SVEA: A Small-scale Benchmark for Validating the Usability of Post-hoc Explainable AI Solutions in Image and Signal Recognition - Technical Appendix

Sam Sattarzadeh, Mahesh Sudhakar, Konstantinos N. Plataniotis
Department of Electrical & Computer Engineering, University of Toronto
sam.sattarzadeh@mail.utoronto.ca

1. Baseline Model

We trained the baseline model with Stochastic Gradient Descent (SGD) optimizer at a constant learning rate of 0.01 and Categorical Cross-Entropy loss function. Since both the training and test set of the MNIST-1D dataset are equally balanced, no weighting is applied in the loss function. Fig. 1 depicts the normalized confusion matrix computed for this model. Similar to the CNNs trained on the original MNIST dataset, misclassification occurs the most between the digits "4" and "9" in our baseline model. However, in general, the 1-dimensional patterns in the MNIST-1D dataset are easier to distinguish rather than the MNIST dataset.

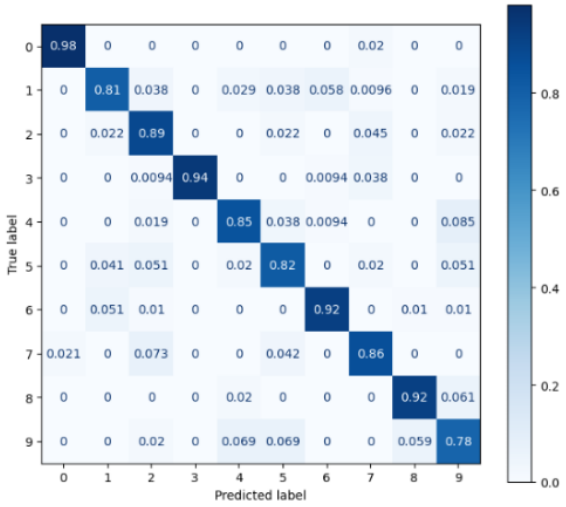

Figure 1. Result of the Normalized Confusion Matrix on the MNIST-1D test set.

2. Keep and Retrain (KAR)

In addition to the "Drop" and "Increase" rates reported in the main manuscript, we carried out the **K**eep **a**nd **R**etrain (KAR) experiment (which is the inverse of Remove and Retrain (ROAR)), that verifies the explanations generated by measuring the degradation in the performance of the model, while preserving the inputs that are considered to be most important [2]. The intuition for this experiment is that if a concrete attribution method highlights the pieces of information that are the most important for the model's decision procedure, the model can achieve a classification accuracy close to (and even in some cases, higher than) its original test accuracy when retrained only with the highlighted information.

We employed the KAR framework to evaluate the correctness of the generated ROEs as specified in the main manuscript. Similar to our proposed experiment, we generated two synthetic training and test set of the ROEs corresponding to the MNIST-1D data for each attribution method. Table 1 replicates the previously generated results alongside the KAR results obtained for the selected explanation methods. As mentioned in [2], the vulnerable point of this experiment is that KAR does not discriminate attribution methods remarkably. Our proposed method relies on training a simpler network (e.g., linear classifier). Since training such models is more complex, the ROE Understandability Test is a more robust metric in distinguishing between well-performing and weaker attribution methods.

3. Additional Qualitative Results

Due to space limitations in the main manuscript, the results for two attribution methods in our benchmark (Score-CAM [6], and XGrad-CAM [1]) are presented in the supplementary materials. Figure 2 contains additional examples of the generated explanation for inputs correctly predicted by the baseline model. In terms of visual clarity, both the qualitative results in the main manuscript and Fig. 2 indicate that perturbation-based methods (RISE [3] and SISE [4]) and the XAI method Integrated Gradients [5] that addresses the limitations of the gradient-based analysis, generate the clearest visual explanations that provide comprehensive usability. This insight is further validated by the quantitative results reported in the main manuscript and extended in Table 1.

Metric	Vanilla Gradient	XGrad-CAM	Score-CAM	Grad-CAM	Grad-CAM++	SISE	Integrated Gradient	RISE
ROE Test (%)	32.7	39.3	42.1	47.6	59.0	61.5	65.1	66.4
Drop%	29.54	24.69	27.03	26.18	17.81	12.19	9.85	7.64
Increase%	29.6	37.7	28.9	36.4	33.9	38.5	40.4	41.0
KAR (%)	72.7	72.0	73.4	70.9	82.0	86.3	87.8	88.5

Table 1. Comparison of the results of the Keep and Retrain (KAR) experiment applied on the state-of-the-art visual explanation methods, along with the results achieved by Drop%, Increase%, and our proposed experiment (ROE Understandability Test).

