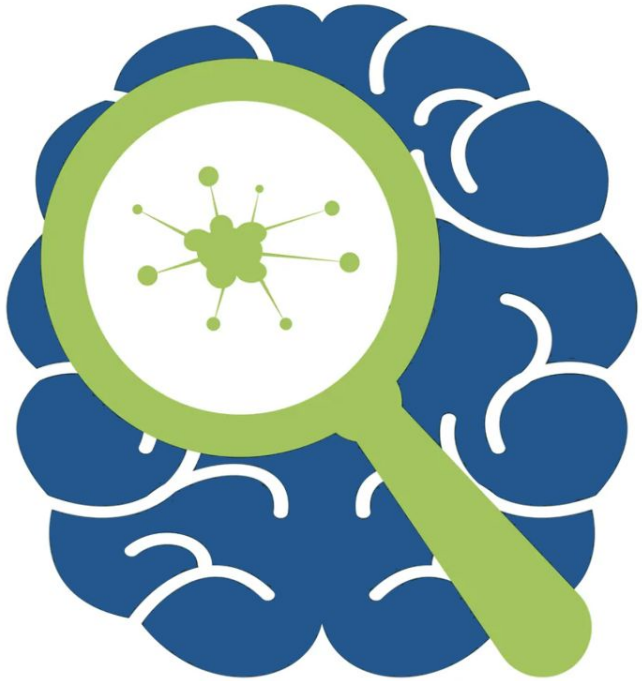
INF-8605 | SPRING 2023
INTERPRETABILITY IN DEEP LEARNING

