

Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

TVCG 2018 Survey



Fred Hohman
[@fredhohman](#)



Minsuk Kahng



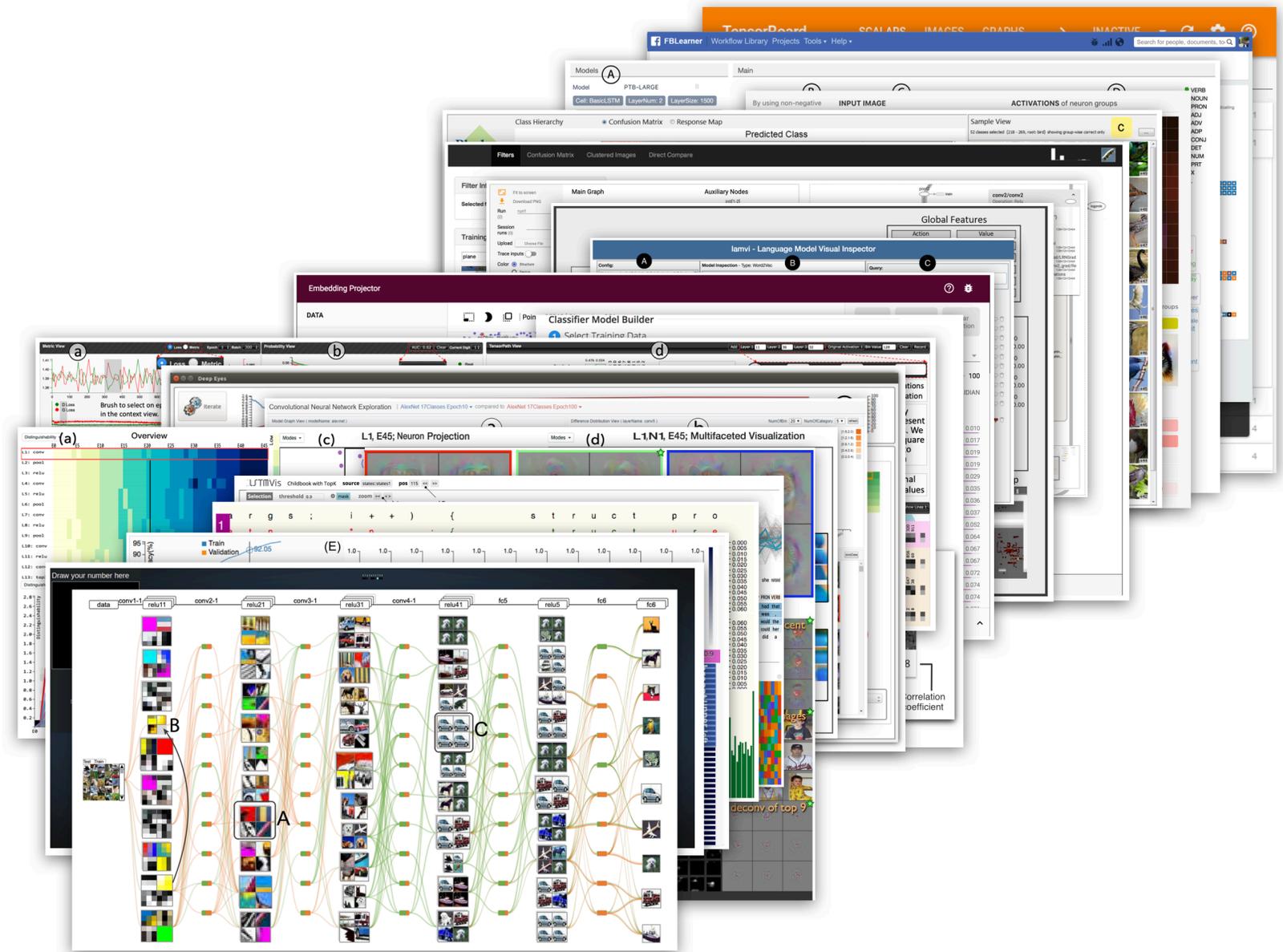
Robert Pienta



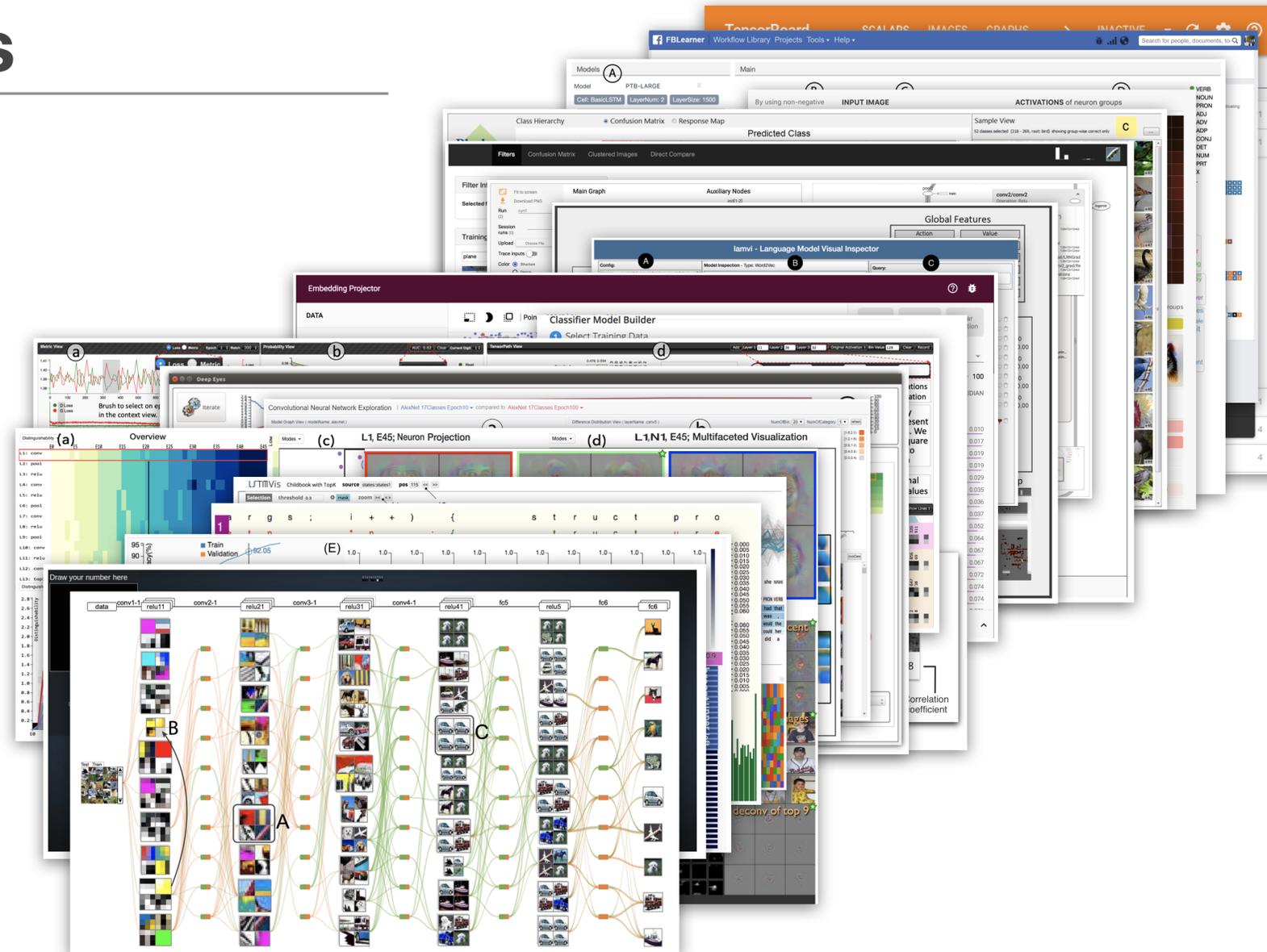
Polo Chau



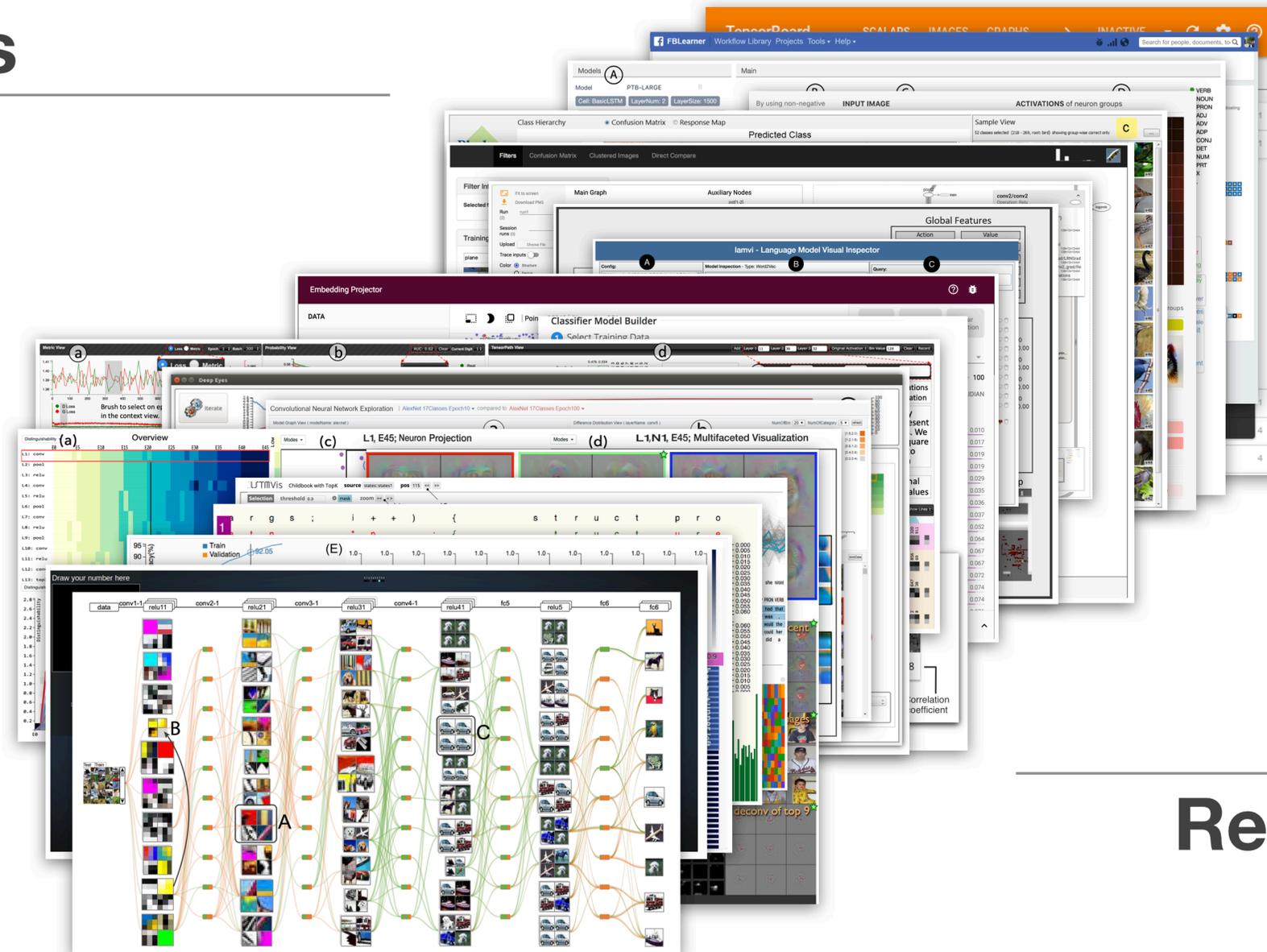
Visual Analytics in Deep Learning



Research Trends



Research Trends



Research Directions

WHY

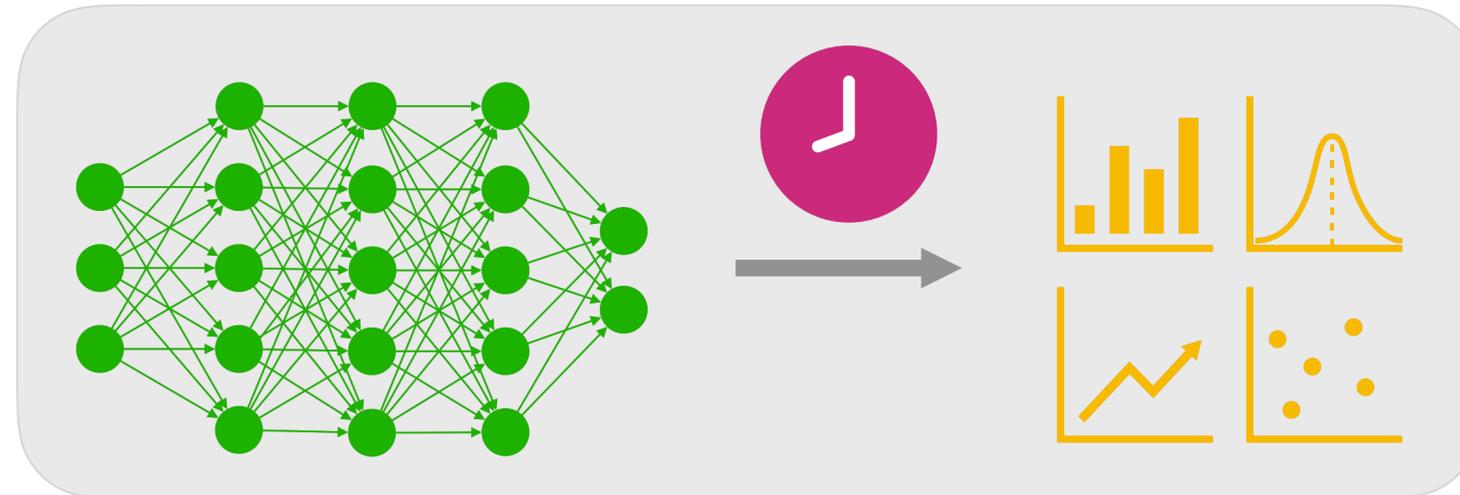
Why would one want to use visualization in deep learning?

WHAT

What data, features, and relationships in deep learning can be visualized?

WHEN

When in the deep learning process is visualization used?



WHO

Who would use and benefit from visualizing deep learning?

HOW

How can we visualize deep learning data, features, and relationships?

WHERE

Where has deep learning visualization been used?

Visual Analytics in Deep Learning

Interrogative Survey Overview

WHY

Why would one want to use visualization in deep learning?

- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts

WHAT

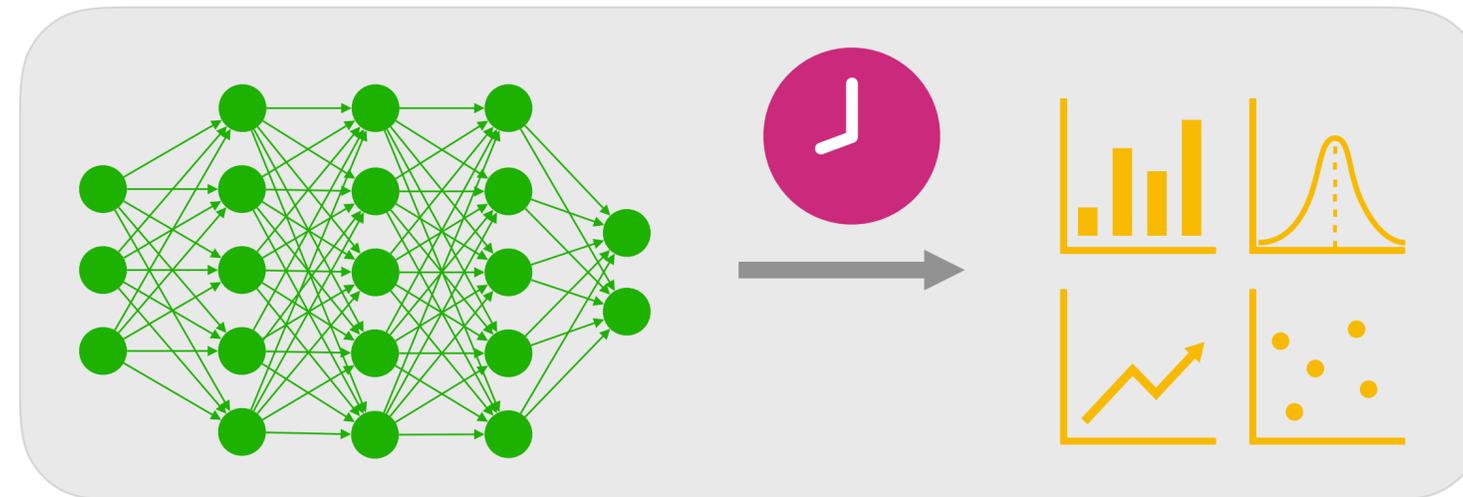
What data, features, and relationships in deep learning can be visualized?

- Computational Graph & Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information

WHEN

When in the deep learning process is visualization used?

- During Training
- After Training



WHO

Who would use and benefit from visualizing deep learning?

- Model Developers & Builders
- Model Users
- Non-experts

HOW

How can we visualize deep learning data, features, and relationships?

- Node-link Diagrams for Network Architecture
- Dimensionality Reduction & Scatter Plots
- Line Charts for Temporal Metrics
- Instance-based Analysis & Exploration
- Interactive Experimentation
- Algorithms for Attribution & Feature Visualization

WHERE

Where has deep learning visualization been used?

- Application Domains & Models
- A Vibrant Research Community

WHY

- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts

WHAT

- Computational Graph & Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information

WHEN

- During Training
- After Training

WHO

- Model Developers & Builders
- Model Users
- Non-experts

HOW

- Node-link Diagrams for Network Architecture
- Dimensionality Reduction & Scatter Plots
- Line Charts for Temporal Metrics
- Instance-based Analysis & Exploration
- Interactive Experimentation
- Algorithms for Attribution & Feature Visualization

WHERE

- Application Domains & Models
- A Vibrant Research Community

