A Survey on
Visualization for Explainable Classifiers

Yao MING

THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Introduction

Explainable Classifiers

Visualization for Explainable Classifiers

Conclusion

Introduction

Motivation

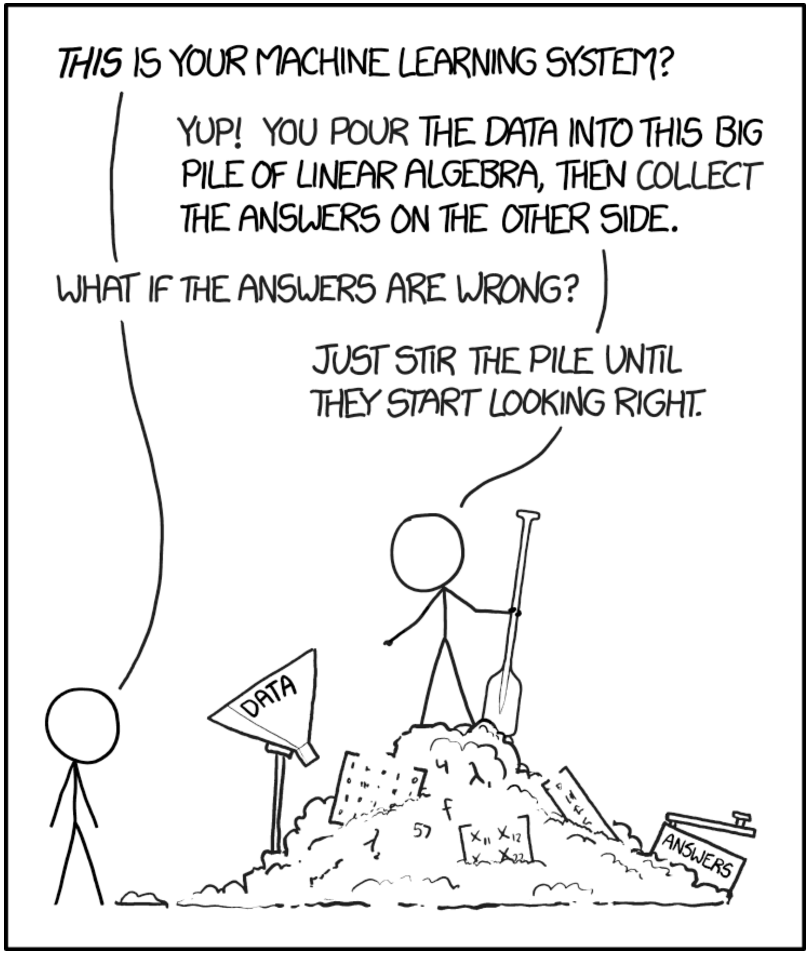
Concepts

Explainable Classifiers

Visualization for Explainable Classifiers

Conclusion

Motivation



https://xkcd.com/1838/

Does this matter?

Motivation

A study from Cost-Effective HealthCare (CEHC) (Cooper et al. 1997)

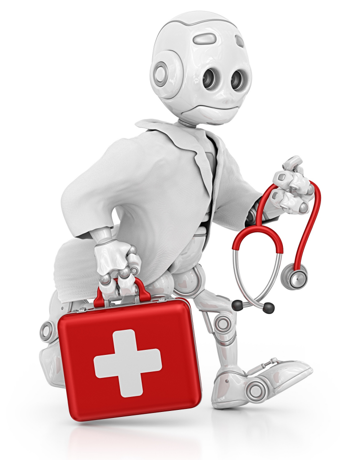
Predicting the **probability of death** (POD) for patients with pneumonia

If HighRisk(x):

admit to hospital

Else:

treat as outpatient



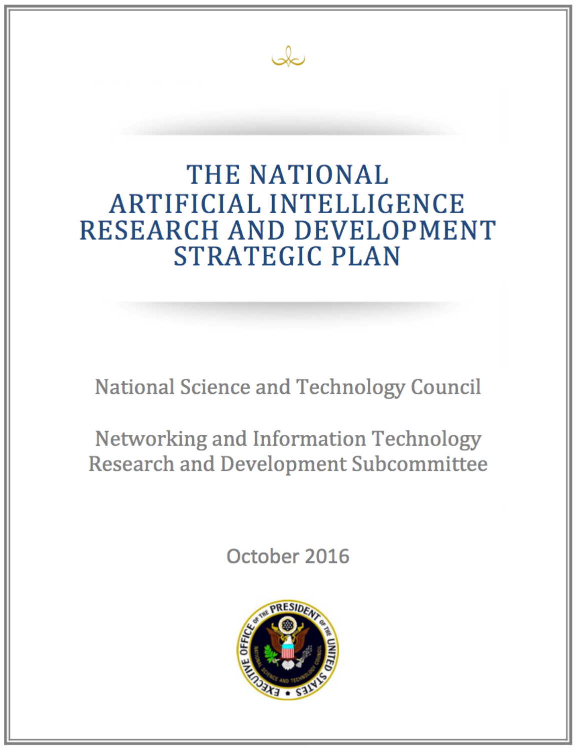
The rule-based model learned:

HasAsthma(x) => LowerRisk(x)

High risk --> aggressive treatment

We want the system to be explainable sometime!

Motivation


THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

National Science and Technology Council

Networking and Information Technology Research and Development Subcommittee

October 2016

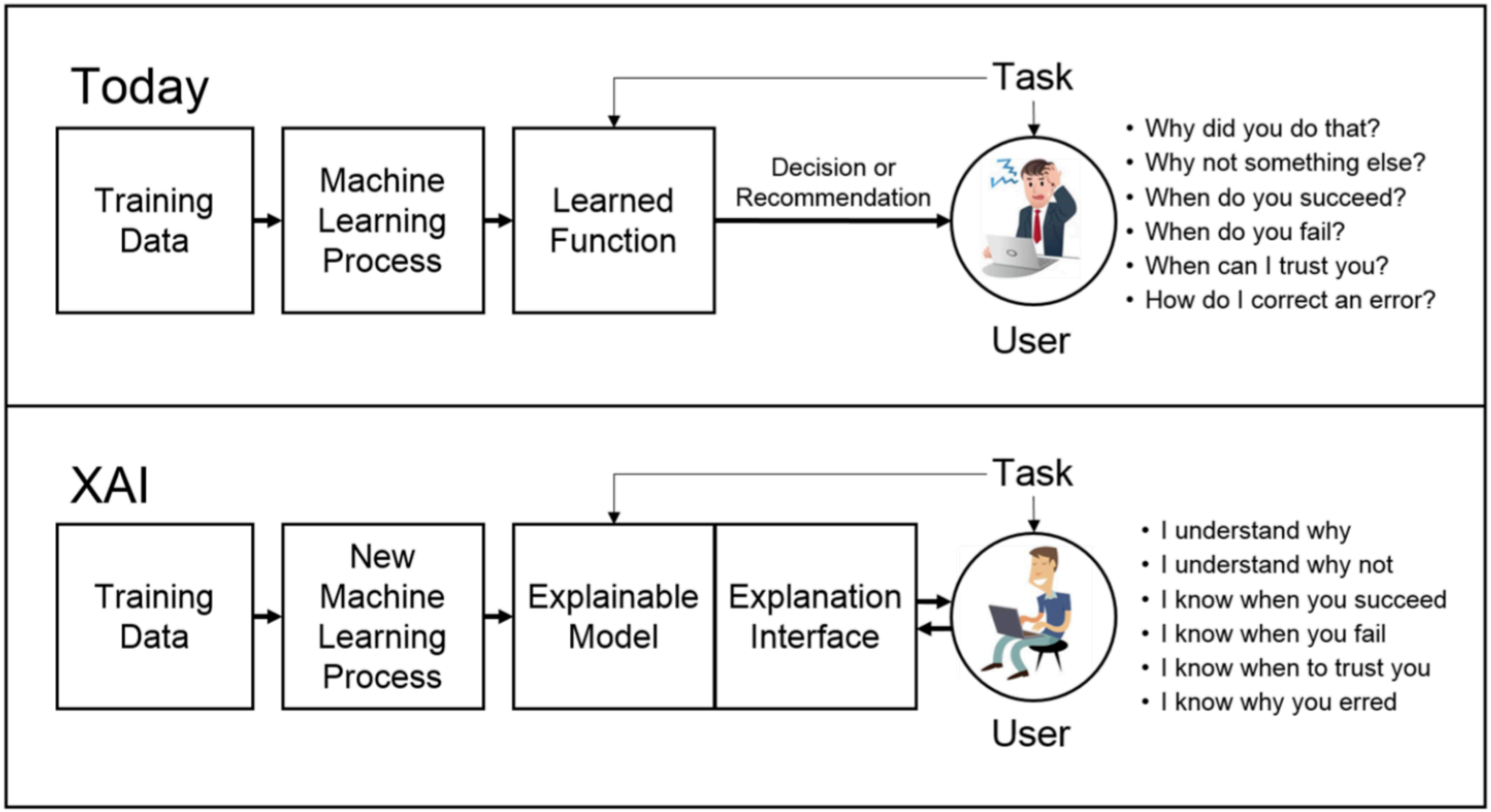
“

Strategy 2: Developing Effective Methods for AI-Human Collaboration

Better visualization and user interfaces are additional areas that need much greater development to help humans understand large-volume modern datasets and information coming from a variety of sources.

