ConceptEvo: Interpreting Concept Evolution in Deep Learning Training

bit.ly/conceptevo

Georgia Tech

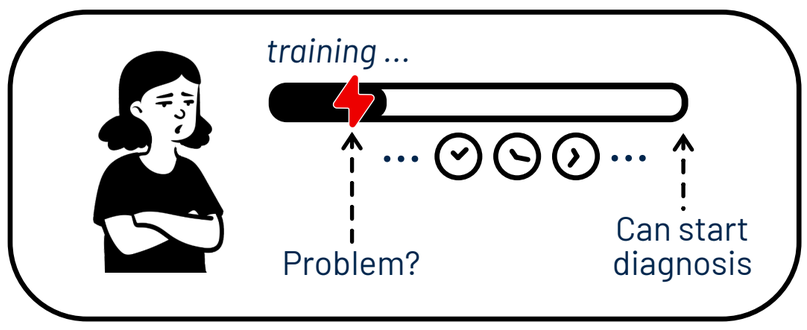
Haekyu Park *haekyu@gatech.edu*

Seongmin Lee · Benjamin Hoover · Austin Wright · Omar Shaikh

Rahul Duggal · Nilaksh Das · Kevin Li · Judy Hoffman · Polo Chau

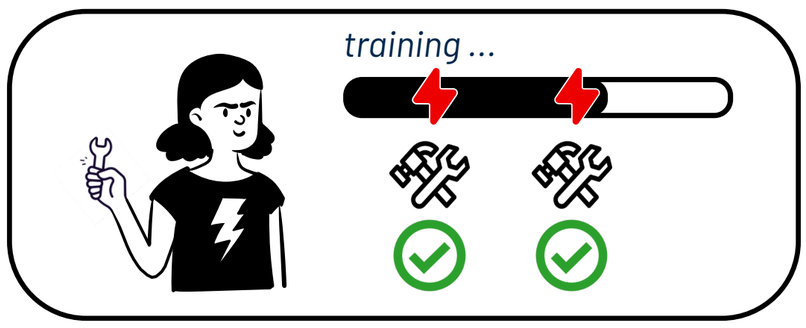
Why Interpreting Deep Learning Training?

Challenge



- Hard to discover **training issues** in real-time
- Missed opportunities for timely intervention

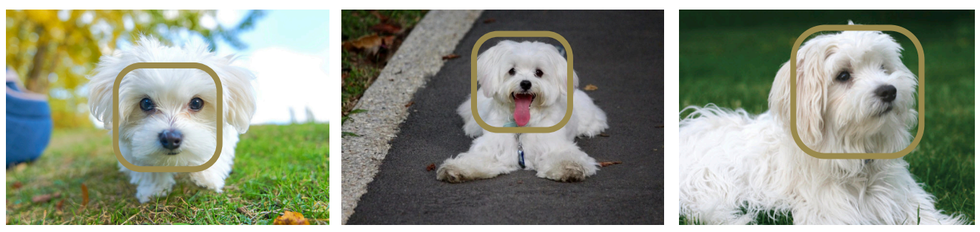
Solution



- Assess training progress and identify potential issues
- How **concepts** learned in the model **evolve** over time

Concepts in Deep Neural Networks

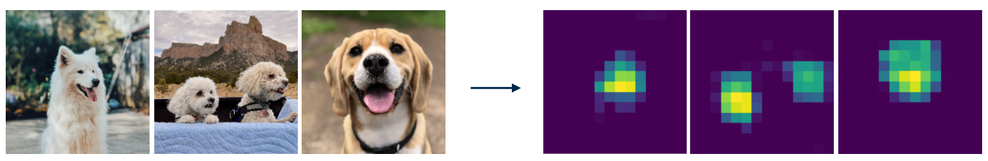
contribute to the model behaviors
e.g., important for the prediction of ***Maltese dog***



Furry head concept

Concepts of **Neurons**?

