New Jersey Institute of Technology
Deep Learning
CS677
Nov 22, 2020

Project 2: Explainability

By-

- Sean Walsh, ucid: sw522
- Jonathan VIdal, ucid: jkv5
- Nayem Paiker, ucid: nrp66

Experts in the field of AI and Deep Learning have always struggled to find a common ground between the high accuracy and the interpretability of a model. Correctly interpreting a model's output is extremely important because it gives insight into how a model can be improved, and supports understanding of the process being modeled. Due to the simplicity of interpretation, simpler models are preferred although those can be less accurate than other complex models. To address this problem, this paper discusses a framework called SHAP, SHapley, Additively exPlanations, an unified framework for interpreting predictions.

SHapley refers to Shapley value which is a solution concept in cooperative game theory. The theory states that in such a game you can assign a unique distribution among players that has contributed to the total surplus generated by the group. The SHAP explanation method computes Shapley values by replacing the concept of players with feature values of a data. The goal of SHAP is to explain the prediction of the model by computing the contribution of each feature that went into that decision.

SHAP assigns each feature of the dataset an importance value for a particular prediction. The novel components of SHAP model include the identification of a new class of additive feature importance measures with a set of desirable properties, the theoretical results show an unique solution in the current class. SHAP brings clarity to the growing space of methods by bringing the perspective of viewing any explanation of a model's prediction is introduced as a model itself and is called the explanation model.which helps to define the additive feature attribution methods. The new SHAP values demonstrate being better and clearer outputs

Additive Feature Attribution methods: For a simple model, the best explanation is the model itself, since it is easy to understand. But complex models, such as Ensemble methods or Deep Networks, are different and unlike simple models, the model itself is not the best explanation. An explanation model is needed to interpret the approximation of the original model. Additive feature attribution methods have an explanation model which is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \qquad (1)$$

where $z' \varepsilon \{0, 1\}^M$, M is the number of simplified input features, and φi ∈ R. Methods with this explanation model attribute an effect ϕ_i to each feature and summing the effects all feature attributions approximates the output f(x) of the original model. Any current methods fit this description.

LIME: Lime interprets individual model predictions based on locally approximating the model around a given prediction. The local linear model that LIME adheres to equation

(1). LIME refers to simplified inputs x' as 'interpretable inputs' and mapping and $x = h_x(x')$ converts the binary vector of interpretable inputs into original input space. To find ϕ , LIME minimizes the following objective function:

$$\varepsilon = argmin\, L(f,\, g,\, \pi_{x'}) + \Omega(g) \quad (2)$$

The faithfulness of the explanation model $g(z')$ to the original model $f(h_x(z'))$ is enforced through loss L over a set of samples in the simplified input space weighted by the local kernel $\pi_{x'}$ and Ω penalizes the complexity of g.

DeepLIFT: It is a recursive prediction explanation method, for which the mapping $x = h_x(x')$ converts binary values into the original inputs, where 1 indicates that an input takes its original value, and 0 indicates that it takes the reference value. e. The reference value, though chosen by the user, represents a typical uninformative background value for the feature. DeepLIFT uses a "summation-to-delta" property that states:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta o} = \Delta O \quad (3)$$

where o = f(x) is the model output, Δo = f(x) − f(r), $\Delta x_i = x_i - r_i$ Δxi = xi − ri , and r is the reference input. If we let $\Omega_i = C_{\Delta x_i \Delta o}$ and $\Omega_0 = f(r)$, then DeepLIFT's explanation model matches equation 1 and is thus another additive feature attribution method.

Layer-Wise Relevance Propagation: , this method is equivalent to DeepLIFT with the reference activations of all neurons fixed to zero. Thus, $x = h_x(x')$ converts binary values into the original input space, where 1 means that an input takes its original value, and 0 means an input takes the 0 value. Layer-wise relevance propagation's explanation model, like DeepLIFT's, matches equation 1.