”

Motivation



The concept of XAI. DARPA, Explainable AI Project 2017

Classification

Classification: Identifying any observation $\boldsymbol{x} \in \mathcal{X}$ as a class $y \in \mathcal{Y}$, $\mathcal{Y} = \{1,2,\ldots,K\}$, given a training set $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$

Classification Model (Classifier): An algorithm f, learned from $\mathcal{D}$, specified by parameters θ, output is a vector representing a probability distribution:

$$\boldsymbol{y} = f_\theta(\boldsymbol{x}),$$

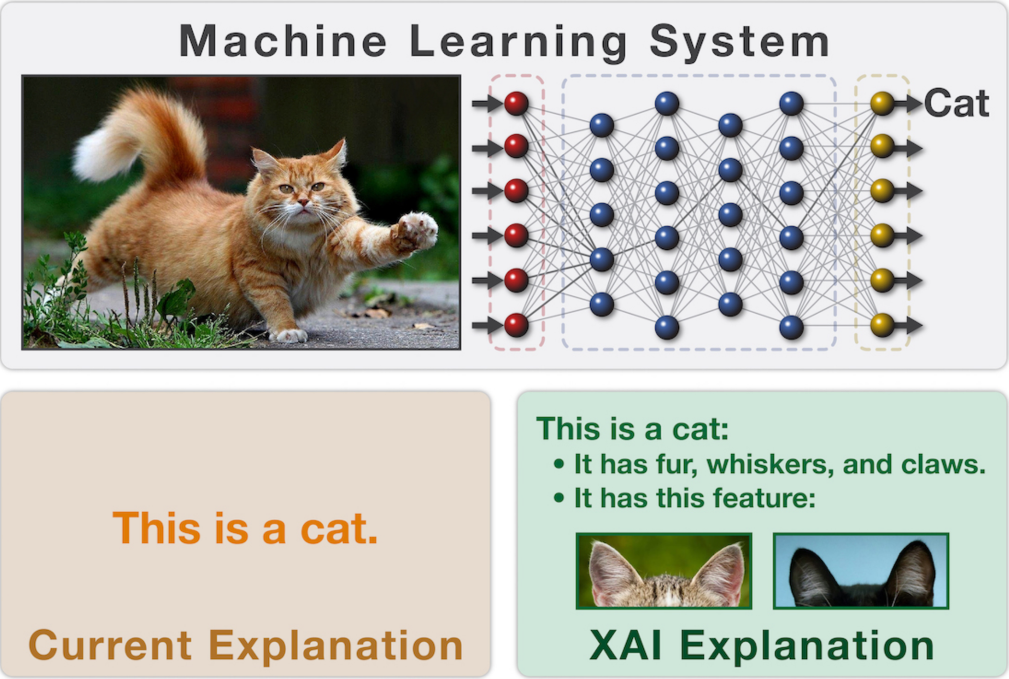
where $\boldsymbol{y} = (y_i) \in \mathbb{R}^K$, $y_i = p(\,y = i \mid \boldsymbol{x}, \mathcal{D}\,)$.

$$\boldsymbol{x} \longrightarrow \boxed{f_\theta} \longrightarrow \boldsymbol{y}$$

What is explainability?

The **explainability** of a classifier: The ability to explain the reasoning of its predictions so that humans can understand. (Doshi-Velez and Kim 2017)

Aliases in literature: interpretability, intelligibility



DARPA, Explainable AI Project 2017

Why explainable?

The Curiosity of Humans

- What has the classifier learned from the data?

Limitations of Machines

- Human knowledge as a complement

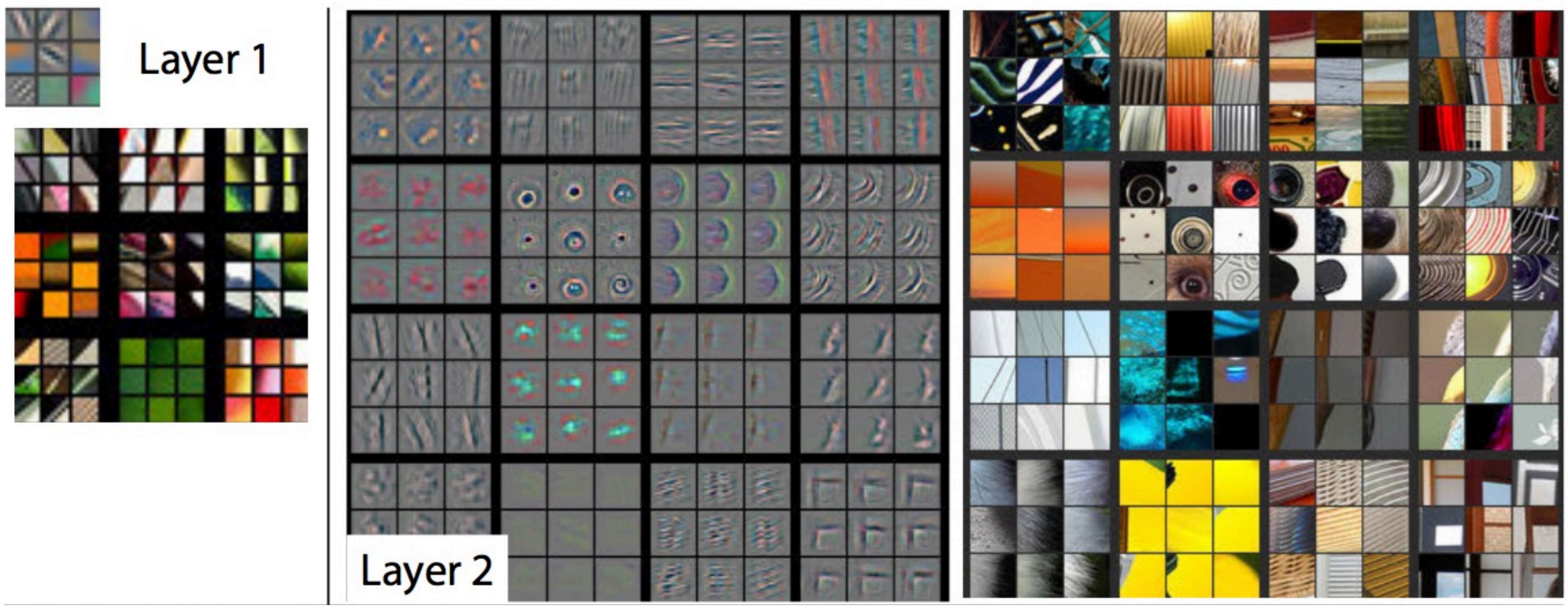
Moral and Legal Issues

- The "right to explanation"
- Fairness (non-discrimination)

Why explainable?

The Curiosity of Humans

- What has the classifier learned from the data?

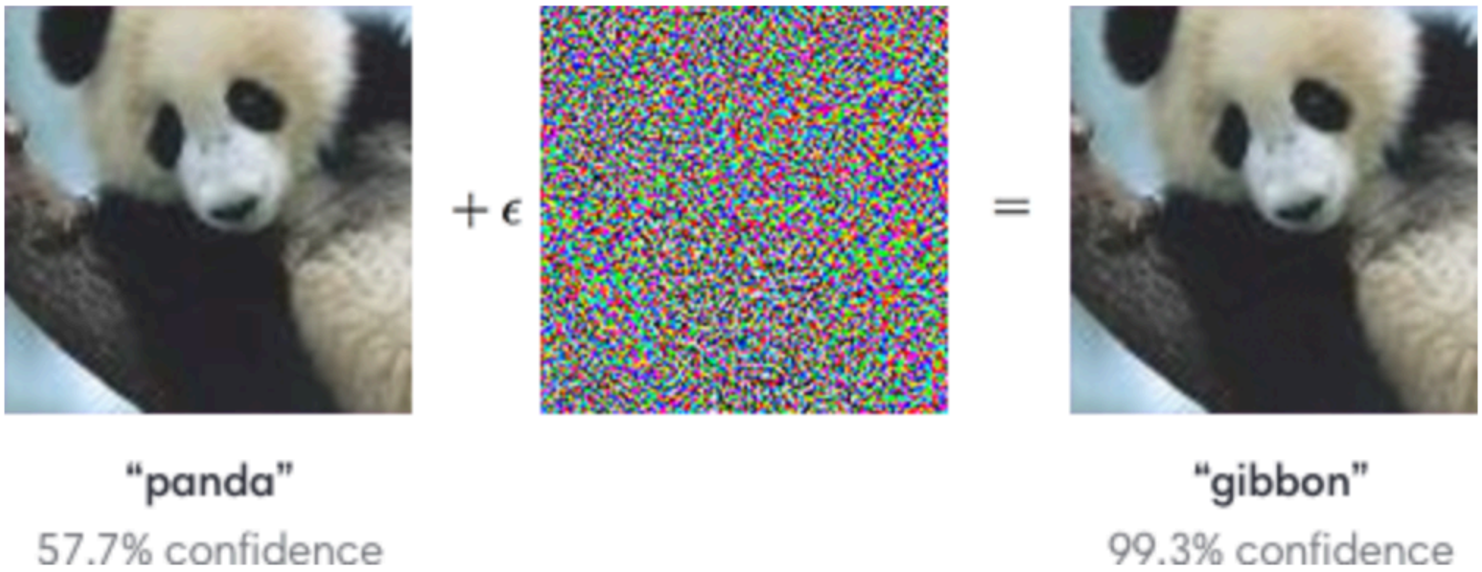


Zeiler and Fergus 2014

Why explainable?

