Summary of Paper: Building Blocks of Interpretability

By: Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, Alexander Mordvintsev

Presented by: Jennifer Fang [Week 01]

Department of Computer Science: University of Virginia

@ https://qdata.github.io/deep2Read/

The Building Blocks of Interpretability

Goal: Explore the interfaces that arise when you combine interpretability techniques.

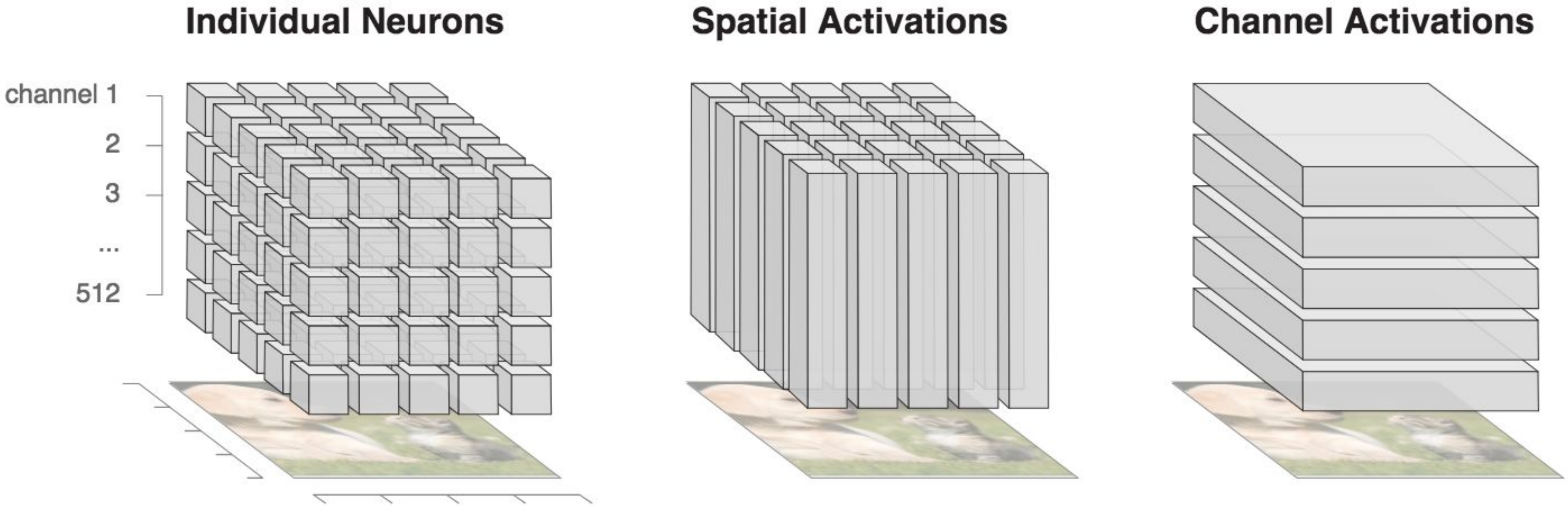
Examples discussed: activation + feature visualization, attribution by spatial position, attribution by channels, and matrix factorization.

Vocabulary

- **Semantic**: relating to language
- **Canonical**: general rule by which something is judged
- **Reify**: make concrete or real
- **Neural Network**: each layer is assembled of series of neurons
  - NN can process data consecutively; first layer is connected with inputs; then layers increase
  - 1st layer: respond to simple shapes like edges
  - Higher layers: respond to complex structures like paws/nose
  - Top layer: neurons respond to complex, abstract concepts aka different animals
  - Output = predicts what the object most likely is

Interpretability key terms

- **Activation**: the amount a neuron fires
  - Neuron is activated if NN decides the info the neuron is receiving is relevant
- **Attribution:** how the NN assembles pieces to arrive at the decision/why such decisions were made
  - Explains the relationships *between* neurons
- **Feature visualization**: tool to answer questions: what is a neuron looking for?



