G–LIME: Statistical Learning for Local Interpretations of Deep Neural Networks Using Global Priors (Abstract Reprint)

Xuhong Li[1], **Haoyi Xiong**[1], **Xingjian Li**[1], **Xiao Zhang**[2], **Ji Liu**[1], **Haiyan Jiang**[1], **Zeyu Chen**[1], **Dejing Dou**[1]

[1]Baidu Inc., Beijing, China
[2]Tsinghua University, Beijing, China

Abstract Reprint. This is an abstract reprint of a journal article by Li, Xiong, Li, Zhang, Liu, Jiang, Chen, and Dou (2023).

Abstract

To explain the prediction result of a Deep Neural Network (DNN) model based on a given sample, LIME [1] and its derivatives have been proposed to approximate the local behavior of the DNN model around the data point via linear surrogates. Though these algorithms interpret the DNN by finding the key features used for classification, the random interpolations used by LIME would perturb the explanation result and cause the instability and inconsistency between repetitions of LIME computations. To tackle this issue, we propose G-LIME that extends the vanilla LIME through high-dimensional Bayesian linear regression using the sparsity and informative global priors. Specifically, with a dataset representing the population of samples (e.g., the training set), G-LIME first pursues the global explanation of the DNN model using the whole dataset. Then, with a new data point, -LIME incorporates an modified estimator of ElasticNet-alike to refine the local explanation result through balancing the distance to the global explanation and the sparsity/feature selection in the explanation. Finally, G-LIME uses Least Angle Regression (LARS) and retrieves the solution path of a modified ElasticNet under varying -regularization, to screen and rank the importance of features [2] as the explanation result. Through extensive experiments on real world tasks, we show that the proposed method yields more stable, consistent, and accurate results compared to LIME.

References

Li, X.; Xiong, H.; Li, X.; Zhang, X.; Liu, J.; Jiang, H.; Chen, Z.; and Dou, D. 2023. G-LIME: Statistical learning for local interpretations of deep neural networks using global priors. *Artificial Intelligence*, 314: 103823.