清華大學
Tsinghua University

Visual Analytics for Explainable Machine Learning

Shixia Liu (刘世霞)

Tsinghua University (清华大学)

Recent ML Systems Achieve Superhuman Performance

AlphaGo beats Go human champ

Deep Net outperforms humans in image classification

IMAGENET

Autonomous search-and-rescue drones outperform humans

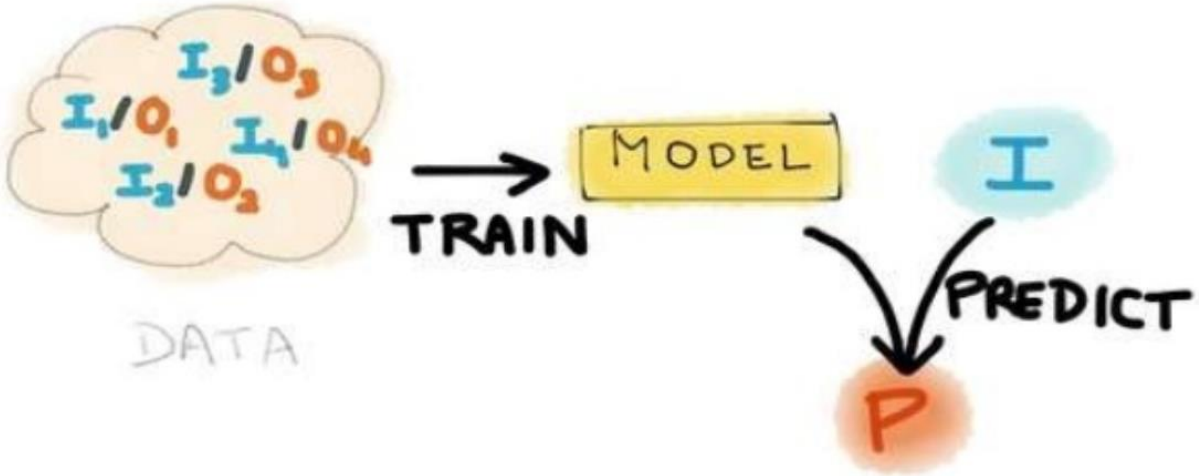
DeepStack beats professional poker players

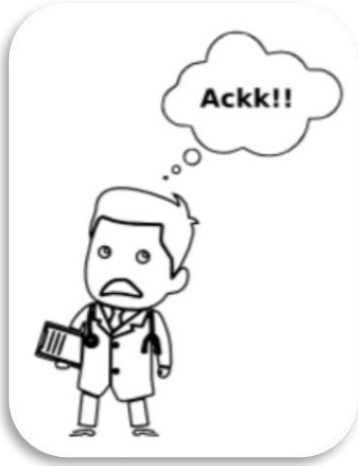
Computer out-plays humans in "doom"

IBM's Watson destroys humans in jeopardy

Deep Net beats human at recognizing traffic signs

Machine Learning as a Black Magic or Black Box





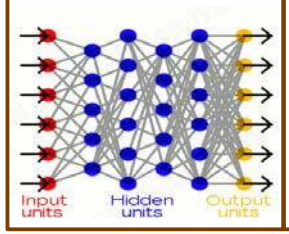
Most machine learning processes are NP-processes

Explainable AI – What Are We Trying To Do?

Today

Training Data

Learning Process



Learned Function

This is a cat
(p = .93)

Output

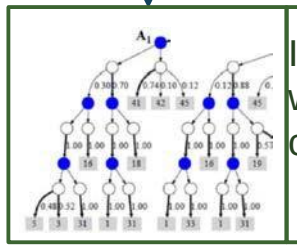
User with a Task

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow

Training Data

New Learning Process



Explainable Model

It has fur, whiskers, and claws.

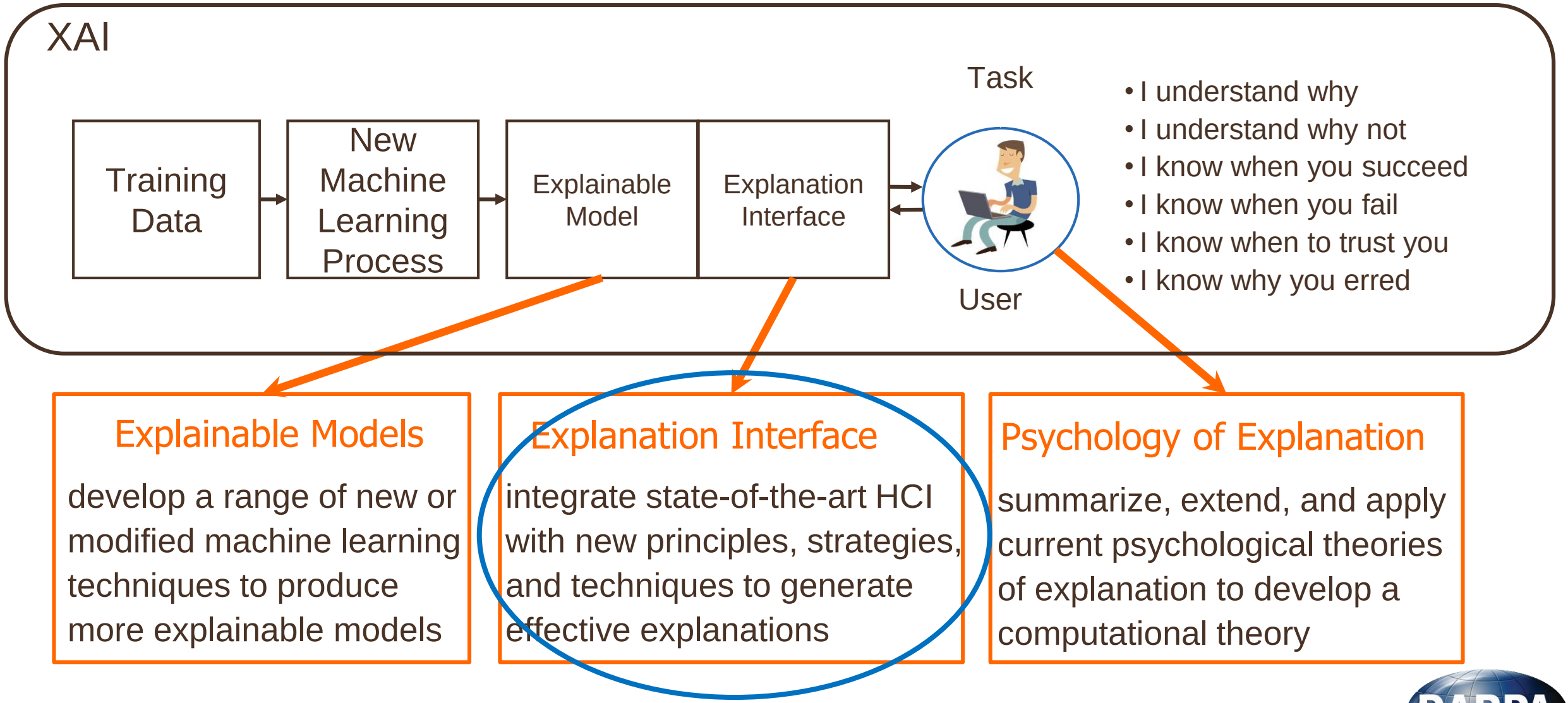
Explanation Interface

User with a Task

- I understand why
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

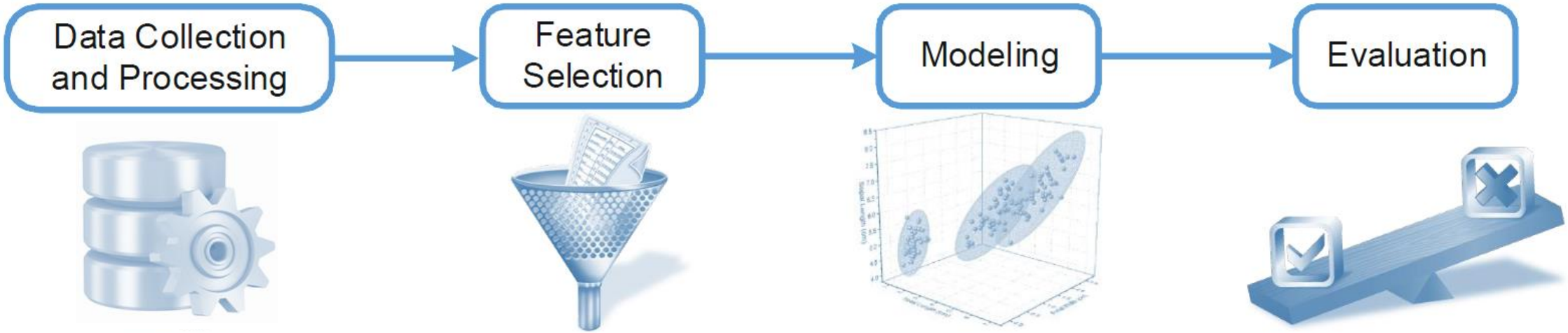
Explainable Artificial Intelligence - Darpa

DARPA - Explainable Artificial Intelligence (XAI)



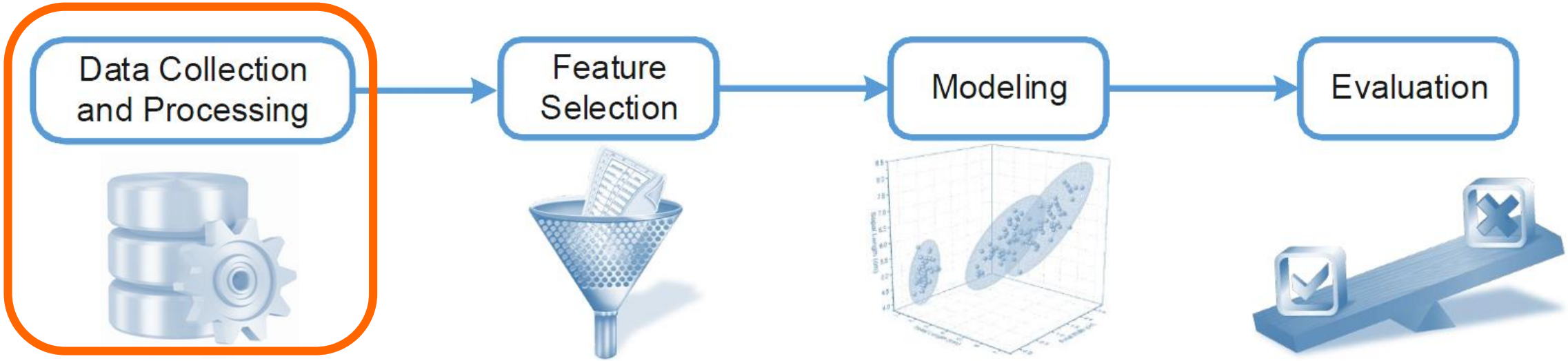
Machine Learning Pipeline

Tasks
- Preprocess and clean the data
- Select and construct appropriate features
- Select an appropriate model family
- Optimize model hyperparameters
- Postprocess machine learning models
- Critically analyze the results obtained

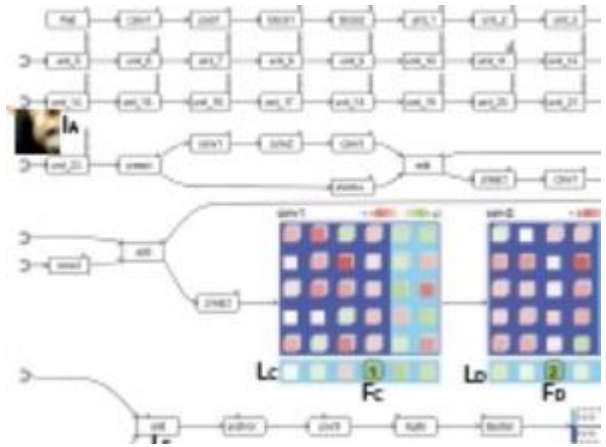


Solution = Data + ML Expertise + Computation

Data is the King of Machine Learning!



LabelInspect, IEEE VIS 2018 (TVCG)

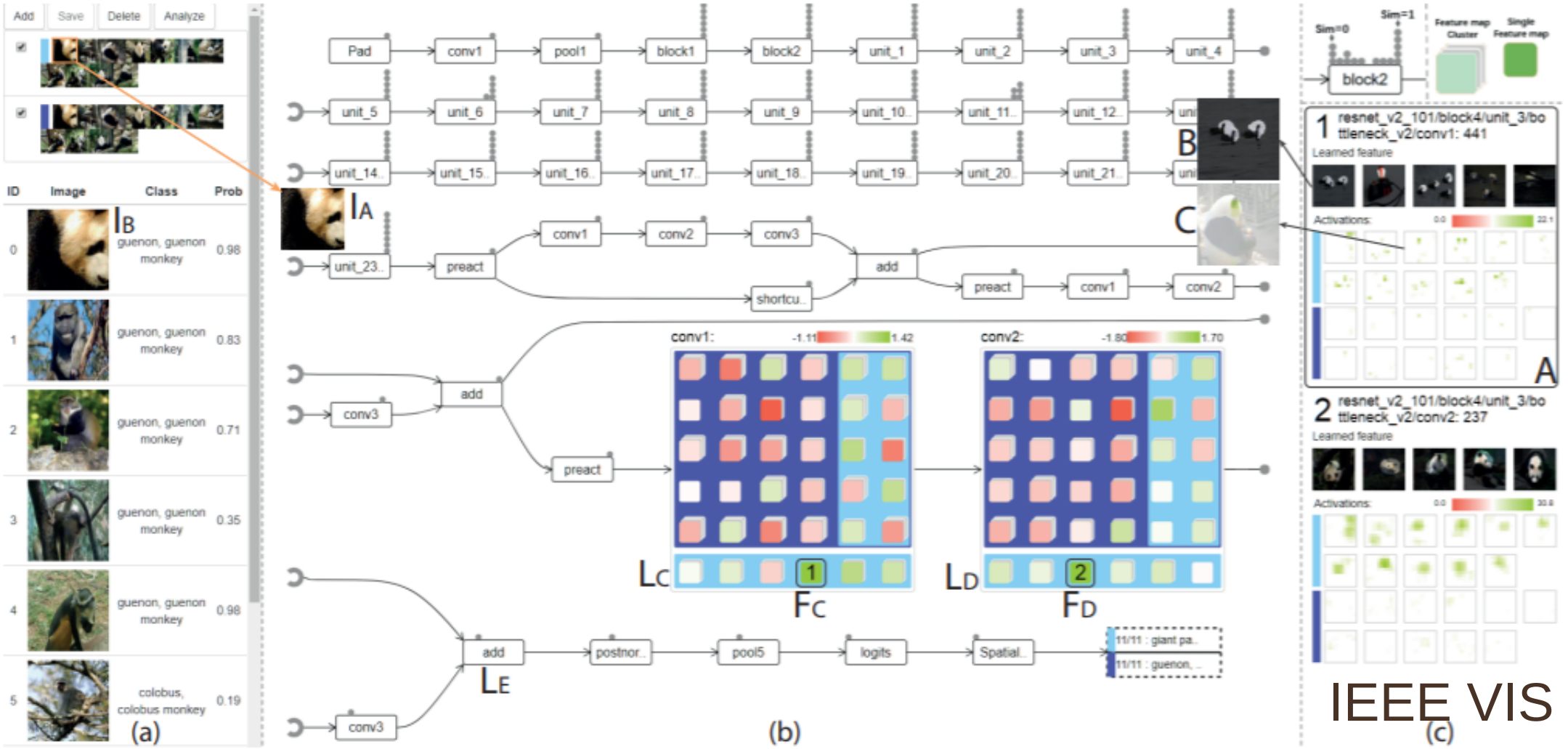


AEVis, IEEE VIS 2018 (TVCG)