Research Paper Presentation

Demystifying black-box DNN training processes through Concept-Monitor

Mohammad Ali Khan, Tuomas Oikarinen, Tsui-Wei Weng



Anders Sildnes
Rafael Adolfo Nozal Cañadas
16 - June - 2023

Presentation structure

Motivation for this paper

What is Concept-Monitor

Case studies

- (I) Monitoring standard training
- (III) Adversarial training

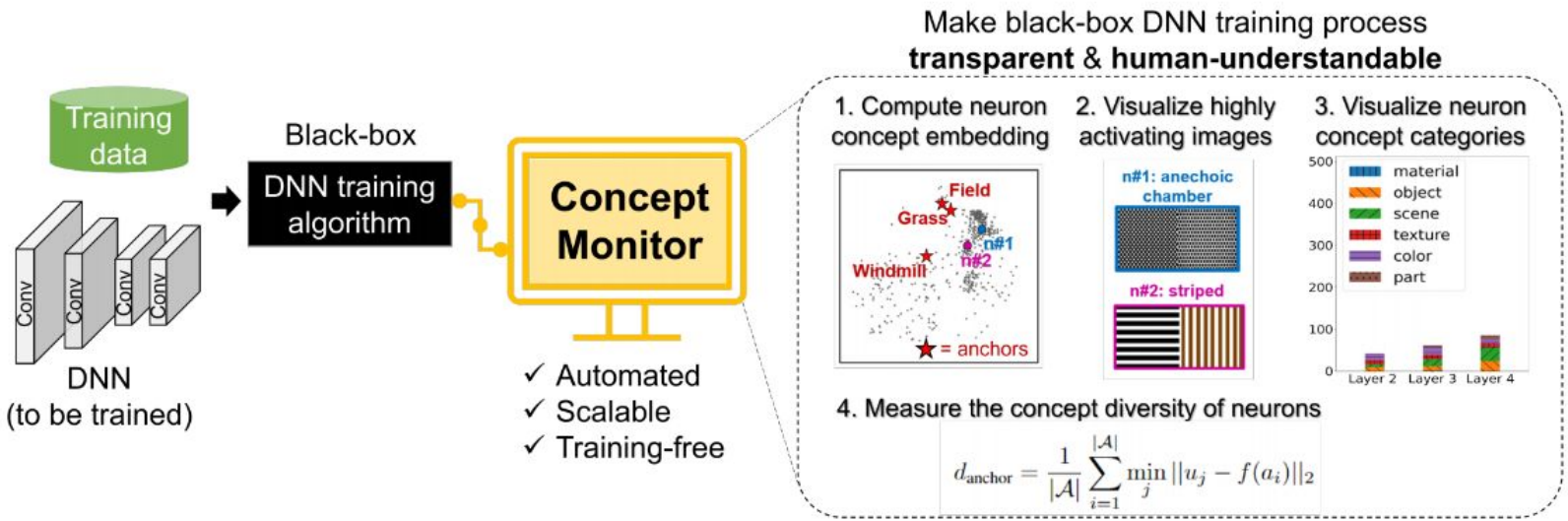
Conclusions

Introduction / Motivation

- Interpretability of Deep Neural Networks (DNN) is difficult and training processes are even more difficult to interpret.

- By understanding model training we can:
  - Improve algorithms
  - Help debug DNNs

What is Concept Monitor?



- **Concept-Monitor** is an automated tool to visualize the training process, i.e., it can visualize *neuron to concept mappings*

- It works **during training**, unlike other frameworks that are *a-posteriori*, after a model has been trained

Step 1: How to compute a neuron score?

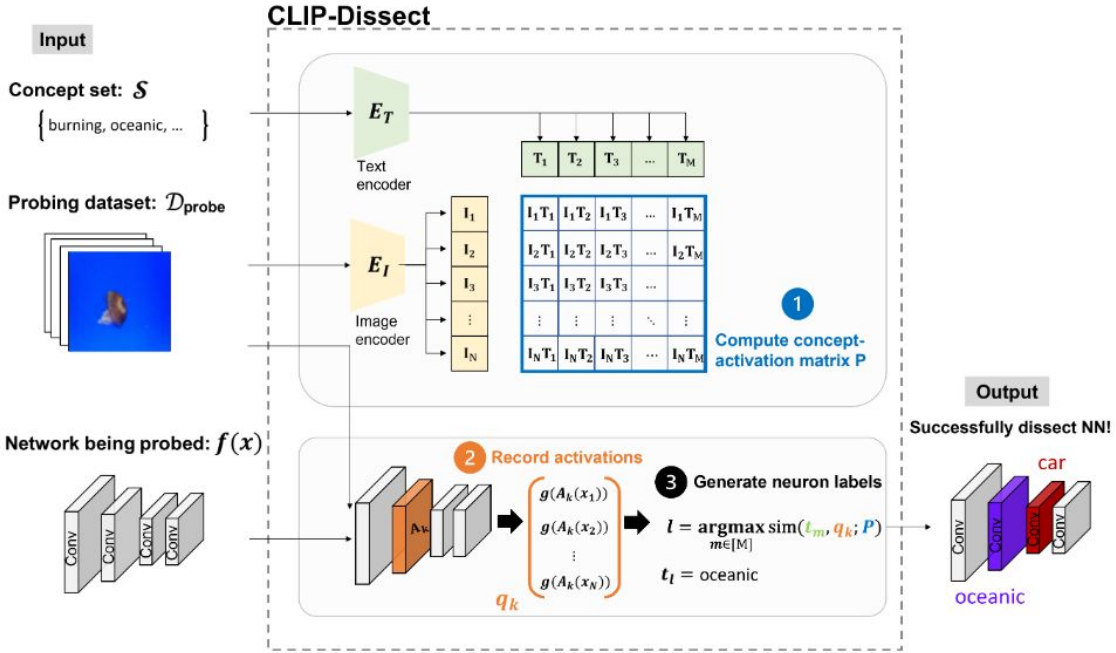


Figure: CLIP-dissect metric (Oikarinen & Weng 2023)

Step 2: Input into unified embedding space

Use dimensionality reduction (UMAP) down to a 2D space.

D-anchor: average euclidean distance to nearest *concepts of interest* for all neurons

- Bad training = High d-anchor, neurons are far away from concepts.
- Good training = Low d-anchor, neurons cover the concepts well.

