Contribution ID: **10**　　Type: **not specified**

Investigating Neural Network Training on a Feature Level using Conditional Independence

There are still unresolved questions regarding the changes in learned representations of deep models during the training process. Gaining a better understanding of this process can assist in validating the training. To achieve this goal, previous research has analyzed training in the mutual information plane. We base our analysis on a method founded on Reichenbach's common cause principle. By employing this method, we examine whether the model utilizes information in human-defined features. Given a set of such features, we investigate the changes in relative feature usage throughout the training process. Our analysis includes multiple tasks, e.g., melanoma classification as a real-world application. We discover that as training progresses, models focus on features containing information relevant to the task, resulting in a form of representation compression. Importantly, we also find that the chosen features can differ between training from scratch and fine-tuning a pre-trained network.

Primary author: PENZEL, Niklas (CVG Jena)

Co-authors: Dr REIMERS, Christian (Max Planck Institute for Biogeochemistry); DENZLER, Joachim; BODESHEIM, Paul (Friedrich-Schiller-Universität Jena)

Session Classification: Poster session