Analyzing the Noise Robustness of Deep Neural Networks

Mengchen Liu*, Shixia Liu*, Hang Su†, Kelei Cao*, Jun Zhu†

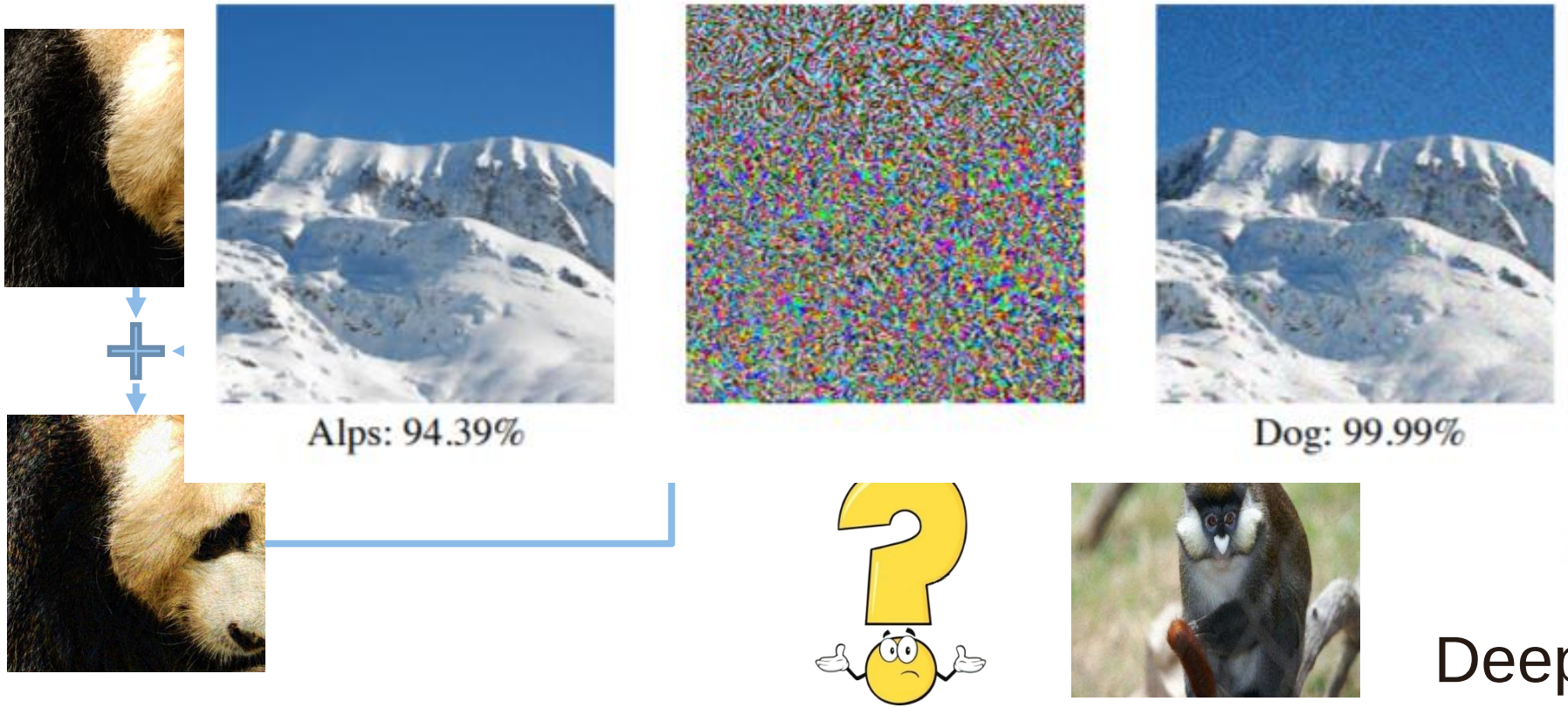
*School of Software, TNList Lab, State Key Lab for Intell. Tech. Sys., Tsinghua University
†Dept. of Comp. Sci.Tech., TNList Lab, State Key Lab for Intell. Tech. Sys., CBICR Center, Tsinghua University



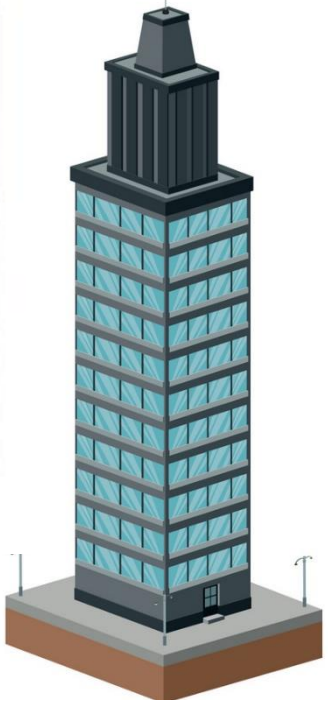
IEEE VIS 2018

Adversarial Examples

- Intentionally designed to mislead a deep neural network (DNN) into making incorrect prediction



Robustness?



Deep neural networks

Technical Challenges

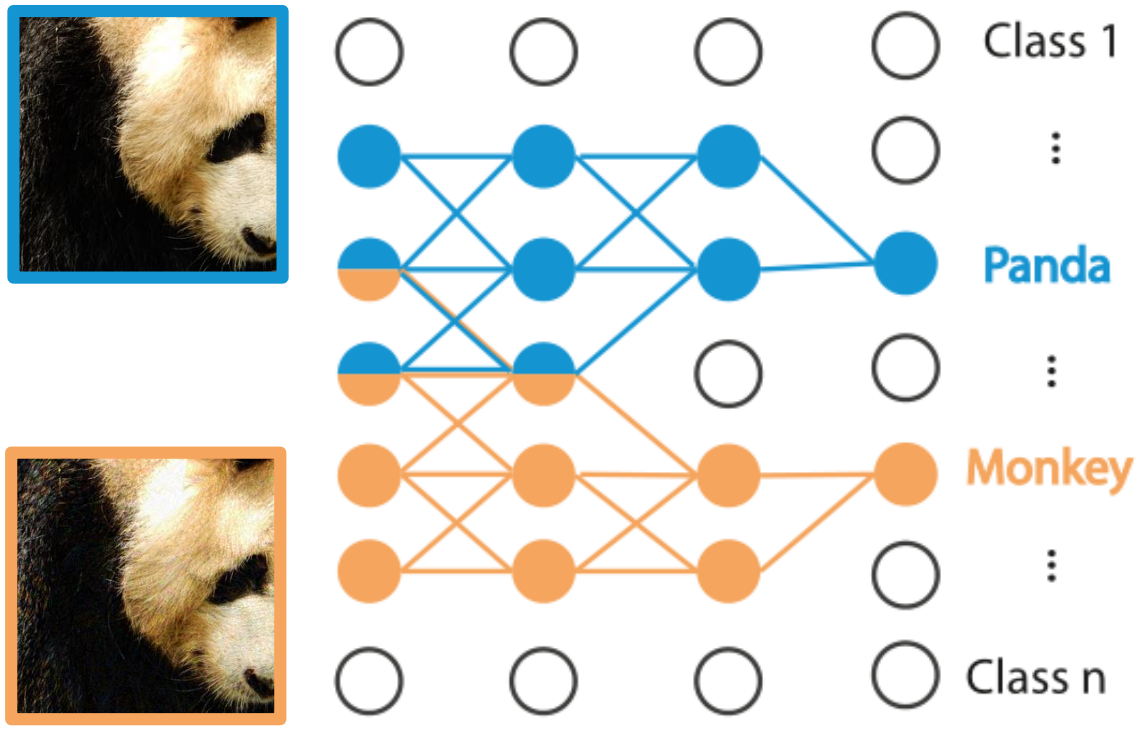
- Extract the datapath for adversarial examples
- Datapath visualization

A datapath

Hundreds of layers

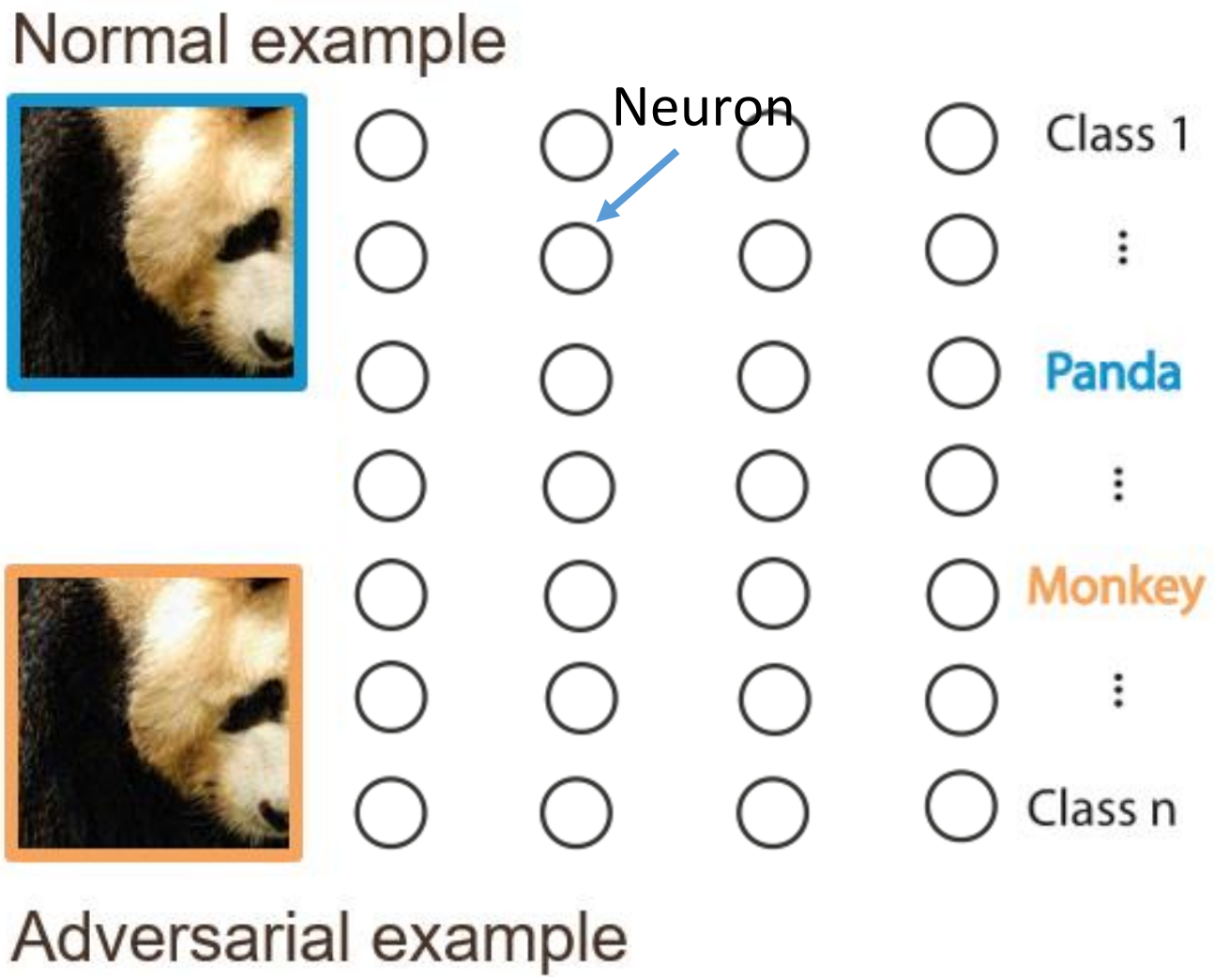
Millions of neurons

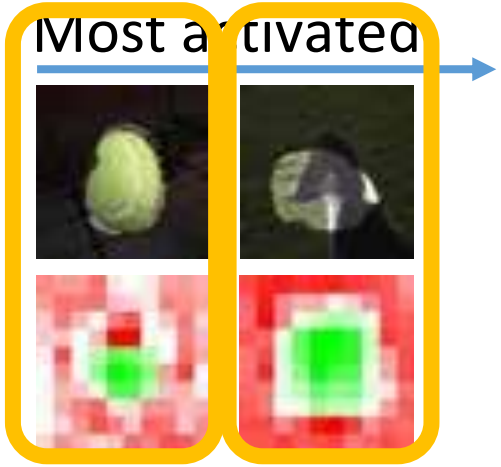
Millions of connections



Datapath Extraction - Motivation

- Current metho
  - Most activate
- Problem
  - Misleading re
    highly recogr
- Reason
  - Neurons hav
  - Gap betweer





Datapath Extraction - Formulation

- The critical neurons for a prediction: the neurons that highly contributed to the final prediction
- Subset selection
  - Keep the original prediction by selecting a minimized subset of neurons

$$N^{opt} = \arg\min_{N_s \subseteq N} (p(x) - p(x; N_s))^2 + \lambda |N_s|^2$$

- Extend to a set of images X

$$N^{opt} = \arg\min_{N_s \subseteq N} \sum_{x_k \in X} (p(x_k) - p(x_k; N_s))^2 + \lambda |N_s|^2$$

Datapath Extraction - Solution

- Directly solving: time-consuming
  - NP-complete
  - Large search space due to the large number of neurons in a CNN

Quadratic approximation

Divide-and-conquer-based search space reduction

An accurate approximation in smaller search space

$$N^{opt} = \arg\min_{N_s \subseteq N} \sum_{x_k \in X} \left(p(x_k) - p(x_k; N_s)\right)^2 + \lambda |N_s|^2$$

Datapath Extraction – Search Space Reduction

- Original problem: 57.78 million dims

Split into layers

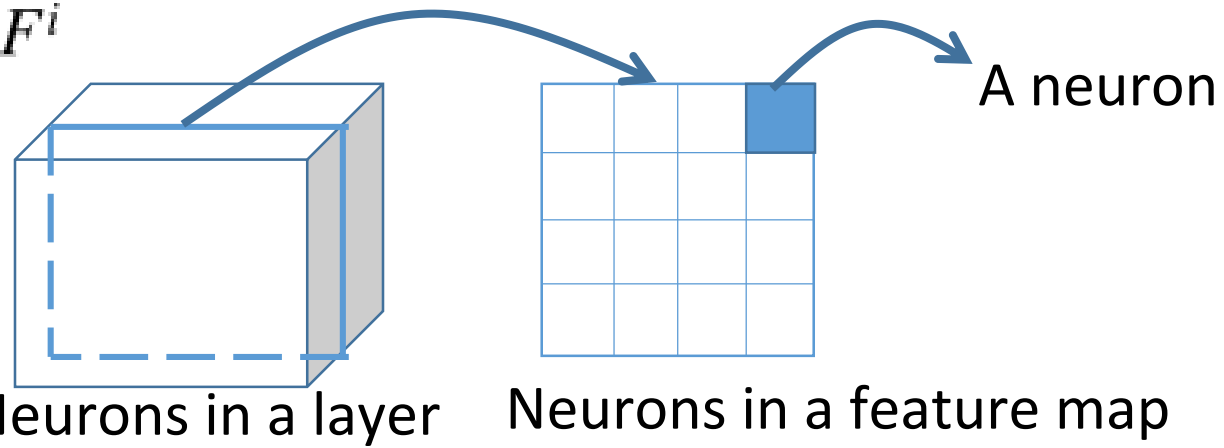
$$N_{opt}^{i} = \arg\min_{N_s^i \subseteq N^i} \left(p(x) - p(x; N_s^i \cup N^{-i})\right)^2 + \lambda^i |N_s^i|^2,$$

2k ~ 1.44 million dims

Split into feature maps

$$F_{opt}^{i} = \arg\min_{F_s^i \subseteq F^i} \left(p(x) - p(x; F_s^i \cup F^{-i})\right)^2 + \lambda^i |F_s^i|^2$$

64 ~ 2k dims



Datapath Extraction – Quadratic Approximation

$$F^i_{opt} = \arg\min_{F^i_s \subseteq F^i} (p(x) - p(x; F^i_s \cup F^{-i}))^2 + \lambda^i |F^i_s|^2$$ Still NP

Reformulate

$$\mathbf{z}^i_{opt} = \arg\min_{z^i_j \in \{0,1\}} (p(x) - p(x; \mathbf{z}^i))^2 + \lambda^i (\sum_j z^i_j)^2,$$

$\mathbf{z}^i = [z^i_1, \ldots, z^i_n]$

z^i_j whether the *j*-th feature map in layer *i* is critical

Discrete to continuous

$$\mathbf{z}^i_{opt} = \arg\min_{z^i_j \in [0,1]} (p(x) - p(x; \mathbf{z}^i))^2 + \lambda^i (\sum_j z^i_j)^2,$$

Needs to calculate $\partial p / \partial z^i_j$ by BP in each iteration

Taylor decomposition

$$\mathbf{z}^i_{opt} = \arg\min_{z^i_j \in [0,1]} \mathbf{z}^i (Q^i + \lambda^i I)(\mathbf{z}^i)^T - 2q\mathbf{q}^i \cdot \mathbf{z}^i,$$

Quadratic optimization

$$Q[j,k] = (\mathbf{a}_j \cdot \frac{\partial p}{\partial \mathbf{a}^i_j}) \times (\mathbf{a}_k \cdot \frac{\partial p}{\partial \mathbf{a}^i_k})$$

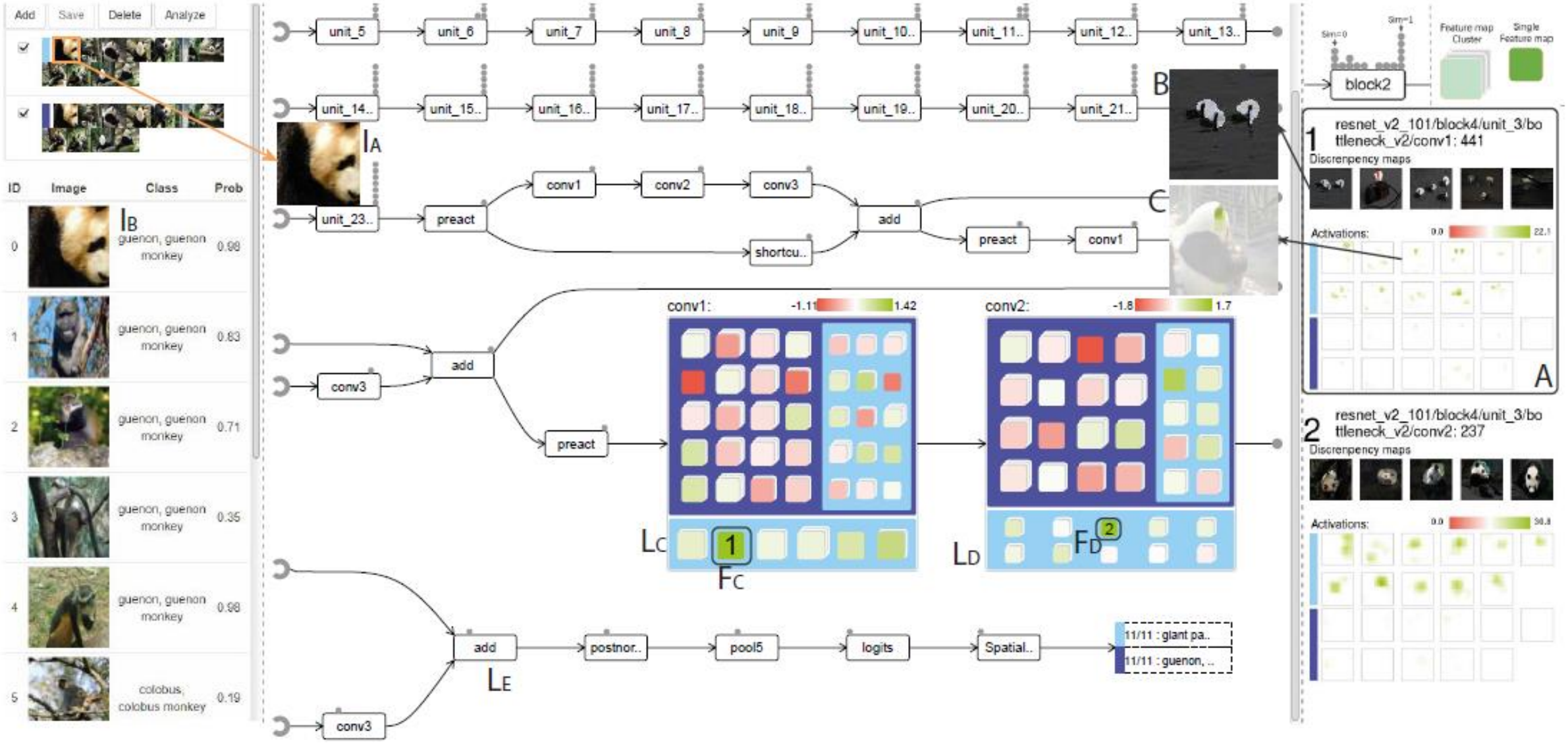
1. Bridge the gap between activation and prediction
2. Each element in *Q* approximately models the interaction between feature map *j* and feature map *k*

