

The Confusion is Real: GRAPHIC – A Network Science Approach to Confusion Matrices in Deep Learning

Author:

Johanna S. Fröhlich

Co-authors:

Bastian Heinlein, Jan U. Claar, Hans Rosenberger, Vasileios Belagiannis, Ralf R. Müller

Background. Neural-network classifiers are increasingly deployed in decision-relevant pipelines such as risk scoring, fraud detection, claims routing, and the triage of regulatory or policy documents. Standard reporting for such classifiers covers metrics like accuracy and F1. These metrics do not show how the model's internal class structure forms during training, or which classes are often confused with each other. For example, a classifier may correctly assign firms to industry sectors based on a text description, but rely mainly on writing style cues. Class-level confusion matrices can expose such structural failures, but become hard to visualize and interpret for datasets with many classes.

Aims. We present GRAPHIC (Graph-based Representation and Analysis of Predictions and Hidden-layer Interpretability via Confusion), an interpretability framework that produces class-level confusion graphs for every hidden layer of a trained network. The goal is to make class structure visible both for individual layers and as it evolves over the course of training.

Method. For each chosen hidden layer, we use the activations from the trained network and train a small linear classifier on them, recording its confusion matrix. Each matrix is interpreted as a directed weighted graph, to which standard measures from network science for community detection and to assess homophily can be applied. The pipeline can be parallelized, which makes it well-suited to HPC clusters.

Results. So far, the method has been validated on image classification tasks, revealing, e.g., a color-driven bias in tree classification in CIFAR-100. Ongoing work extends this to the economic domain by analyzing financial news categorization on a class-level.

Conclusion. GRAPHIC offers a structured, per-layer view of how a classifier confuses classes, complementing scalar metrics like accuracy. Large classifiers used in credit, risk, and operations decisions are often trained in environments like the HPC. Where these decisions carry real-world consequences, trust matters as much as accuracy, and GRAPHIC can be folded directly into the training pipeline to provide a class-level view of model behavior.