Limitations of Machines

- Human knowledge as a complement
- Robustness of the model



Adversarial examples attack

(https://blog.openai.com/adversarial-example-research/)

Why explainable?

Moral and Legal Issues

- The "right to explanation”

  The EU general data protection regulation (GDPR 2018) Recital 71:

  In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to **obtain an explanation of the decision** reached after such assessment and to challenge the decision.

- Fairness (non-discrimination)
  - Classification systems for loan approval.
  - Resume filter for hiring.

Introduction

Explainable Classifiers

Interpretable Architecture

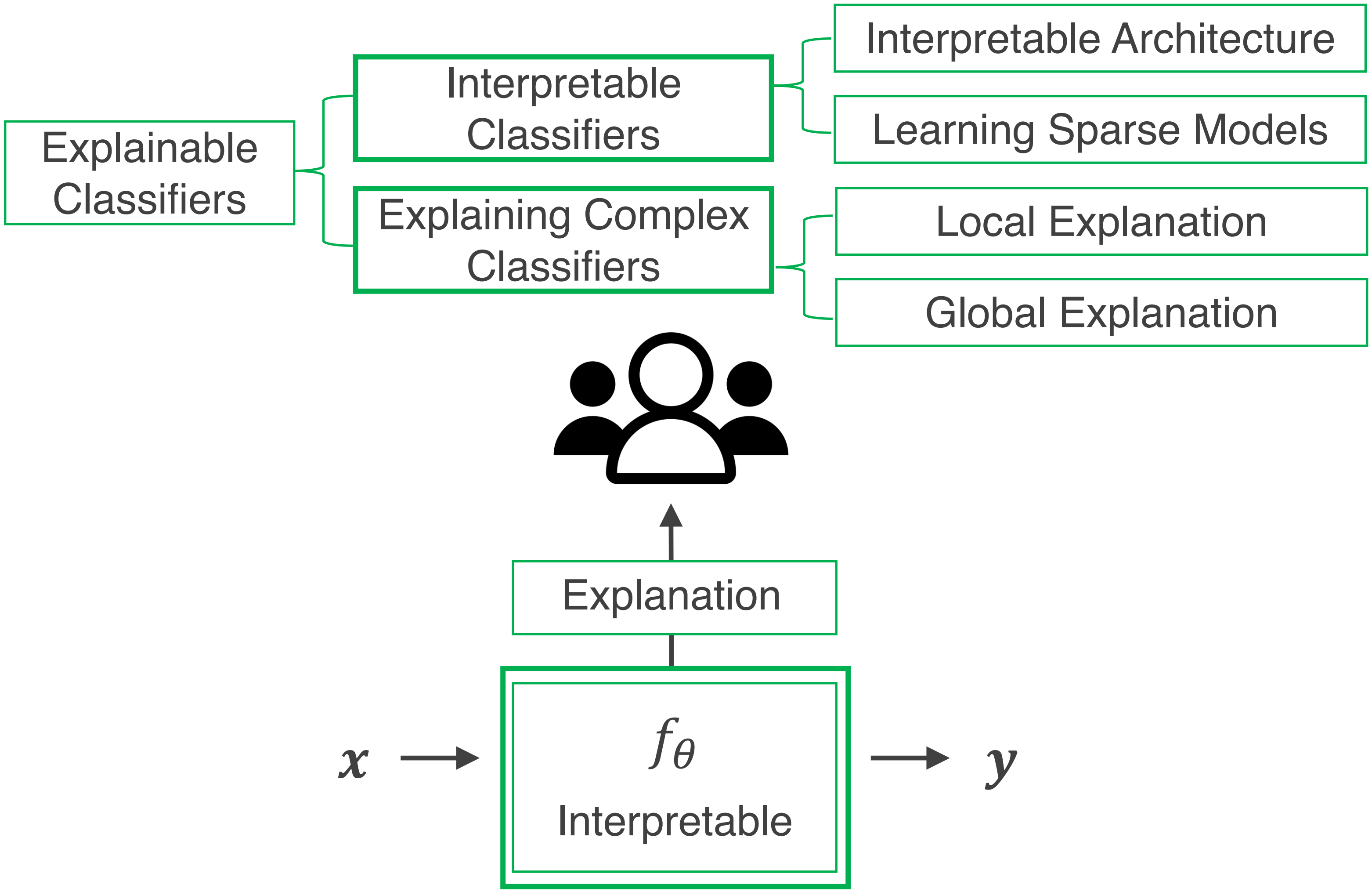
Explaining Complex Classifiers

Visualization for Explainable Classifiers

Conclusion

Explainable Classifier

Two strategies to provide explainability:



Interpretable Classifiers

Classifiers that are **commonly recognized** as understandable, and hence need little effort to explain them

$$\boldsymbol{x} \longrightarrow \boxed{f_\theta} \longrightarrow \boldsymbol{y}$$

Interpretable architecture:

- f consists of computation blocks that are easy to understand
- E.g., decision trees

Learning sparse models:

- $|\theta|$ is smaller so that it is easy to understand
- E.g., simplification

Interpretable Classifiers

Classifiers that are **commonly recognized** as understandable, and hence need little effort to explain them

$$x \longrightarrow \boxed{f_\theta} \longrightarrow y$$

Categories		Related Papers	Remarks
Interpretable Classifiers	Interpretable Architecture	Decision Trees [7], Rule Lists [27, 59], Rule Sets [60]	rule-based
		Linear Models [6]	linear
		kNNs [12, 22]	instance-based
	Learning Sparse Models	Decision Trees [43], Sparse SVMs [11], Sparse CNNs [29]	simplification
		Sparsity by Bayesian [56], Integer Models [55, 58]	direct-sparsity

Not as explainable as they seemed to be!

Interpretable Classifiers

Interpretable Architecture – Classic Methods

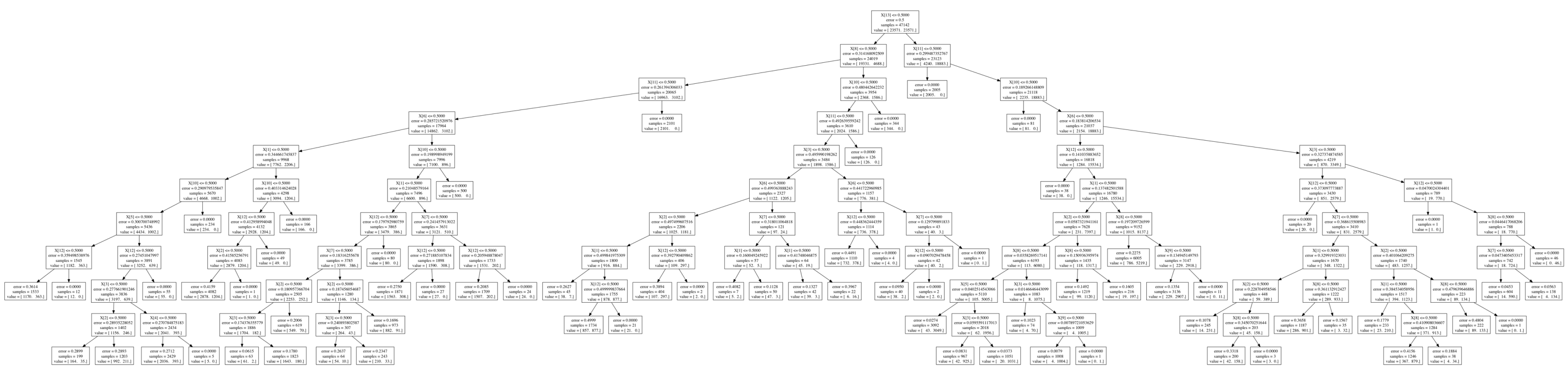
kNN (instance-based)

t is classified as Y because a, b, and c are similar to t.

