– Supplemental Document – Efficient Shapley Explanation For Features Importance Estimation Under Uncertainty

Xiaoxiao Li[⋆], Yuan Zhou[†], Nicha C. Dvornek[†⋆], Yufeng Gu[◊], Pamela Ventola[‡] and James S. Duncan[⋆*†]

[⋆] Biomedical Engineering, Yale University, New Haven, CT, USA
[*] Electrical Engineering, Yale University, New Haven, CT, USA
[†]Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA
[‡] Child Study Center, Yale School of Medicine, New Haven, CT, USA
[◊] College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou, China

Appendix

Investigation on Repeat Sampling Times

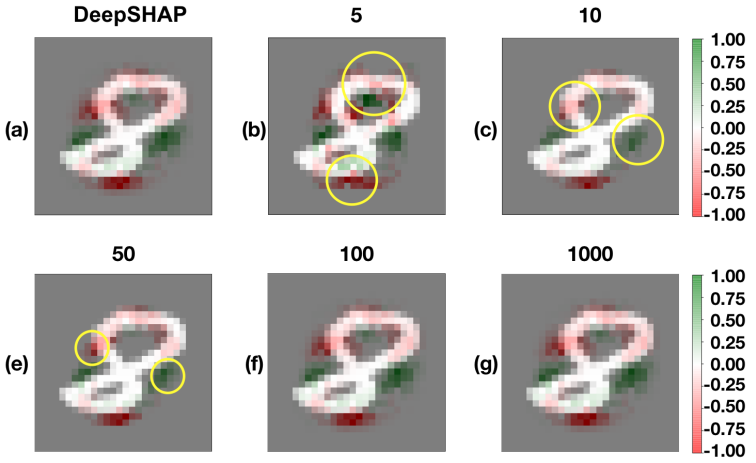

Fig. 1: Feature importance estimations for a digit '8' with reference '0's. (a) shows the overlaid Shapley values generated by DeepSHAP, which is used as ground truth for sanity check. (b) - (g) show the overlaid Shapley values using DistDeepSHAP with different repeating times (denoted on each subfigure). Compared to (a), the differences in (b) - (g) are pointed out by yellow circles, where the larger circle indicates the larger difference.

As we use bootstrap sampling, the Shapley value calculation times will be linearly proportional to the sampling times. However, this computational cost

can be controlled by the user and the method is still faster compared with the original Shapley value explanation as it is shown in Eq. (1) [8,6]. Here, we investigated the suitable repeating times in our proposed DistDeepSHAP, where we used the Shapley values estimated by DeepSHAP [6] as the ground truth. The repeating times were varied from 5 to 1000. We showed the example of estimating feature scores of digit '8' with reference '0's in Fig. 1. The number of repeating times was denoted on each subfigure generated by our DistDeepSHAP. The difference between the Shapley value estimations generated by DistDeepSHAP and those of DeepSHAP were circled in yellow, where the larger circle indicated the larger difference. When sample time $R \geq 100$, the Shapley value estimation results were similar to DeepSHAP (Fig. 1 (a)). The results presented in the main body of this work use repeating time 100 for the trade-off between accurate estimation and time cost. For sampling 100 instances on single CPU for our MNIST experiment, the running time is 1.5s. Even so, our algorithm can be programmed in parallel.