Author	Year	WHY				WHO			WHAT					HOW				WHEN		WHERE	
		Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational Units	Neurons in High-dimensional Space	Aggregated Information	Node-link Diagrams for Network Architecture	Dimensionality Reduction & Scatter Plots	Line Charts for Temporal Metrics	Instance-based Analysis & Exploration	Interactive Experimentation	Algorithms for Attribution & Feature Visualization	During Training	After Training
Abadi, et al.	2016	■	■	■		■	■					■			■				■	■	arXiv
Bau, et al.	2017	■		■		■				■						■		■		■	CVPR
Bilal, et al.	2017	■	■			■				■						■		■		■	TVCG
Bojarski, et al.	2016	■	■			■				■						■		■		■	arXiv
Bruckner	2014	■	■			■			■	■			■			■		■		■	MS Thesis
Carter, et al.	2016	■			■	■	■	■		■	■	■	■			■	■			■	Distill
Cashman, et al.	2017	■	■			■	■			■	■					■				■	VADL
Chae, et al.	2017	■	■			■				■					■	■				■	VADL
Chung, et al.	2016	■	■			■			■	■	■	■	■	■	■	■				■	FILM
Goyal, et al.	2016	■						■		■						■	■	■		■	arXiv
Harley	2015	■			■			■	■	■			■			■	■			■	ISVC
Hohman, et al.	2017	■		■	■			■								■	■	■		■	CHI
Kahng, et al.	2018	■	■			■	■		■	■	■	■	■	■	■	■				■	TVCG
Karpathy, et al.	2015	■				■	■			■	■	■		■		■				■	arXiv
Li, et al.	2015	■				■	■			■	■	■		■		■				■	arXiv
Liu, et al.	2017	■	■			■			■	■	■	■	■			■				■	TVCG
Liu, et al.	2018	■	■			■			■	■	■	■	■	■	■					■	TVCG
Ming, et al.	2017	■		■		■				■					■					■	VAST
Norton & Qi	2017	■	■		■	■	■	■								■	■			■	VizSec
Olah	2014	■			■			■		■			■		■	■				■	Web
Olah, et al.	2018	■			■	■	■	■	■	■	■	■				■	■	■		■	Distill
Pezzotti, et al.	2017	■	■			■				■	■	■	■	■	■					■	TVCG
Rauber, et al.	2017	■	■	■		■				■	■	■	■	■	■					■	TVCG
Robinson, et al.	2017	■				■	■			■	■	■			■					■	GeoHum.
Rong, et al.	2016	■	■			■	■			■		■			■					■	ICML VIS
Smilkov, et al.	2016	■				■				■	■	■	■	■	■					■	NIPS Workshop
Smilkov, et al.	2017	■	■		■			■	■	■		■	■	■	■					■	ICML VIS
Strobelt, et al.	2017	■	■			■	■			■	■	■	■	■	■					■	TVCG
Tzeng & Ma	2005	■				■			■	■		■	■	■	■					■	VIS
Wang, et al.	2018	■	■	■		■			■	■	■	■	■	■	■					■	TVCG
Webster, et al.	2017				■			■								■	■			■	Web
Wongsuphasawat, et al.	2018		■			■			■			■			■						TVCG
Yosinski, et al.	2015	■			■	■	■	■	■	■	■	■			■	■	■			■	ICML DL
Zahavy, et al.	2016	■	■			■				■	■	■	■	■	■					■	ICML
Zeiler, et al.	2014	■	■			■			■	■								■		■	ECCV
Zeng, et al.	2017	■		■		■			■	■					■					■	VADL
Zhong, et al.	2017	■	■			■				■	■	■	■	■	■					■	ICML VIS
Zhu, et al.	2016	■				■	■	■				■				■	■	■		■	ECCV