Classic Shapley Value Estimation: Shapley regression values are feature importances for linear models in the presence of multicollinearity. This method requires retraining the model on all feature subsets S ⊆ F, where F is the set of all features. It assigns an importance value to each feature that represents the effect on the model prediction of including that feature. e. To compute this effect, a model $f_{SU\{i\}}$ is trained with that feature present, and another model f_s is trained with the feature withheld. Then, predictions from the two models are compared on the current input $f_{SU\{i\}}(x_{SU\{i\}}) - f_s(x_s)$. The preceding differences are computed for all possible subsets S ⊆ F \ {i} and shapley values are then computed and used as feature attributions.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{SU\{i\}}(x_{SU\{i\}}) - f_S(x_S) \right] \quad (4)$$

Since the explanation model form of Shapley sampling values is the same as that for Shapley regression values, it is also an additive feature attribution method. Quantitative input influence is a broader framework that addresses more than feature attributions. However, as part of its method it independently proposes a sampling approximation to Shapley values that is nearly identical to Shapley sampling values. It is thus another additive feature attribution method.

Simple Properties Uniquely Determine Additive Feature Attributions:

Property 1 (Local accuracy): This is the first property. . When approximating the original model f for a specific input x, local accuracy requires the explanation model to at least match the output of f for the simplified input x'.

$$f(x) = g(x^i) = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \quad (5)$$

The explanation model $g(x^i)$ matches the original model f(x) when $x = h_x(x')$.

Property 2 (Missingness): Missingness is the second property. s. If the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact.

$$x'_i = 0 \Rightarrow \phi_i = 0 \quad (6)$$

Missingness constraints features where $x'_i = 0$ to have no attributed impact.

Property 3 (Consistency): This is the third property, which states that t if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.) Let $f_x(z') = f(h_x(z'))$ and $z' \setminus i$ denote setting $z'_i = 0$. For any two models f and f' , if

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \quad (7)$$

for all inputs $z^i \varepsilon \{0, 1\}^M$, then $\phi_i(f', x) \geq \phi_i(f, x)$.

Theorem 1: Only one possible explanation model g follows equation 1 and satisfies all 3 properties.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{|M|!} [f_x(z') - f_x(z' \setminus i)] \quad (8)$$

where $|z^i|$ is the number of non-zero entries in z^i , and $z^i \subseteq x^i$ represents all z^i vectors where the non-zero entries are a subset of the non-zero entries in x'.

SHAP (SHapley Additive exPlanation) Values: As mentioned before, SHAP values, as unified measure of feature importance, are the Shapley values of a conditional expectation function of the original model and are the solution to equation 8.

Model-Agnostic Approximations:

Kernel SHAP (Linear LIME + Shapley values): Linear LIME model, which is an additive feature attribution method, uses a linear linear explanation model to locally approximate f, and the local is measured in the simplified binary input space. Shapley values are the only solution to equation 2 which satisfies local accuracy. If the LIME choice for these parameter values are made heuristically, using the choice of Loss function L, weighting kernel $\pi_{x'}$ and regularization term Ω, equation 2 cannot recover the shapley values. Due to that the local accuracy and consistency are violated, which leads to some unexpected unintuitive behavior in certain circumstances. In order to avoid heuristical choosing of parameters in equation 2, and to find Loss function L, weighting kernel $\pi_{x'}$ and regularization term Ω to recover shapley values, the theorem described below can be followed.

Theorem 2: Under the definition of Additive Feature Attribution methods, the specific forms of Loss function L, weighting kernel $\pi_{x'}$ and regularization term Ω that makes solutions of equation 2 consistent with local accuracy and consistency are:

$$\Omega(g) = 0$$

$$\pi_{x'}(z') = \frac{(M-1)}{(M \text{ choose } |z'|)|z'|(M-z')}$$

$$L(f, g, \pi_{x'}) = \sum_{x' \varepsilon Z} [f(h_x^{-1}(z')) - g(z')]^2 \pi_{x'}(z')$$

where $|z'|$ is the number of non-zero elements in z'.