$\mathbf{a}_j$: activation vector of the j-th feature map

$$\mathbf{q}^i = [\mathbf{a}^i_1 \cdot \frac{\partial p}{\partial \mathbf{a}^i_1}\Big|_{\mathbf{a}}, \cdots, \mathbf{a}^i_n \cdot \frac{\partial p}{\partial \mathbf{a}^i_n}\Big|_{\mathbf{a}}]$$

$$Q = (\mathbf{q}^i)^T \mathbf{q}^i$$

Datapath Visualization



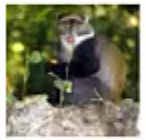
+noise

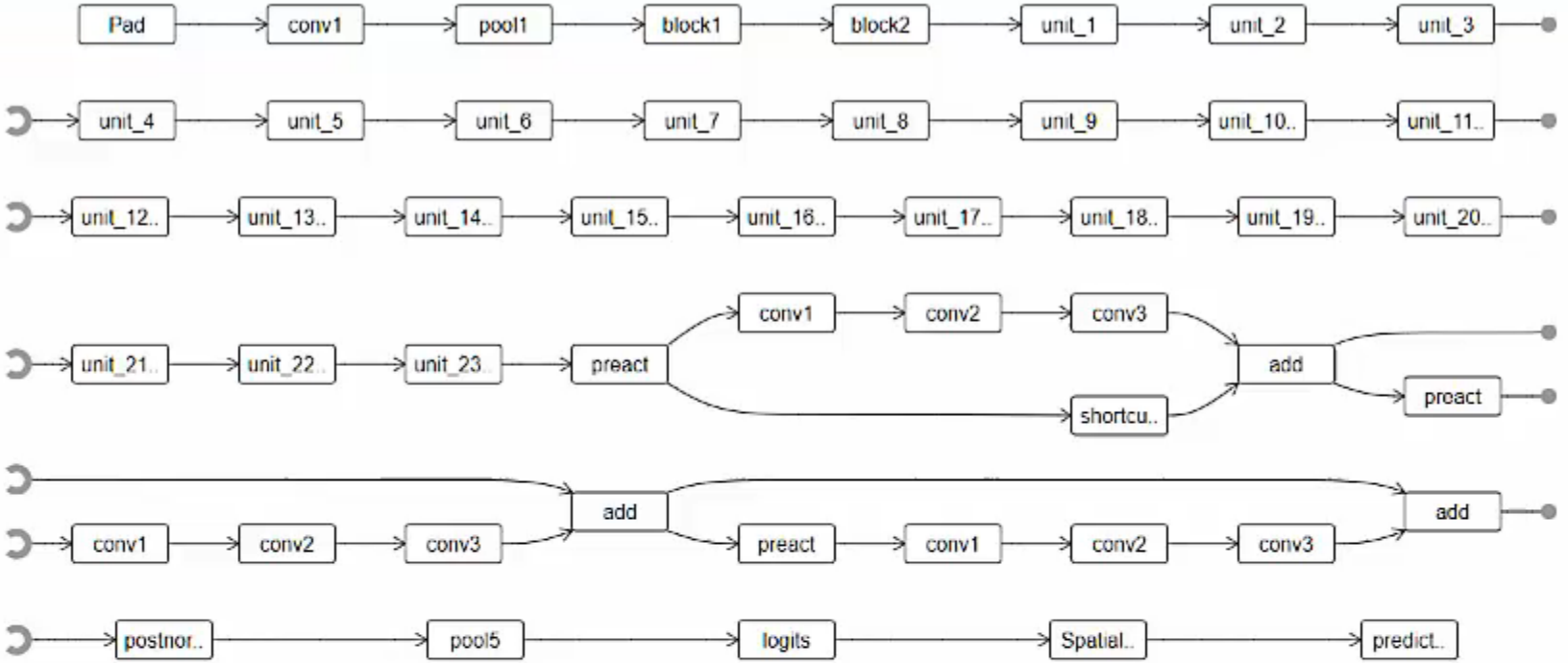
classified as a

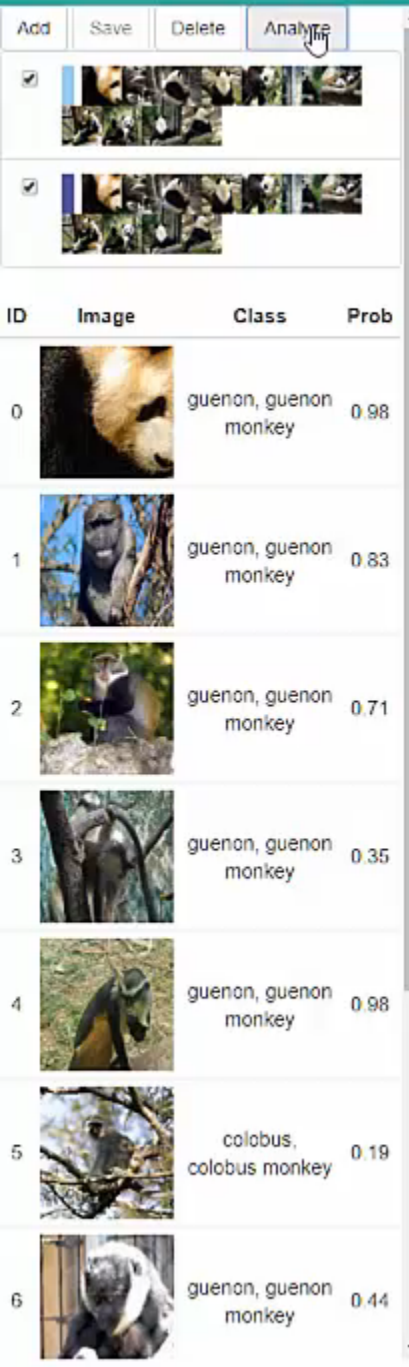
Why?

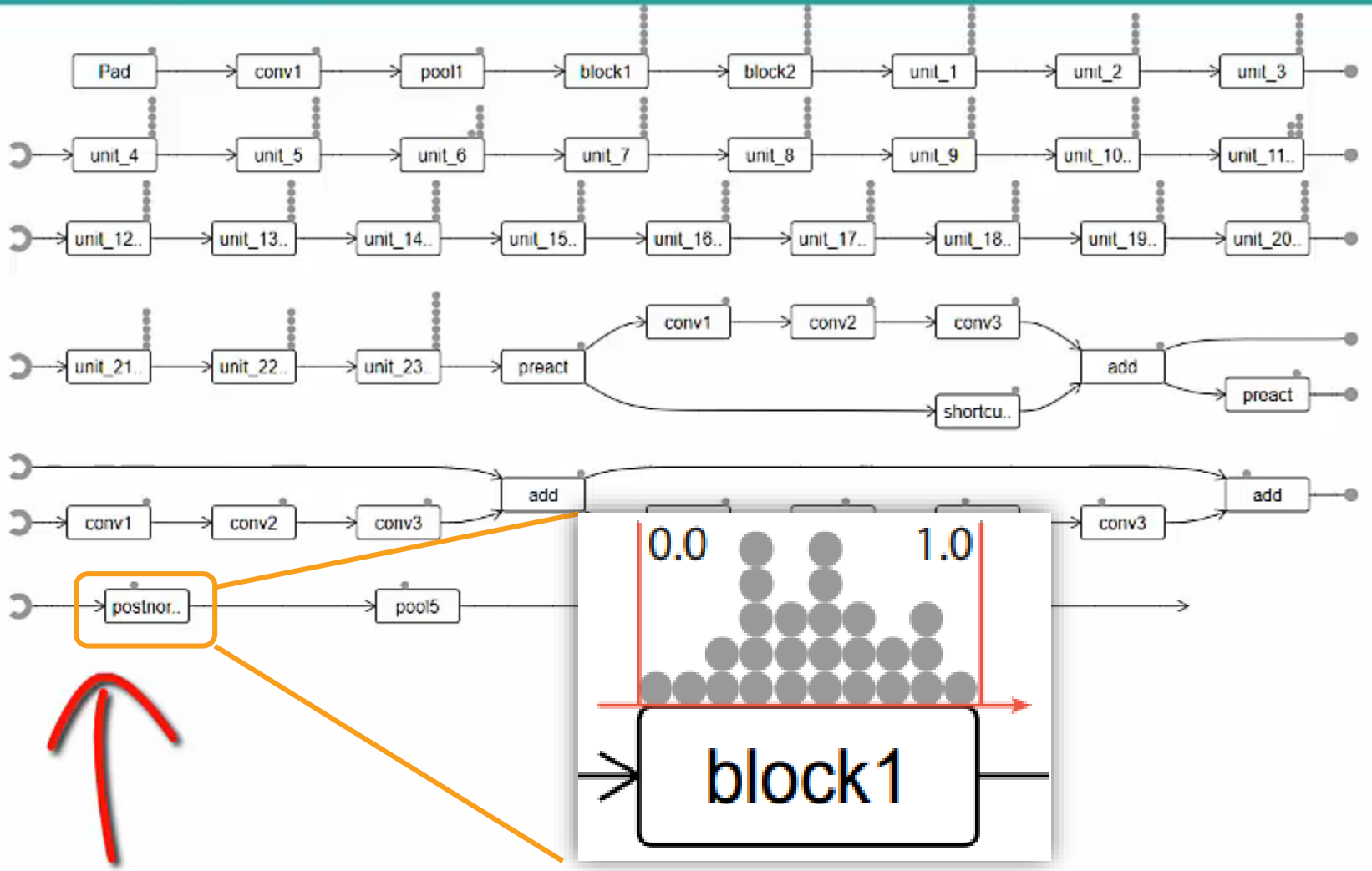
Analyzing the Noise Robustness of Deep Neural Networks

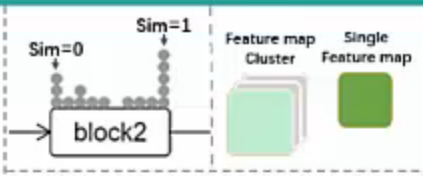
Add | Save | Delete | Analyze

ID	Image	Class	Prob
0		guenon, guenon monkey	0.98
1		guenon, guenon monkey	0.83
2		guenon, guenon monkey	0.71
3		guenon, guenon monkey	0.35
4		guenon, guenon monkey	0.98
5		colobus, colobus monkey	0.19
6		guenon, guenon monkey	0.44





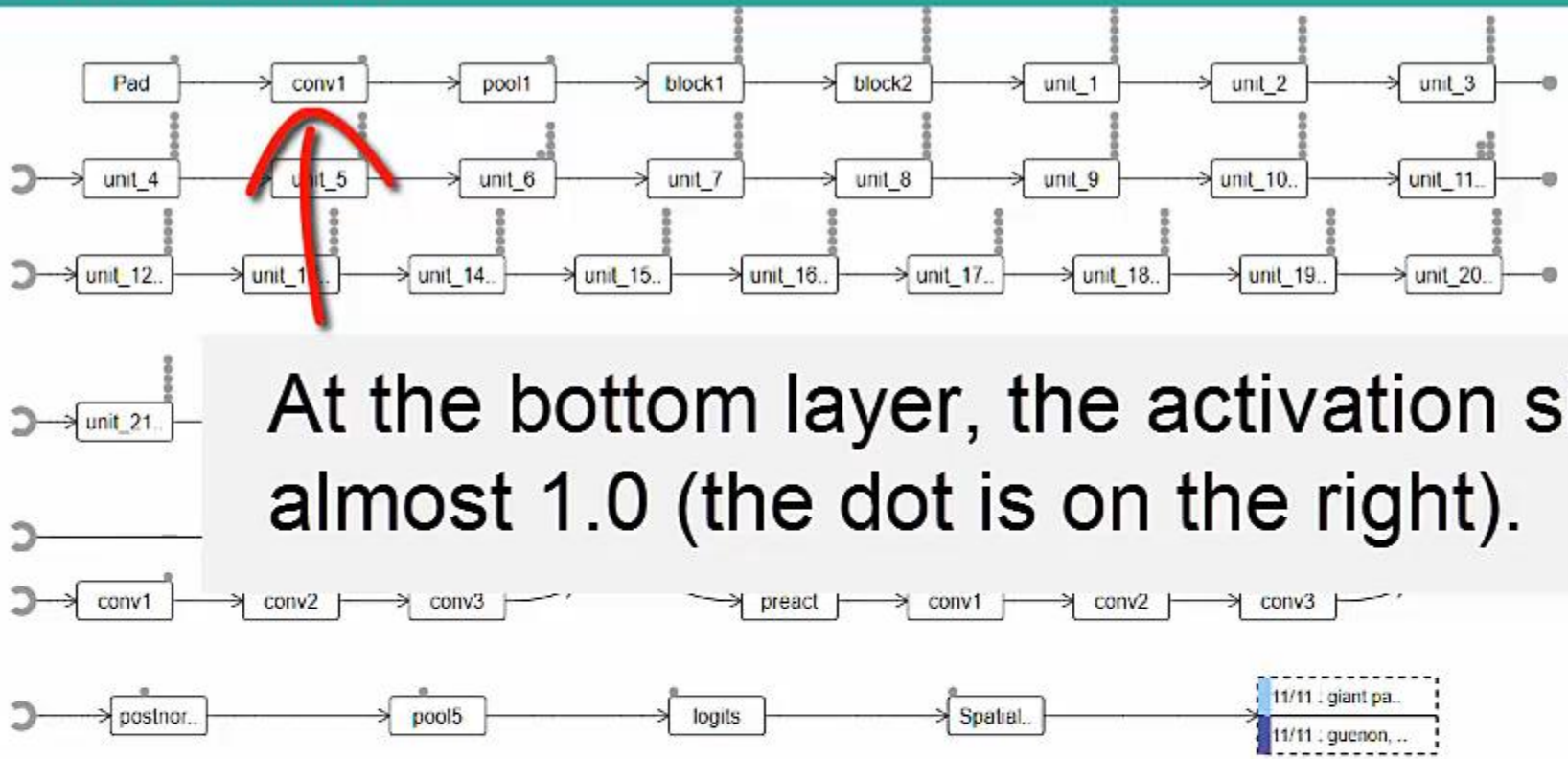


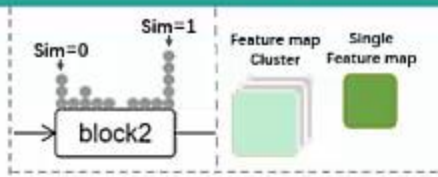


The dot plots encode the activation similarity between these two datapaths.