“We see that in poor training the neurons don’t diverge much along training.”

Case study (I): Monitoring standard training

ResNet-18 is a pre-trained convolutional neural network of 18 layers

Places365 scene recognition dataset

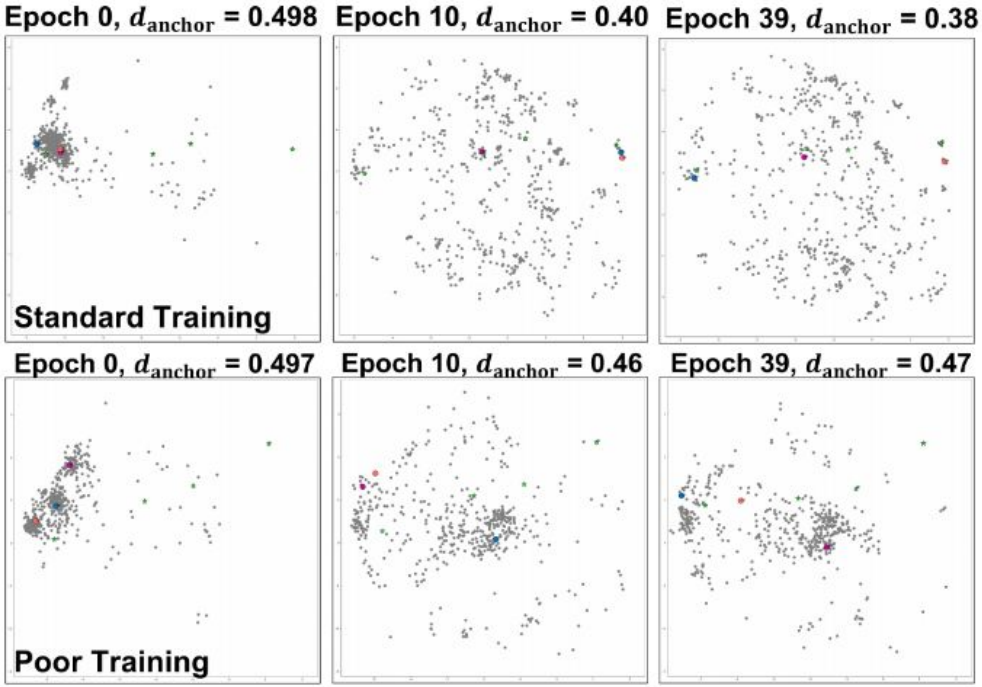
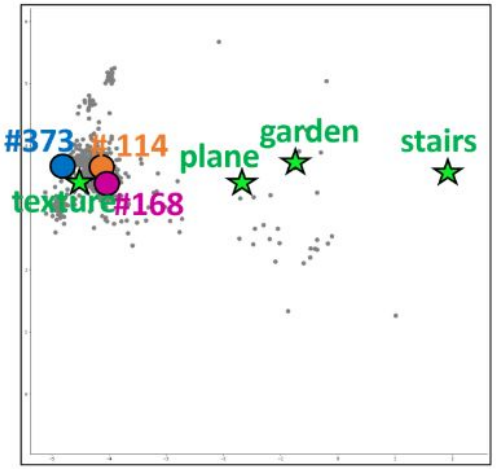
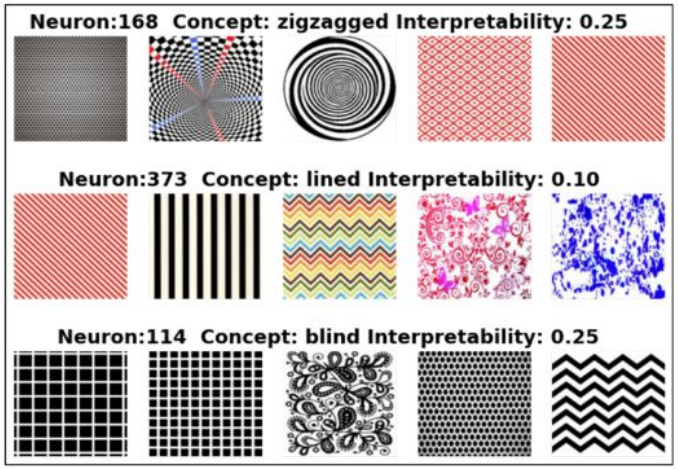


Figure 3. Investigating poor training using unified embedding space. We see that in poor training the neurons don't diverge much along training. The high d_{anchor} also indicates inability to represent useful concepts.