It is important to note that $\pi_{x'}(z') = \infty |z'| \varepsilon \{0, M\}$, which enforces $\phi_0 = f_x(\Theta)$ and $f(x) = \sum_{i=0}^{M} \phi_i$

Model-Specific Approximations:

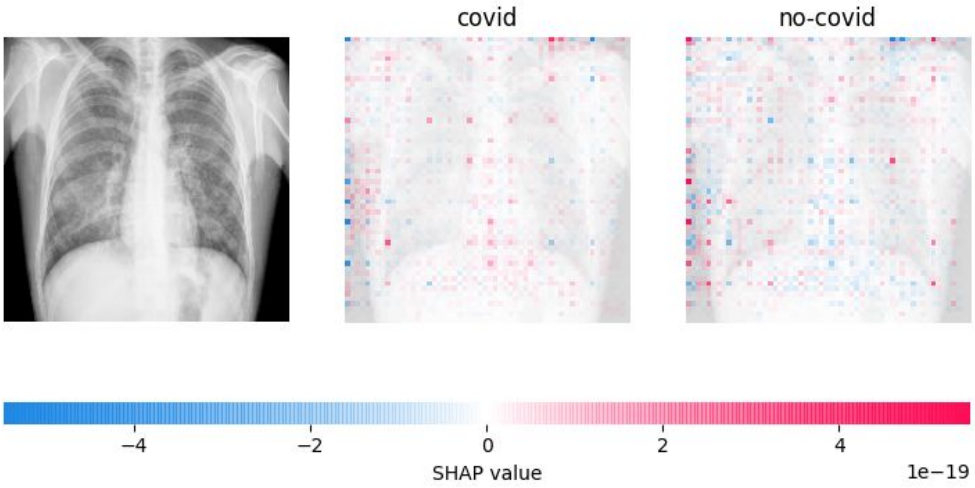
Kernel SHAP is a model-agnostic method to approximate SHAP values using ideas from LIME and Shapley values. It uses a specially-weighted local linear regression to estimate SHAP values for any model. While this model improves the estimation of SHAP values, other four model-type-specific approximation methods can be developed faster as follows.

Linear SHAP and Low-Order SHAP:

For linear models, Linear SHAP, SHAP values can be approximated directly from the model's weight coefficients. In smaller values of the number of simplified input features, Low-Order SHAP, it is efficient to choose an approximation of the conditional expectorations.

Max SHAP and Deep SHAP:
Max SHAP and Deep SHAP are the novel approximation methods. With a permutation formulation of Shapley values, Max SHAP is optimized by calculating the probability on each input value. This increases the maximum value over the next input value. In addition, the connection between the Shapley values and DeepLIFT improves computation performance on deep network models if the model is linear. Since DeepLIFT is an additive feature attribution method and Shapley value represents only the attribution value, DeepLIFT becomes a compositional approximation of SHAP value. Deep SHAP computes by smaller components of the networks.

SHAP Experiment on COVID 19 X-ray Image
After training the model using the covid-cxr codebase we are able to leverage those weights by re-creating the model and analyze it using SHAP. Since this is a multi-layered CNN we need to choose a particular layer to analyze since each one would be analyzing different types of features and in this case we have chosen layer 7. SHAP comes with several “explainers” and one must be chosen based on the data and model. Given that we are dealing with images the gradient explainer makes the most sense. An image gradient is a directional change in the intensity or color in an image. This will help us identify which areas of the image were identified as the highest importance when making the prediction. The explainer output can then be mapped over the original image to allow a human, even one with no particular knowledge of the model, to easily see what features of the image were prioritized. For this particular example, a medical professional can use their own expertise to determine if they agree with where it is focusing and make a judgement on whether they have confidence in the results.



Predictions for two input images are explained in the plot above. Red pixels represent positive SHAP values that increase the probability of the class, while blue pixels represent negative SHAP values that reduce the probability of the class. Given that we used a binary classifier we see the two outputs.

Conclusion:

To conclude, this paper introduced SHAP which stands for SHapley Additive exPlanations. It is the most popular model agnostic technique that is used to explain predictions. SHAP is more popular and reliable technique as it offers local accuracy as the explanation model should match the original model, missingness means that features missing in the original input have no impact, and consistency as simplified input's contribution increases or stays the input's attribution should not decrease. In addition, we present different estimation methods for SHAP value with examples and visualizations.