DNNViz: Training Evolution Visualization for Deep Neural Networks

Gil Clavien
University of Fribourg, Switzerland

Can a visualization tool:

1. Give us an insight about information distribution into a deep neural network?
2. Help us validate neural network architecture?

Many visualization tools are available

Features usually missing:

- ★ Visualization during the neural network training.
- ★ Visualization of the activations on the whole dataset at once.
- ★ Dynamic aggregation of inputs.
- ★ Neural network model-agnostic.

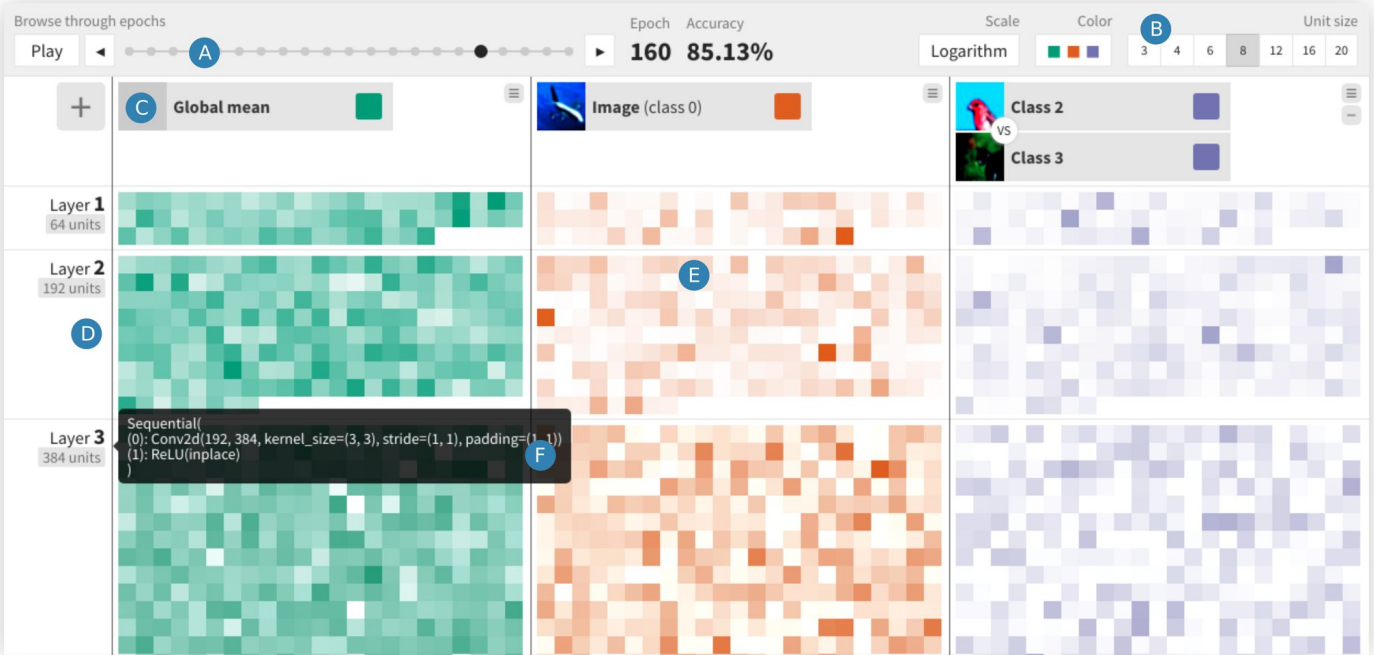
DNNViz combine different approaches

We develop a tool that:

- ★ inspect and displays **activations** on the entire network at once;
- ★ **aggregates activations** depending on the input class;
- ★ aggregates all the activations;
- ★ visualizes activations **all along** the neural network **training**;
- ★ compare the activations of the network.

DNNViz

A) Epoch selection panel. B) Visualization option panel. C) Column Input. D) Layers' information. E) Activation unit visualization. F) Layers' metadata.