Limits: lack close instances to t

Decision Tree (rule-based)

Seem to be interpretable

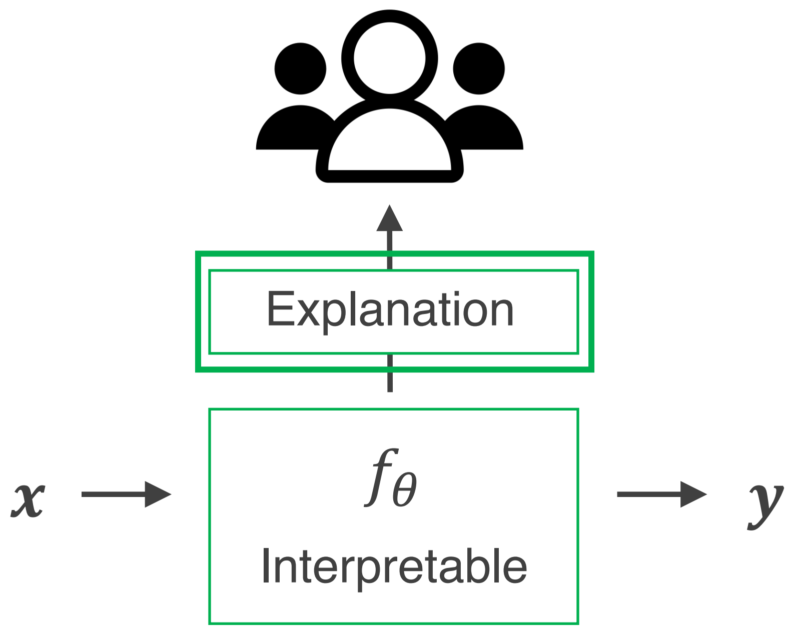


Limits: performance V.S. explainability

Explainable Classifier

Two strategies to provide explainability:

- **Interpretable Classifiers**
- **Explaining Complex Classifiers**



Explaining Complex Classifiers

What are explanations of classifiers?

Cognitive Science (Lombrozo 2006)**:**
Explanations are characterized as arguments that demonstrate all or a subset of the **causes** of the **explanandum** (the subject being explained), usually following deductions from natural laws or empirical conditions.

What is the explanandum?

1. The prediction of the classifier. (**Local explanation**)
   - Why is x classified as y?
2. The classifier itself. (**Global explanation**)
   - What has the classifier learned in general?

A summary of local explanations on $\mathcal{X}$

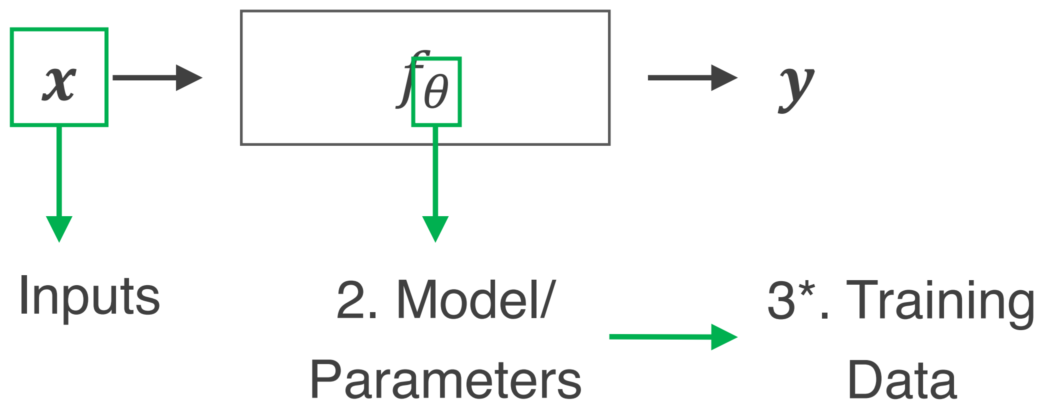
Explaining Complex Classifiers

What is explanations?

Cognitive Science (Lombrozo 2006)**:**

Arguments … of the **causes** of the **explanandum** …

What are the causes of the prediction(s) of a classifier?



Model-aware / Model-unaware

Explaining Complex Classifiers

Categories			Related Papers	Remarks
Explanations of Classifiers	Local	Model-unaware	Sensitivity Analysis [50, 28, 51]	gradient-based
			LIME [46]	model induction
			Generate Visual Explanations [19]	extra labels
		Model-aware	De-convolution [65],	CNN
			Layer-wise Propagation [4],	CNN
			Prediction Difference [66],	Image
			Output Decomposition [36],	LSTM
			Direct Mapping [21]	RNN
	Global	Unaware	Greedy-pick [46], Top-k [65]	sampling
		Model-aware	Partition Hidden Space [14, 44],	NN
			Activation maximization [13, 50],	CNN
			Network Dissection [5]	CNN

Local explanations

Sensitivity Analysis - Why is x classified as y?

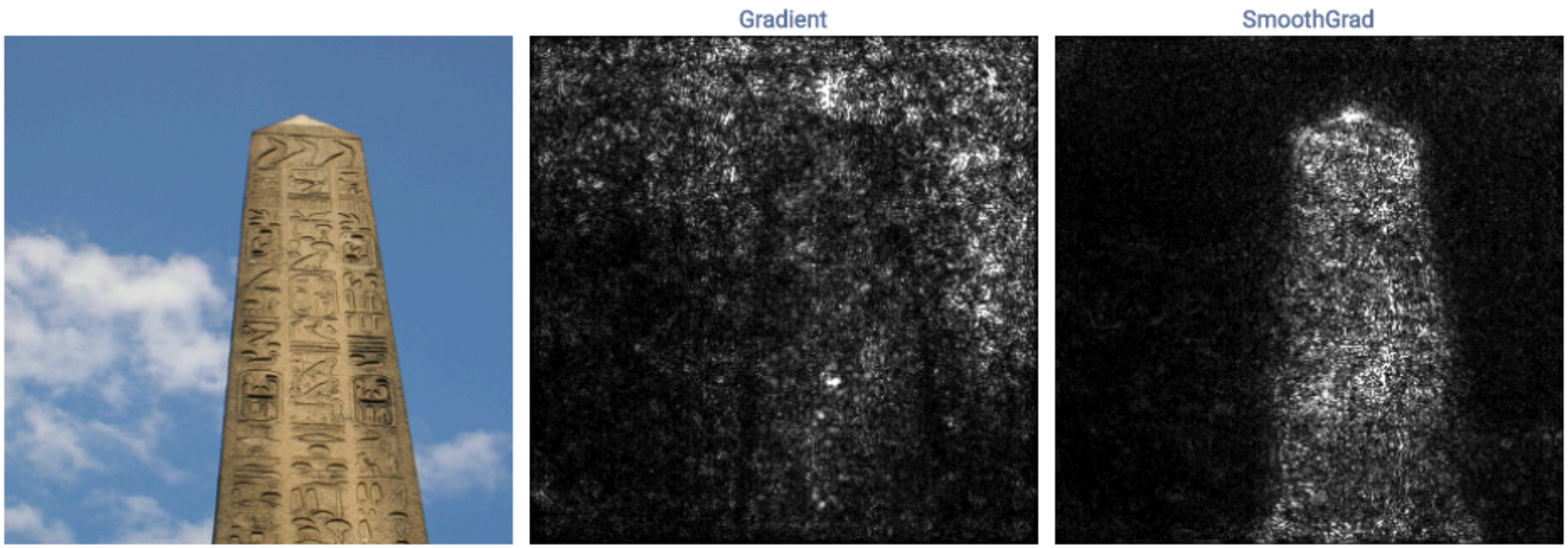
Gradients (ImageNet 2013)
(Simonyan et al. 2014)

$\frac{\partial y_i}{\partial x}(\boldsymbol{x}_{test})$

1. Too noisy!
2. High grad => important?

Local explanations

Sensitivity Analysis - Why is x classified as y?

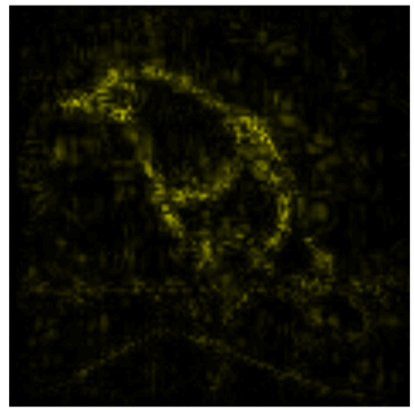


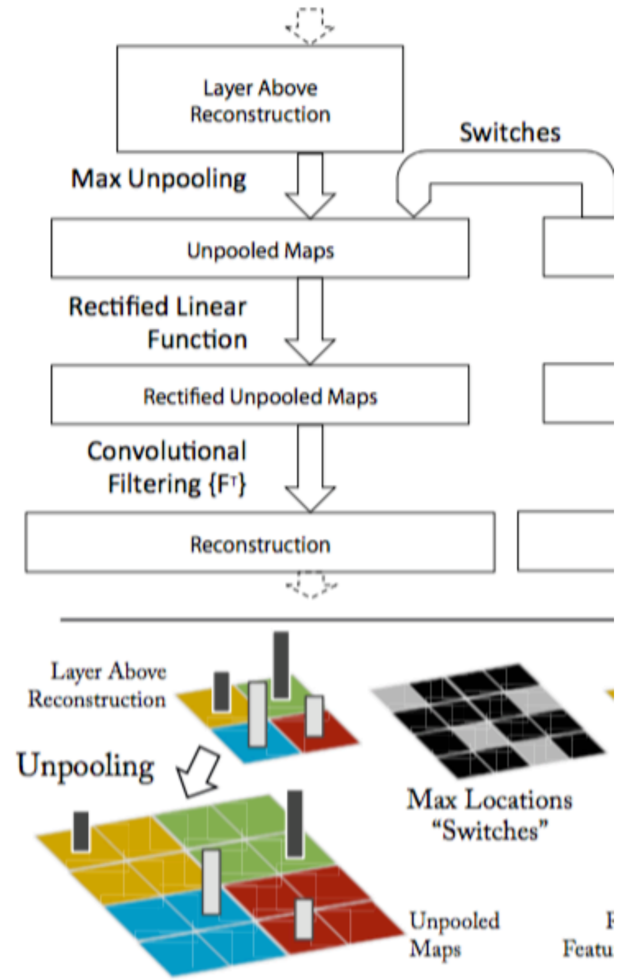
SmoothGrad (Smilkov et al. 2017)
Sampling noisy images and average the gradient map $\frac{1}{n}\sum_{j=1}^{n}\frac{\partial y_i}{\partial \boldsymbol{x}}(\boldsymbol{x}_{test}+\mathcal{N}(0,\sigma^2))$

