Gorokhovatskyi, O., Peredrii, O. (2024). Recursive Division Explainability as a Factor of CNN Quality. In: Babichev, S., Lytvynenko, V. (eds) Lecture Notes in Data Engineering, Computational Intelligence, and Decision-Making, Volume 1. ISDMCI 2024. Lecture Notes on Data Engineering and Communications Technologies, vol 219. Springer, Cham.

Abstract: The black-box nature of artificial neural network models and the lack of the interpretability about the decisions being made by them is well-known problem. In this paper, we present the result of the research related to the possibility to measure numerically the explainability of already trained (and probably used in production) CNN model using methods based on the perturbations of input image. The level of explainability is measured as the average IOU over the binarized classification explanations provided by RISE, Grad-CAM, and our recursive division (RD) method, proposed earlier, and ground-truth labels of the dataset. The evaluation of explainability is done for five networks having different architectures and accuracies and trained to solve cat/dog image classification problem for Oxford-IIIT Pet Dataset with 37 classes. It has been shown that CNN with high accuracy does not necessarily have good IOU explainability. It is clear though that the most accurate model can have not very good IOU measure as well as the model that is not very accurate may have good IOU value. The decision about which model is better is split between different methods we used for comparison. The question about measuring the explainability becomes more acute when ground-truth labels are missing for the dataset, so qualitative evaluations with IOU are not possible. For this case the proposed recursive division explainability (RDE) value may be used as some indicator of the interpretability of the model. It is also shown that the infrerentce of RD is better compared to other methods. Future research in this field may relate to the better binarization of the results of known methods in order to evaluate explainability as well as improvements of RD approach in order to make high-quality visual explanations.

Keywords: image classification, explanation, explainability, perturbation, recursive division (RD), CNN quality, complementary images