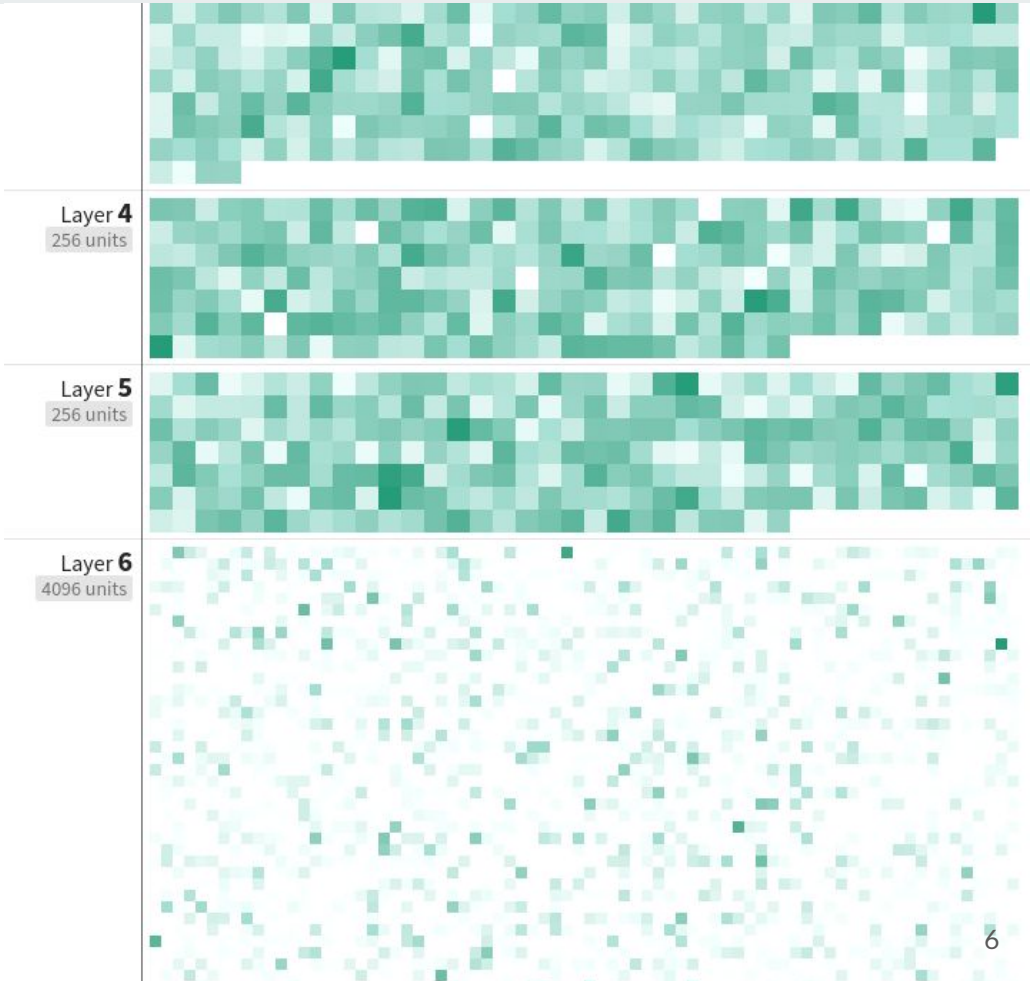
Activations

Displayed informations:

- ★ Metadata
- ★ Layers and layers' metadata
- ★ Activations

Two type of activations:

- ★ Compressed feature map;
- ★ Single neuron.



