XPlainIT - A demonstrator for explaining deep learning models trained over structured data

Susan Mckeever, Ihsan Ullah, Andre Rios
CeADAR Ireland's Center for Applied AI, Technological University Dublin, Dublin, Ireland.
Email: {susan.mckeever,ihsan.ullah,andre.rios}@tudublin.ie

***Abstract*—The XplainIT demonstrator is about enabling users to understand and visualise the decisions of deep learning models. The novelty of our work is the application of deep learning networks (1-D CNN) to structured data - and the subsequent use of Layer-Wise Relevance Propagation (LRP) as an explainability approach for structured data input. LRP is a highly visual technique typically used for image input. Our XplainIT demonstrator shows it is applicable, useful, and intuitive for structured data input - thus widening its scope to many modeling scenarios. In addition to LRP, XPlainIT applies established explainability techniques, SHAP and LIME, to allow users to understand and compare the capabilities across multiple explainability techniques. XPlainIT provides trained models for two business case scenarios - fraud detection and customer churn. Users of the demonstrator can interact with local instance level and global dataset level explanations, across three explanation approaches.**

I. Background

The Explainability of neural networks has grown in parallel with the growth in complexity and depth of networks. With the re-emergence of neural networks for deep learning (DL) networks, explainability approaches such as the work of Zeiler and Fergus [1], [2], [3] focused on the visualisation and understanding of mid and high-level features learned by a network for computer vision. Since then, several explainability techniques have been introduced, such as Local interpretable model-agnostic explanations (LIME) [4], Shapley additive explanation (SHAP) [5] (for other types of data e.g. structured data) and class activation map (CAM), gradient weighted-CAM (Grad-CAM) [6], layer-wise relevance propagation (LRP) [7] (input as image).

LIME and SHAP have been widely applied to explain predictions from traditional ML algorithms, for all types of data. Both techniques use an algorithm that trains over the output layer of the neural network to explain network learning. For DL models that process image input, Gradient-Weighted CAM (Grad-CAM) [6] is a generalization of CAM that can target any layer and introduces gradient information to CAM. Gradient information is combined with class activation maps to visualize the importance of each input. The majority of explainability approaches that explore DL networks with input as an image, however, focus on the high-level layers in the process, resulting in coarser visualization. The explainability techniques used in vision are sometimes used for other application areas. Assaf and Schumann [8], for example, presented a demonstrator to explain a deep CNN and MLP with Grad-Cam for Multivariate Time Series Predictions.

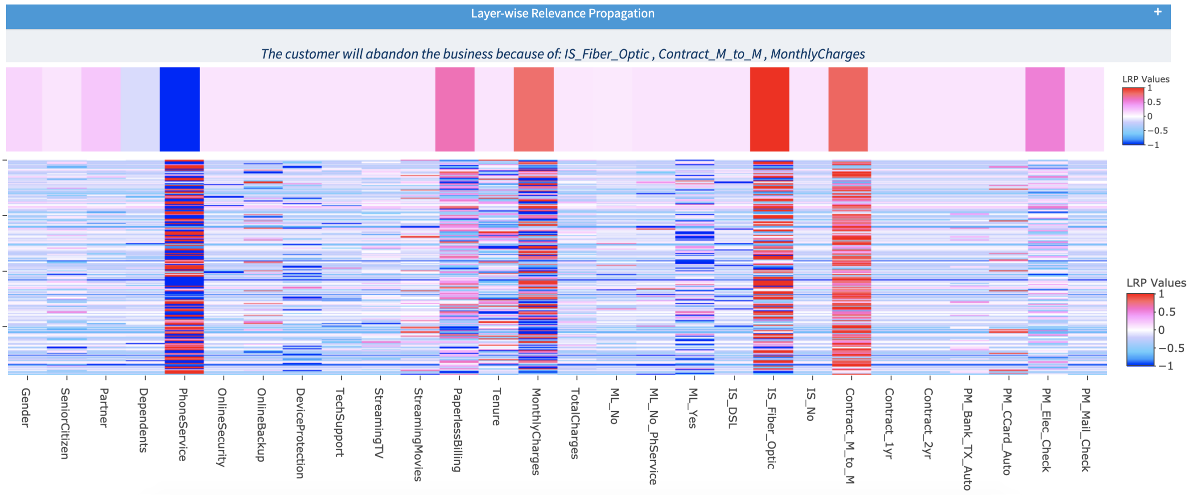
LRP works similar to back-propagation, propagating the relevance/likelihood from output to input pixel of the input layer. The relevance values are either positive (in favor) or negative (against) the decision. LRP is mainly applied to the explanation of DL based vision or text systems. In our system, we apply LRP to the explanation of deep models trained over structured data. Many real-world datasets are in the form of structured data (such as customer loan applications, or survey form data) and there are no known research works or implementations that use 1-D CNN and LRP with structured data. The proposed 1DCNN learns discriminative features for our two model scenarios: prediction of whether telecoms customer churn and classification of credit card transactions as fraudulent or not. LRP is applied to highlight the discriminative features for each model in the form of a heatmap. In addition, we apply SHAP and LIME to our models to enable users to use and compare multiple explainability approaches.