Author	Year	WHY				WHO			WHAT					HOW				WHEN		WHERE	
		Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational Units	Neurons in High-dimensional Space	Aggregated Information	Node-link Diagrams for Network Architecture	Dimensionality Reduction & Scatter Plots	Line Charts for Temporal Metrics	Instance-based Analysis & Exploration	Interactive Experimentation	Algorithms for Attribution & Feature Visualization	During Training	After Training
Abadi, et al.	2016	■	■	■		■	■					■			■				■	■	arXiv
Bau, et al.	2017	■		■		■				■						■		■		■	CVPR
Bilal, et al.	2017	■	■			■				■						■		■		■	TVCG
Bojarski, et al.	2016	■	■			■				■						■		■		■	arXiv
Bruckner	2014	■	■			■			■	■			■			■		■		■	MS Thesis
Carter, et al.	2016	■			■	■	■	■		■	■	■	■			■	■			■	Distill
Cashman, et al.	2017	■	■			■	■			■	■					■				■	VADL
Chae, et al.	2017	■	■			■				■					■	■				■	VADL
Chung, et al.	2016	■	■			■			■	■	■	■	■	■	■	■				■	FILM
Goyal, et al.	2016	■						■		■						■	■	■		■	arXiv
Harley	2015	■			■			■	■	■			■			■	■			■	ISVC
Hohman, et al.	2017	■		■	■			■								■	■	■		■	CHI
Kahng, et al.	2018	■	■			■	■		■	■	■	■	■	■	■	■				■	TVCG
Karpathy, et al.	2015	■				■	■			■	■	■			■	■				■	arXiv
Li, et al.	2015	■				■	■			■	■	■			■	■				■	arXiv
Liu, et al.	2017	■	■			■			■	■	■	■	■			■				■	TVCG
Liu, et al.	2018	■	■			■			■	■	■	■	■		■	■				■	TVCG
Ming, et al.	2017	■		■		■				■					■					■	VAST
Norton & Qi	2017	■	■		■	■	■	■								■	■			■	VizSec
Olah	2014	■			■			■		■			■		■	■				■	Web
Olah, et al.	2018	■			■	■	■	■	■	■	■	■				■	■	■		■	Distill
Pezzotti, et al.	2017	■	■			■				■	■	■	■	■	■					■	TVCG
Rauber, et al.	2017	■	■	■		■				■	■	■	■	■	■					■	TVCG
Robinson, et al.	2017	■				■	■			■	■	■			■					■	GeoHum.
Rong, et al.	2016	■	■			■	■			■		■			■					■	ICML VIS
Smilkov, et al.	2016	■				■				■	■	■	■	■	■					■	NIPS Workshop
Smilkov, et al.	2017	■	■		■			■	■	■	■	■	■	■	■	■				■	ICML VIS
Strobelt, et al.	2017	■	■			■	■			■	■	■	■	■	■					■	TVCG
Tzeng & Ma	2005	■				■			■	■			■		■					■	VIS
Wang, et al.	2018	■	■	■		■			■	■	■	■	■	■	■					■	TVCG
Webster, et al.	2017				■			■								■	■			■	Web
Wongsuphasawat, et al.	2018		■			■			■			■			■						TVCG
Yosinski, et al.	2015	■			■	■	■	■	■	■	■	■	■			■	■	■		■	ICML DL
Zahavy, et al.	2016	■	■			■				■	■	■	■	■	■					■	ICML
Zeiler, et al.	2014	■	■			■			■	■									■	■	ECCV
Zeng, et al.	2017	■		■		■			■	■					■					■	VADL
Zhong, et al.	2017	■	■			■				■	■	■	■	■	■					■	ICML VIS
Zhu, et al.	2016	■				■	■	■				■				■	■	■		■	ECCV

Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau



FBLeaver Workflow Library Projects Tools Help

ActiVis: Visualization of Deep Neural Networks #15782570

Search for people, documents, to: Q

COMPUTATION GRAPH

Operator node (grey square), Blob node (white circle), Blob node w/ activation (orange circle)

Graph nodes: Concat, Dropout, dropout_0, FC, fc_0, Relu, FC, Softmax

INSTANCE SELECTION

Left column shows correctly classified instances. Right column shows misclassified instances, with border colors indicating predicted classes.

DESC, ENTY, ABBR, HUM, NUM, LOC

NEURON ACTIVATION

Each row represents a group of instances. Each column is a neuron. Columns sorted by activation strength for Neuron idx.

By class: DESC, ENTY, ABBR, HUM, NUM, LOC

By user-defined filters: Contain 'Where', Contain 'located', Contain 'How many', Contain 'How'

By instance ID: #94, #30, #108

PROJECTED

Scatter plot of instance activations

Activations for the instance #108: What is the highest dam in the U.S. ?

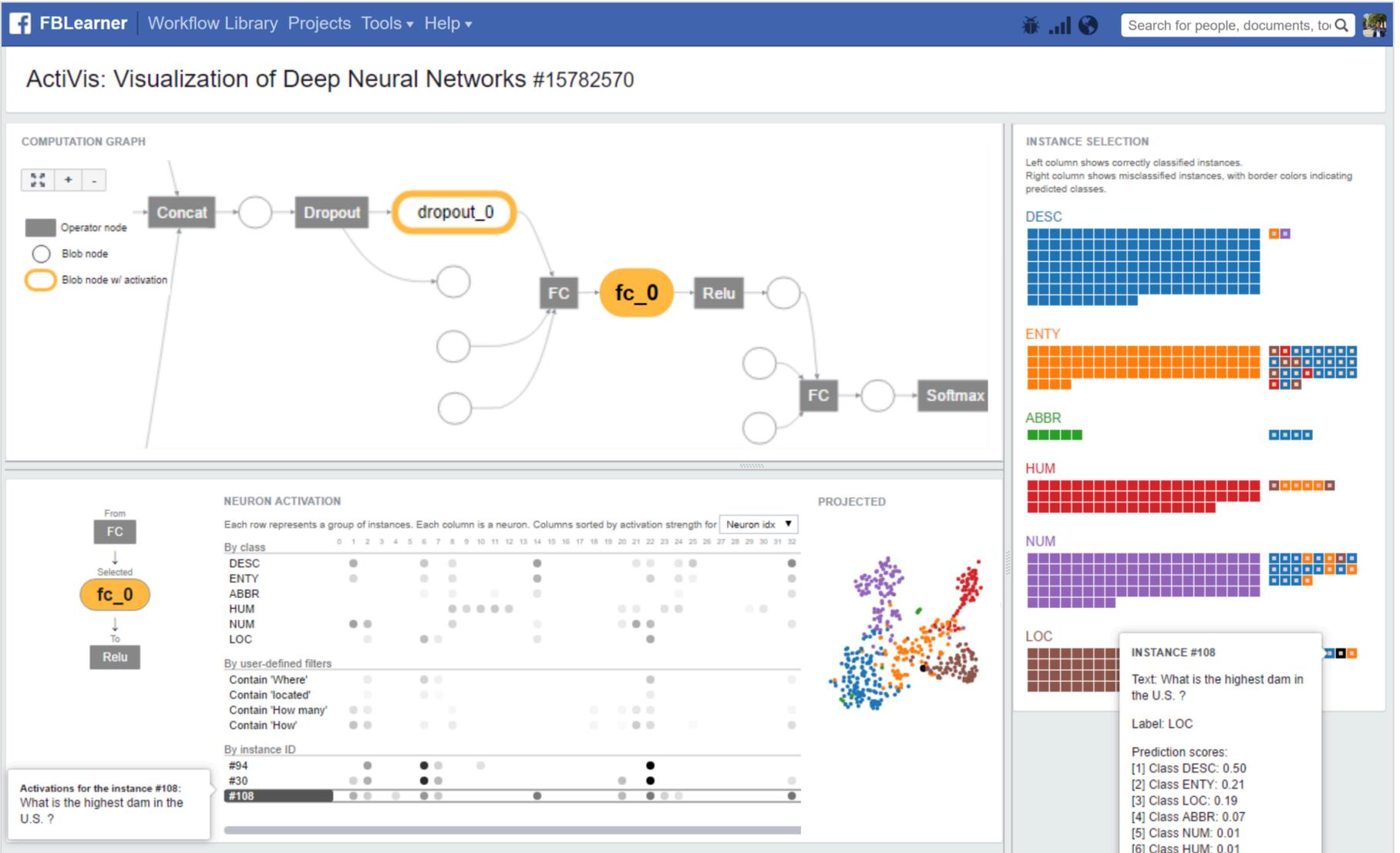
INSTANCE #108
Text: What is the highest dam in the U.S. ?
Label: LOC
Prediction scores:
[1] Class DESC: 0.50
[2] Class ENTY: 0.21
[3] Class LOC: 0.19
[4] Class ABBR: 0.07
[5] Class NUM: 0.01
[6] Class HUM: 0.01

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Example

Kahng, et al. 2018



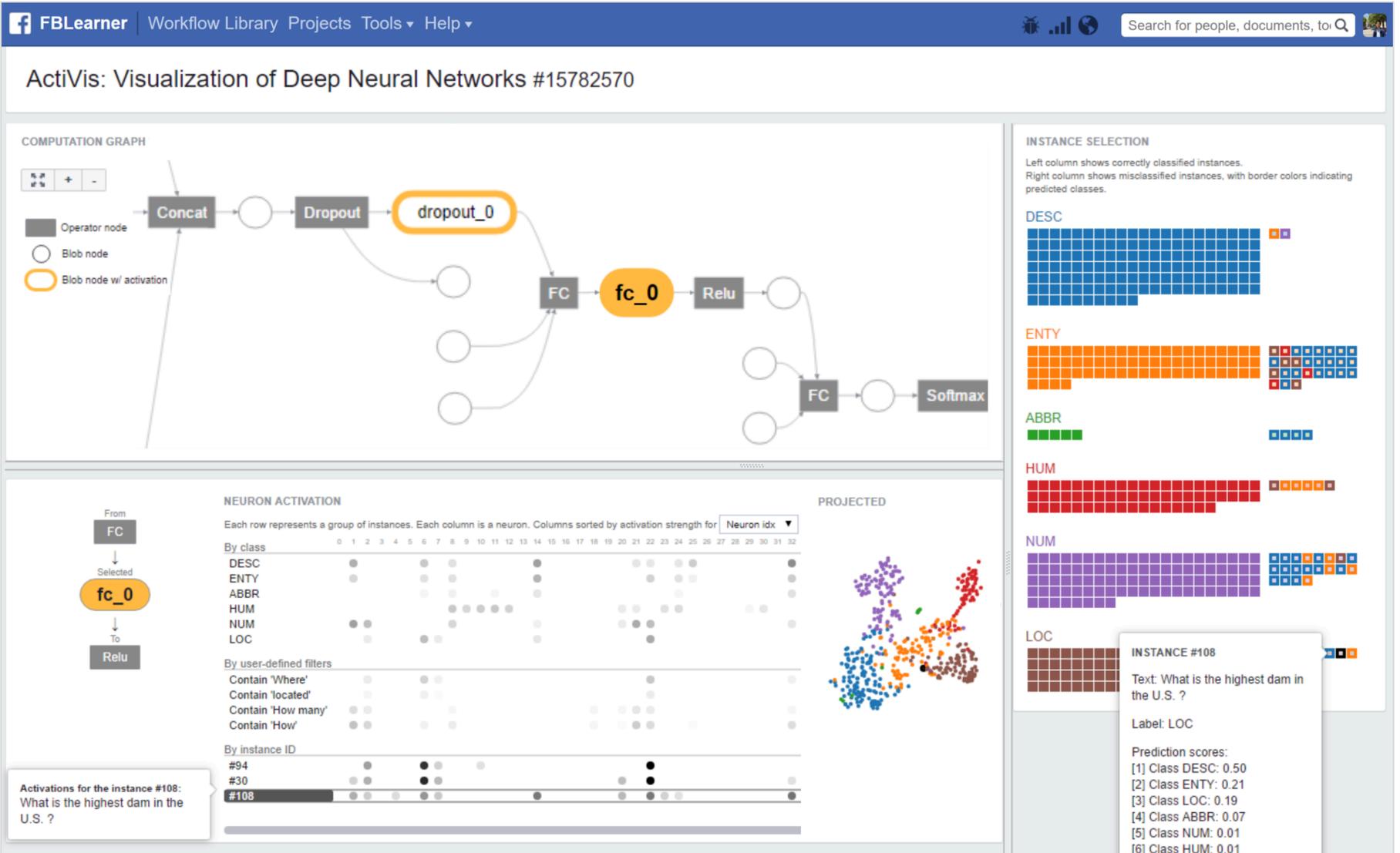
- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts
- Model Developers
- Model Users
- Non-experts
- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information
- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Example

Kahng, et al. 2018



- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts
- Model Developers
- Model Users
- Non-experts
- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information
- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Example

Kahng, et al. 2018

WHY

- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts
- Model Developers
- Model Users
- Non-experts
- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information
- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots

The screenshot displays the ActiVis interface for a deep neural network model. The top navigation bar includes 'FBLearner', 'Workflow Library', 'Projects', 'Tools', and 'Help'. The main content area is titled 'ActiVis: Visualization of Deep Neural Networks #15782570'.