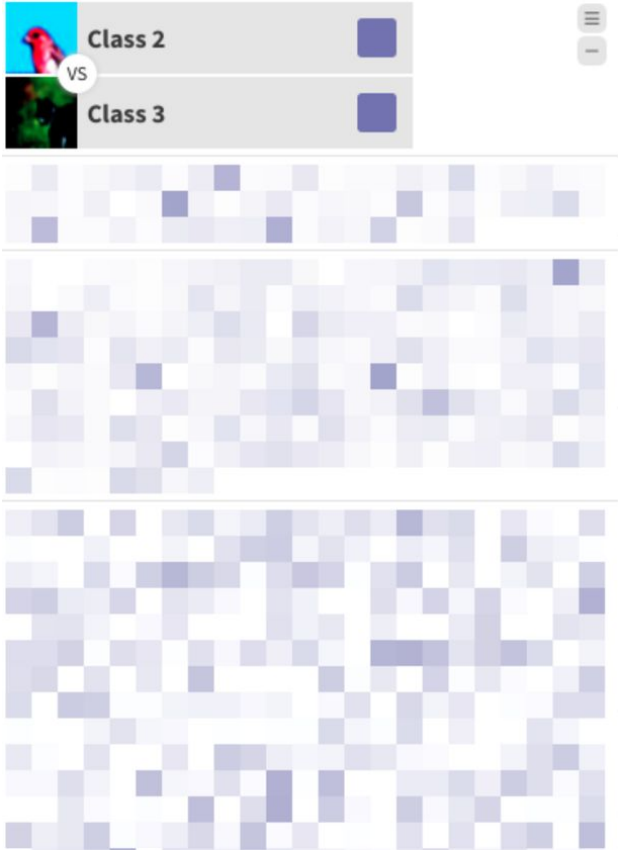
Two column modes

Normal mode

Select one inupt and display its activations.

Difference mode

Select two input and display the difference of their activations.



Use cases

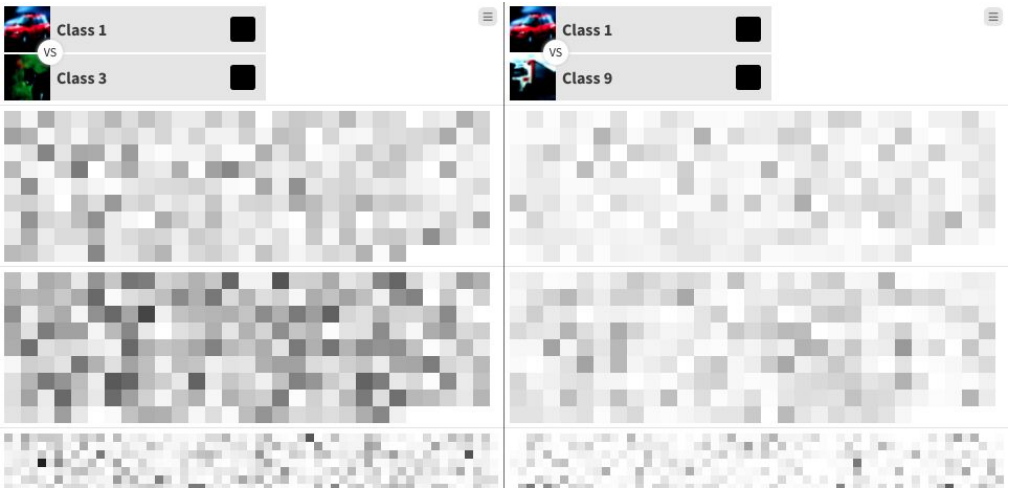
Real case scenario

- ★ Class similiarity detection;
- ★ Network pruning;
- ★ Dataset tampering detection.

Educational task

- ★ Understanding why and how a training doesn't produce the expected results;
- ★ Understanding the link between the classes and their features;
- ★ Comparing the distribution of the activations on the network.

Class similiarity detection



Class-based visualization in difference mode.

First layers of an AlexNet model.