Individual Neurons

Spatial Activations

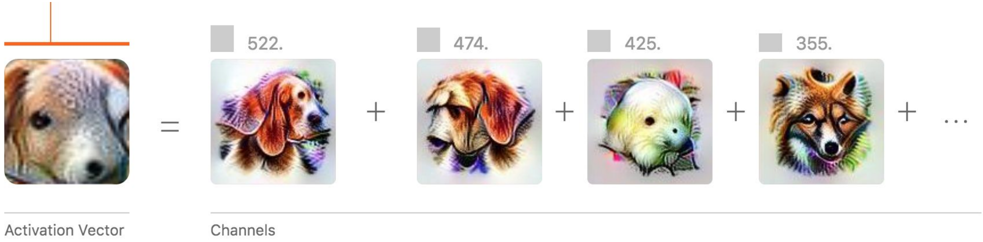
Channel Activations

Activations + feature visualization

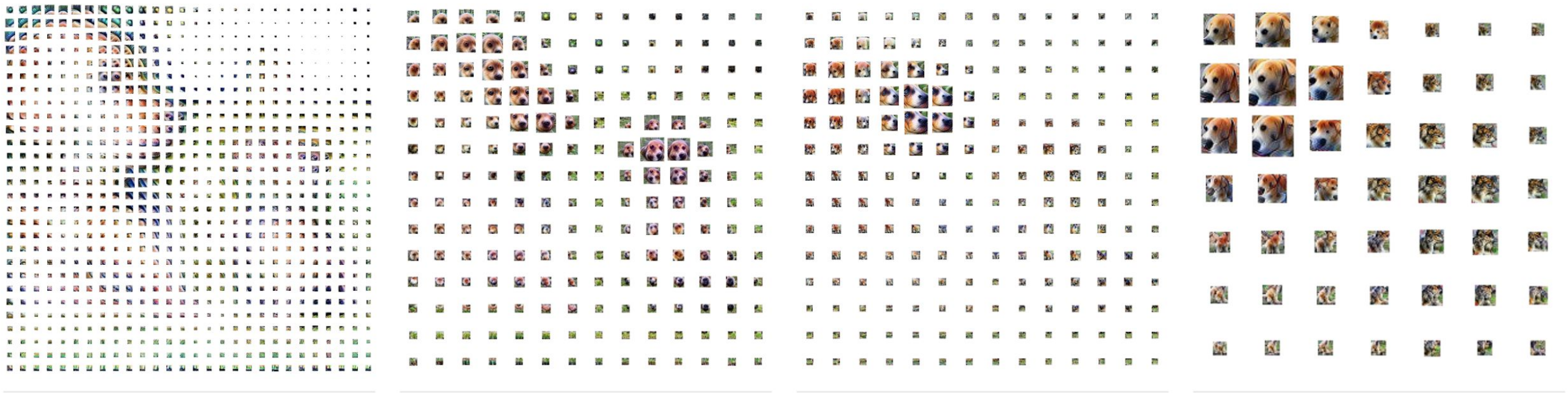
Activations: usually in the form of abstract vectors

After adding feature visualization → transform into

Semantic dictionary: pair neuron activation with visualization of that neuron → sort by magnitude; bring meaning to hidden layers of NN



Scaled by magnitude aka how strong was the detection



MIXED3A

MIXED4A

MIXED4D

MIXED5A

Attribution with saliency maps

Most common interface = saliency map; a simple heatmap that highlights pixels of input that caused the output classification

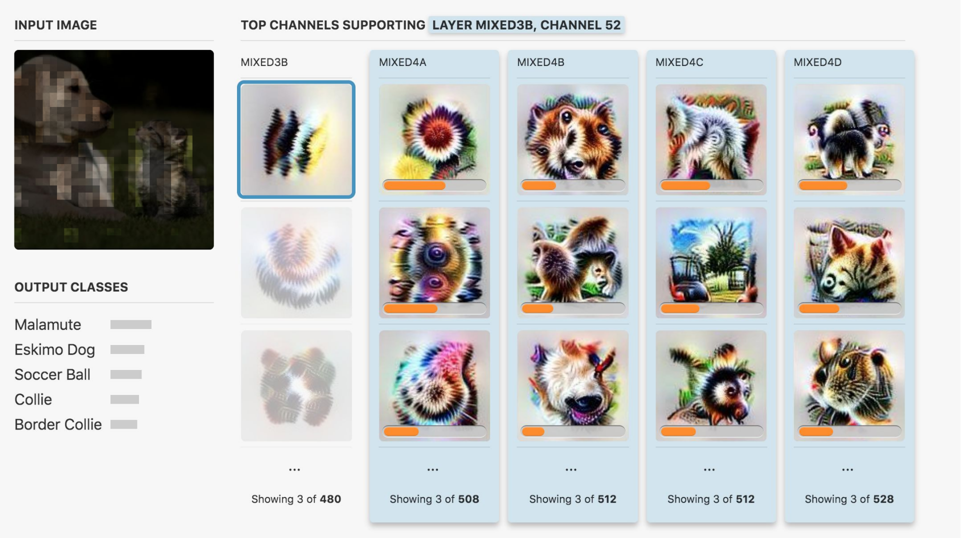
Applications to interpretability: apply it to the hidden layers of the NN; ask whether the high-level idea detected at each position was important

