Demo Proposal for ASRU 2015

What the DNN heard? Dissecting the DNN for a better insight

Khe Chai Sim
Computer Science Department
National University of Singapore
simkc@comp.nus.edu.sg

Abstract

Deep Neural Network (DNN) has been well received as a powerful machine learning model in a wide range of pattern classification tasks. Despite its superior performance in handling complex real-world problems, DNNs have been used pretty much as a *black box*, without offering much insights in terms of how and why high quality classification performance has been achieved. To address this problem, a novel DNN interpretation technique was proposed in [1], where activity patterns are used to project the hidden units of the DNN onto a meaningful 2-dimensional *hidden activity space* using the t-distribution Stochastic Neighbour Embedding (t-SNE) [2]. The projected points are displayed with colours to reflect the activation values for the purpose of visualisation.

Figure 1 shows the distribution of the hidden unit projected onto the 2-dimensional *hidden activity space*, for a DNN model trained on the Aurora 4 multi-condition data with 7 hidden layers (2048 hidden units in each layer). All the hidden units across different layers are projected onto the same hidden activity space. Figure 2 shows the interpretable regions in the hidden activity space with respect to the phone attribute. Hidden units within a specific phone region have a higher probability of being *active* with respect to that phone.

This demo will showcase a DNN visualization tool based on the above visualization technique. A screenshot of the visualization tool is depicted in Figure 3. The tool can be used to display the changes in the activity pattern over time for all the hidden units in different hidden layers.

References

[1] Khe Chai Sim, "On Constructing And Analysing An Interpretable Brain Model For The DNN Based On Hidden Activity Patterns," to appear in ASRU 2015.

[2] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 219, no. 1, pp. 1–48, 2008.

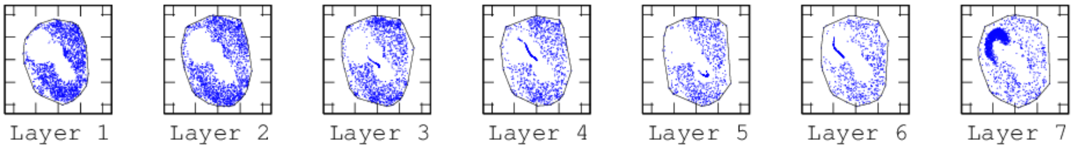


Figure 1: Projection of hidden units onto the *hidden activity space* with respect to the phone attribute.

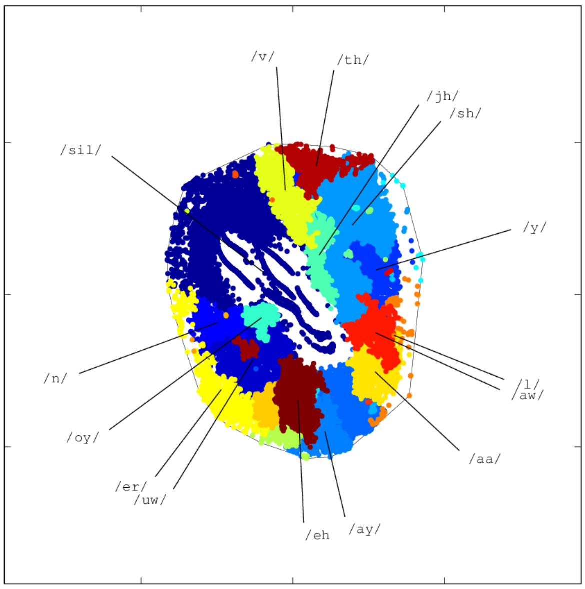
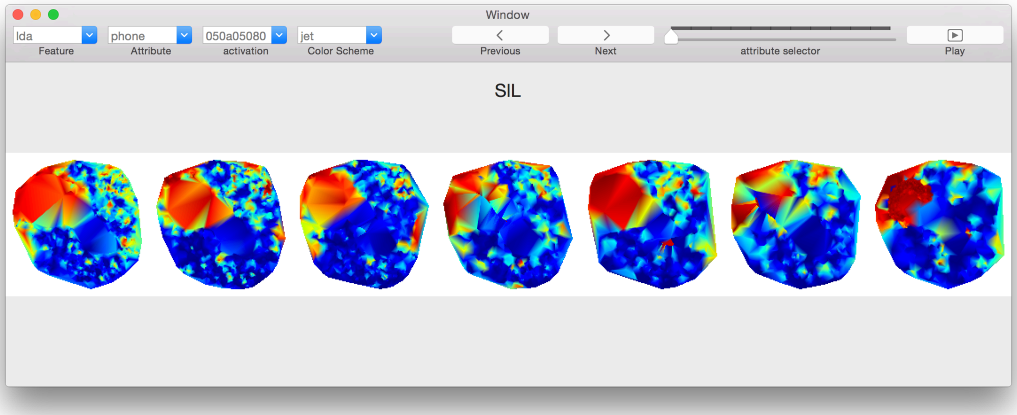


Figure 2: Interpretable regions of the hidden activity space with respect to the phone attribute.


Feature: lda

Attribute: phone

activation: 050a05080

Color Scheme: jet

Figure 3: A screenshot of the graphical user interface for DNN visualization and interpretation.