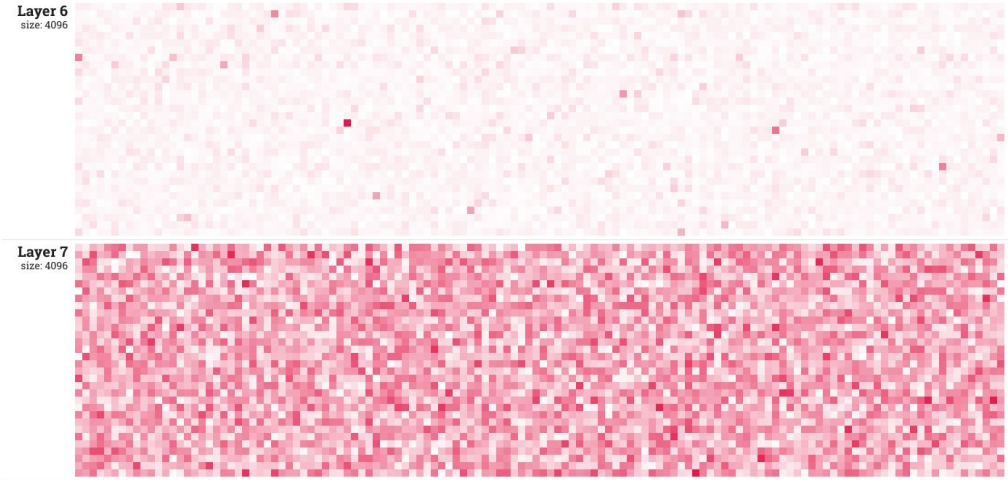
The darker the units are, the stronger the difference between the two classes is.

Network pruning

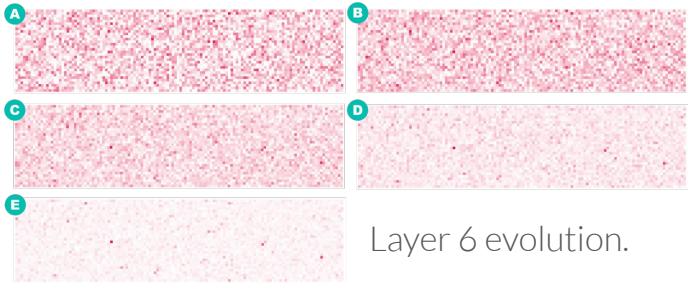
Network pruning: removing unimportant weights from a trained network.

General-based visualization shows network's space "use".

Confirmation of known results: the last fully-connected layers are wasted.



Global activations, Alexnet, CIFAR-10, Layers 6 and 7.



Layer 6 evolution.

Tampering

Adversarial attack: particular manipulation not detectable by human but changing the classification.

Dataset tampering: one pixel modification produced to confuse the neural network

Tampering detection

Train and capture activations from **two datasets** (CIFAR-10 and Tampered CIFAR-10).

Try to detect the differences with the tool.

Normal dataset

Tampered dataset

Behind the scene

DNNViz: two sides project

Backend

A DeepDIVA service that collects, process and saves all the activations of a neural network.

Produce a **JSON output**.

Frontend

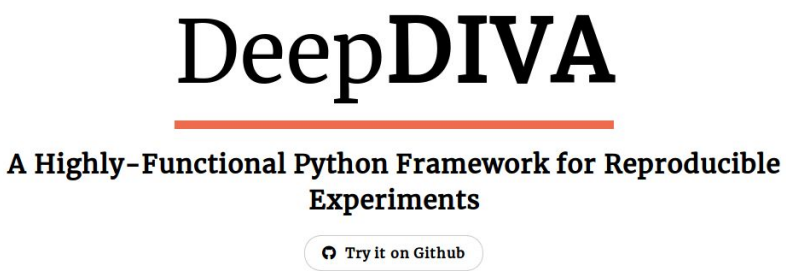
A data visualization tool that consume **JSON output**.

Display the interactive visualization to users.

DeepDIVA integration

DeepDIVA:

- ★ An open-source Python deep-learning framework.
- ★ Offers out-of-the-box deep learning interaction.
- ★ Allows creating custom tasks.



Feel free to use the tool

Backend Task: DeepDIVA
https://diva-dia.github.io/DeepDIVAweb

Visualization tool: DNNViz
https://github.com/DIVA-DIA/DeepDIVA-DNNViz

Questions?