Visual Analytics for Interpretability on Deep Neural Networks

Haekyu Park Fred Hohman Nilaksh Das Caleb Robinson Polo Chau

haekyu.com / haekyu@gatech.edu

Try at: haekyu.com

Research Summary

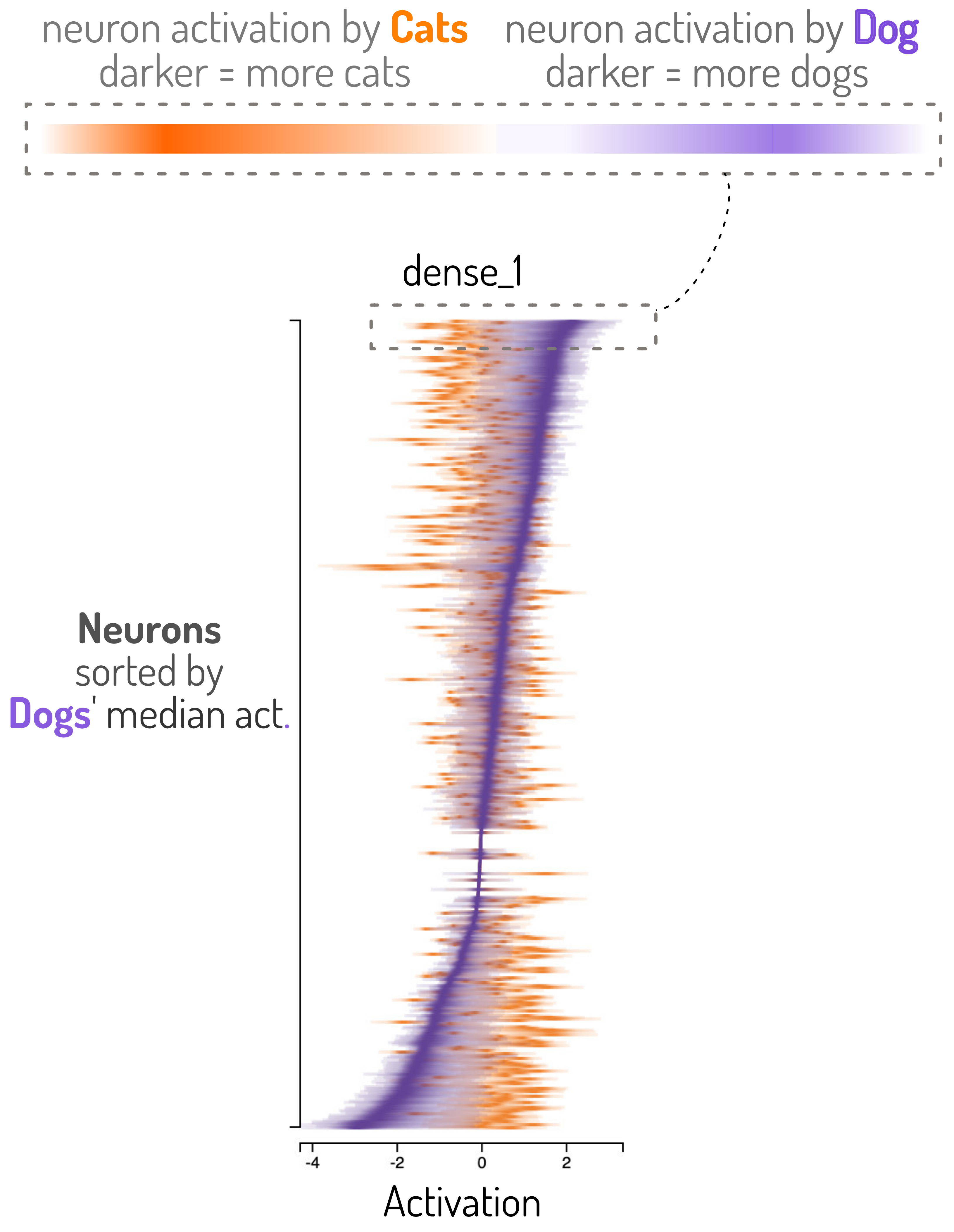
Deep Neural Networks are widely used to solve high-stakes problems (e.g., medical diagnosis, self-driving cars), but often as **"black boxes"**. **Without deep understanding of how the models work, people do not know how to fix them when they fail.**

I am developing **visual analytic techniques for interpreting convolutional neural networks**, both of which are based on neuron activations from all image input data. First, NeuralDivergence helps users explore how the neurons react to inputs. Second, Summit summarizes and visualizes how the learned features interact to make predictions.