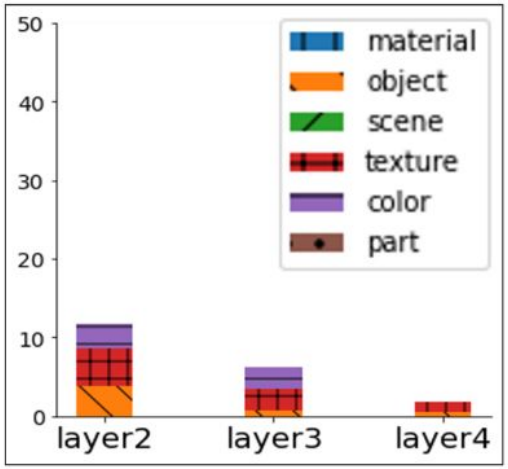
Epoch 0, d_{anchor} = 0.498



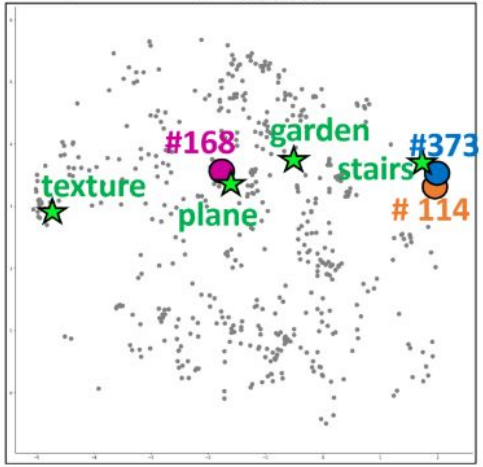
Visualize highly activating images



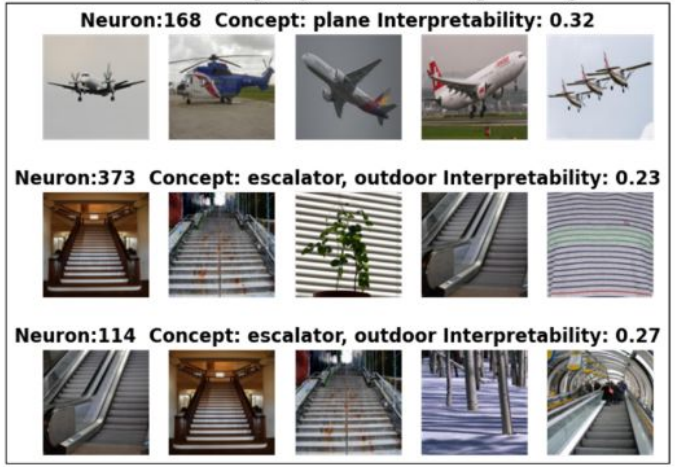
Visualize neuron concept categories



Epoch 10, d_{anchor} = 0.402



Visualize highly activating images



Visualize neuron concept categories

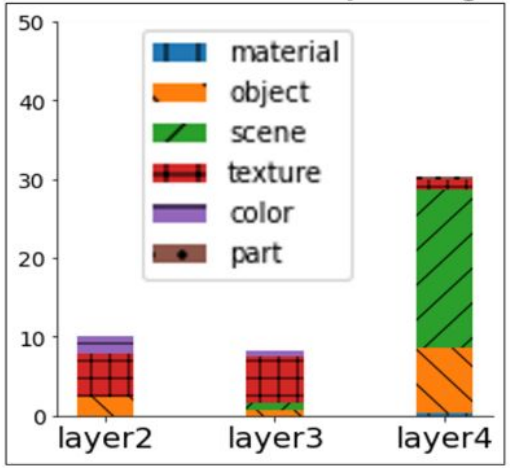


Figure 2. **Case study (I): Monitoring standard training**. We use Concept-Monitor to analyze a standard trained Resnet-18 model on Places365 dataset. We visualize the training at three different epochs, specifically tracking the trajectories of neurons #114 (orange circle),#168 (pink circle) and #373 (blue circle) of layer 4. The 1st column plots our unified embedding space, where each gray dot represents a neuron in layer 4 and green stars represent anchor words. The tracked neurons are coloured differently for visualization. The 2nd column shows the highly activating images of the tracked neurons along with their similarity to the closest concept. Finally the 3rd column shows the percentage of interpretable neurons in layer 2-4 and which category they belong to.

Case study (III): Adversarial Training (ARM)

(Madry et al. 2017)