Add | Save | Delete | Analyze

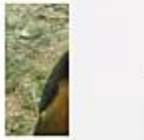
ID	Image	Class	Prob
0		guenon, guenon monkey	0.98
1		guenon, guenon monkey	0.83
2		guenon, guenon monkey	0.71
3		guenon, guenon monkey	0.35
4		guenon, guenon monkey	0.98
5		colobus, colobus monkey	0.19
6		guenon, guenon monkey	0.44

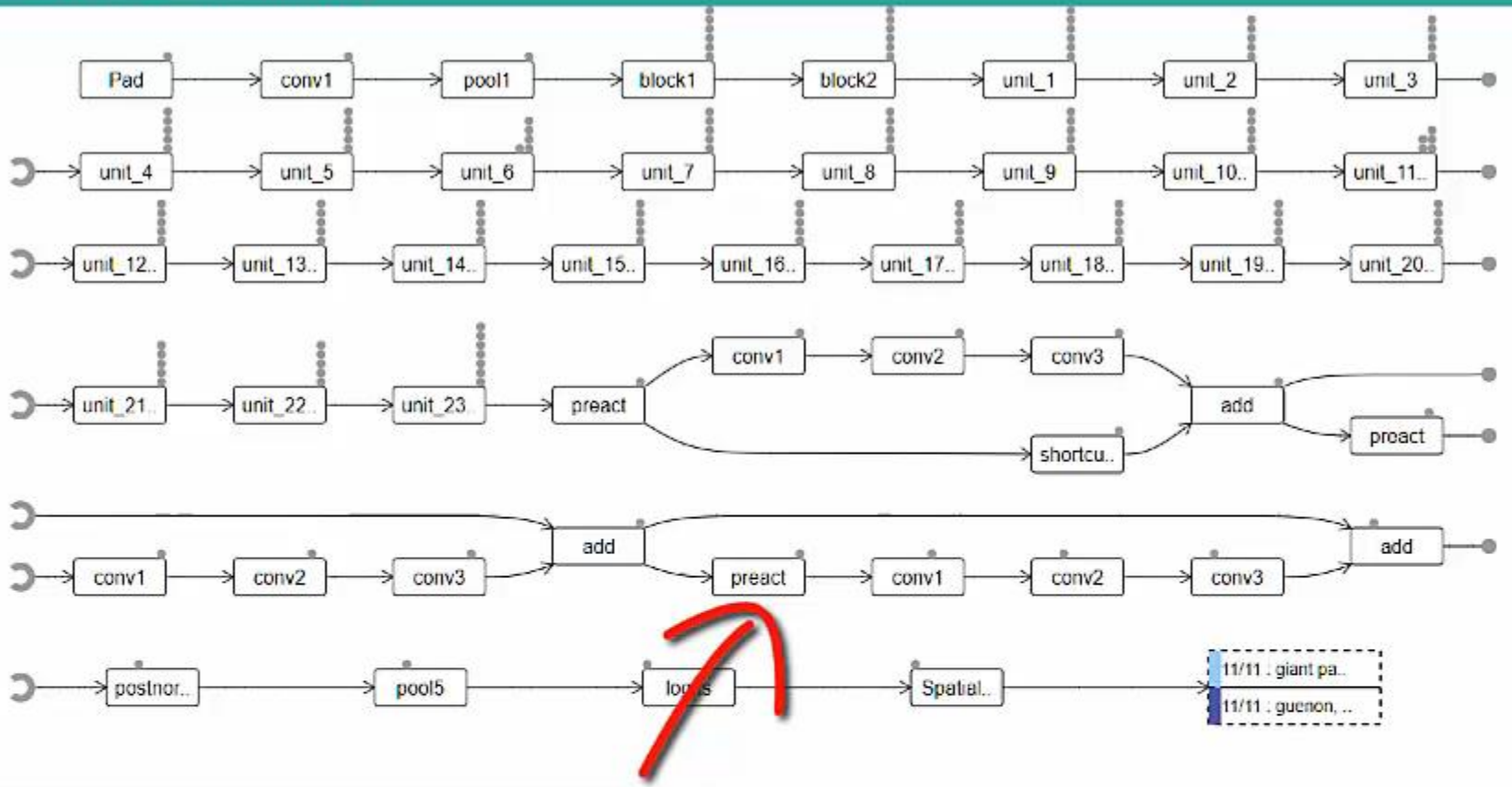


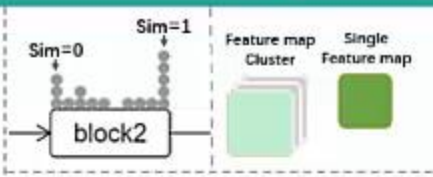


At the bottom layer, the activation similarity is almost 1.0 (the dot is on the right).

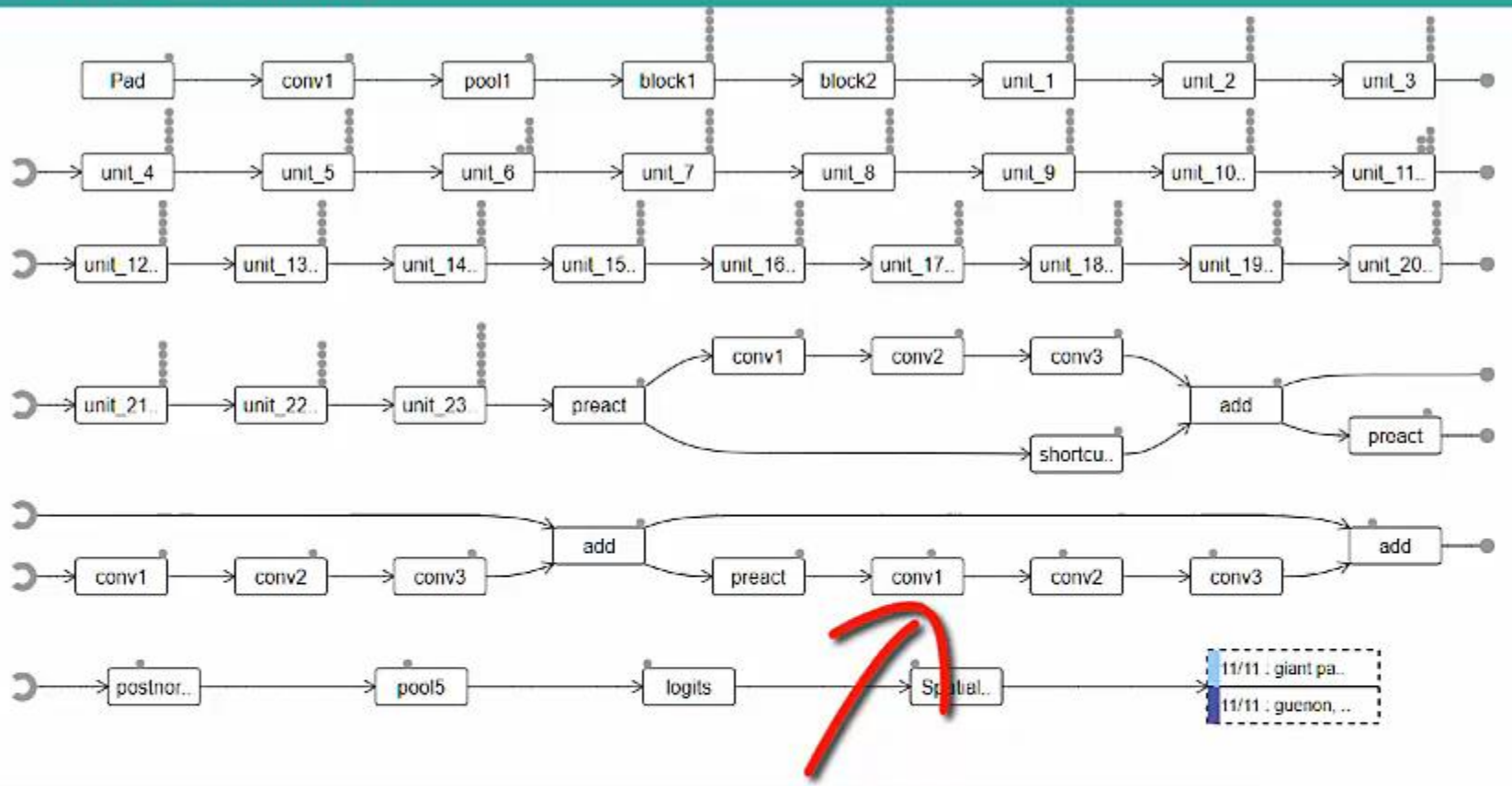
Add Save Delete Analyze

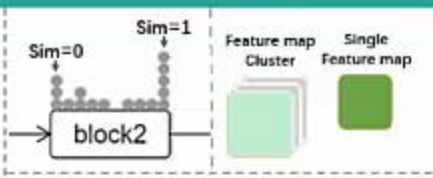
ID	Image	Class	Prob
0		guenon, guenon monkey	0.98
1		guenon, guenon monkey	0.83
2		guenon, guenon monkey	0.71
3		guenon, guenon	0.35
4			
5		colobus, colobus monkey	0.19
6		guenon, guenon monkey	0.44





This similarity remains to be 1.0 until this layer.

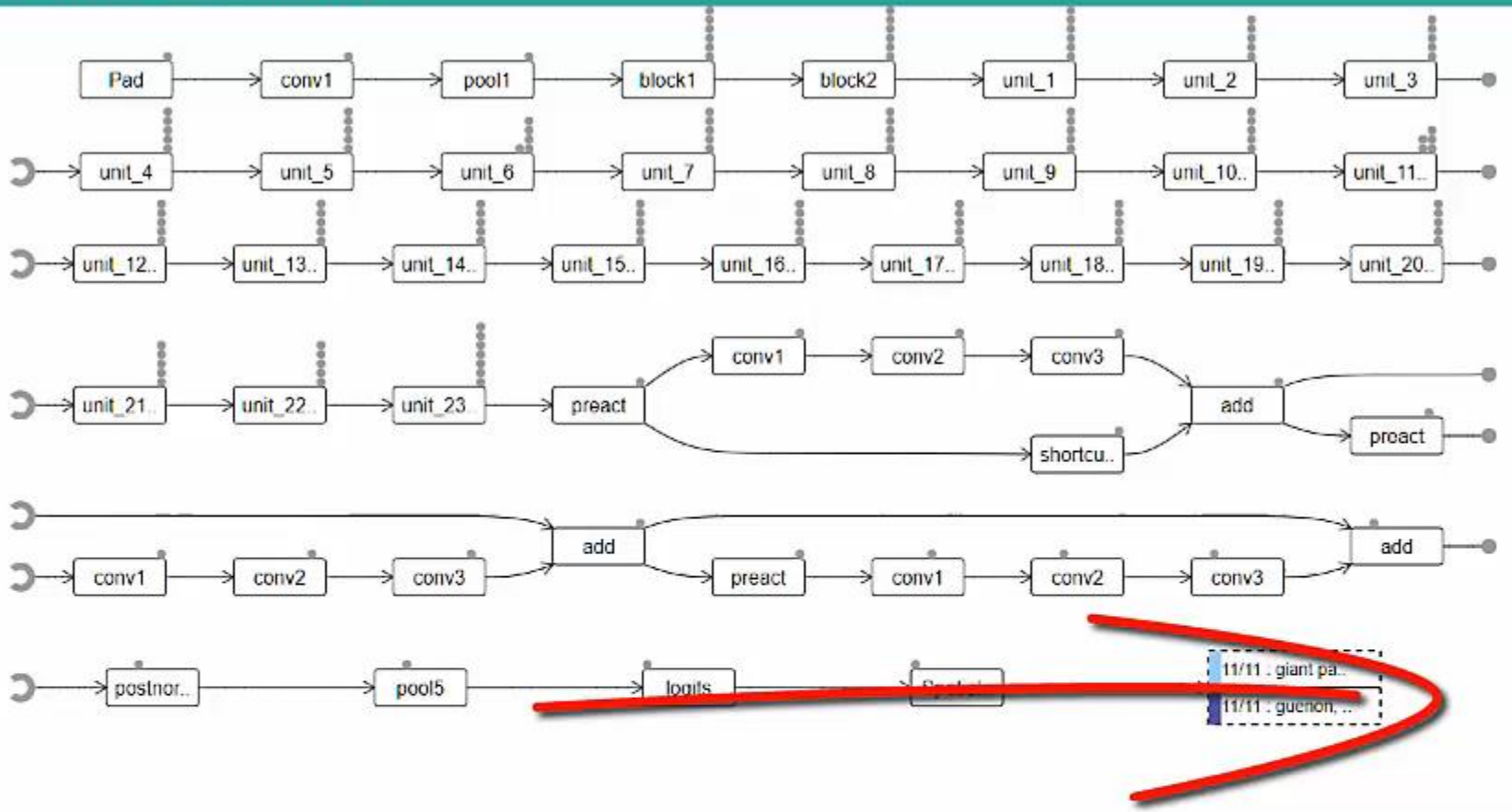


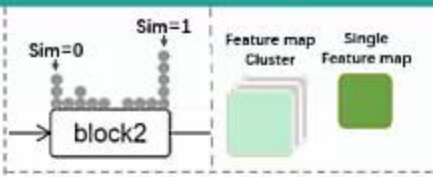


A "diverging point" appears, where the activation similarity largely decreases.

Add | Save | Delete | Analyze

ID	Image	Class	Prob
0		guenon, guenon monkey	0.98
1		guenon, guenon monkey	0.83
2		guenon, guenon monkey	0.71
3		guenon, guenon monkey	0.35
4		guenon, guenon monkey	0.98
5		colobus, colobus monkey	0.19
6		guenon, guenon monkey	0.44

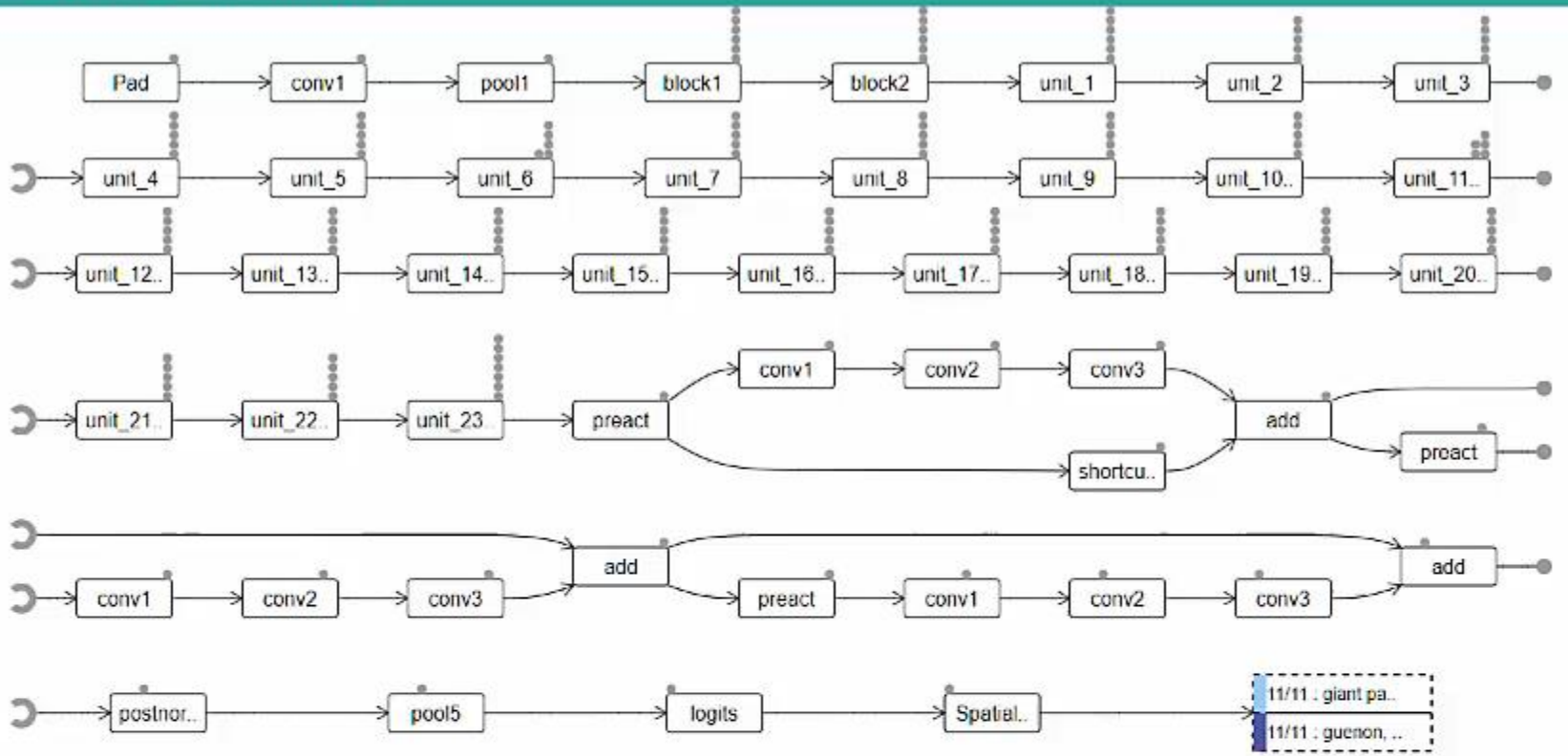


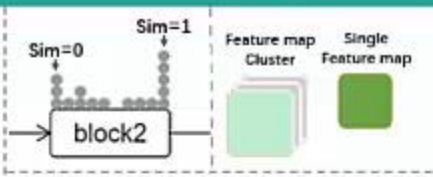


In the following layers, the activation similarity continues decreasing to 0.