- Perform attribution from each spatial position of each hidden layer shown to the 1,000 output classes
- Visualize 1000-dim vector using dimensionality reduction
- Product: multi-directional saliency map

Attribution with channel attribution

Slice the cube by channels instead of spatial locations

- How much did each detector contribute to output?



Problems with these 2 approaches

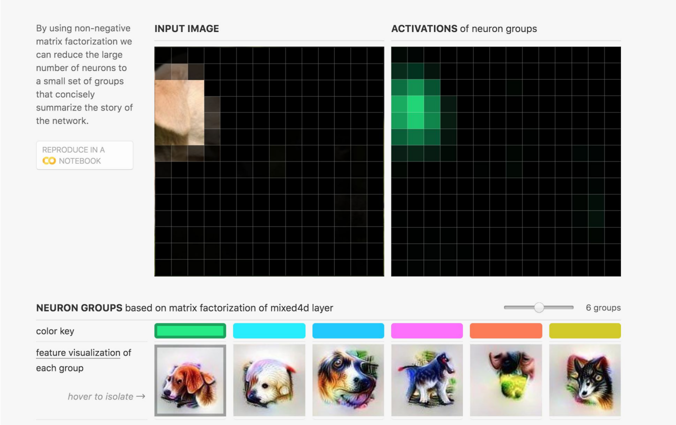
1. Easy to end up with too much info
    a. Would take hours to understand the large # of channels that slightly impact the output
2. Both aggregations are lossy and can miss important parts
    a. Could avoid loss by working with individual neurons aka not aggregating
    b. But that defeats the purpose

Make things human-friendly

Problem: find meaningful ways of breaking up activations

Answer: *matrix factorization*

Reduce large # of neurons into a small set of groups that summarize the NN

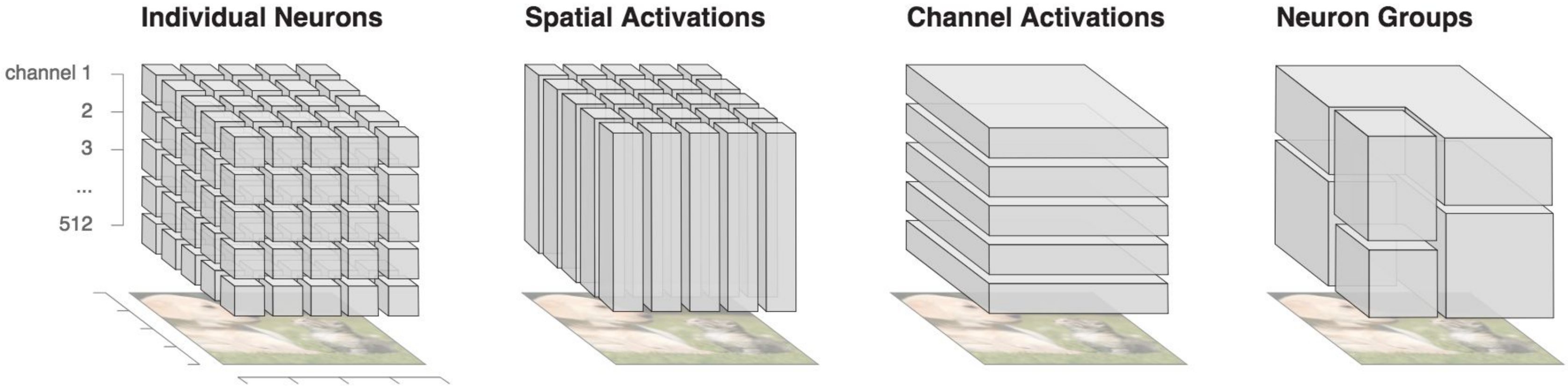


By using non-negative matrix factorization we can reduce the large number of neurons to a small set of groups that concisely summarize the story of the network.