COMPUTATION GRAPH: A node-link diagram showing the flow of data through the network. Nodes include 'Concat', 'Dropout', 'dropout_0', 'FC', 'fc_0', 'Relu', 'FC', and 'Softmax'. A legend identifies 'Operator node' (grey rectangle), 'Blob node' (white circle), and 'Blob node w/ activation' (orange circle). The 'fc_0' node is highlighted in orange.

NEURON ACTIVATION: A matrix where rows represent groups of instances and columns represent neurons (indexed 0-32). The matrix is sorted by activation strength for a selected neuron. A 'PROJECTED' scatter plot on the right shows the high-dimensional space of neurons. A legend on the left shows filters: 'By class' (DESC, ENTY, ABBR, HUM, NUM, LOC), 'By user-defined filters' (Contain 'Where', 'located', 'How many', 'How'), and 'By instance ID' (#94, #30, #108).

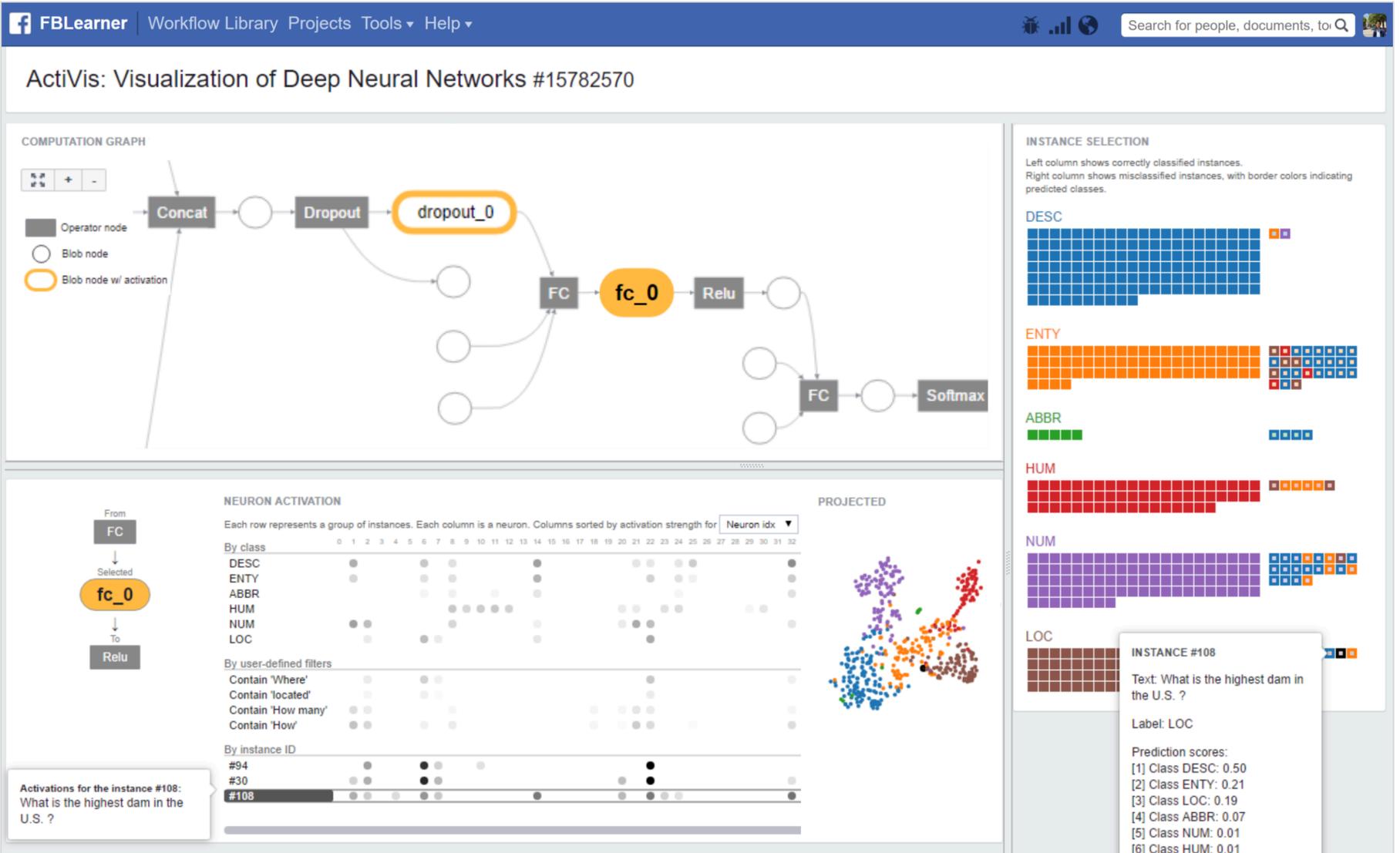
INSTANCE SELECTION: A panel showing classification results for various classes: DESC, ENTY, ABBR, HUM, NUM, and LOC. Each class has a grid of colored squares representing instance classifications. A tooltip for 'INSTANCE #108' shows the text 'What is the highest dam in the U.S.?' and the predicted label 'LOC'. Prediction scores are listed: [1] Class DESC: 0.50, [2] Class ENTY: 0.21, [3] Class LOC: 0.19, [4] Class ABBR: 0.07, [5] Class NUM: 0.01, [6] Class HUM: 0.01.

ACTIVATIONS FOR THE INSTANCE #108: A small box at the bottom left shows the path from the 'FC' node to the selected 'fc_0' node and then to the 'Relu' node.

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Example

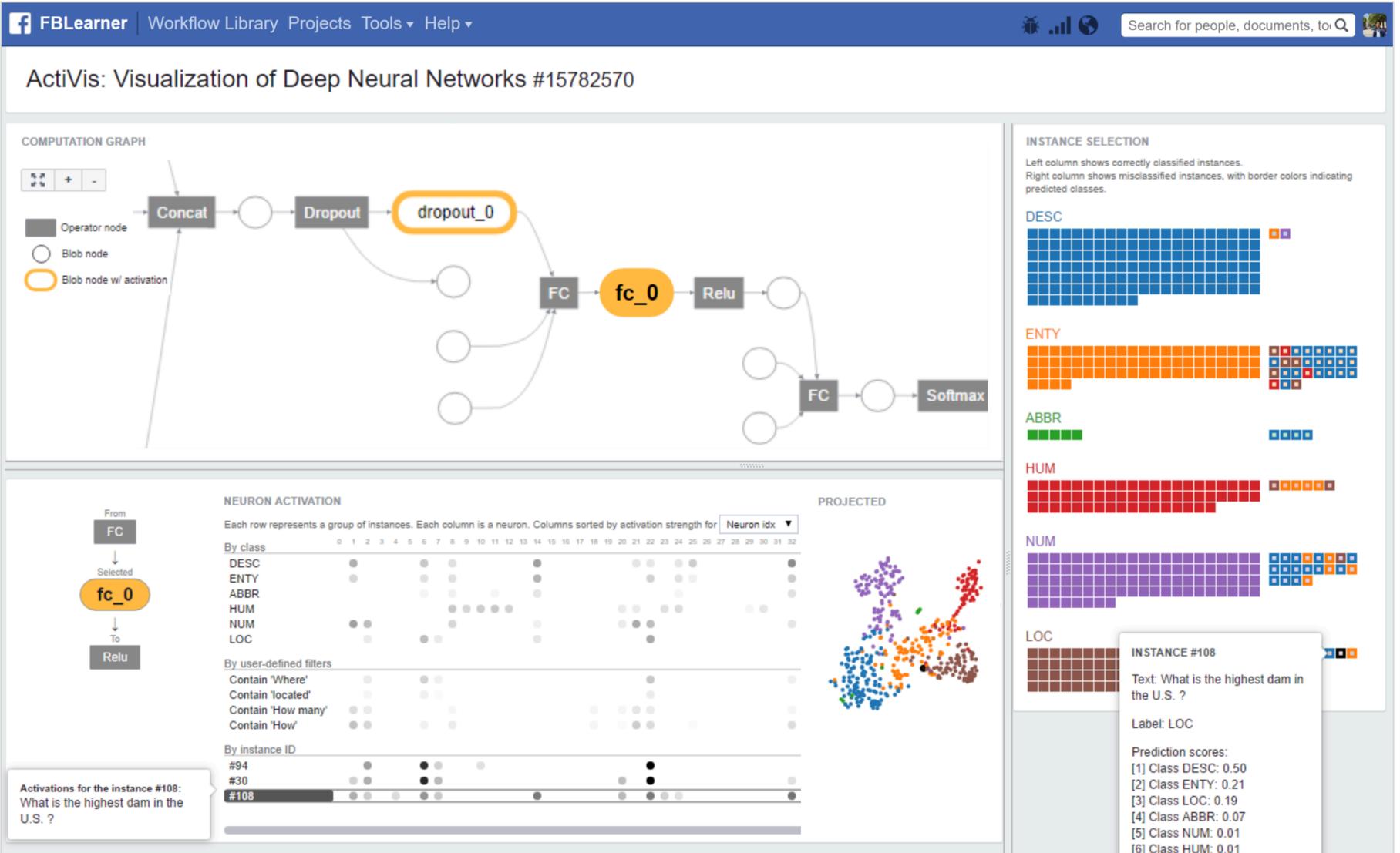


- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- WHO**
- Model Developers
- Model Users
- Non-experts
- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information
- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots
- Line Charts for Temporal Metrics
- Instance-based Analysis & Exploration
- Attribution & Feature Visualization
- Interactive Experimentation