Limit: Expensive; Non-deterministic

Local model-aware explanations

Utilizing the structure of the model - CNN





De-convolution (Zeiler and Fergus 2014):
Inverse operations of different layers

Pros:

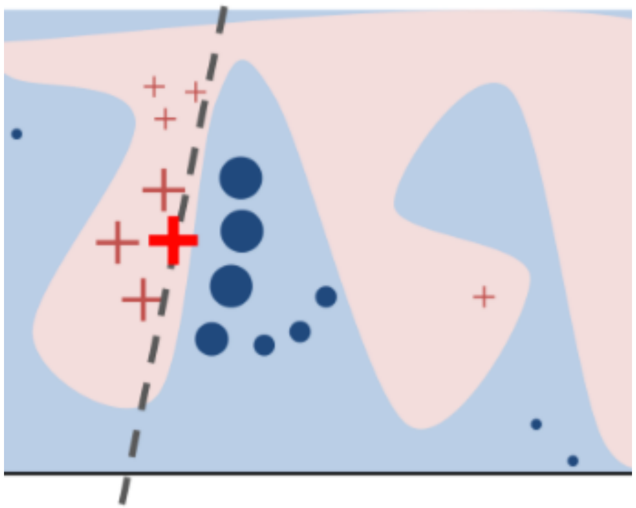
- Can apply to neurons
- Better explanations

Cons:

- Only for layer-wise, invertible models
- No relations

Local model-unaware explanations

Model Induction



(a) Husky classified as wolf (b) Explanation

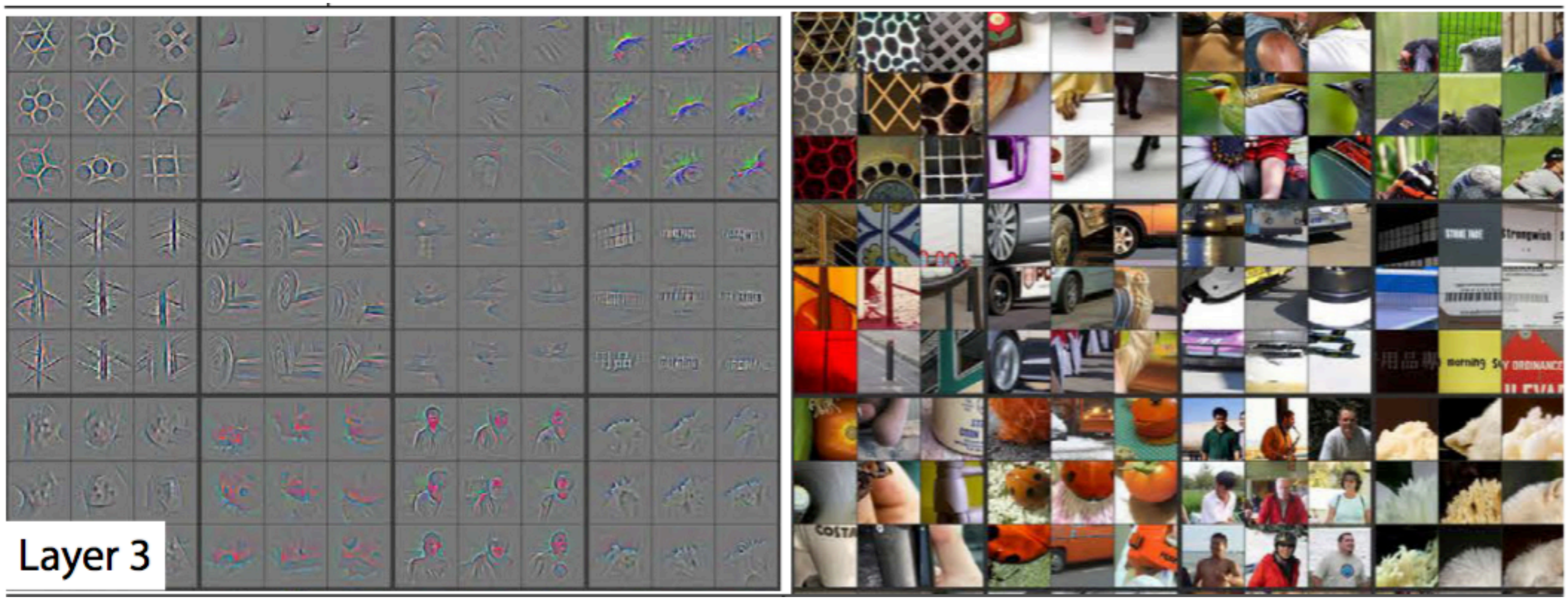
Locally approximate a complex classifier using a simple one (linear)
0-1 explanation (Ribeiro et al. 2016)

Limits: 1. induction of a simple one is by random sampling local points;
2. expensive
3. generating image patch require extra efforts

Global model-unaware explanations

Sampling local explanations

1. Select top-k instances with max activations (Zeiler and Fergus 2014)

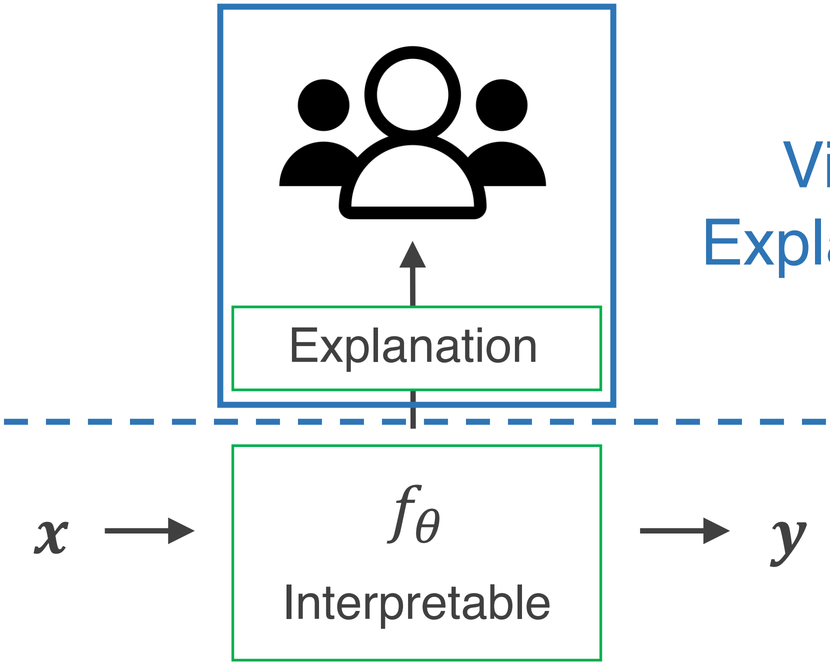


2. Select local explanations that greedily covers the most important features (Ribeiro et al. 2016)

Limit to the data; special case; expensive

Explainable Classifiers

The lack of human in the study!