Add | Save | Delete | Analyze

ID	Image	Class	Prob
0		guenon, guenon monkey	0.98
1		guenon, guenon monkey	0.83
2		guenon, guenon monkey	0.71
3		guenon, guenon monkey	0.35
4		guenon, guenon monkey	0.98
5		colobus, colobus monkey	0.19
6		guenon, guenon monkey	0.44



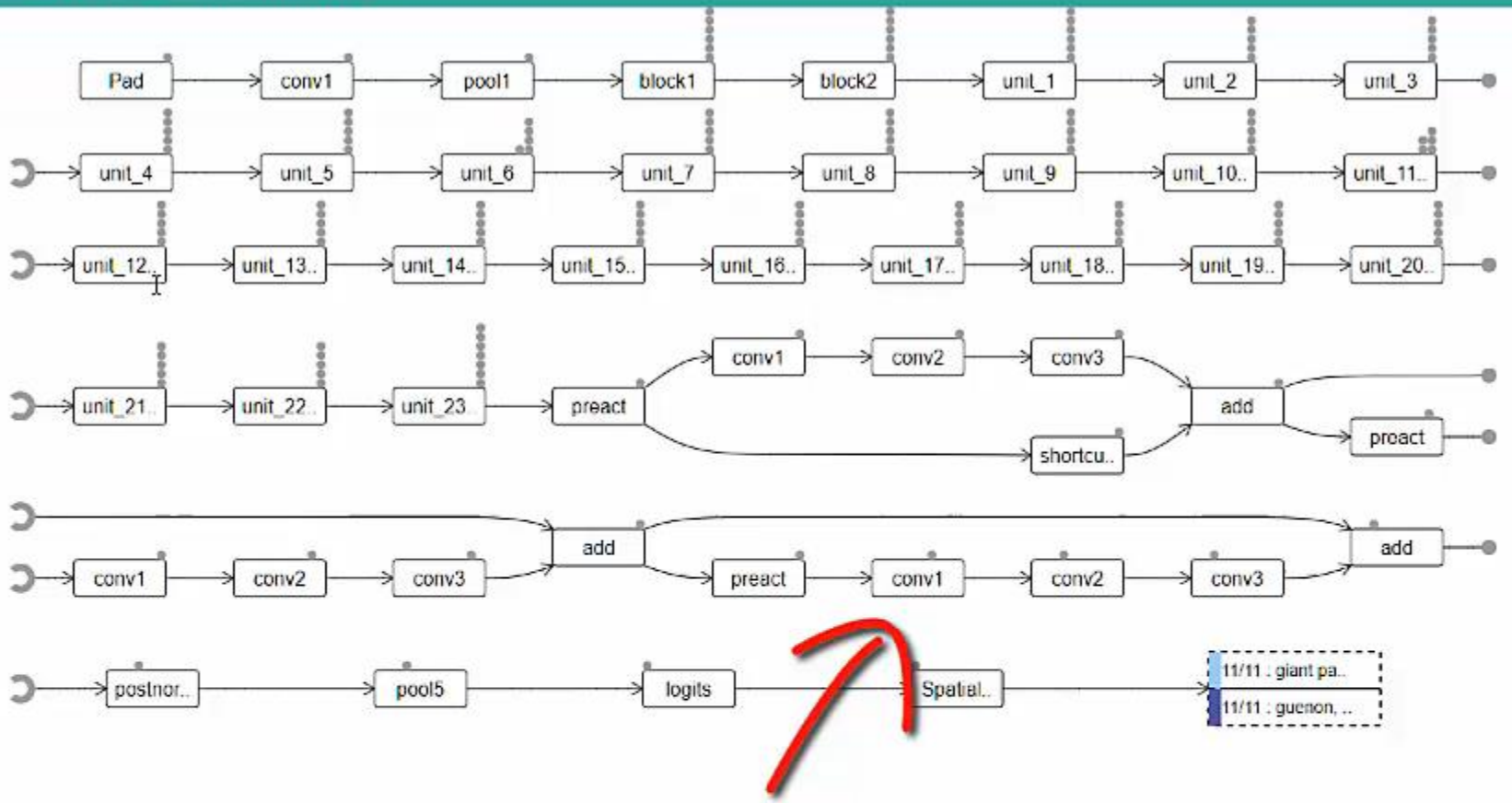


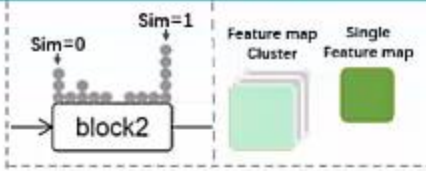
Results in the misclassification.

Analyzing the Noise Robustness of Deep Neural Networks

Add | Save | Delete | Analyze

ID	Image	Class	Prob
0		guenon, guenon monkey	0.98
1		guenon, guenon monkey	0.83
2		guenon, guenon monkey	0.71
3		guenon, guenon monkey	0.35
4		guenon, guenon monkey	0.98
5		colobus, colobus monkey	0.19
6		guenon, guenon monkey	0.44

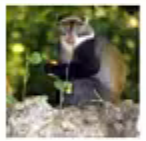


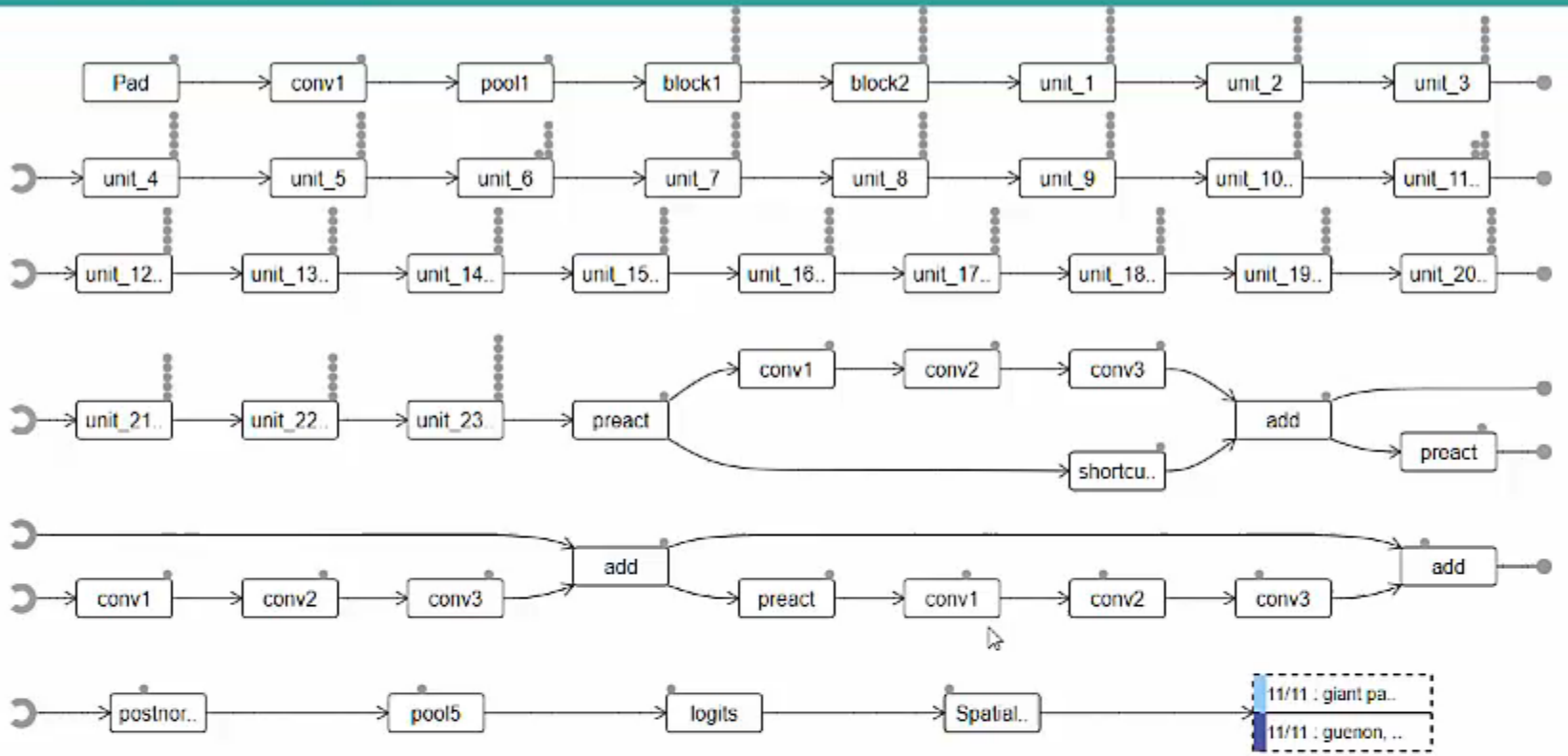


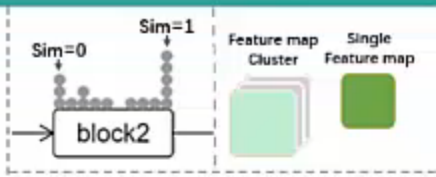
Analyze which feature map is critical for the divergence.

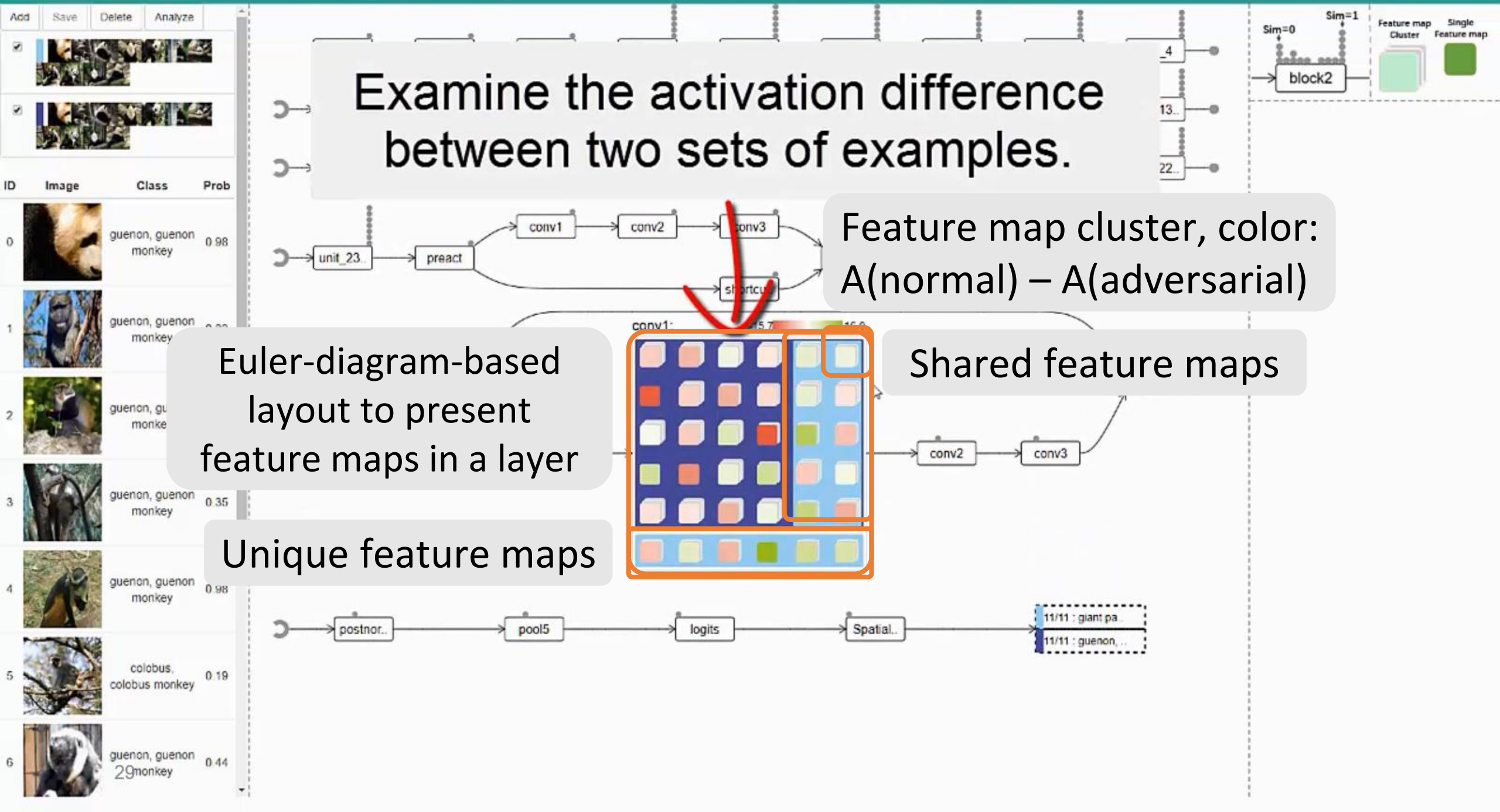
Analyzing the Noise Robustness of Deep Neural Networks

Add | Save | Delete | Analyze

ID	Image	Class	Prob
0		guenon, guenon monkey	0.98
1		guenon, guenon monkey	0.83
2		guenon, guenon monkey	0.71
3		guenon, guenon monkey	0.35
4		guenon, guenon monkey	0.98
5		colobus, colobus monkey	0.19
6		guenon, guenon monkey	0.44



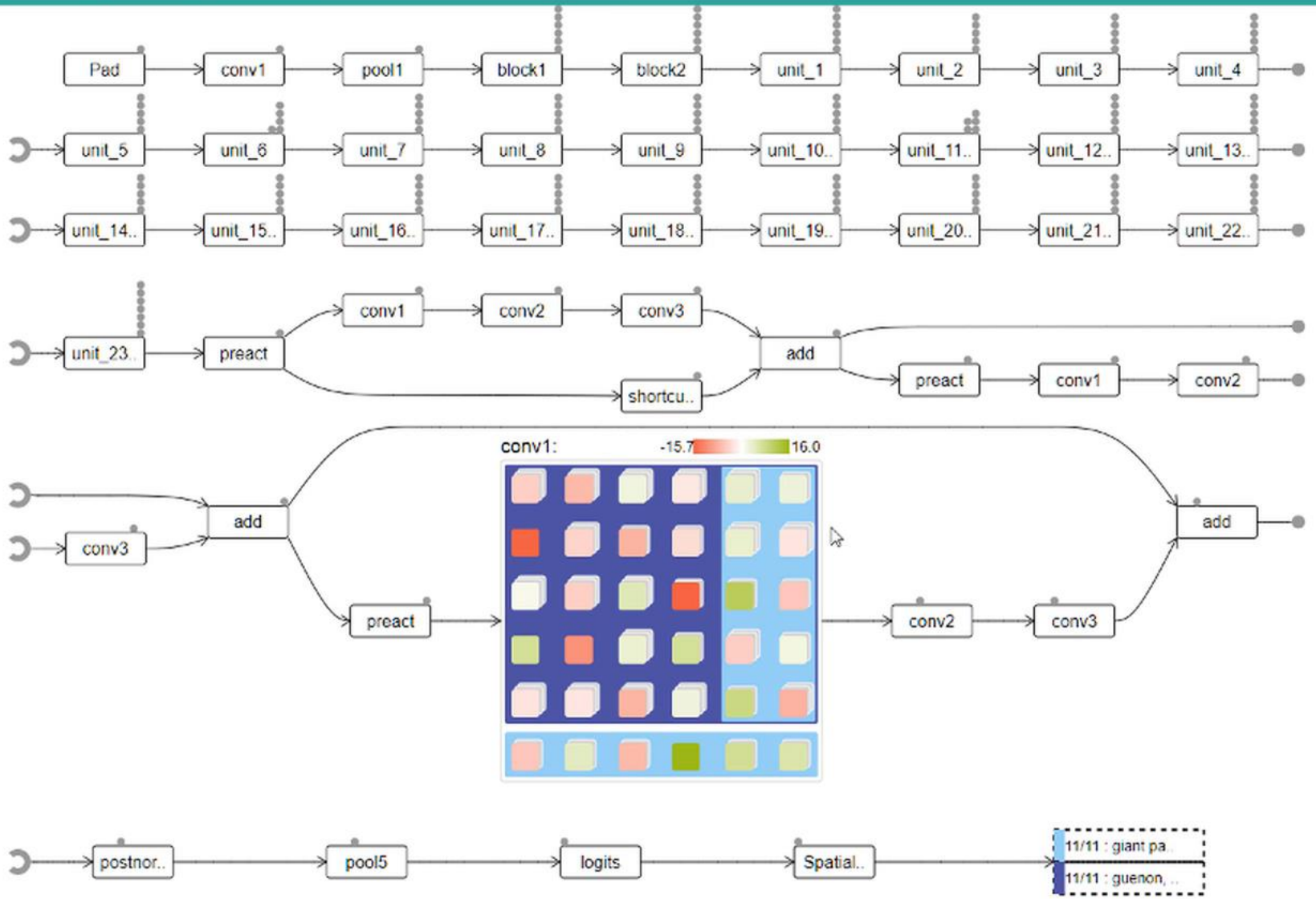


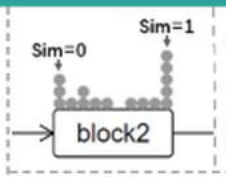

Add
Save
Delete
Analyze
ID
Image
Class
Prob
0
guenon, guenon monkey
0.98
1
guenon, guenon monkey
2
guenon, gu monke
3
guenon, guenon monkey
0.35
4
guenon, guenon monkey
0.98
5
colobus, colobus monkey
0.19
6
guenon, guenon 29monkey
0.44
Examine the activation difference between two sets of examples.
unit_23
preact
conv1
conv2
conv3
shortcut
Feature map cluster, color: A(normal) – A(adversarial)
Euler-diagram-based layout to present feature maps in a layer
Shared feature maps
Unique feature maps
conv1:
conv2
conv3
postnor..
pool5
logits
Spatial..
11/11 : giant pa..
11/11 : guenon, ..
Sim=0
Sim=1
block2
Feature map Cluster
Single Feature map