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Example



Comparing & Selecting Models
Teaching Deep Learning Concepts

- Model Developers
- Model Users

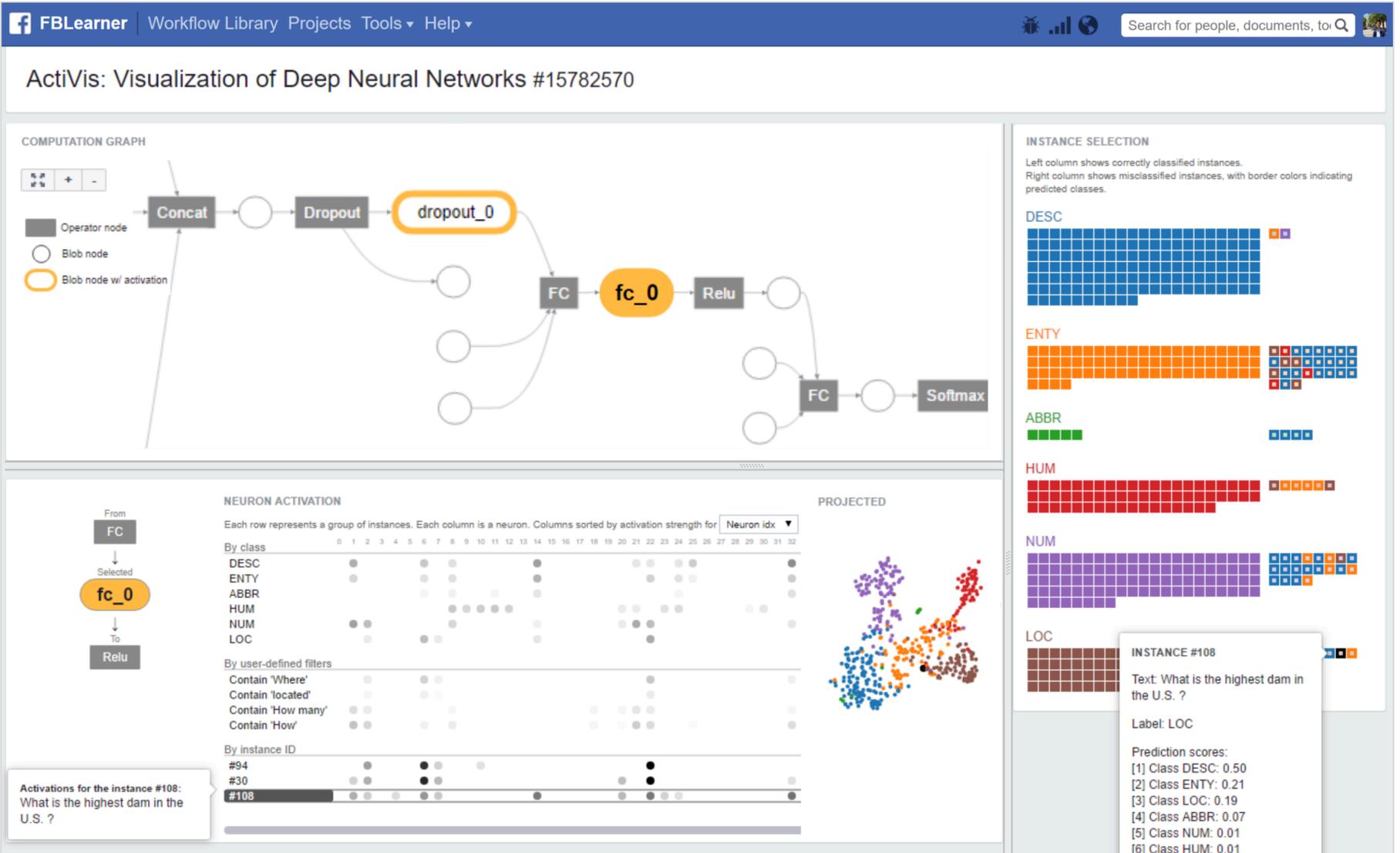
WHAT

- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High Dimensions
- Aggregated Information
- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots
- Line Charts for Temporal Metrics
- Instance-based Analysis & Exploration
- Attribution & Feature Visualization
- Interactive Experimentation
- During Training
- After Training

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Example



- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High Dimensions
- HOW**
- Aggregated Information

- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots
- Line Charts for Temporal Metrics
- Instance-based Analysis & Exploration
- Attribution & Feature Visualization
- Interactive Experimentation

- During Training
- After Training
- Publication Venue

TVCG

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Example

- Dimensionality Reduction & Scatter Plots
- Line Charts for Temporal Metrics
- Instance-based Analysis & Exploration
- Attribution & Feature Visualization
- Interactive Experimentation

WHEN

During Training

After Training

Publication Venue

TVCG

The screenshot displays the ActiVis interface for a deep neural network model. The top navigation bar includes 'FBLearner', 'Workflow Library', 'Projects', 'Tools', and 'Help'. The main title is 'ActiVis: Visualization of Deep Neural Networks #15782570'. The interface is divided into several panels:

- COMPUTATION GRAPH:** Shows the flow of data through the network. Nodes include 'Concat', 'Dropout', 'dropout_0', 'FC', 'fc_0', 'Relu', 'FC', and 'Softmax'. A legend identifies 'Operator node' (grey), 'Blob node' (white), and 'Blob node w/ activation' (orange).
- NEURON ACTIVATION:** A matrix where rows represent groups of instances and columns represent neurons (0-32). It is sorted by activation strength for a selected neuron. A 'PROJECTED' scatter plot shows the resulting data distribution.
- INSTANCE SELECTION:** A grid of colored squares representing instance classifications for classes: DESC (blue), ENTY (orange), ABBR (green), HUM (red), NUM (purple), and LOC (brown).
- INSTANCE #108:** A detailed view of a specific instance with the text 'What is the highest dam in the U.S.?' and its predicted label 'LOC'. It also shows prediction scores for all classes.

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Example

- Instance-based Analysis & Exploration
- Attribution & Feature Visualization
- Interactive Experimentation

During Training
 After Training
WHERE

TVCG

Publication Venue

FBLearner Workflow Library Projects Tools Help Search for people, documents, to

ActiVis: Visualization of Deep Neural Networks #15782570

COMPUTATION GRAPH

INSTANCE SELECTION

Left column shows correctly classified instances. Right column shows misclassified instances, with border colors indicating predicted classes.

DESC

ENTY

ABBR

HUM

NUM

LOC

INSTANCE #108
 Text: What is the highest dam in the U.S. ?
 Label: LOC
 Prediction scores:
 [1] Class DESC: 0.50
 [2] Class ENTY: 0.21
 [3] Class LOC: 0.19
 [4] Class ABBR: 0.07
 [5] Class NUM: 0.01
 [6] Class HUM: 0.01

NEURON ACTIVATION

Each row represents a group of instances. Each column is a neuron. Columns sorted by activation strength for Neuron idx.

By class

Class	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	
DESC																																		
ENTY																																		
ABBR																																		
HUM																																		
NUM																																		
LOC																																		

By user-defined filters

- Contain 'Where'
- Contain 'located'
- Contain 'How many'
- Contain 'How'

By instance ID

#94																																		
#30																																		
#108																																		

PROJECTED

Activations for the instance #108: What is the highest dam in the U.S. ?

8 Survey Highlights

8 Survey Highlights

1. Model Interpretation
2. Expert Tool Focus
3. Instance-based Analysis
4. Expanding Audience
5. Furthering Interpretability
6. Human-in-the-loop
7. Evaluating Explanations
8. Protecting Against Attacks

Research trends

Research directions

Research Trend

3. Instance-based Analysis

33 / **38** works use **instance-based analysis**

Neural networks lack global explanations

Instance-based analysis enables local explanations

Author	Year	WHY				WHO			WHAT				HOW				WHEN		WHERE		
		Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational Units	Neurons in High-dimensional Space	Aggregated Information	Node-link Diagrams for Network Architecture	Dimensionality Reduction & Scatter Plots	Line Charts for Temporal Metrics	Instance-based Analysis & Exploration	Interactive Experimentation	Algorithms for Attribution & Feature Visualization	During Training	After Training
Abadi, et al.	2016	●	●	●		●	●								●				●	●	arXiv
Bau, et al.	2017	●	●	●		●										●		●		●	CVPR
Bilal, et al.	2017	●	●			●										●		●		●	TVCG
Bojarski, et al.	2016	●	●			●										●		●		●	arXiv
Bruckner	2014	●	●			●									●		●		●		MS Thesis
Carter, et al.	2016	●		●		●	●	●								●		●		●	Distill
Cashman, et al.	2017	●	●			●	●									●		●		●	VADL
Chae, et al.	2017	●	●			●										●		●		●	VADL
Chung, et al.	2016	●	●			●									●	●	●		●		FILM
Goyal, et al.	2016	●				●										●		●	●	●	arXiv
Harley	2015	●		●		●									●		●		●		ISVC
Hohman, et al.	2017	●	●	●		●										●		●	●	●	CHI
Kahng, et al.	2018	●	●			●	●								●	●				●	TVCG
Karpathy, et al.	2015	●				●	●									●		●		●	arXiv
Li, et al.	2015	●				●	●									●		●		●	arXiv
Liu, et al.	2017	●	●			●									●		●		●		TVCG
Liu, et al.	2018	●	●			●									●		●		●		TVCG
Ming, et al.	2017	●	●			●										●		●		●	VAST
Norton & Qi	2017	●	●	●		●	●	●								●		●		●	VizSec
Olah	2014	●		●		●									●		●		●		Web
Olah, et al.	2018	●		●		●	●	●								●		●	●	●	Distill
Pezzotti, et al.	2017	●	●			●									●	●				●	TVCG
Rauber, et al.	2017	●	●	●		●									●					●	TVCG
Robinson, et al.	2017	●				●	●									●		●		●	GeoHum.
Rong, et al.	2016	●	●			●	●									●		●		●	ICML VIS
Smilkov, et al.	2016	●				●									●		●			●	NIPS Workshop
Smilkov, et al.	2017	●	●	●		●									●	●		●		●	ICML VIS
Strobel, et al.	2017	●	●			●	●								●		●			●	TVCG
Tzeng & Ma	2005	●				●									●		●			●	VIS
Wang, et al.	2018	●	●	●		●									●	●				●	TVCG
Webster, et al.	2017	●		●		●										●		●		●	Web
Wongsuphasawat, et al.	2018	●				●									●		●			●	TVCG
Yosinski, et al.	2015	●		●		●	●	●								●		●	●	●	ICML DL
Zahavy, et al.	2016	●	●			●									●		●			●	ICML
Zeiler, et al.	2014	●	●			●														●	ECCV
Zeng, et al.	2017	●	●			●										●		●		●	VADL
Zhong, et al.	2017	●	●			●									●	●		●		●	ICML VIS
Zhu, et al.	2016	●				●	●	●								●		●	●	●	ECCV

4. Expanding Audience

Note: list current as of early 2018.

		Venue
VIS, HCI Conferences	TVCG	IEEE Transactions on Visualization and Computer Graphics
	VAST	IEEE Conference on Visual Analytics Science and Technology
	InfoVis	IEEE Information Visualization
	CHI	ACM Conference on Human Factors in Computing Systems

4. Expanding Audience

Note: list current as of early 2018.

		Venue
ML, DL Conferences	NeurIPS	Conference on Neural Information Processing Systems
	ICML	International Conference on Machine Learning
	CVPR	Conference on Computer Vision and Pattern Recognition
	ICLR	International Conference on Learning Representations

4. Expanding Audience

Note: list current as of early 2018.

