Attributing Learned Concepts in Neural Networks to Training Data

Nicholas Konz, Charles Godfrey, Madelyn Shapiro, Jonathan Tu, Henry Kvinge, Davis Brown

Pacific Northwest National Laboratory, Duke University, Thomson Reuters Labs, University of Washington

Why care?

Models seem to represent their important hidden features linearly as directions (the 'linear representation hypothesis').

We measure these *concepts* with linear probes, and ask the questions:

1. Which examples in the model's training data were important for learning these concepts?

2. How robust is the formation of these concepts?

We approach this by **attributing concept probe predictions back to the base model's training set.**

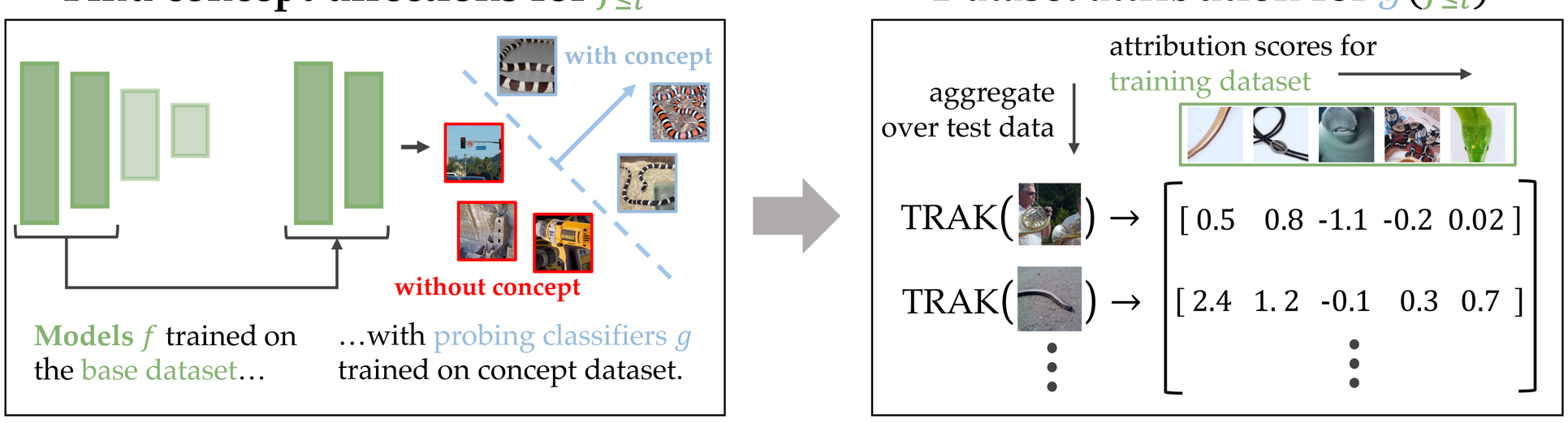
We perform data attribution for learned hidden-layer concept directions.

Concept learning is *convergent*: robust to training example removal, and consistent across different training runs.

Concepts of Interest

- ***Snakes*** *(ImageNet snake classes)*
- ***High-Low Frequency:***
  Transitions from high to low spatial frequencies

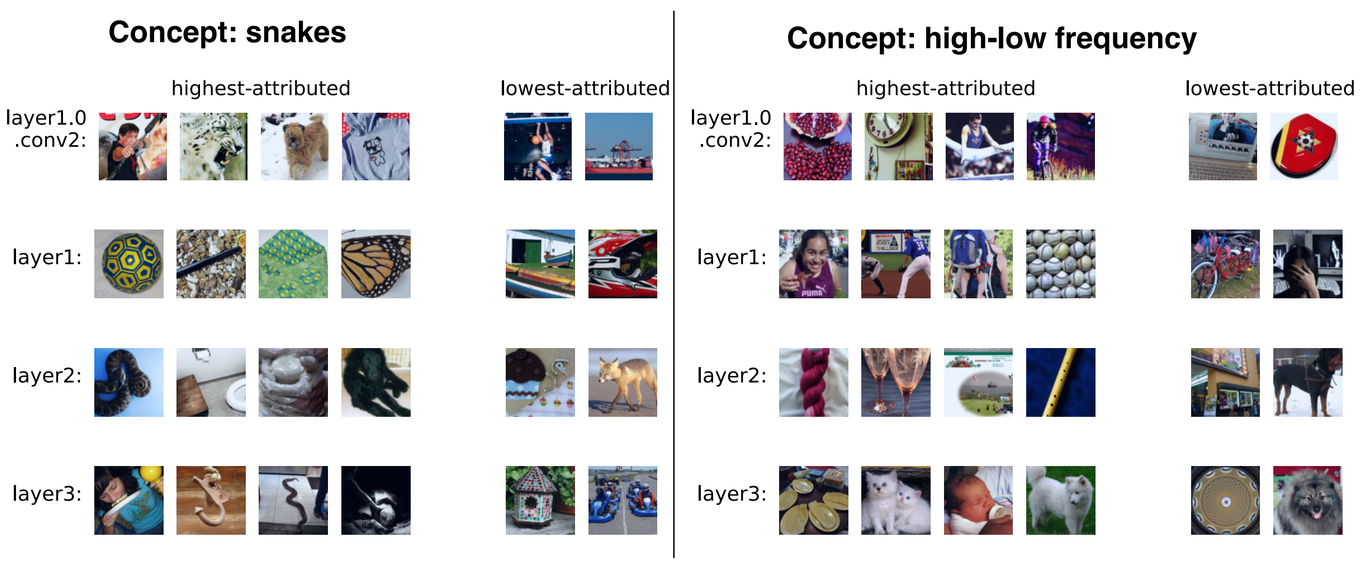
Schematic of our approach for hidden feature attribution