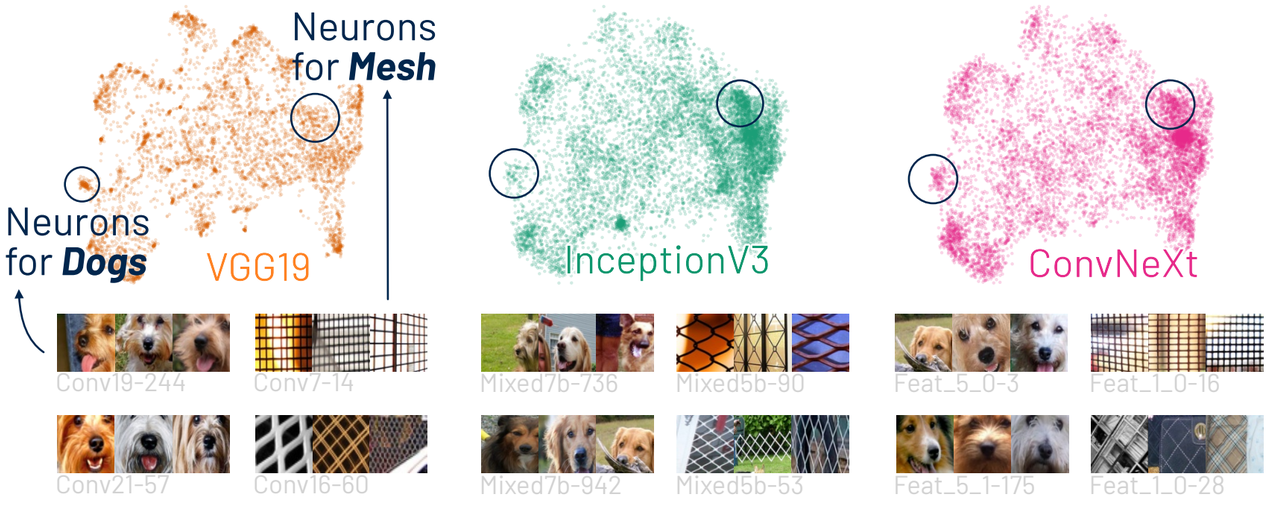
- Neurons as a concept detector, exhibiting **strong activation** in response to **specific features**
- e.g.,) A neuron in an InceptionV1