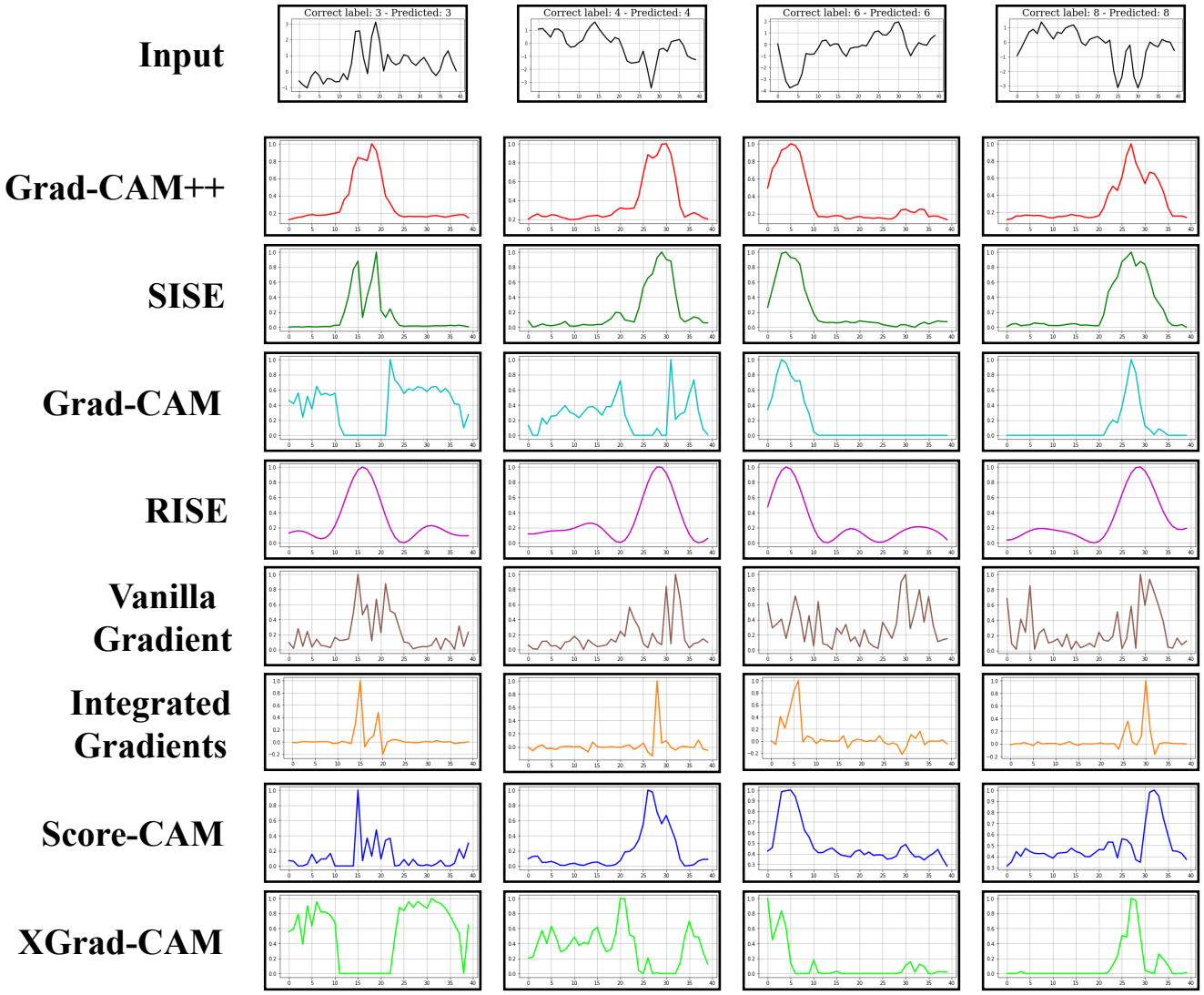

Figure 2. Additional qualitative results on the test samples of the MNIST-1D dataset.

References

[1] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

[2] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9737–9748. Curran Associates, Inc., 2019.

[3] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[4] Sam Sattarzadeh, Mahesh Sudhakar, Anthony Lem, Shervin Mehryar, K. N. Plataniotis, Jongseong Jang, Hyunwoo Kim, Yeonjeong Jeong, Sangmin Lee, and Kyunghoon Bae. Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation, 2020.

[5] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[6] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.