NEURON GROUPS based on matrix factorization of mixed4d layer

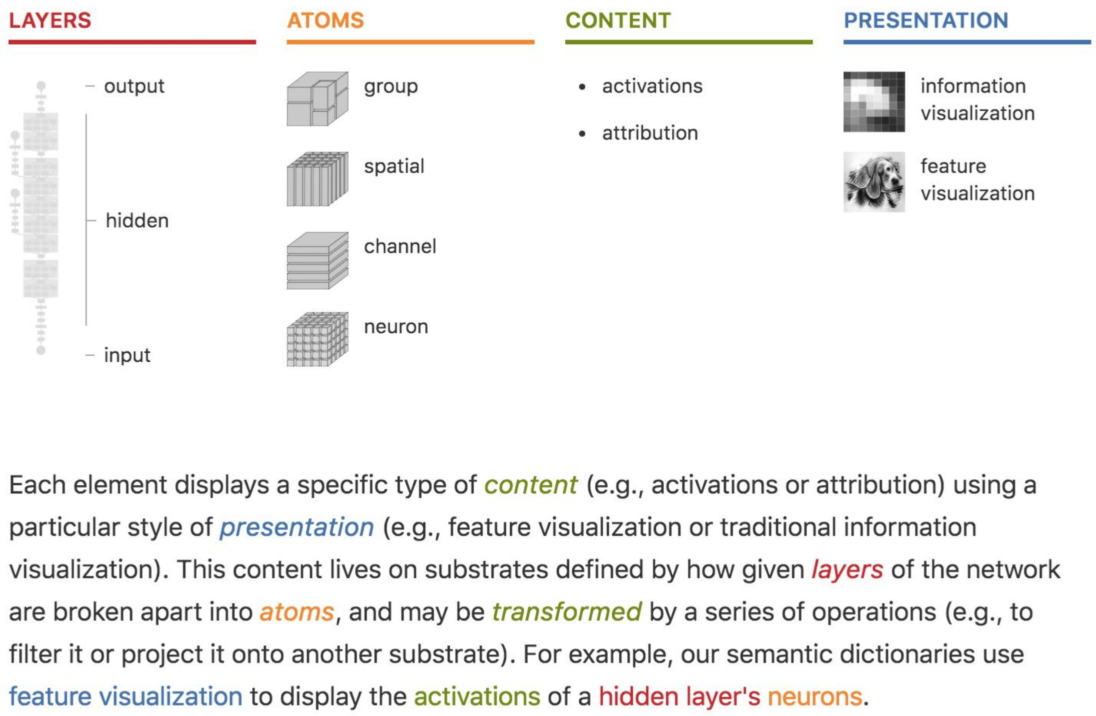
Group for the floppy ear characterization

Each image requires a unique grouping



In addition to naturally slicing a hidden layer's cube of activations into neurons, spatial locations, or channels, we can also consider more arbitrary groupings of locations and channels.

How to make combinations of techniques



Each element displays a specific type of *content* (e.g., activations or attribution) using a particular style of *presentation* (e.g., feature visualization or traditional information visualization). This content lives on substrates defined by how given *layers* of the network are broken apart into *atoms*, and may be *transformed* by a series of operations (e.g., to filter it or project it onto another substrate). For example, our semantic dictionaries use feature visualization to display the activations of a hidden layer's neurons.

High-level summary of the paper

Trustworthiness

1. Do neurons have relatively consistent meaning across different input images AND is that meaning made real by feature visualization?
2. Does attribution make sense and do we trust any current attribution methods?
    a. Many current techniques are unreliable or fundamentally flawed
    b. Function's output could be result of non-linear interactions btwn inputs

Conclusion: There is interesting work to be done in the future to build powerful, trustworthy interfaces for interpretability. Interpretability will be powerful for enabling human management and enabling AI to be fair and safe.

Future work:

- Inspect influences of the dataset
  - Ex: which datasets caused the floppy ear detector to fire
  - Which data sets caused detectors to increase labrador classification
- Make interfaces for interpretability trustworthy
- Learn from human feedback
- Interfaces for comparing multiple models