II. XPlainIT Architecture

The XPlainIT demonstrator contains two architectural stages. The first stage is the off-line stage of creating the trained models: data pre-processing and training of the 1-D CNN. Pre-processing involves converting categorical features to binary, normalization, augmentation of data using SMOTE [9] to balance minority classes, and stratified partitioning of data into training and testing sets of 80% and 20%, respectively. The proposed 1-D CNN with 8 layers (input (1x28), convolution (25 kernels), activation (ReLu), convolution (50 kernels), convolution (100 kernels), fully connected (2200 kernels), fully-connected (2 Neurons), and softmax layer) is trained over two structured datasets (Telecom Customer Churn Detection (TCCD[1]) and Credit-card fraud detection (CCFD[2]) datasets). The second stage of the XPlainIT demonstrator is the interactive demo stage, where XAI techniques (LRP, SHAP, and LIME) are used to visualize the important features from the two trained models produced in the first stage. Our principal focus is on LRP, with SHAP and LIME explainability techniques also implemented to give alternative features lists for comparison and a basis for comparing model explanations and generation speed. The user can view statistical information about the underlying datasets, such as class balance and types of features. Global (over the whole dataset) and local (single

[1] https://www.kaggle.com/bandiatindra/telecom-churn-prediction

[2] https://www.kaggle.com/mlg-ulb/creditcardfraud



The customer will abandon the business because of: IS_Fiber_Optic , Contract_M_to_M , MonthlyCharges

Fig. 1. Visualizing LRP heatmaps for Local (1^{st} Heatmap) and global (2^{nd} Heatmap) in Telecom Churn testing set.

record) LRP-based interpretable heatmaps are provided. The demonstrator enables the user to interactively explore and switch between global/local models, classes, and explainability techniques for the two business scenarios. The user can also interact to enter in new fraud/ churn examples, view the real-time decision and associated local explanation using the three explainability techniques. The following section explains each explanation type.

III. Experiments, Visualization, & Explanation

We used the TCCD and CCFD datasets as datasets well-known in the data science field and representative of two common real-world problems. The TCCD dataset contains 19 named features, expanded to 28 after categorical feature conversion. There are 73.42% not-churn and 26.58% customer churn instances. The CCFD dataset contains 32 anonymous features (for privacy). The dataset is highly imbalanced with 99.83% non-fraud vs 0.17% fraudulent samples. We show sample diagrams from the Telecom Churn model in the next sections.

A. Local Interpretation

With local interpretation, XPlainIt visualizes features relevant to the model decision for an individual record level. The user selects a specific record from the test set and a heatmap is generated, showing the relative weighting of features for that decision. In addition, a textual explanation of the model decision is provided, as shown in Figure 1 1^{st} row. The confidence level for each decision is also shown. In Figure 1 2^{nd} row, you can see an example that shows that the top three features contributing most to a predicted customer churn are MonthlyCharges, contract m-t-m (month-to-month), and IS_Fiber_Optic (internet service Fiber optic). Based on domain knowledge, it is evident that if the contract is monthly, with high charges, and the customer is of not appropriate age (e.g. elderly) with fiber optic, are indeed liable to churn.

B. Global Interpretation

A global analysis heatmap explains the behaviour of the model over the whole or subset of the testing set, such as considering and analysing only true positives (TP), or true negatives (TN) samples. Figure 1 3^{rd} row shows a clear pattern of features that have high relevance and impact on the decision of the classifier. Equivalent interactive visualisations are provided for LIME and SHAP in XPlainIT.

C. New Record Analysis

In addition to the existing testing sets, we apply an interactive function in XPlainIT to enable the user to enter in a new fraud or customer churn case. XPlainIT generates the classification decision and important features in the form of a heatmap, in real-time, for the sample customer/ transaction entered.

D. Visualization of LRP vs SHAP vs LIME

SHAP and LIME are two established XAI techniques, which we use for comparison and demonstration in our demonstrator. We provide appropriate visualisation types to support these, such as BoxPlot visualisations for SHAP. The features highlighted by SHAP and LIME for our two models are largely similar to LRP. An important advantage that emerges in our demonstrator is that LRP has a far lower real-time computational cost i.e. LRP (2s) vs LIME (more than 20s) and SHAP (around 60s) on a CPU.

IV. Conclusion

The XPlainIT demonstrator is relevant to any organisation who wants to understand how explainability techniques can be used to explain the decision making of a deep neural network (e.g. 1-D CNN). The use of LRP for structured data has provided intuitive visual explanations, at a significantly faster speed than SHAP or LIME. Such explanations support the identification of features that are critical to a decision versus features that are irrelevant to the model and slow down processing. It can also address the issue of decision transparency at the customer level and the associated GDPR concerns, for example explaining why a customer loan application has been rejected by a DL based system. In the future, it can be applied to other bigger datasets and deeper models.

References

[1] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8689 LNCS, no. PART 1, pp. 818–833, 2014.

[2] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *In CVPR*, 2010.

[3] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2018–2025.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, August 13-17, 2016*, pp. 1135–1144.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, and O. D. Suárez, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," in *PloS one*, 2015.

[8] R. Assaf and A. Schumann, "Explainable deep neural networks for multivariate time series predictions," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2019-August, no. 2, pp. 6488–6490, 2019.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.