CRUNCH Seminars at Brown, Division of Applied Mathematics

Friday - June 12, 2020

Explaining Neural Networks by Decoding Layer Activations

Xuhui Meng

To derive explanations for deep learning models, ie. classifiers, we propose a 'CLAssifier-DECoder' architecture (ClaDec). ClaDec allows to explain the output of an arbitrary layer. To this end, it uses a decoder that transforms the non-interpretable representation of the given layer to a representation that is more similar to training data. One can recognize what information a layer maintains by contrasting reconstructed images of ClaDec with those of a conventional auto-encoder(AE) serving as reference. Our extended version also allows to trade human interpretability and fidelity to customize explanations to individual needs. We evaluate our approach for image classification using CNNs. In alignment with our theoretical motivation, the qualitative evaluation highlights that reconstructed images (of the network to be explained) tend to replace specific objects with more generic object templates and provide smoother reconstructions. We also show quantitatively that reconstructed visualizations using encodings from a classifier do capture more relevant information for classification than conventional AEs despite the fact that the latter contain more information on the original input.