Visualization for
Explainable Classifier

Introduction

Explainable Classifiers

Visualization for Explainable Classifiers

Vis for Exploratory Data Analysis

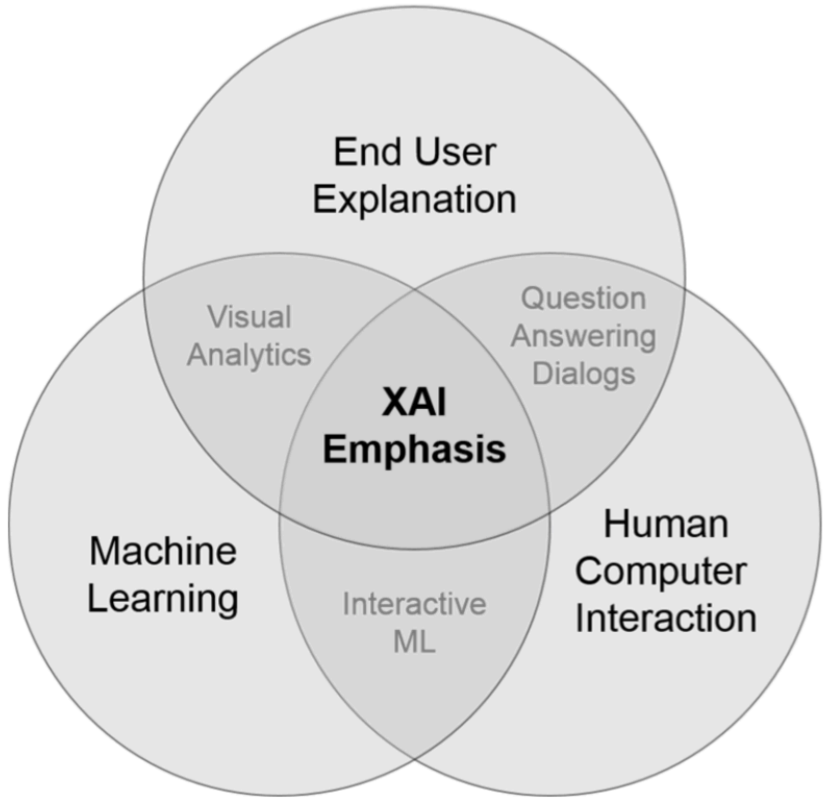
Vis for Model Development

Vis for Operation

Conclusion

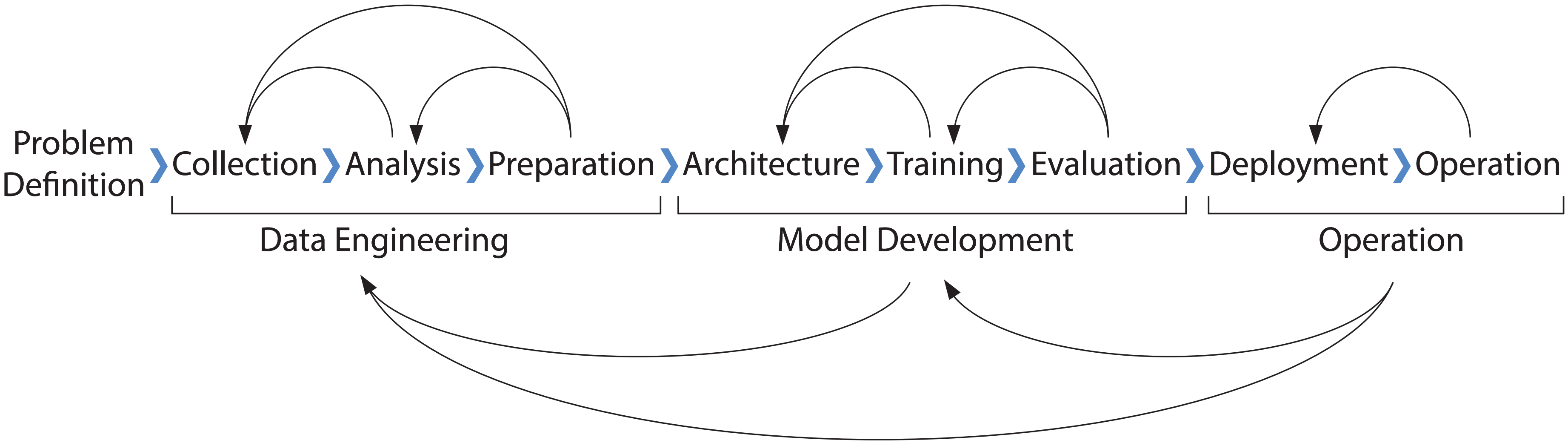
| Visualization for Explainable Classifiers

What role is visualization playing in explainable classifiers?



DARPA, Explainable AI Project 2017

The Life Cycle of a Classifier



What are the problems?

Vis for Exploratory Data Analysis

- What does my dataset look like? Any mislabels?

Vis for Model Development

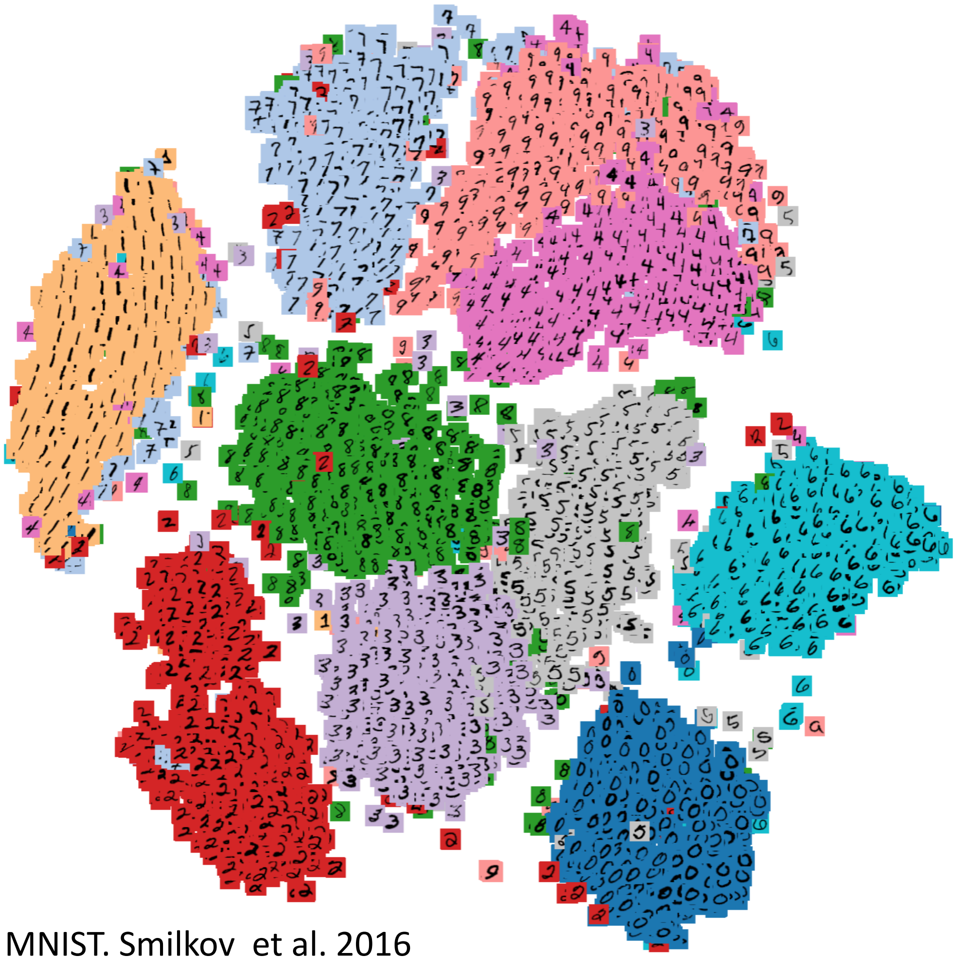
- Architecture: What is the classifier? How to compute?
- Training: How the model gradually improves? How to diagnose?
- Evaluation: What has the model learned from the data?
- Comparison: Which classifier should I choose?

Vis for Operation

- Deploy: How to establish users’ trust?
- Operation: How to identify possible failure?

Visualization for Exploratory Data Analysis

What does my dataset look like?



MNIST. Smilkov et al. 2016

It might be difficult to classify between (3,5) and (4,9)!

Methods:

- PCA
- Multidimensional Scaling
- t-SNE

Augmenting:

- Glyph (Smilkov et al. 2016)
- Color (Wang and Ma 2013)

What are the problems?

Vis for Exploratory Data Analysis

- ~~What does my dataset look like? Any mislabels?~~

Vis for Model Development

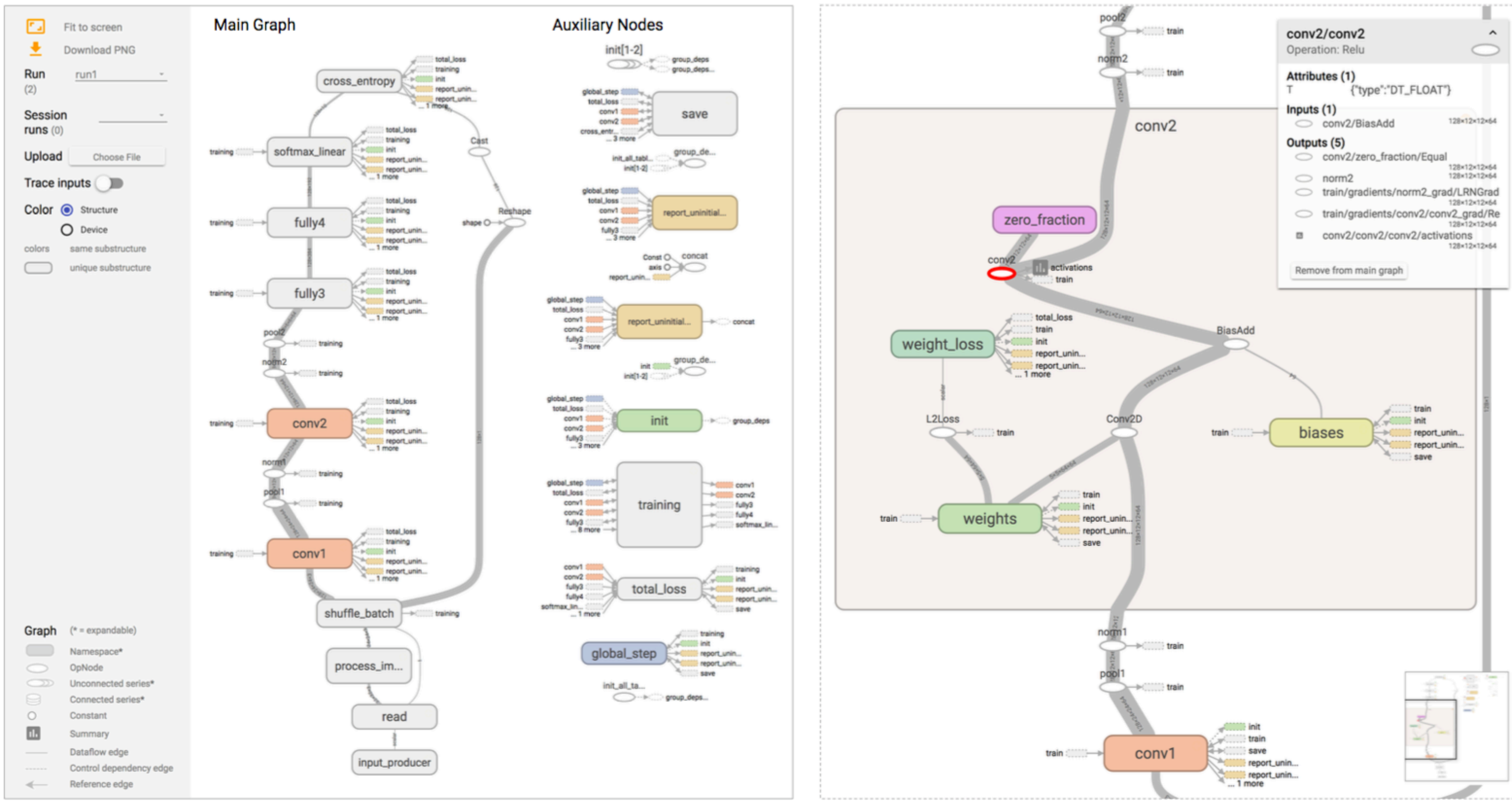
- **Architecture: What is the classifier? How to compute?**
- Training: How the model gradually improves? How to diagnose?
- Evaluation: What has the model learned from the data?
- Comparison: Which classifier should I choose?