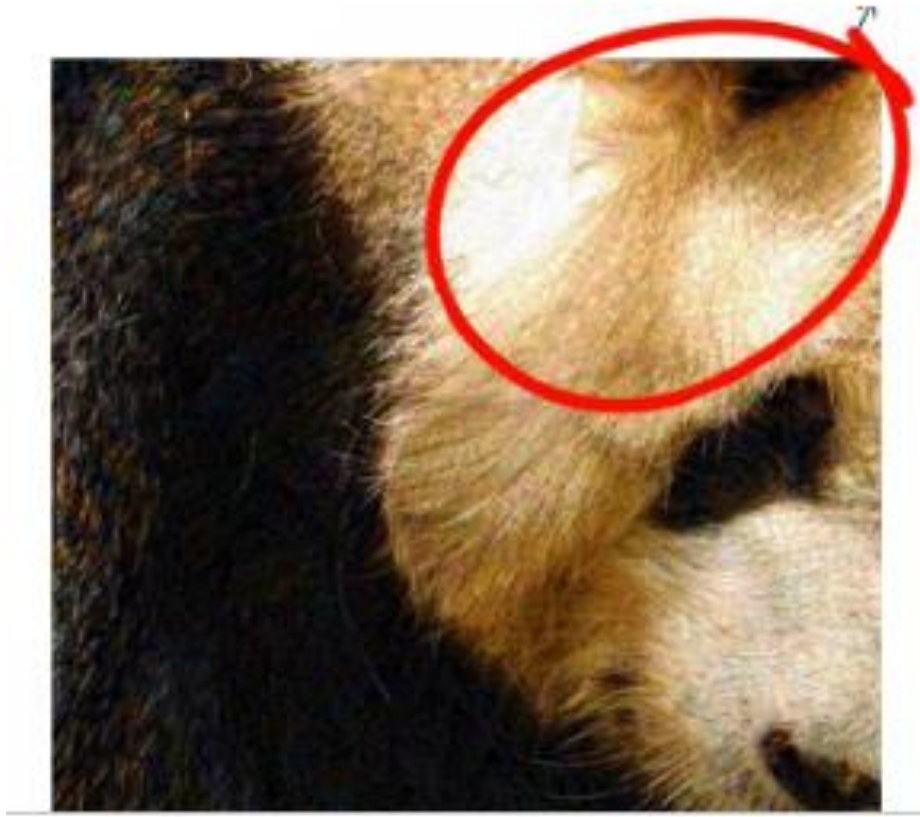
Analyzing the Noise Robustness of Deep Neural Networks

Add | Save | Delete | Analyze

ID	Image	Class	Prob
0		guenon, guenon monkey	0.98
1		guenon, guenon monkey	0.83
2		guenon, guenon monkey	0.71
3		guenon, guenon monkey	0.35
4		guenon, guenon monkey	0.98
5		colobus, colobus monkey	0.19
6		guenon, guenon monkey	0.44

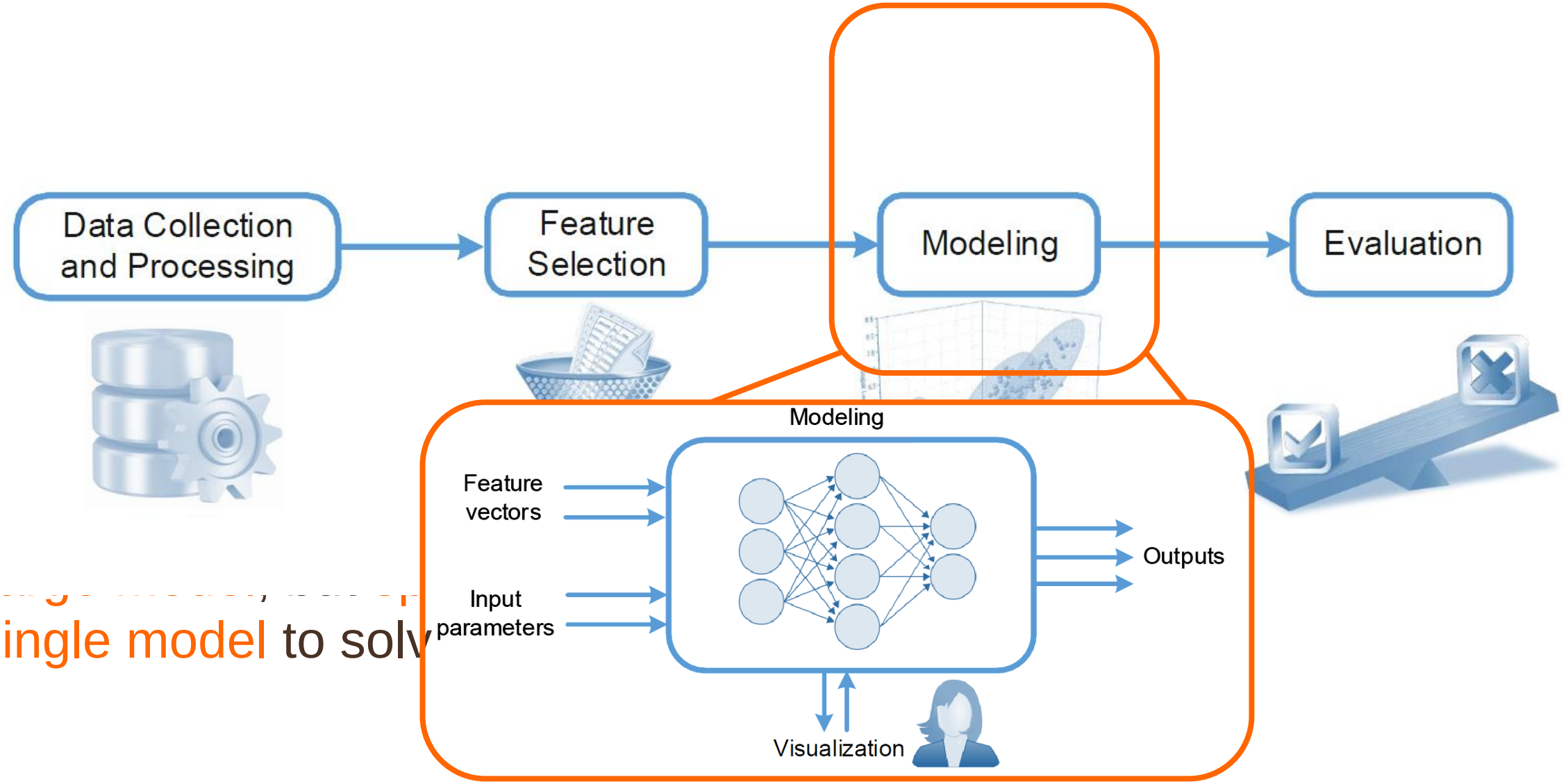







Activations:
0.0
16

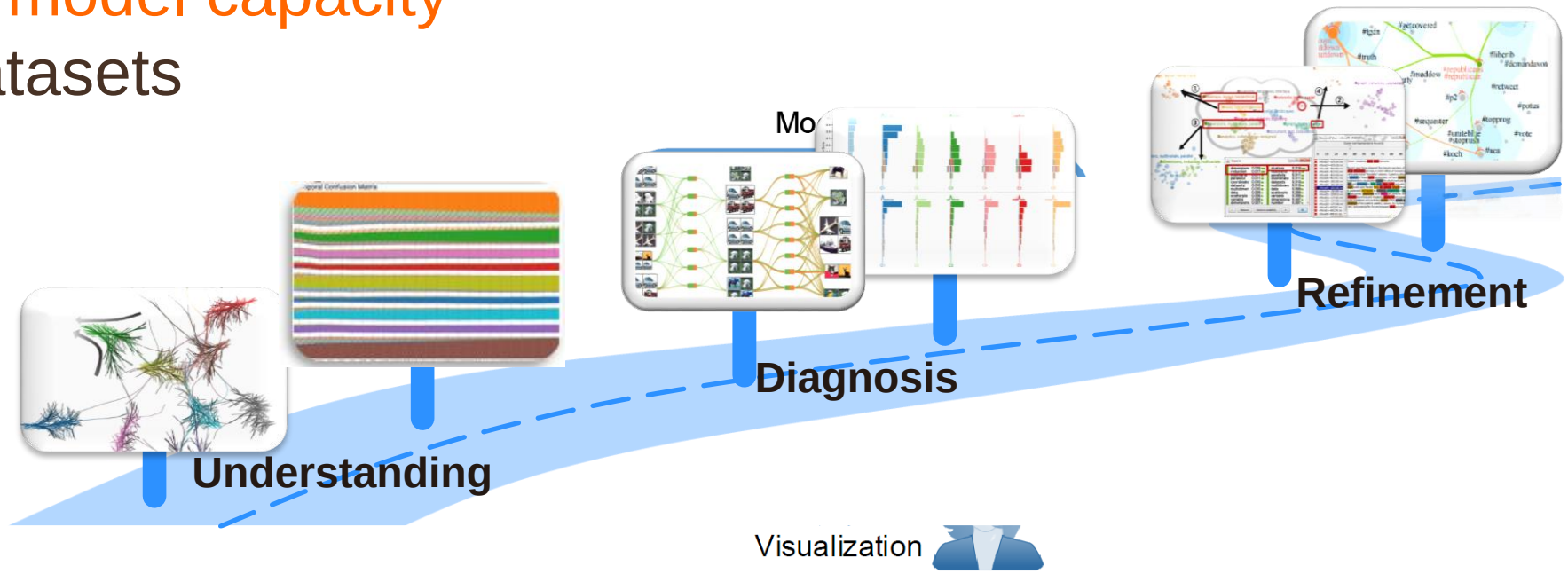
Visualization and Machine Learning pipeline



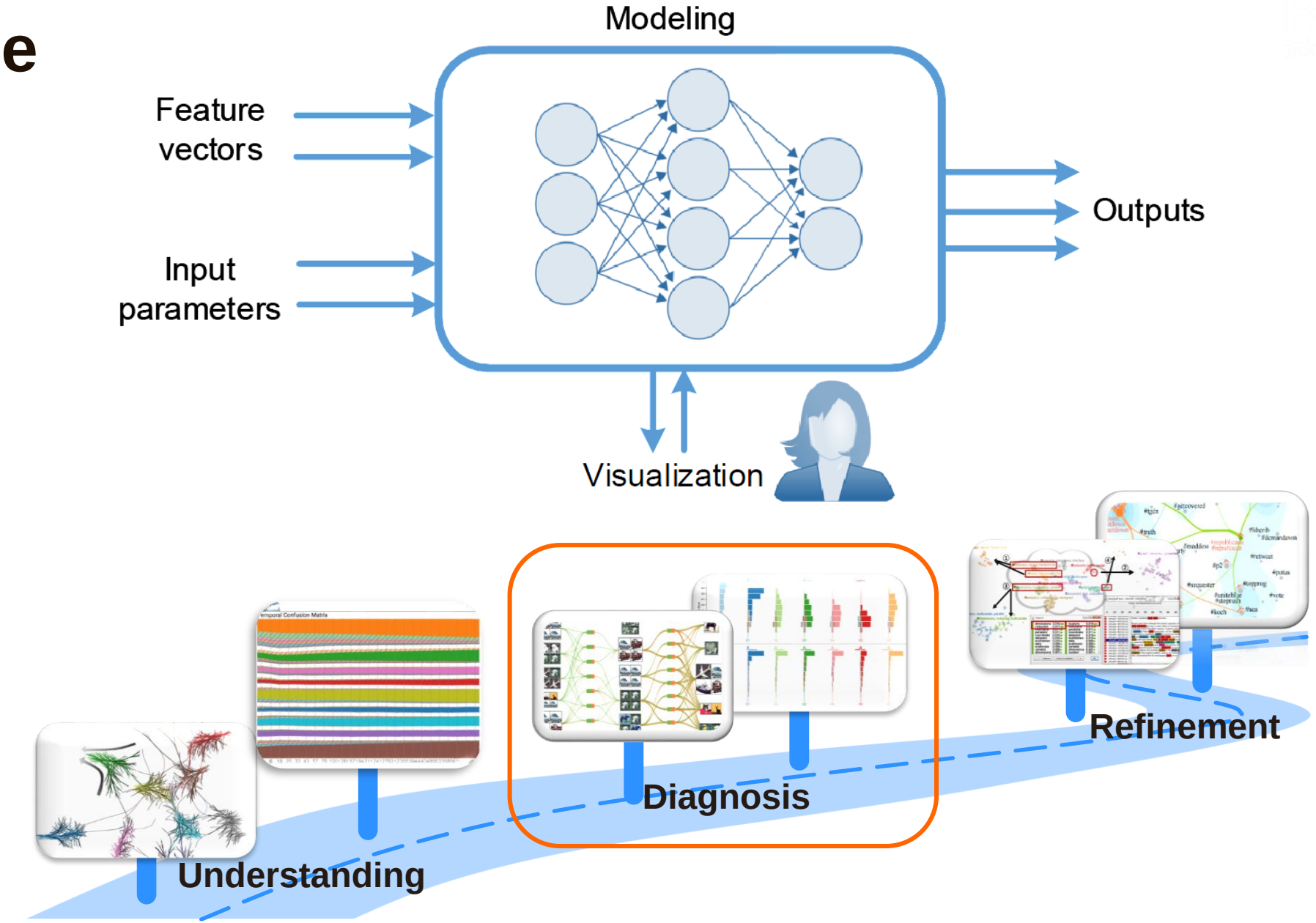
Single model to solv

Outline

Want huge model capacity
for large datasets



Outline


Modeling
Feature vectors
Input parameters
Outputs
Visualization
Understanding
Diagnosis
Refinement

清華大學
Tsinghua University

Analyzing the Training Processes of Deep Generative Models

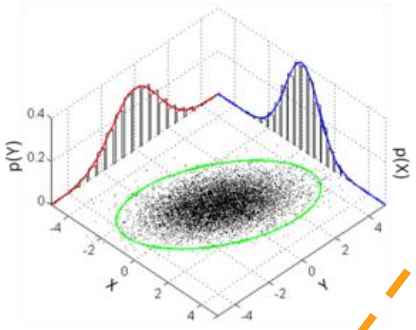
Mengchen Liu, Jiaxin Shi, Kelei Cao, Jun Zhu, Shixia Liu

Tsinghua University

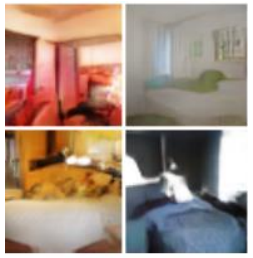
IEEE VAST 2017 (IEEE TVCG)

Deep Generative Models (DGMs)

DGM



Random variables



Z

X

Deep neural network

$p(z|x,k) \propto p(x|z,k)p(z|k)$

Z

X

Smiling face

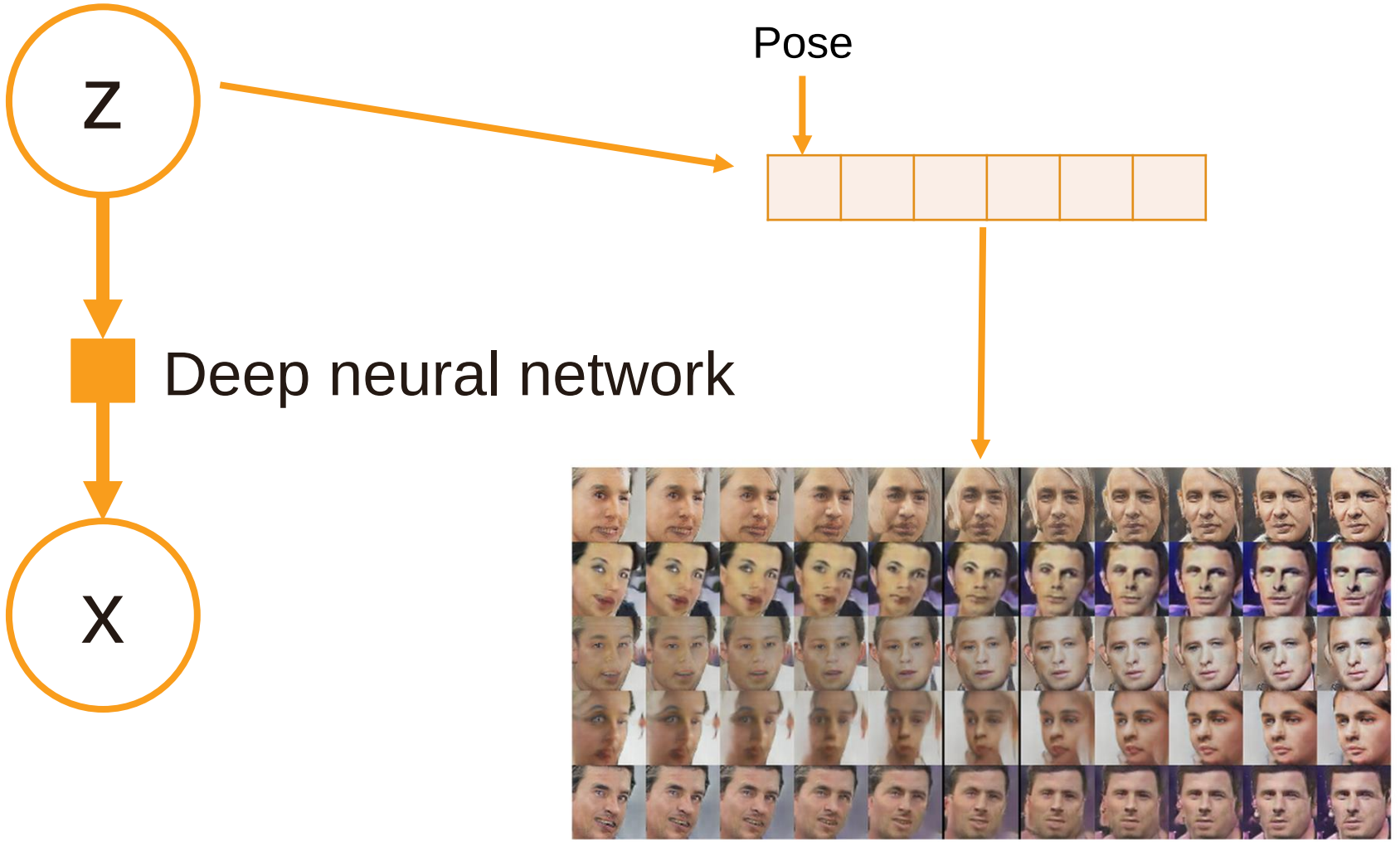
Unsupervised / semi-supervised

Knowledge ?