Given these images → This neuron is activated by *"dog face"* concept

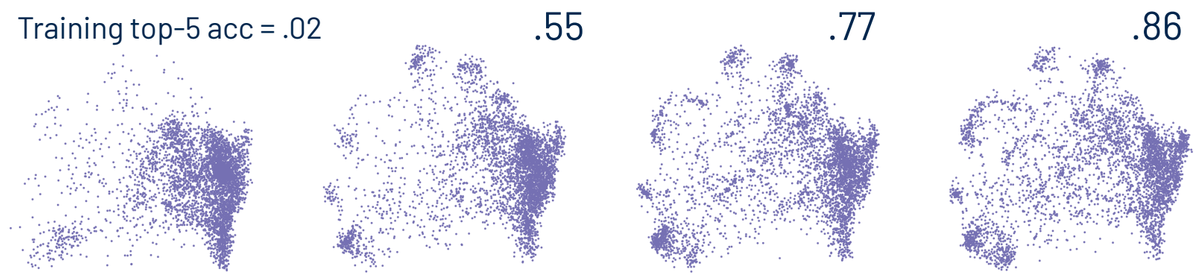
Unified Semantic Space

for concept comparison during training

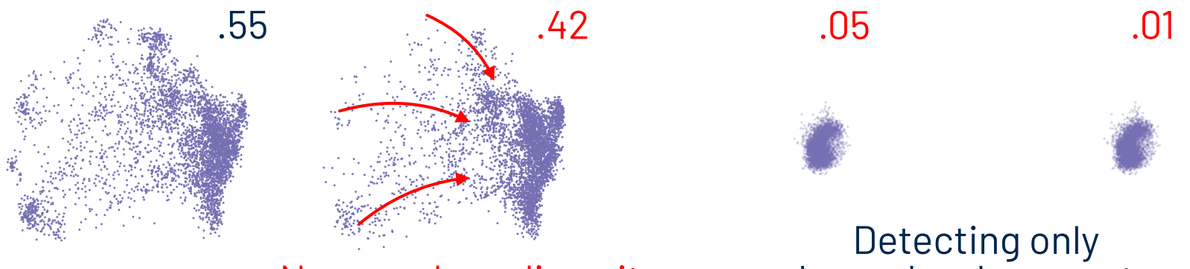


Identify potential training issues

Well trained VGG16



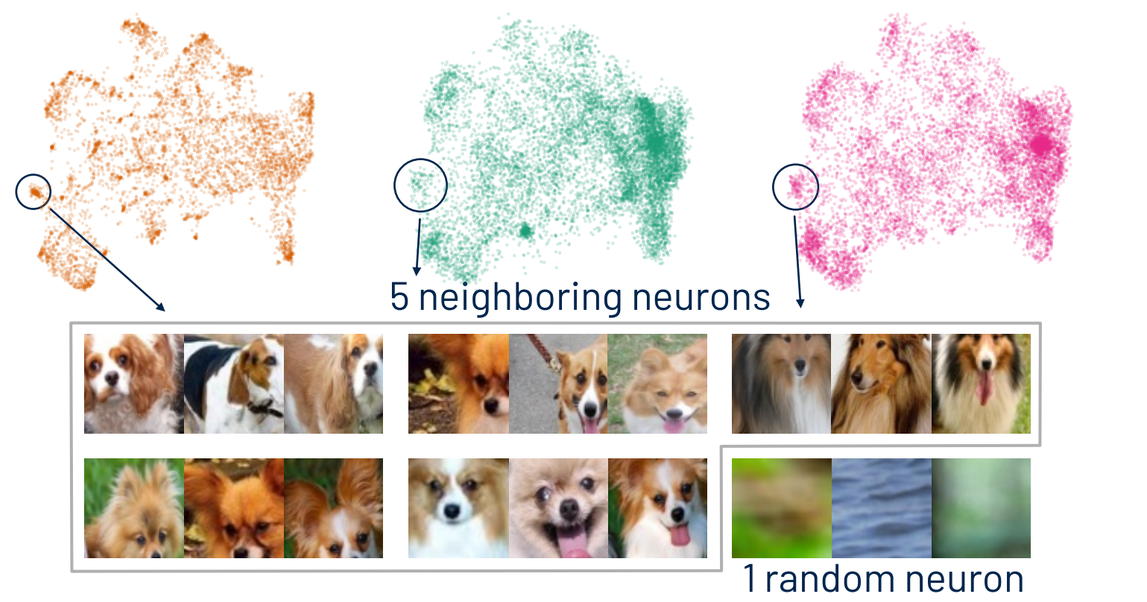
Sub-optimally trained VGG16



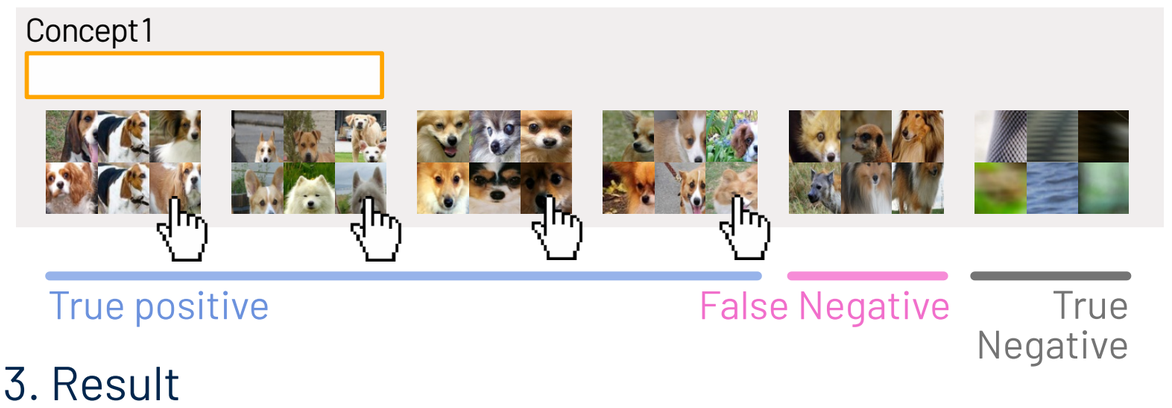
Experiment

How well does ConceptEvo **align concepts** across different model trainings?

1. Sample 5 neighboring neurons, and a random neuron

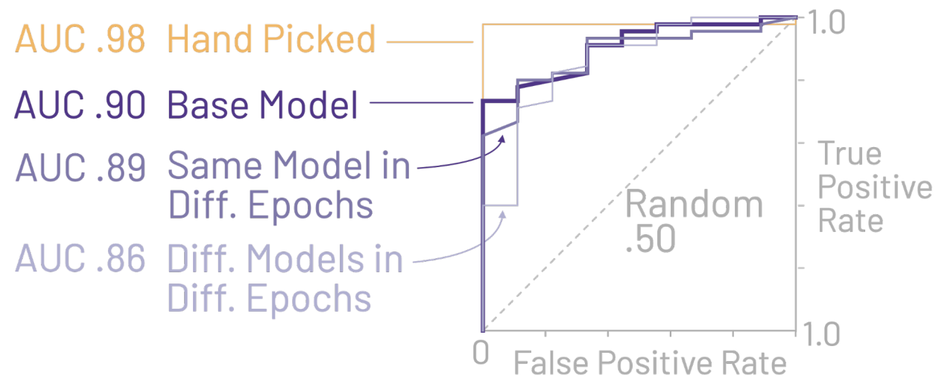


2. Ask people to choose the same concepts
   260 unique Amazon MTurk workers (9 questions each)



3. Result

ROC curve for human estimations for concept alignment



ConceptEvo-found concept groups are highly **discernible and aligned** even across epochs and models