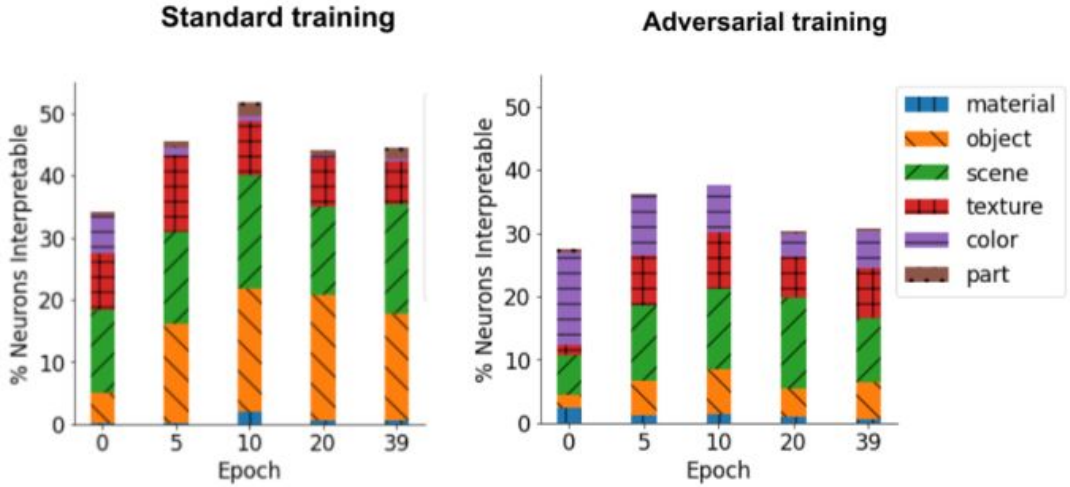


Figure 6. Comparison of the types of concepts learned by standard training compared to adversarial training in the second to last layer (layer4), and differences in types of concepts learned by the models. Note these figures are the network at the end of each epoch, so epoch 0 is after one epoch of training.

Standard training

Concept: **dog** - 15 neurons

Concept: **car** - 11 neurons

Concept: **cat** - 10 neurons

Adversarial training

Concept: **chequered** - 14 neurons

Concept: **red** - 13 neurons

Concept: **striped** - 12 neurons

Conclusions

Good things

Simple method. Define a scoring function and apply it in 4 user cases.

Model agnostic.

Automatic and no need of training, done as the model train.

Interesting

Bad things

Not novel despise them claiming so (Olah et al, 2020, Goh et al., 2021, Nguyen et al., 2016, Cammarata et al. 2020, Elhage et al., 2022).

Limited experiments evaluating the method. Only images, and only a very small subset of images. Ok for proof of concept.

They don't discuss the limitation of their own method, as for example, computationally high?

Euclidean distance in UMAP/t-SNE is difficult to interpret (https://distill.pub/2016/misread-tsne/).

Thank You