image representations by inverting them Mahendran Vedaldi, CVPR, 2015), Constrained optimization (Deep image prior Ulyanov Vedaldi Lempistky,
CVPR, 2018) and posterior probability (Plug & play generative networks: Conditional iterative
generation of images in latent space Nguyen, Yosinksi, Bengio, Dosovitskiy, Clune, CVPR, 2017). These methods basically discuss a loss function which can be used to obtain better results for reconstruction of the image by visualizing the output layer. To me, it is somewhat like GAN's. In the presentation, the speaker Dr. Vadaldi has shown the visualization of the CNN network at different layers. In that, as the model moves away from the input layer, the visualization becomes lesser interpretable. At the final fully connected layer, it becomes almost unidentifiable.

Saliency maps(Silent Features of the image) show the pixels which are critically responsible for the predictions. These maps can be obtained with the help of gradients, gradients prove a local approximation of the model and there are 3 popular methods(as mentioned by the speaker) for it, which include Deconvolution, gradient(backpropagation), and guided backpropagation. The only difference in these methods is how used activation function Relu is reversed. Better channel specificity can be achieved by backpropagation only a few layers. The performance for saliency map methods can be obtained by computing the degree of correlation between results and the ground truth semantic labels like we did in the previous cases to identify objects.

Tutorial on Interpreting and Explaining Deep Models in Computer Vision

An interpretable model helps you to understand and account for the factors that are not included in the model and account for the context of the problem when taking actions based on model predictions. Improving generalization and performance. Now the question comes, Why interpretability is necessary? What benefits we can have if our model is interpretable?. Interpretability verifies the classifier if it is working as per the expectations or not, as sometimes decisions can e costly and dangerous like autonomous car crashes as it recognizes a wall as some other object. Similarly, in the case of medical fields where if AI plays a role, then it becomes very crucial for the model to give accurate results. Thus if the designer of the model knows what the model is failing in, it can improve classifiers to make some accurate results. Also, there are times when we come across a decision that has never been expected from a given model, so then it becomes a necessity to interpret what the model is actually learning.

While predicting a particular class, observing the trained parameter class in the feature space, shows that points lying nearby those feature points, this method of interpreting is known as Class Prototypes. When the points in the feature space are very close to the decision boundary, the model still predicted correct classifier, this can be further understood by observing the heatmaps of the predicted image, we will come to know that the prediction made will have majority bright colors like white and yellow that resemble more activated function, this method is known as an individual explanation. Whereas in sensitivity analysis, the relevance of the input feature is given by the square partial derivative. Sensitivity explains the importance of pixels in an image by reducing or enlarging it. This method explains a variation of the function, not the function value itself. It doesn't highlight the object in the image. For the neural network explanation of the predictions, LayerWise Relevance Propagation(LRP) is one of the popular methods. It explains nonlinear classifiers based on generic theory related to Tayler decomposition and deep Taylor decomposition which is applicable to any neural network with monotonous activation, Bag of Words models, Fishers Vectors, SVMs, etc. In short, it basically tells which pixel contributes to how much to the prediction. There are some more methods that discuss the explanation towards a prediction done by the model which we have already discussed above like deconvolution, backpropagation gradient, etc.

Now that we have soo many models which one to choose and when that is a question. For that, these are methods can get compared with the ground truth and analyze the minimum error obtained. Conservation of the input features should be proportional to the number of explainable evidence at the output. If a neural network is certain about a prediction, input featured are either relevant(positive) or irrelevant(zero). If there is an object then the explanation scores should be continuous. The model must agree with the explanation that if the input features are relevant to the attributes, removing them should reduce evidence at the output. In fact, there are some methods too which do not require any experiments to perform unlike the other methods which require testing empirically, directly from the equation we can deduce some properties. Like in the example of LRP-$\alpha\beta$

propagation rule which satisfies the property of an explanation. Here sensitivity analysis and gradient*input have some crucial limitations as well. The suitable LRP-$\alpha\beta$ propagation rule can be seen as performing a deep Taylor decomposition for deep ReLU nets. The deep Taylor decomposition allows us to consistently extend the framework to new models and new types of data.