Venue

Workshops

VADL	IEEE VIS Workshop on Visual Analytics for Deep Learning
HCML	CHI Workshop on Human Centered Machine Learning
IDEA	KDD Workshop on Interactive Data Exploration & Analytics
	ICML Workshop on Visualization for Deep Learning
WHI	ICML Workshop on Human Interpretability in ML
	NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning
	NIPS Interpretable ML Symposium
FILM	NIPS Workshop on Future of Interactive Learning Machines
	ACCV Workshop on Interpretation and Visualization of Deep Neural Nets
	ICANN Workshop on Machine Learning and Interpretability

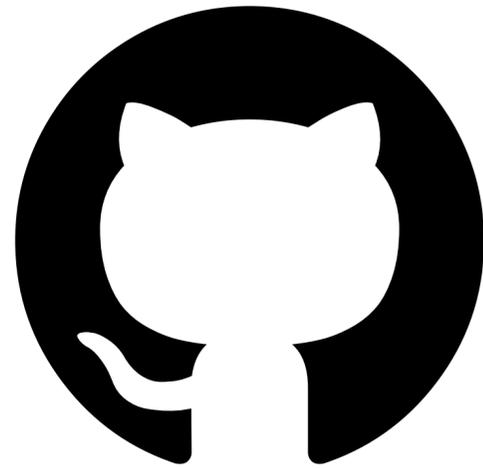
Online

Distill	Distill: Journal for Supporting Clarity in Machine Learning
arXiv	arXiv.org e-Print Archive

Research Trend

4. Expanding Audience

Note: list current as of early 2018.



Top venues highly value
open source

Research Direction

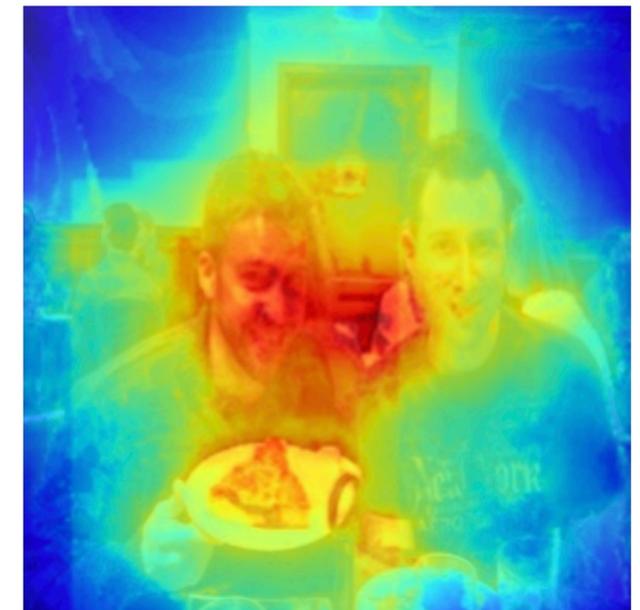
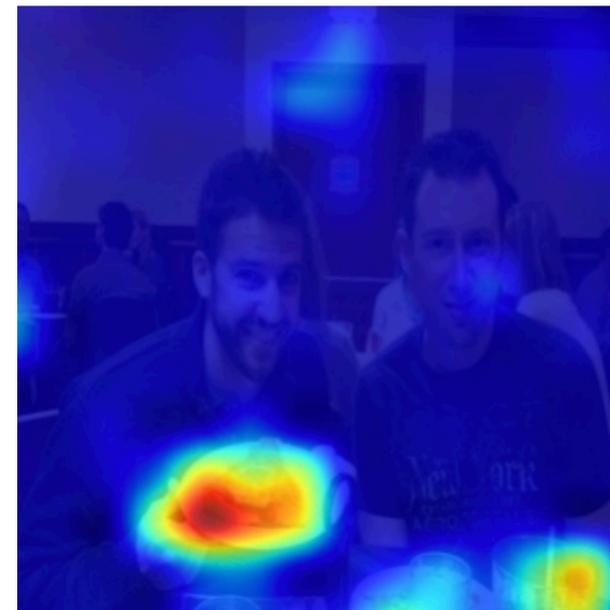
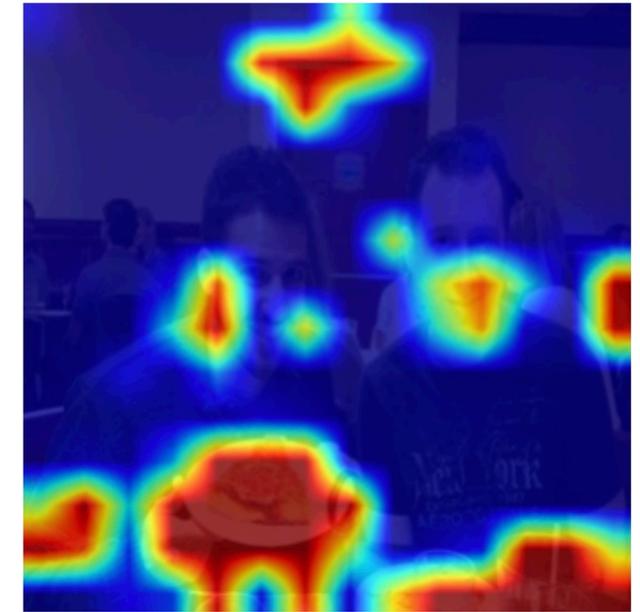
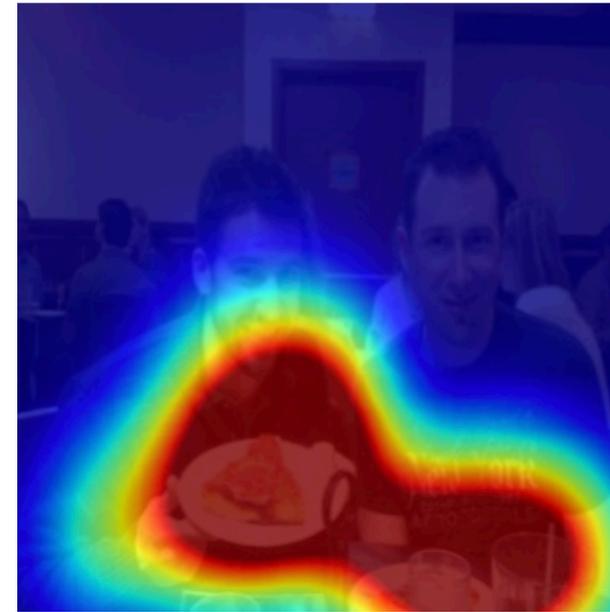
5. Furthering Interpretability

Q: What are they doing?

A: eating

Attention

Das, Agrawal, et al. 2016

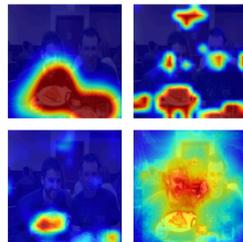


Research Direction

5. Furthering Interpretability

Attention

Das, Agrawal, et al. 2016



Saliency

Smilkov, et al. 2017

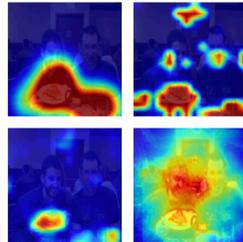


Research Direction

5. Furthering Interpretability

Attention

Das, Agrawal, et al. 2016



Saliency

Smilkov, et al. 2017



Feature visualization

Olah, et al. 2017

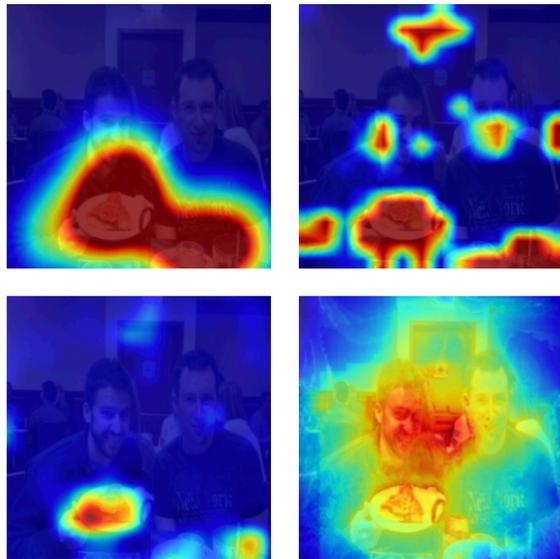


Research Direction

5. Furthering Interpretability

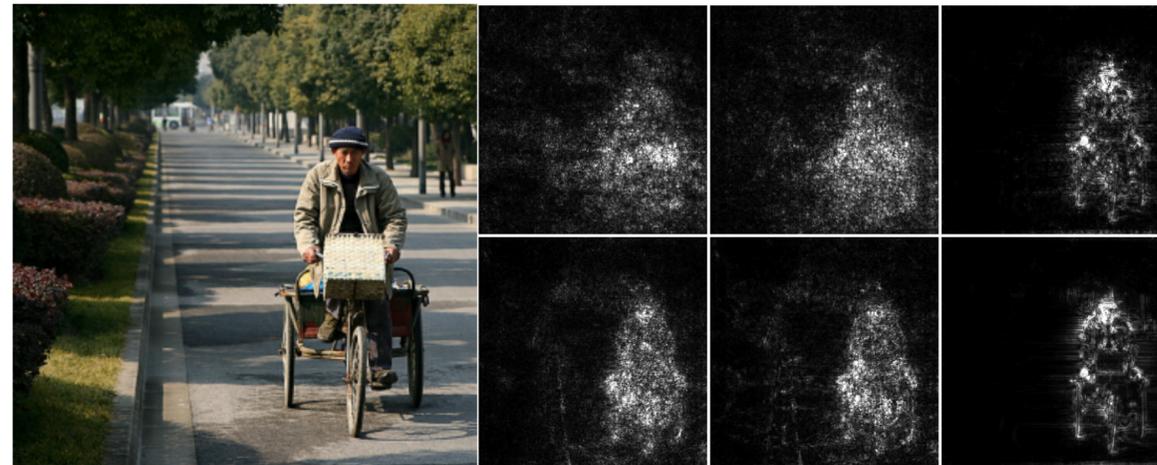
Attention

Das, Agrawal, et al. 2016



Saliency

Smilkov, et al. 2017



Feature visualization

Olah, et al. 2017



Research Direction

5. Furthering Interpretability

The screenshot shows the Distill journal website with a dark blue header containing the logo and navigation links (ABOUT, PRIZE, SUBMIT). The main content area lists several articles:

- Distill Update 2018** (Aug. 14, 2018) - An Update from the Editorial Team. Labeled as EDITORIAL.
- Differentiable Image Parameterizations** (July 25, 2018) - A powerful, under-explored tool for neural network visualizations and art. Accompanied by a 2x4 grid of colorful, abstract images.
- Feature-wise transformations** (July 9, 2018) - A simple and surprisingly effective family of conditioning mechanisms. Accompanied by a diagram showing a 'FILM generator' receiving input from a landscape image and outputting parameters to a neural network architecture consisting of sub-networks, normalization layers, and FILM layers.
- The Building Blocks of Interpretability** (March 6, 2018) - Interpretability techniques are normally studied in isolation. We explore the powerful interfaces that arise when you combine them — and the rich structure of this combinatorial space. Accompanied by an image of a dog and a kitten next to a heatmap visualization.
- Using Artificial Intelligence to** (Dec. 4, 2017) - Partially visible at the bottom.