Vis for Operation

- Deploy: How to establish users' trust?
- Operation: How to identify possible failure?

Visualization for Model Development

Architecture: How to explain the computation of a model?



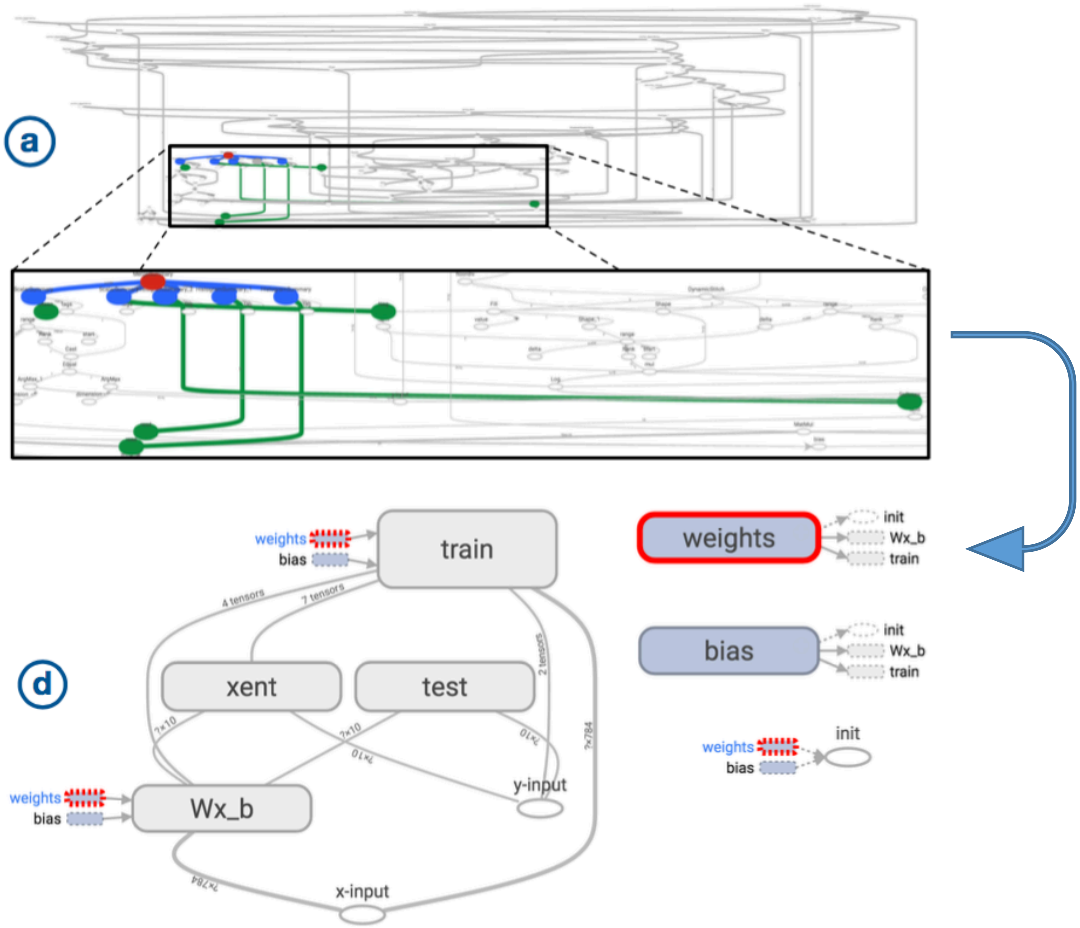
(a) (b)

#Global

Data Flow Graph (TensorBoard). Wongsuphasawat et al. 2017

Visualization for Model Development

Architecture: How to explain the computation of a model?



Data Flow Graph (TensorBoard). Wongsuphasawat et al. 2017

Visualization for Model Development

Architecture: How to explain the computation of a model?

What are the specific tasks?

1. Show an **overview** of the high-level components and their **relationships**
2. Recognize **similarities and differences** between components
3. Examine the **nested structure** of a high-level component
4. Inspect **details** of individual operations

What are the challenges?

C1. Mismatch between graph topology and semantics

A group of operations ⇔ A component?

C2. Graph heterogeneity

Different importance: inference > gradients/optimizations > logger/summary

C3. Interconnected Nodes

Connections between important nodes and less important nodes mess the graph

Data Flow Graph (TensorBoard). Wongsuphasawat et al. 2017

Visualization for Model Development

Architecture: How to explain the computation of a model?

Tasks:

1. Show an **overview** of the high-level components and their **relationships**
2. Recognize **similarities and differences** between components
3. Examine the **nested structure** of a high-level component
4. Inspect **details** of individual operations

Challenges:

C1. Mismatch between graph topology and semantics

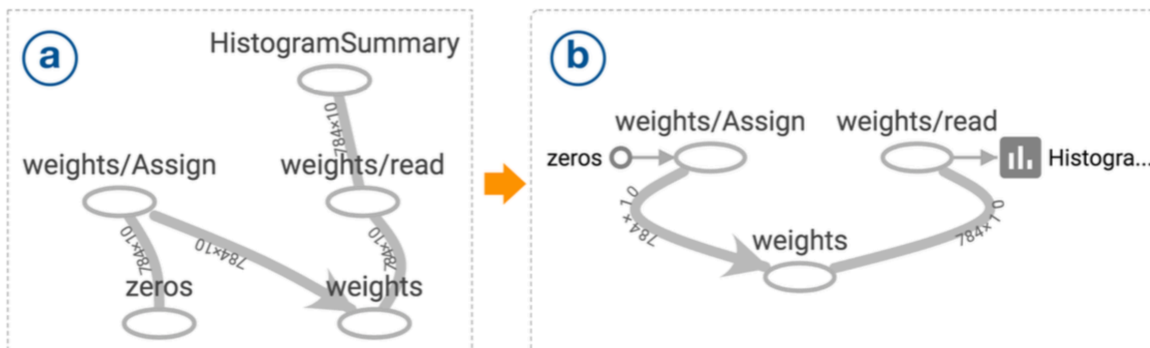
A group of operations ⇔ A

C2. Graph heterogeneity

Different importance: infere

C3. Interconnected Nodes

Connections between impo



Extract non-critical operations (**C2**)

Data Flow Graph (TensorBoard). Wongsuphasawat et al. 2017

HKUST

Visualization for Model Development

Architecture: How to explain the computation of a model?