$z = f(x)$

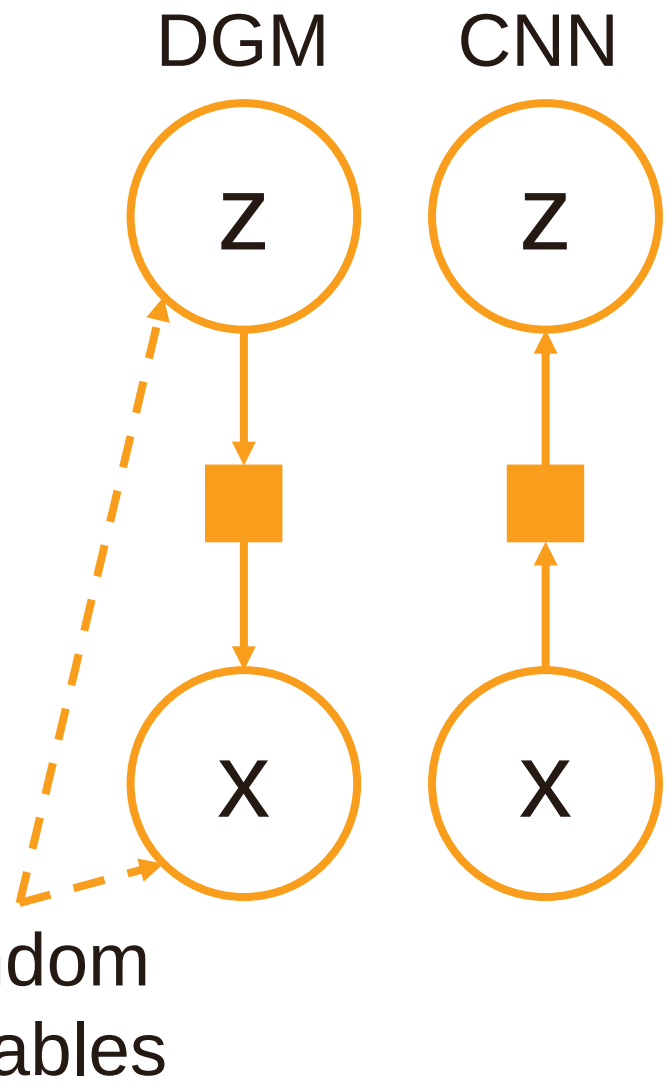
VAE
GAN

Deep Generative Models (DGMs)



Training a DGM is Hard

- DGM often involves both deterministic functions and random variables
  - CNN: deterministic functions (e.g., convolution)
- DGM involves a top-down generative process and a bottom-up Bayesian inference process
  - CNN: a bottom-up process: input at the bottom layer -> high-level features -> outputs



Challenges

- Handle a large amount of time series data
  - Millions of activations/gradients/weights in a DGM
- Identify the root cause of a failed training process
  - Errors may arise from multiple possible sources: abnormal training samples, inappropriate network structures, and lack of numerical stability in the library
  - Even when we can determine that the error is caused by the network structure, it is often difficult to locate the specific neurons

Our Solution

- A blue noise polyline sampling algorithm
  - Selects polyline samples the with blue-noise properties
  - Preserves outliers and reduce visual clutter

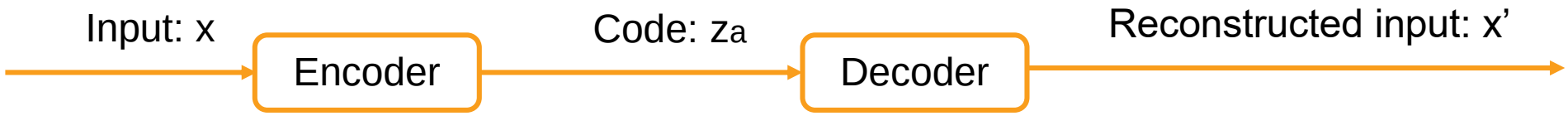
- A credit assignment algorithm
  - Discloses how other neurons contribute to the output of the neuron of interest

DGMTracker

- Better understand and diagnose the training process of a DGM

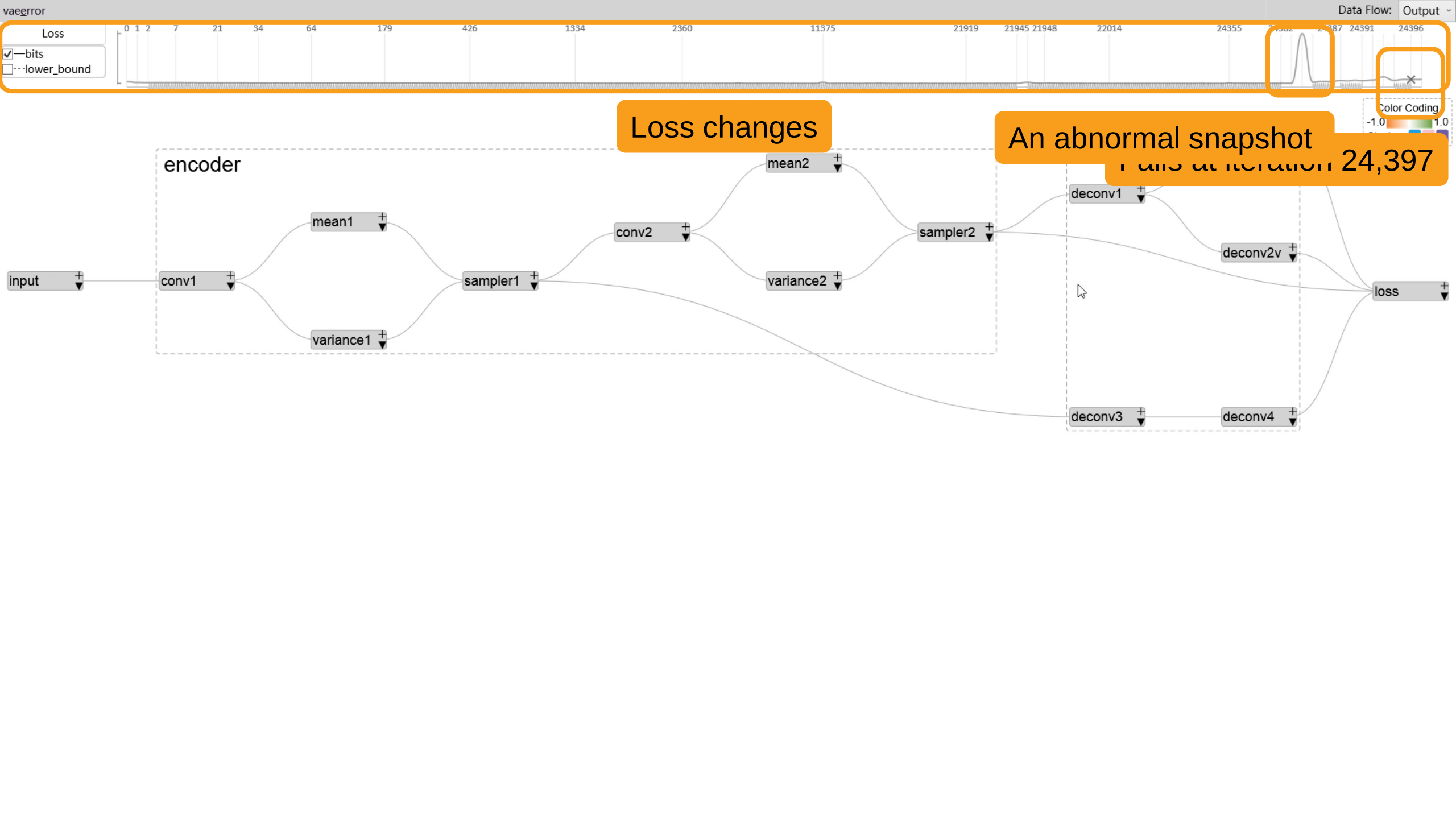
Case Study: Debugging a Failed Training Process of a Variational Autoencoder (VAE)

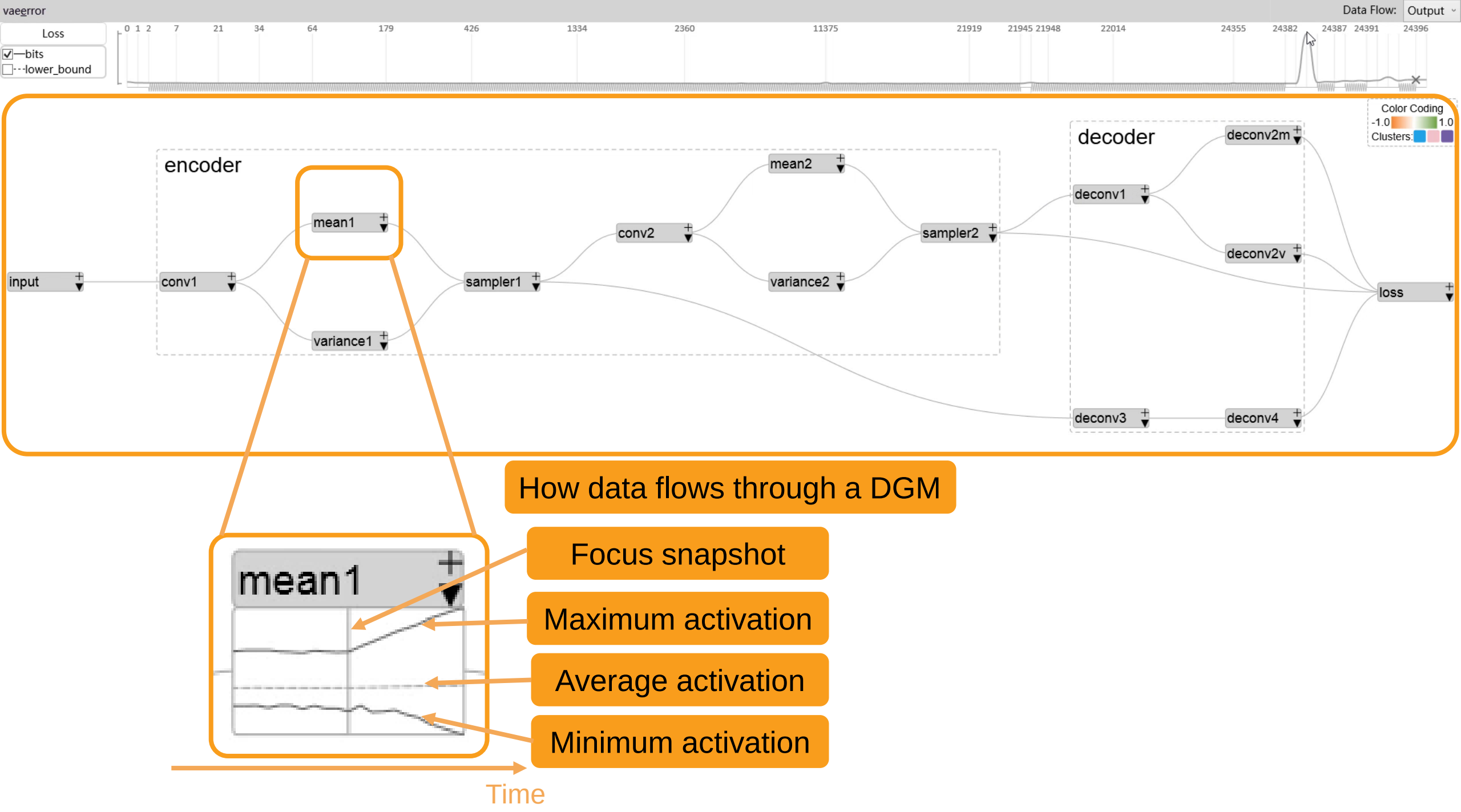
- Autoencoder
  - Reconstruct their input with minimum information loss

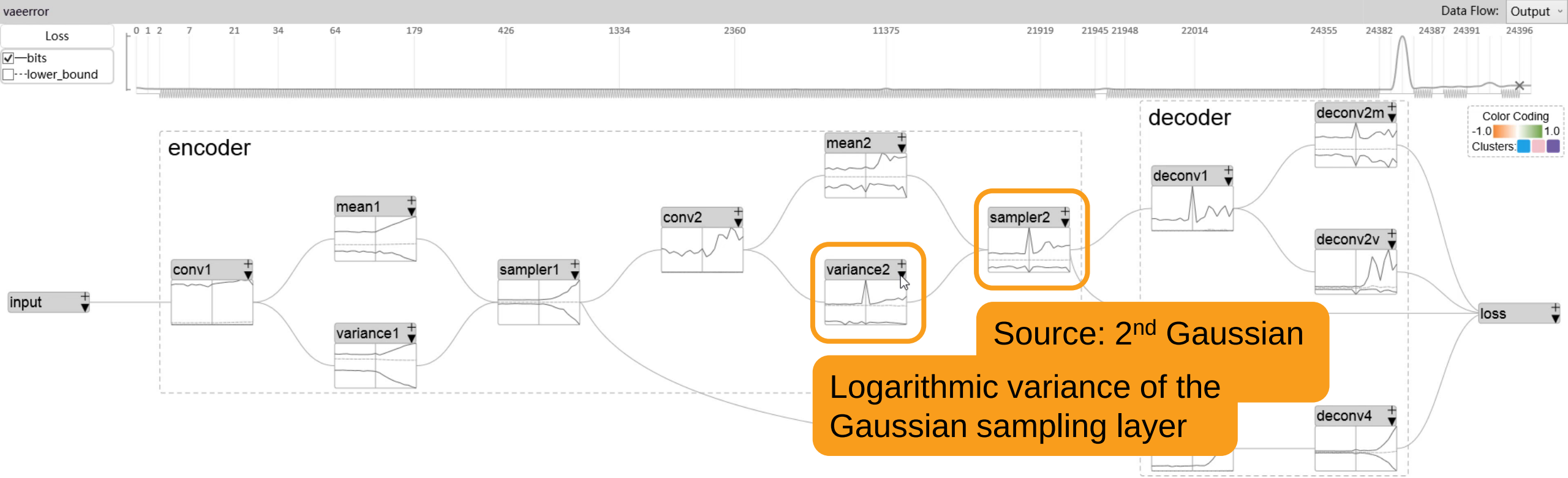


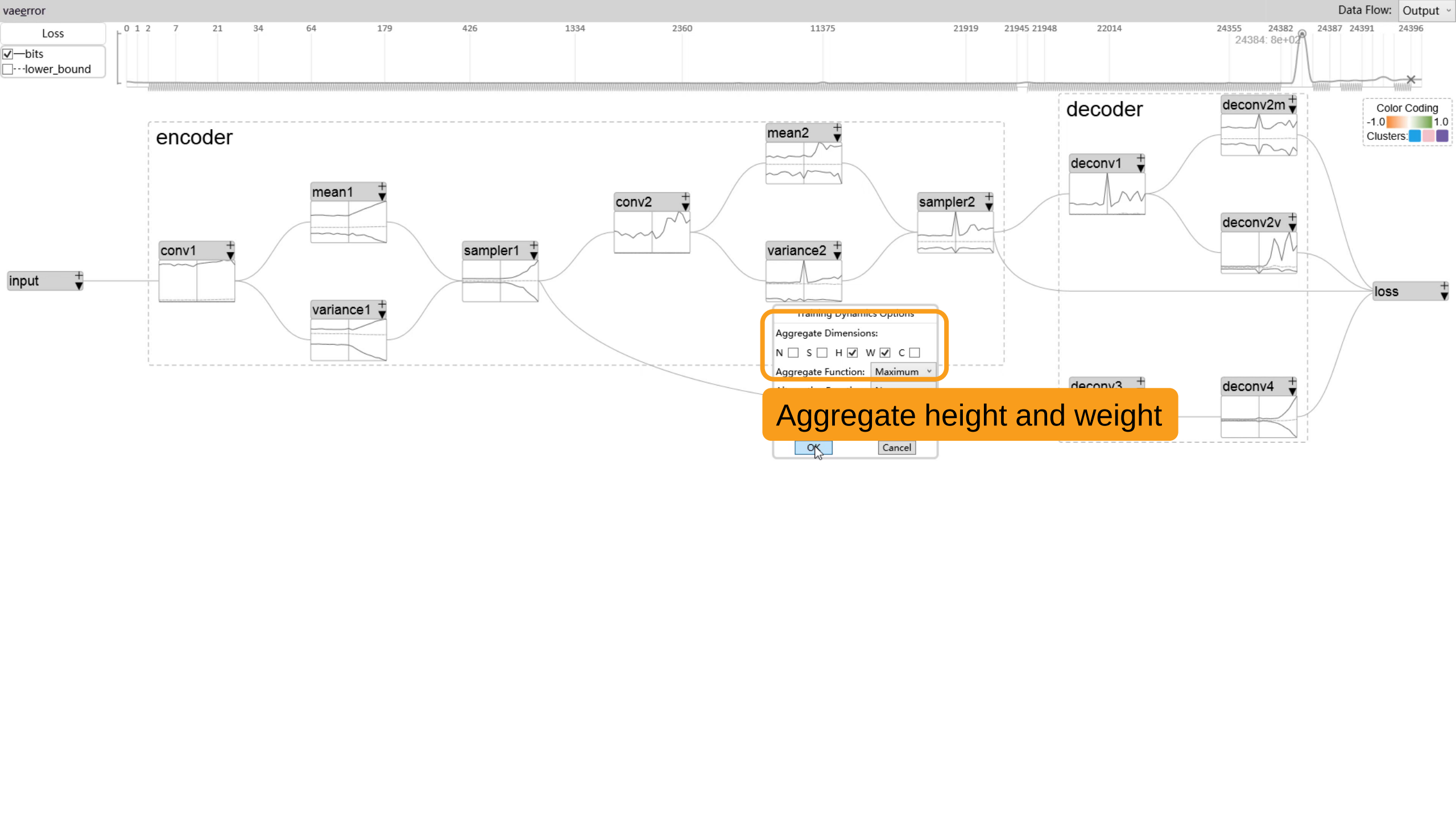
- Variational autoencoder
  - Probabilistic version of an autoencoder
  - z_v : a vector of random variables
  - z_a : a vector of real numbers