Exploration

(NeuralDivergence, PacificVis19)

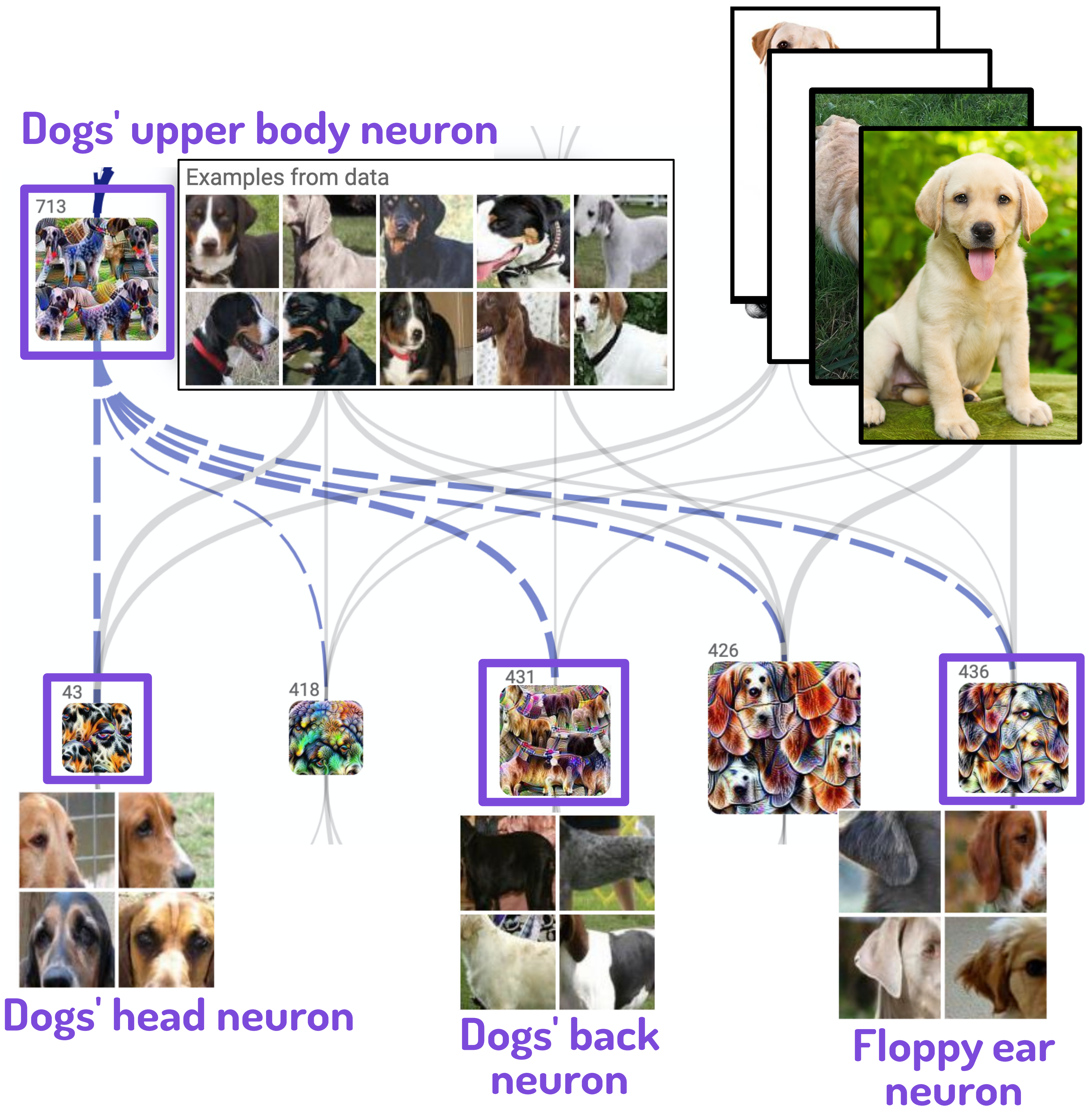
For easier observation of **neuron activation patterns**, we **compress activations** of all images for all neurons in all layers into a distribution representation as a horizontal density bar.



Understanding

(Summit, Vis19)

We provide a **visual summary** of a model's **recognition process** for a class as a graph, where nodes are highly activated neurons for the class and the edges are interaction between the neurons.


Dogs' upper body neuron
Examples from data
713
43
418
431
426
436
Dogs' head neuron
Dogs' back
neuron
Floppy ear
neuron