The fluid representations of networks estimating liquid viscosity

Jan Jaap R. van Assen[1], Shin'ya Nishida[1], Roland W. Fleming[2]
[1]NTT Communication Science Laboratories
[2]Justus Liebig University Giessen

Liquids exhibit incredibly diverse behaviours across a large spectrum of possible scenes. Despite this, we are very well able to estimate the sliminess or viscosity of a liquid with high constancy [1]. Here we used DNNs (Deep Neural Networks) to gain new insights from image-based models that can visually estimate viscosity. We created a training set of 100,000 computer-generated animations of liquids with 16 different viscosities interacting with 10 classes of scene with different kinds of liquid interactions, e.g. pouring or dipping. The network consists of three convolution layers and one fully connected layer using a 'slow-fusion' design that gradually integrates the temporal domain [2]. We gathered perceived viscosity ratings from 16 observers for a subset of stimuli that were excluded from training. Not only does our network make humanlike estimation errors, it predicts average human judgments better than most individual observers, in terms of both error and correlation.

Here we concentrate on differences in the internal representations that emerge in such networks, which we investigated using representational similarity analysis (RSA), based on a large range of different predictors (e.g. colour saturation, motion energy, GIST). We found that many distinct stimulus characteristics—not just viscosity, which the network was trained on—are represented in the final fully-connected layer with 4096 nodes. We retrained only this final layer with 15 nodes (minimum nodes with comparable prediction performance) and find that in this case viscosity dominated the internal representation. Figure 1 shows t-SNE plots of 160 stimuli in activation space of the 4096-node and 15-node versions of the network. The capacity of the final layer strongly influences the encoded features. In the 4096-nodes case, scene specific clusters emerge, each with their own viscosity dimension, whereas with 15 nodes a single viscosity dimension across scenes emerges.

In addition, we trained 100 instances of the same network with the same training parameters. The only differences between the networks are the random initialization and the random training order of the stimuli. We compared the activation patterns across networks using RSA, creating network similarity matrices. Using an expectation-maximization algorithm we find that there are eight clusters of networks with unique neural patterns and yet all predict viscosity with very similar performance. This suggests that in our tested parameter space there are at least eight distinct solutions to the viscosity estimation problem.

These findings show that DNNs are a very powerful tool for extracting informative features for a given task, but the malleable nature of these large parameter spaces requires caution when making inferences from DNNs to the human visual system.

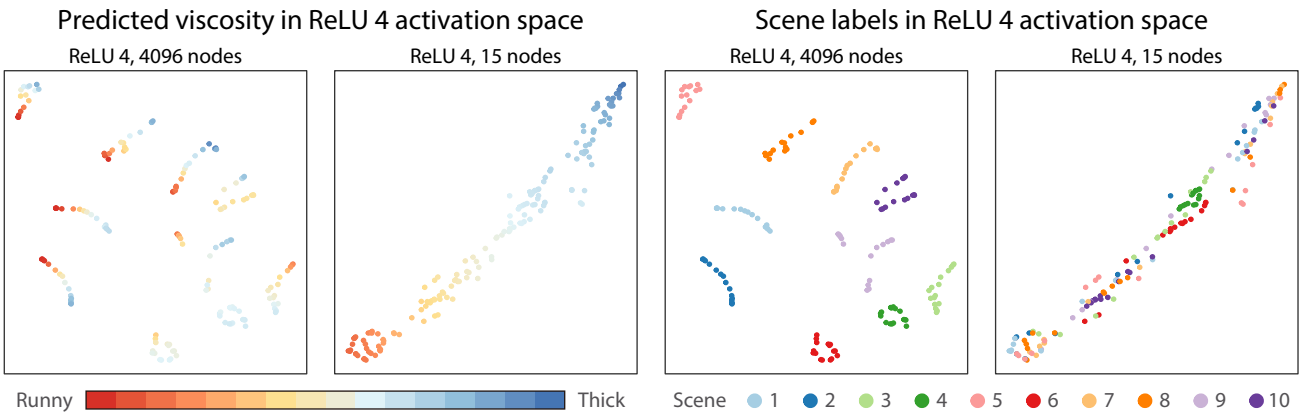

Figure 1: t-SNE plots showing 160 stimuli in the activation space of the final layer of the network. One layer was trained with 4096 nodes and the other layer with 15 nodes. Both predicted viscosity and scene categories are colour coded.

[1] Van Assen, J. J. R., Barla, P., & Fleming, R. W. (2018). Visual features in the perception of liquids. *Current Biology*, 28(3), 452-458.
[2] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).