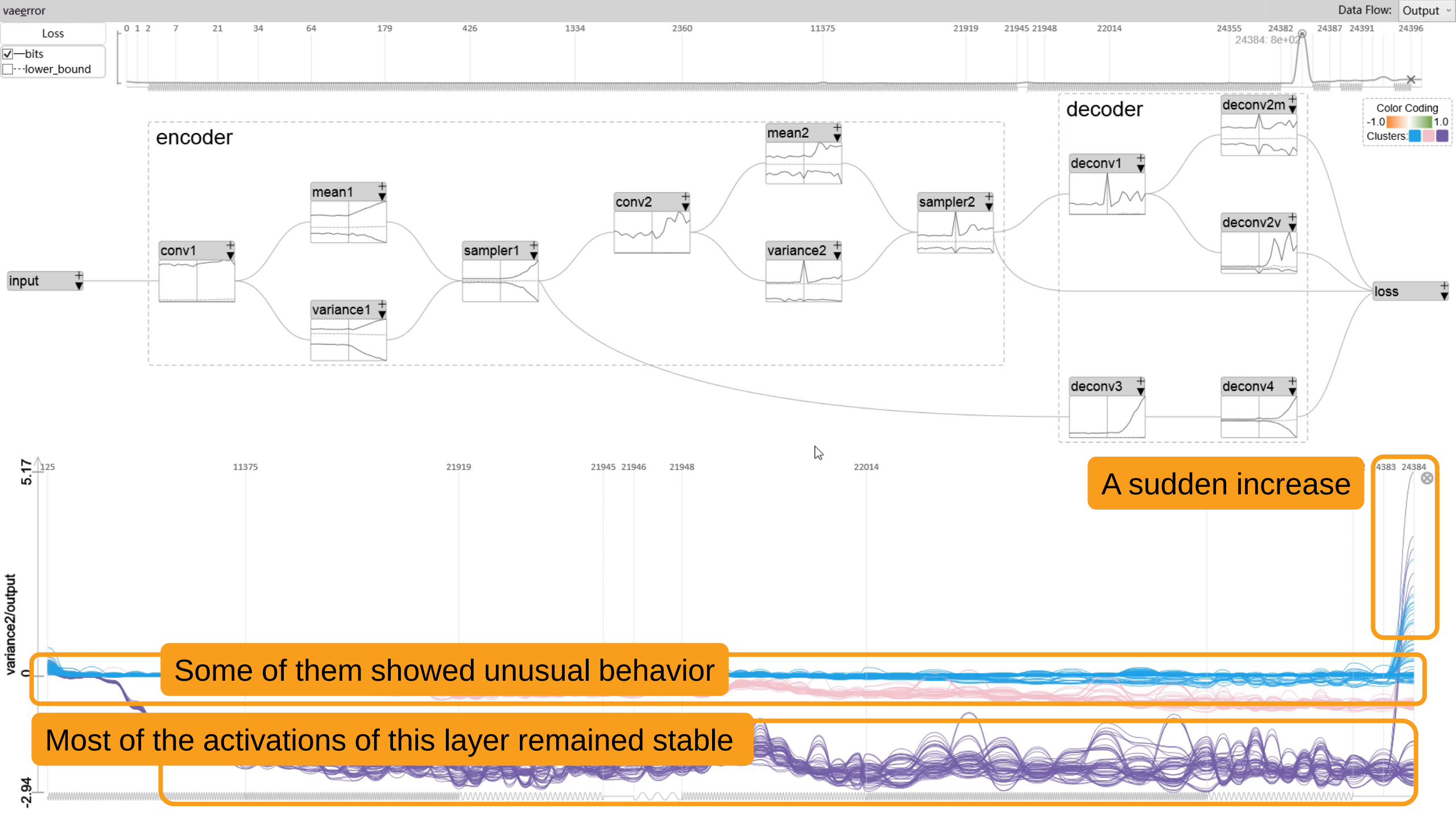
Dataset: CIFAR10 dataset
Loss = NaN (10k-30k iterations)
An example case: fails at 24,397

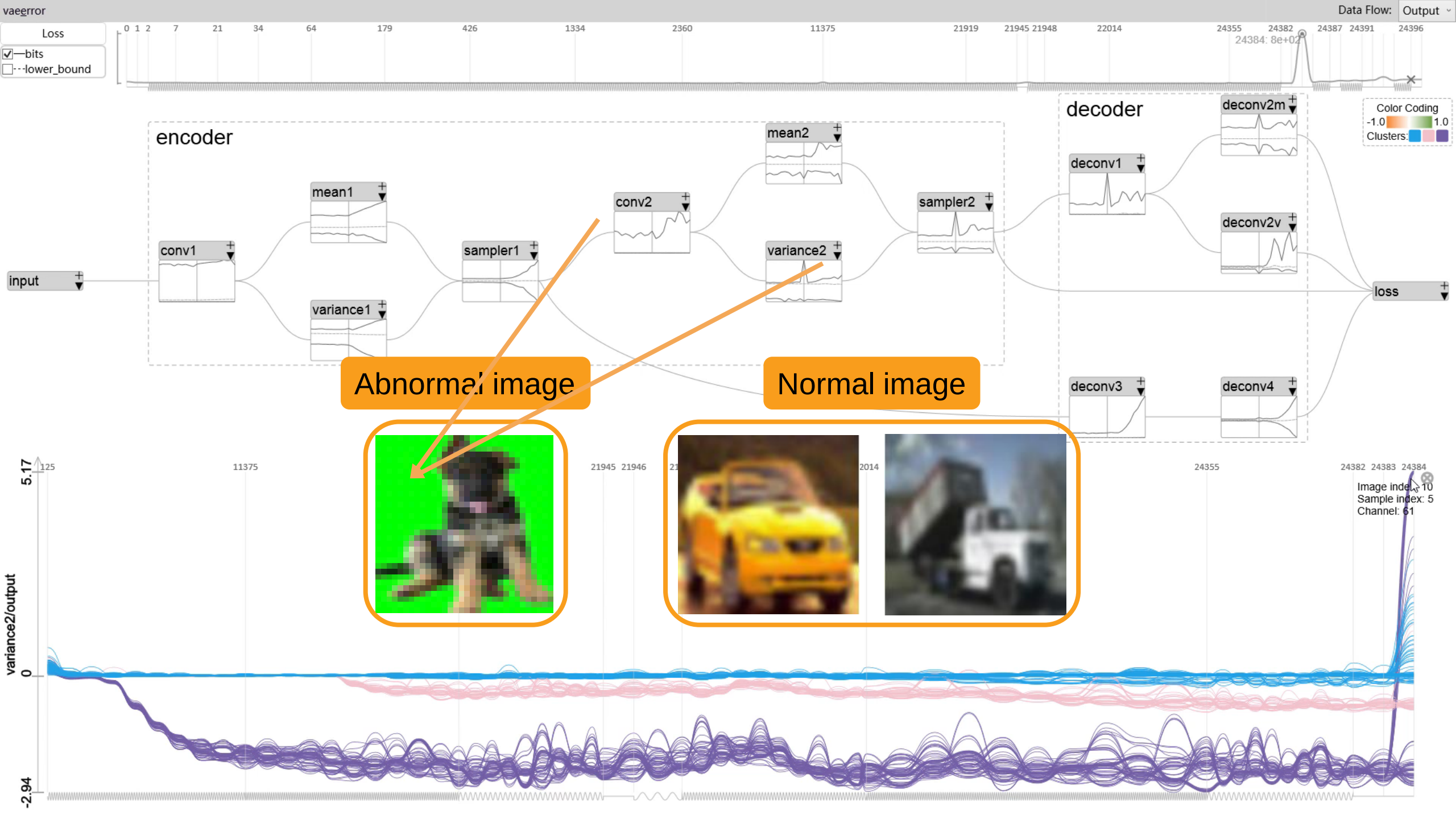

vaeerror
Data Flow:
Output
Loss
bits
lower_bound
0 1 2 7 21 34 64 179 426 1334 2360 11375 21919 21945 21948 22014 24355 24391 24396
Loss changes
An abnormal snapshot
Color Coding
-1.0
1.0
encoder
input
conv1
mean1
variance1
sampler1
conv2
mean2
variance2
sampler2
deconv1
deconv2v
deconv3
deconv4
loss


vaeerror
Data Flow:
Output
Loss
bits
lower_bound
0 1 2 7 21 34 64 179 426 1334 2360 11375 21919 21945 21948 22014 24355 24382 24387 24391 24396
Color Coding
-1.0
1.0
Clusters:
encoder
decoder
input
conv1
mean1
variance1
sampler1
conv2
mean2
variance2
sampler2
deconv1
deconv2m
deconv2v
deconv3
deconv4
loss
How data flows through a DGM
Focus snapshot
Maximum activation
Average activation
Minimum activation
Time


vaeerror
Data Flow: Output
Loss
bits
lower_bound
0 1 2 7 21 34 64 179 426 1334 2360 11375 21919 21945 21948 22014 24355 24382 24387 24391 24396
encoder
decoder
Color Coding
-1.0 1.0
Clusters:
input
conv1
mean1
variance1
sampler1
conv2
mean2
variance2
sampler2
deconv1
deconv2m
deconv2v
deconv4
loss
Source: 2nd Gaussian
Logarithmic variance of the Gaussian sampling layer

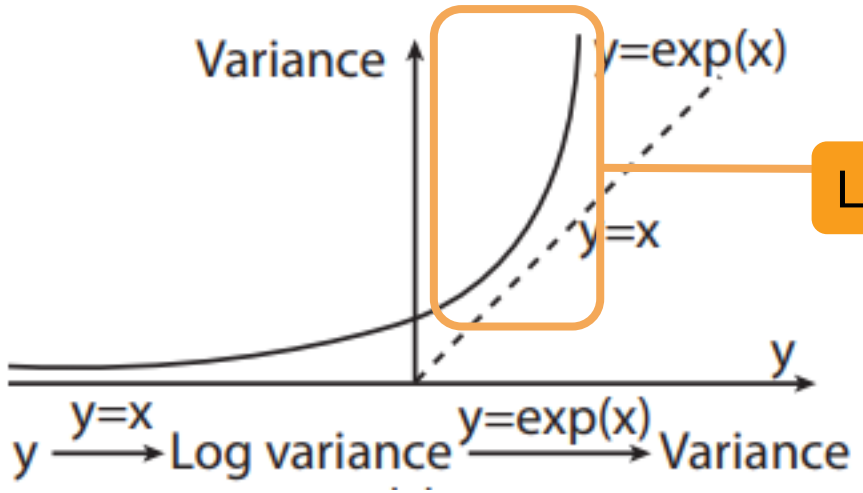

vaeerror
Data Flow:
Output
Loss
bits
lower_bound
0 1 2 7 21 34 64 179 426 1334 2360 11375 21919 21945 21948 22014 24355 24382 24384: 8e+02 24387 24391 24396
encoder
decoder
Color Coding
-1.0
1.0
Clusters:
input
conv1
mean1
variance1
sampler1
conv2
mean2
variance2
sampler2
deconv1
deconv2m
deconv2v
deconv3
deconv4
loss
Training Dynamics Options
Aggregate Dimensions:
N S H W C
Aggregate Function:
Maximum
OK
Cancel
Aggregate height and weight


vaeerror
Data Flow: Output
Loss
bits
lower_bound
0 1 2 7 21 34 64 179 426 1334 2360 11375 21919 21945 21948 22014 24355 24382 24387 24391 24396
24384: 8e+02
encoder
decoder
Color Coding
-1.0 1.0
Clusters:
input
conv1
mean1
variance1
sampler1
conv2
mean2
variance2
sampler2
deconv1
deconv2m
deconv2v
deconv3
deconv4
loss
5.17
0
-2.94
variance2/output
125 11375 21919 21945 21946 21948 22014 24383 24384
A sudden increase
Some of them showed unusual behavior
Most of the activations of this layer remained stable


vaeerror
Data Flow: Output
Loss
bits
lower_bound
encoder
decoder
input
conv1
mean1
variance1
sampler1
conv2
mean2
variance2
sampler2
deconv1
deconv2m
deconv2v
deconv3
deconv4
loss
Color Coding
-1.0
1.0
Clusters:
Abnormal image
Normal image
Image index: 10
Sample index: 5
Channel: 61
variance2/output
5.17
0
-2.94

Solution

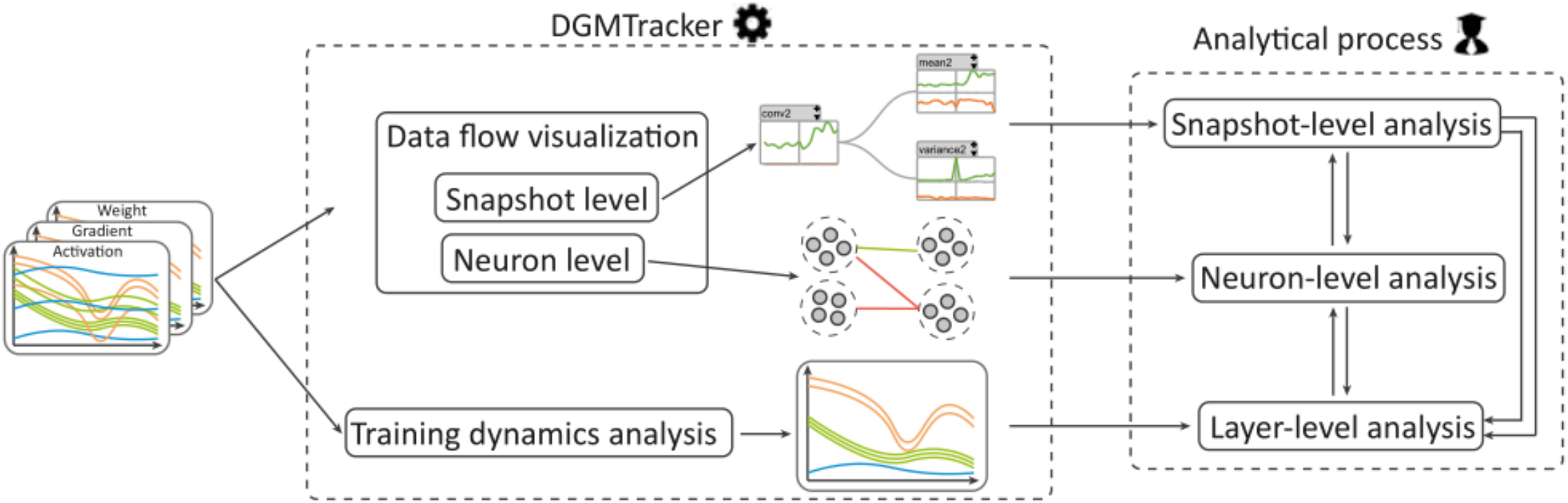
- Trial 1:
  - Replacing  with  , but the training failed again
  - By the same analysis, we find another “bad” image 
- Trial 2:



Much smoother

Large variance → large samples → large increase in loss

DGMTracker


DGMTracker
Analytical process
Weight
Gradient
Activation
Data flow visualization
Snapshot level
Neuron level
Training dynamics analysis
conv2
mean2
variance2
Snapshot-level analysis
Neuron-level analysis
Layer-level analysis

Research Opportunities

- Human-in-the-loop visual analytics for practitioners
  - Existing deep learning models are data-driven
  - Combine human expert knowledge and deep learning techniques through interactive visualization
- Progressive visual analytics of deep learning models
  - The training of many deep learning models is time-consuming
  - Progressive visual analytics techniques are needed

Liu S, Wang X, Liu M, Zhu J. Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics. 2017 Mar 1;1(1):48-56.

Jaegul Choo, Shixia Liu.Visual Analytics for Explainable Deep Learning. IEEE Computer Graphics and Applications, 2018.

Research Opportunities (Cont'd)

- Improving the robustness of deep learning models for secure artificial intelligence
  - Deep learning models are generally vulnerable to adversarial perturbations
  - Incorporate human knowledge to improve the robustness of deep learning models
- Reducing the size of the required training set
  - One-shot learning or zero-shot learning
- Visual analytics for advanced deep learning architectures
  - ResNet and DenseNet

References

- M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. IEEE TVCG, 23(1):91–100, 2017
- M Liu, J Shi, K Cao, J Zhu, S Liu. Analyzing the Training Processes of Deep Generative Models. IEEE TVCG, , 24(1), 77-87, 2018.
- S. Liu, J. Xiao, J. Liu, X. Wang, J. Wu, & J. Zhu. Visual diagnosis of tree boosting methods. IEEE TVCG, 24(1), 163-173, 2018.
- Jaegul Choo, Shixia Liu. Visual Analytics for Explainable Deep Learning. IEEE Computer Graphics and Applications, 2018.
- S. Liu, X. Wang, M. Liu, and J. Zhu. Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics, 1(1):48–56, 2017.

Acknowledgements

- Professors
  - Prof. Jun Zhu (Tsinghua), Prof. Hang Su (Tsinghua)
  - Prof. Jaegul Choo (Korea University)
- Students
  - Mengchen Liu, Kelei Cao, Changjian Chen, Fangxin Ouyang, Jiaxin Shi, Zhen Li, Chongxuan Li

Thanks a lot for your attention!