National University of Singapore
Department of Statistics and Data Science
ST5229 Deep Learning in Data Analytics
Group Project Proposal

Focus article

Title:	Evaluating the Visualization of What a Deep Neural Network Has Learned
Author(s):	Wojciech Samek, Sebastian Lapuschkin, et al.
Name of journal or conference proceedings it was published in:	IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS
Year:	2017
Volume:	28
Pages:	2660-2673
Category:	Application/**Algorithm development** (Please circle)

Short description of problem that focus article aims to address and proposed methodology (maximum of 200 words):

Traditional DNN research typically focuses on enhancing model accuracy, efficiency, or speed. In contrast, this paper addresses the challenge of interpreting DNN decisions by objectively quantifying the quality of pixel-level heatmaps—visual explanations that highlight the image regions driving classification outcomes. The study compares three heatmap generation methods: sensitivity analysis, deconvolution, and layer-wise relevance propagation (LRP). The authors introduce a novel methodology that involves a region perturbation strategy, where image regions are sequentially altered ("flipped") in order of their relevance, and the Area Over the Perturbation Curve (AOPC) is used as a quantitative metric to measure the impact on classifier output. Using the MIT Places dataset and the Caffe reference model for ImageNet, the results demonstrate that relevance heatmaps produced by LRP yield the highest AOPC values. This indicates that LRP more accurately identifies the most critical pixels for the network's decision, offering a robust and time-efficient framework for evaluating heatmap quality.

List of four articles that you plan to include in the literature review (provide full citation):

1.	D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, Mar. 2010
2.	M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.
3.	S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015.
4.	G. Montavon, S. Bach, A. Binder, W. Samek, and K.-R. Müller. (2015). "Explaining nonlinear classification decisions with deep Taylor decomposition." [Online]. Available: http://arxiv.org/abs/1512.02479

Provisional plan of the division of work among group members:

Name	**Task**
Hangao Liang	Sensitivity analysis and application
Roman Buckle	Deconvolution method
Nguyen Tuan Hiep	LRP algorithm: theory and application