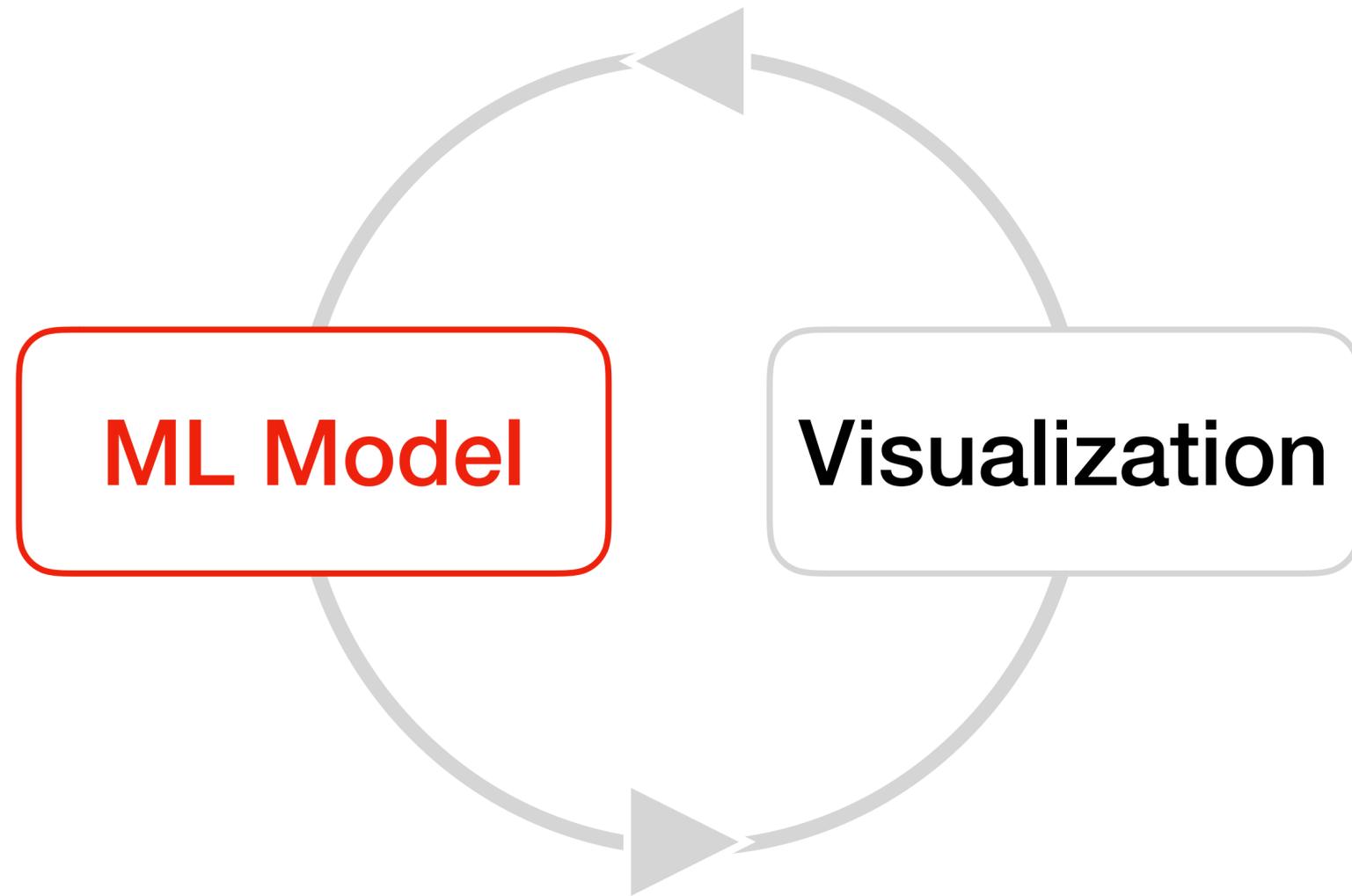
Distill

Journal for Supporting
Clarity in Machine Learning

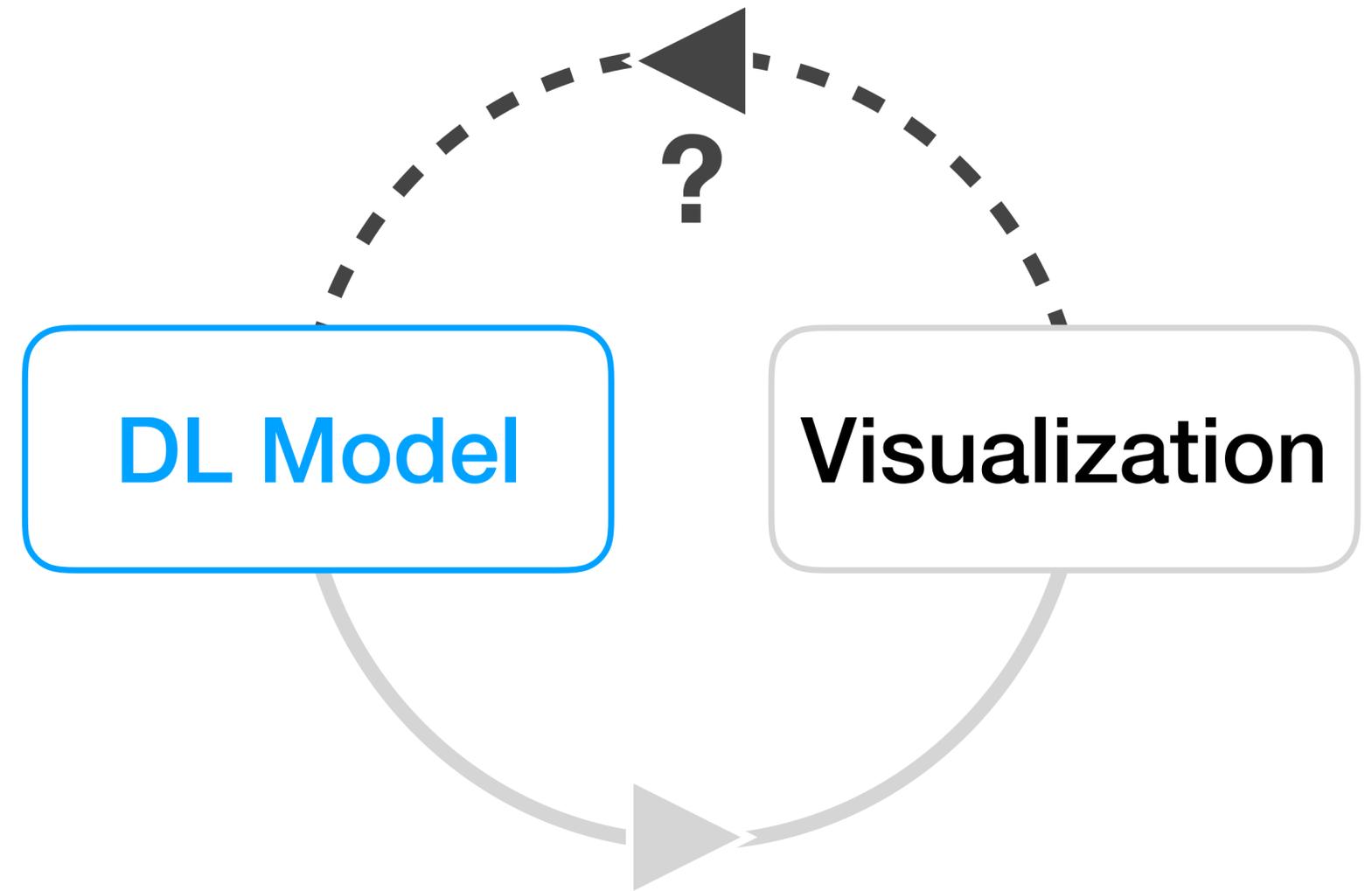
Research Direction

6. Human-in-the-loop

Interactive Machine Learning



Interactive Deep Learning



7. Evaluating Explanations

Towards A Rigorous Science of Interpretable Machine Learning

Finale Doshi-Velez* and Been Kim*

From autonomous cars and adaptive email-filters to predictive policing systems, machine learning (ML) systems are increasingly ubiquitous; they outperform humans on specific tasks [Mnih et al., 2013, Silver et al., 2016, Hamill, 2017] and often guide processes of human understanding and decisions [Carton et al., 2016, Doshi-Velez et al., 2014]. The deployment of ML systems in complex applications has led to a surge of interest in systems optimized not only for expected task performance but also other important criteria such as safety [Otte, 2013, Amodei et al., 2016, Varshney and Alemzadeh, 2016], nondiscrimination [Bostrom and Yudkowsky, 2014, Ruggieri et al., 2010, Hardt et al., 2016], avoiding technical debt [Sculley et al., 2015], or providing the right to explanation [Goodman and Flaxman, 2016]. For ML systems to be used safely, satisfying these auxiliary criteria is critical. However, unlike measures of performance such as accuracy, these crite-

Research Direction

7. Evaluating Explanations

Doshi-Velez,
Kim. 2017



More
specific
and costly



Evaluation

Humans

Tasks

Application-grounded

Yes

Real

Human-grounded

Yes

Simple

Functionally-grounded

No

Proxy

8. Protecting Against Attacks

Benign



“panda” ✓

Perturbation



attack

Attacked



“gibbon” ✗

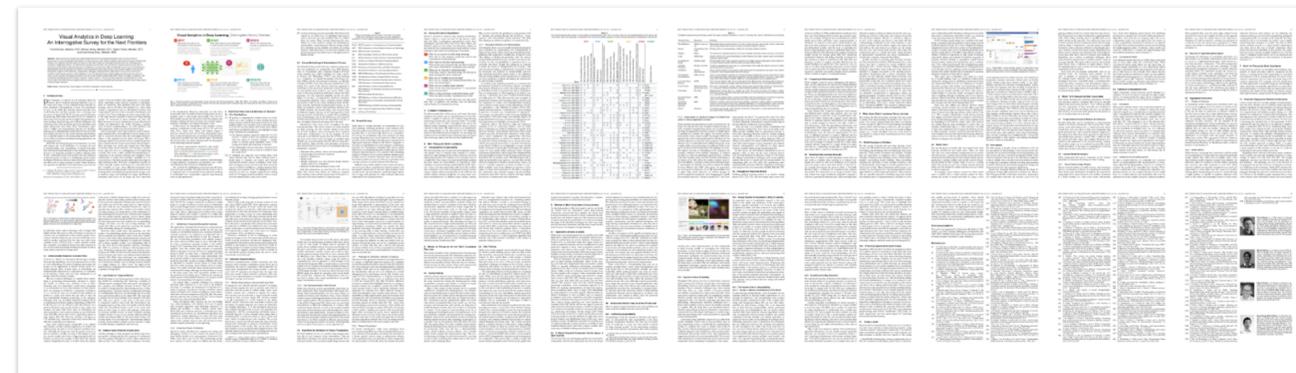
Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

[Fred Hohman](#), [Minsuk Kahng](#), [Robert Pienta](#), [Duen Horng Chau](#)

Deep learning has recently seen rapid development and significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the innate complexity and nonlinear structure of deep neural networks, the underlying decision making processes for why these models are achieving such high performance are challenging and sometimes mystifying to interpret.