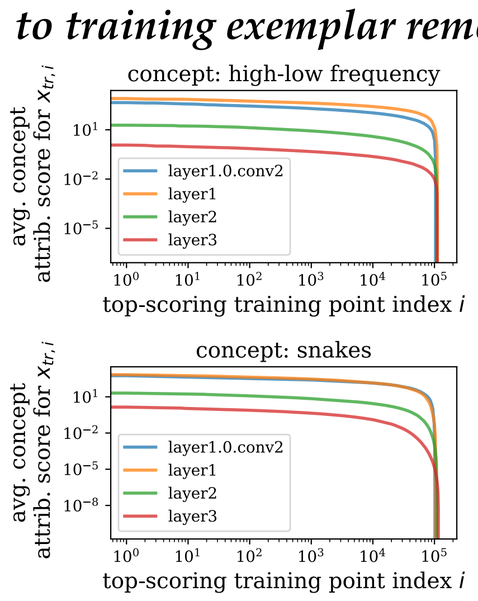
1. Train N models with different random seeds on the training set.
2. Choose a hidden layer i, append a probing classifier g to its output, freeze the weights of $f_{\leq i}$, and train $g \circ f_{\leq i}$ on the concept dataset.
3. Calculate attributions (with e.g., TRAK) for $g \circ f_{\leq i}$ on elements of the test set in terms of the original training data. Aggregate across fixed layers and concepts.

Main Results

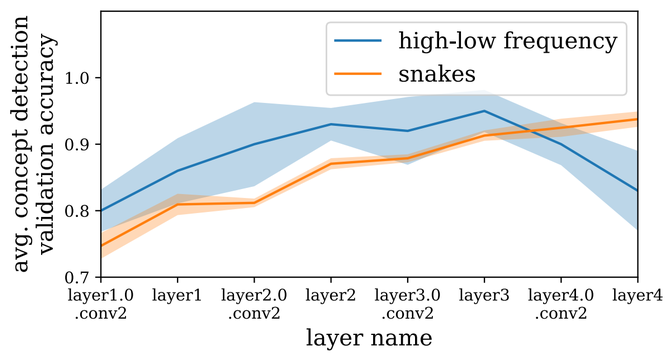
Training set attributions for concept learning

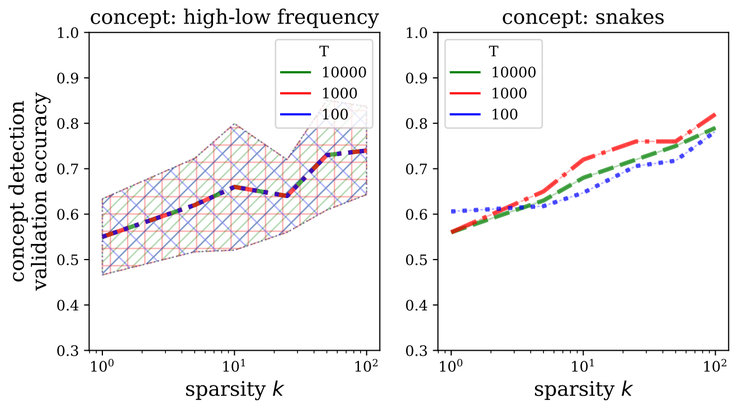


Robustness of concept learning to training exemplar removal



Concept presence at different network layers





Top T concept-attributed training examples removed

For additional information, contact:

Nicholas Konz | nicholas.konz@duke.edu
Davis Brown | davis.brown@pnnl.gov

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

12/6/23