As deep learning spreads across domains, it is of paramount importance that we equip users of deep learning with tools for understanding when a model works correctly, when it fails, and ultimately how to improve its performance. Standardized toolkits for building neural networks have helped democratize deep learning; visual analytics systems have now been developed to support model explanation, interpretation, debugging, and improvement.



[Read the paper.](#)

We present a survey of the role of visual analytics in deep

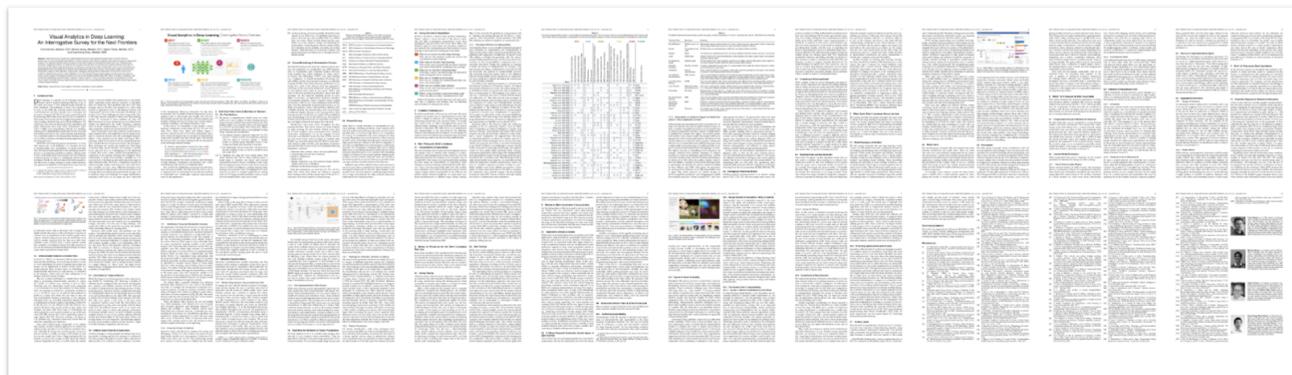
Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

[Fred Hohman](#), [Minsuk Kahng](#), [Robert Pienta](#), [Duen Horng Chau](#)

Deep learning has recently seen rapid development and significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the innate complexity and nonlinear structure of deep neural networks, the underlying decision making processes for why these models are achieving such high performance are challenging and sometimes mystifying to interpret.

As deep learning spreads across domains, it is of paramount importance that we equip users of deep learning with tools for understanding when a model works correctly, when it fails, and ultimately how to improve its performance. Standardized toolkits for building neural networks have helped democratize deep learning; visual analytics systems have now been developed to support model explanation, interpretation, debugging, and improvement.



Read the paper.

We present a survey of the role of visual analytics in deep

[bit.ly/
va-dl-survey](https://bit.ly/va-dl-survey)

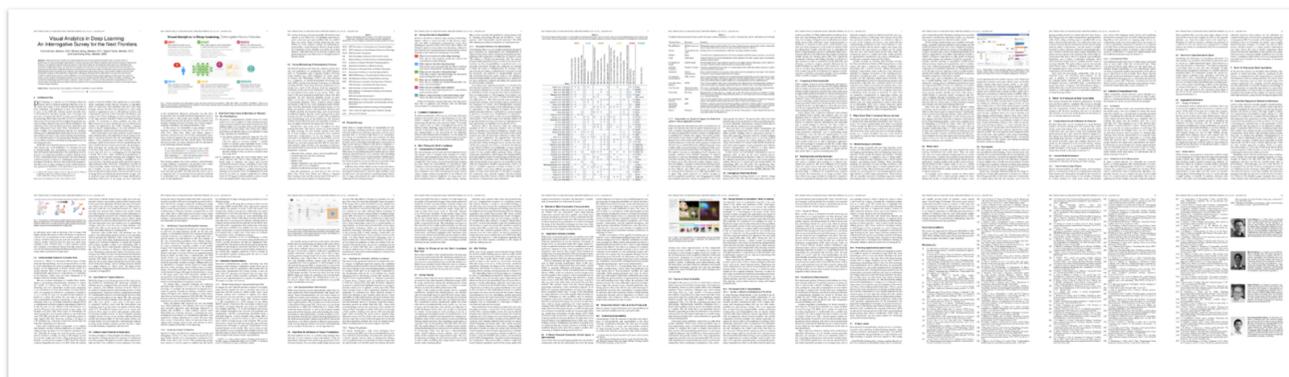
Deep Learning

An Interrogative Survey for the Next Frontiers

Fred Hohman, Minsuk Kahng, Robert Pienta, Duen Horng Chau

Deep learning has recently seen rapid development and significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the innate complexity and nonlinear structure of deep neural networks, the underlying decision making processes for why these models are achieving such high performance are challenging and sometimes mystifying to interpret.

As deep learning spreads across domains, it is of paramount importance that we equip users of deep learning with tools for understanding when a model works correctly, when it fails, and ultimately how to improve its performance. Standardized toolkits for building neural networks have helped democratize deep learning; visual analytics systems have now been developed to support model explanation, interpretation, debugging, and improvement.



Read the paper.

We present a survey of the role of visual analytics in deep learning research, noting its short yet impactful history and summarize the state-of-the-art using a human-centered interrogative framework, focusing on the Five W's and How (**WHY**, **WHO**, **WHAT**, **HOW**, **WHEN**, and **WHERE**), to thoroughly summarize deep learning visual analytics research. We conclude by highlighting research directions and open research problems.

bit.ly/
va-dl-survey

Read the paper

Author	Year	WHY				WHO			WHAT				HOW				WHEN		WHERE		
		Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational Units	Neurons in High-dimensional Space	Aggregated Information	Node-link Diagrams for Network Architecture	Dimensionality Reduction & Scatter Plots	Line Charts for Temporal Metrics	Instance-based Analysis & Exploration	Interactive Experimentation	Algorithms for Attribution & Feature Visualization	During Training	After Training
Abadi, et al.	2016	■	■	■		■	■				■			■					■	■	arXiv
Bau, et al.	2017	■		■		■				■					■		■			■	CVPR
Bilal, et al.	2017	■	■			■				■					■		■			■	TVCG
Bojarski, et al.	2016	■	■			■				■					■		■			■	arXiv
Bruckner	2014	■	■			■			■	■			■		■		■			■	MS Thesis
Carter, et al.	2016	■		■		■	■	■		■	■	■	■			■	■			■	Distill
Cashman, et al.	2017	■	■			■	■			■	■					■				■	VADL
Chae, et al.	2017	■	■			■				■	■			■	■					■	VADL
Chung, et al.	2016	■	■			■			■	■	■	■	■	■	■	■				■	FILM
Goyal, et al.	2016	■					■			■						■	■	■		■	arXiv
Harley	2015	■		■			■		■	■	■			■		■	■			■	ISVC
Hohman, et al.	2017	■		■	■		■				■					■	■	■		■	CHI
Kahng, et al.	2018	■	■			■	■			■	■	■	■	■	■					■	TVCG
Karpathy, et al.	2015	■				■	■			■	■	■	■	■	■					■	arXiv
Li, et al.	2015	■				■	■			■	■	■	■	■	■					■	arXiv
Liu, et al.	2017	■	■			■			■	■	■	■	■	■	■					■	TVCG
Liu, et al.	2018	■	■			■			■	■	■	■	■	■	■					■	TVCG
Ming, et al.	2017	■		■		■				■					■					■	VAST
Norton & Qi	2017	■	■	■		■	■	■								■	■			■	VizSec
Olah	2014	■		■			■				■			■	■	■				■	Web
Olah, et al.	2018	■		■		■	■	■		■	■	■	■	■	■					■	Distill
Pezzotti, et al.	2017	■	■			■				■	■	■	■	■	■					■	TVCG
Rauber, et al.	2017	■	■	■		■				■	■	■	■	■	■					■	TVCG
Robinson, et al.	2017	■				■	■			■	■	■	■	■	■					■	GeoHum.
Rong, et al.	2016	■	■			■	■			■					■					■	ICML VIS
Smilkov, et al.	2016	■				■				■	■	■	■	■	■					■	NIPS Workshop
Smilkov, et al.	2017	■	■	■			■		■	■	■		■	■	■					■	ICML VIS
Strobelt, et al.	2017	■	■			■	■			■	■	■	■	■	■					■	TVCG
Tzeng & Ma	2005	■				■			■	■			■	■						■	VIS
Wang, et al.	2018	■	■	■		■			■	■	■	■	■	■	■					■	TVCG
Webster, et al.	2017			■			■								■	■				■	Web
Wongsuphasawat, et al.	2018		■			■			■				■								TVCG
Yosinski, et al.	2015	■		■		■	■	■		■	■	■	■	■	■					■	ICML DL
Zahavy, et al.	2016	■	■			■				■	■	■	■	■	■					■	ICML
Zeiler, et al.	2014	■	■			■			■	■	■									■	ECCV
Zeng, et al.	2017	■		■		■			■	■					■					■	VADL
Zhong, et al.	2017	■	■			■				■	■	■	■	■	■					■	ICML VIS
Zhu, et al.	2016	■				■	■	■							■	■	■			■	ECCV

bit.ly/
va-dl-survey

Paper table,
with links

Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

Fred Hohman, Minsuk Kahng, Robert Pienta, Duen Horng Chau

Deep learning has recently seen rapid development and significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the innate complexity and nonlinear structure of deep neural networks, the underlying decision-making processes for why these models are achieving such high performance are challenging and sometimes mystifying to interpret.

As deep learning spreads across domains, it is of paramount importance that we equip users of deep learning with tools for understanding when a model works correctly, when it fails, and ultimately how to improve its performance.

Standardized toolkits for building neural networks have helped democratize deep learning; visual analytics systems have now been developed to support model explanation, interpretation, debugging, and improvement.



Read the paper.

We present a survey of the role of visual analytics in deep learning research, noting its short yet impactful history and summarize the state-of-the-art using a human-centered interrogative framework, focusing on the Five W's and How (WHY, WHO, WHAT, HOW, WHEN, and WHERE), to thoroughly summarize deep learning visual analytics research. We conclude by highlighting research directions and open research problems.

This survey helps new researchers and practitioners in both visual analytics and deep learning to quickly learn key aspects of this young and rapidly growing body of research, whose impact spans a diverse range of domains.

Overview of representative works in visual analytics for deep learning. Each row is one work; works are sorted alphabetically by first author's lastname. Each column corresponds to a subsection from the six interrogative questions. A work's relevant subsection is indicated by a colored cell.

Year	Why	Who	What	How	When	Where
2014						
2015						
2016						
2017						
2018						
2019						
2020						
2021						
2022						
2023						
2024						
2025						
2026						
2027						
2028						
2029						
2030						

Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers
 Fred Hohman, Minsuk Kahng, Robert Pienta, Duen Horng Chau
 IEEE Transactions on Visualization and Computer Graphics (TVCG), 2018.

Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

IEEE TVCG 2018

Thanks!

bit.ly/
va-dl-survey



Fred Hohman
@fredhohman



Minsuk Kahng



Robert Pienta



Polo Chau