Tasks:

1. Show an **overview** of the high-level
2. Recognize **similarities and differe**
3. Examine the **nested structure** of a
4. Inspect **details** of individual operatio

Challenges:

C1. Mismatch between graph topolo

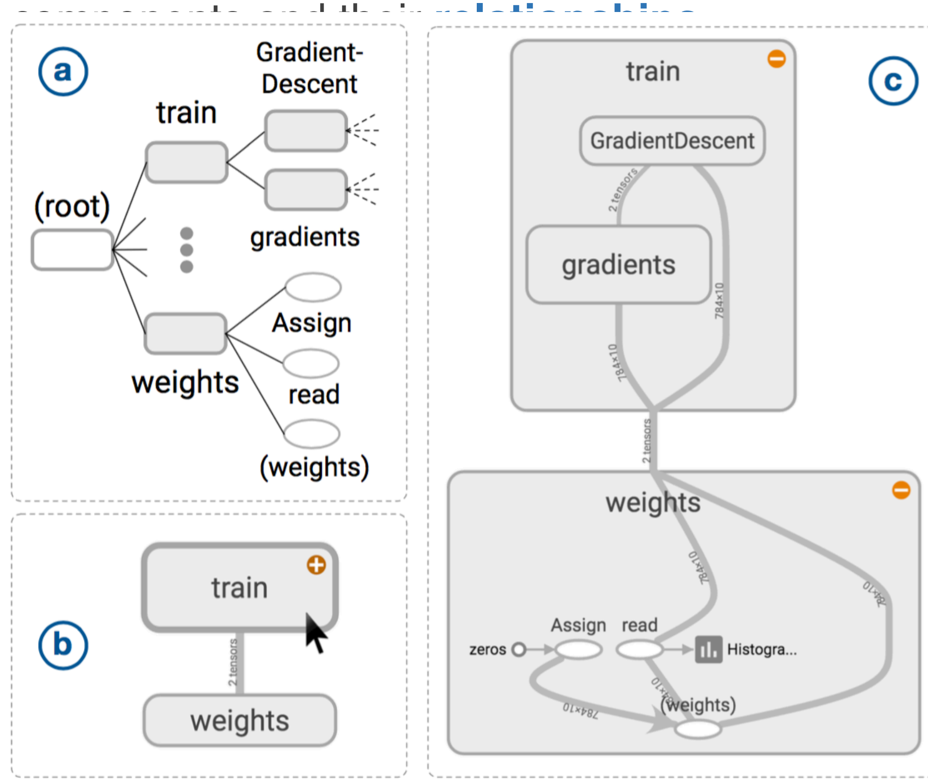
A group of operations ⇔ A component?

~~C2. Graph heterogeneity~~

Different importance: inference > gradier

C3. Interconnected Nodes

Connections between important nodes a



Build hierarchical graph based on namespaces (**C1**)

Data Flow Graph (TensorBoard). Wongsuphasawat et al. 2017

Visualization for Model Development

Architecture: How to explain the computation of a model?

Tasks:

1. Show an **overview** of the high-level compo
2. Recognize **similarities and differences** b
3. Examine the **nested structure** of a high-le
4. Inspect **details** of individual operations

Challenges:

~~C1. Mismatch between graph topology and se~~

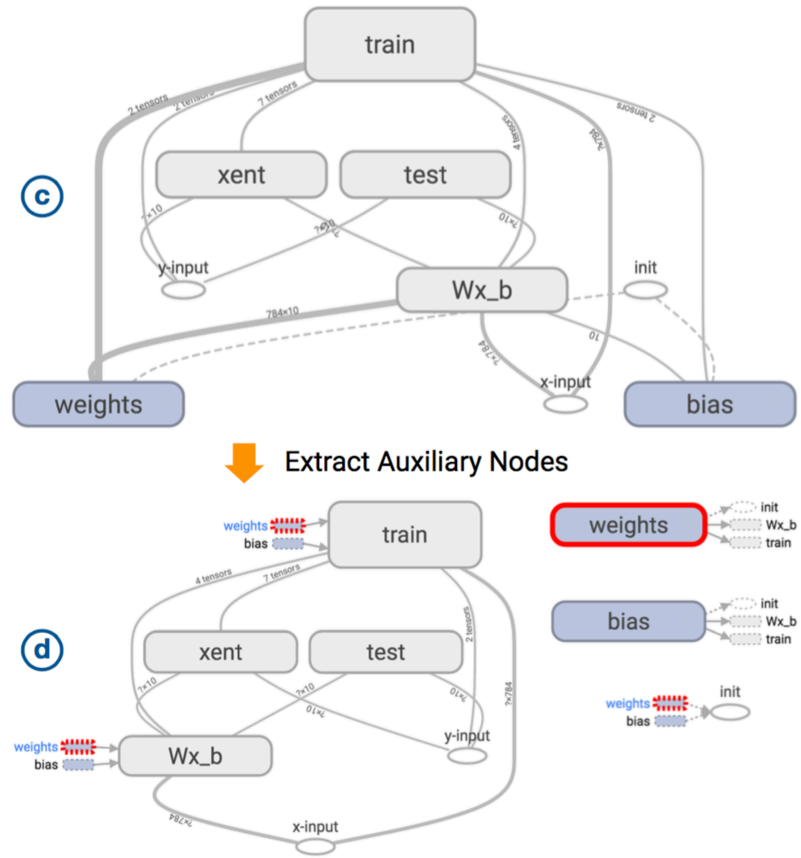
A group of operations ⇔ A component?

~~C2. Graph heterogeneity~~

Different importance: inference > gradients/optim

C3. Interconnected Nodes

Connections between important nodes and less



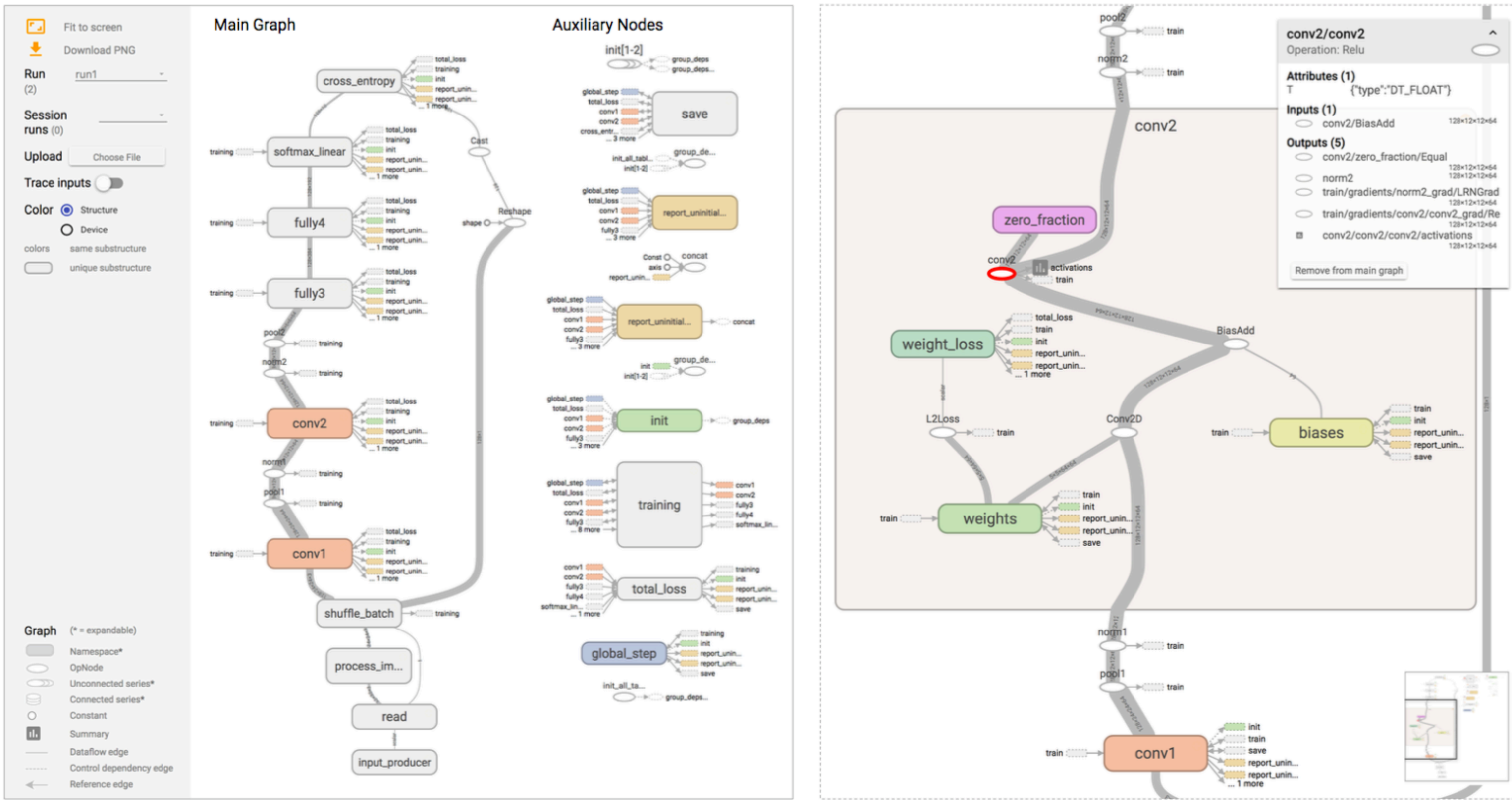
Extract auxiliary nodes from the graph (**C3**)

Data Flow Graph (TensorBoard). Wongsuphasawat et al. 2017

HKUST

Visualization for Model Development

Architecture: How to explain the computation of a model?



(a) (b)

#Global

Data Flow Graph (TensorBoard). Wongsuphasawat et al. 2017

Others: ActiVis (Facebook). Kahng et al. 2017

What are the problems?

Vis for Exploratory Data Analysis

- ~~What does my dataset look like? Any mislabels?~~

Vis for Model Development

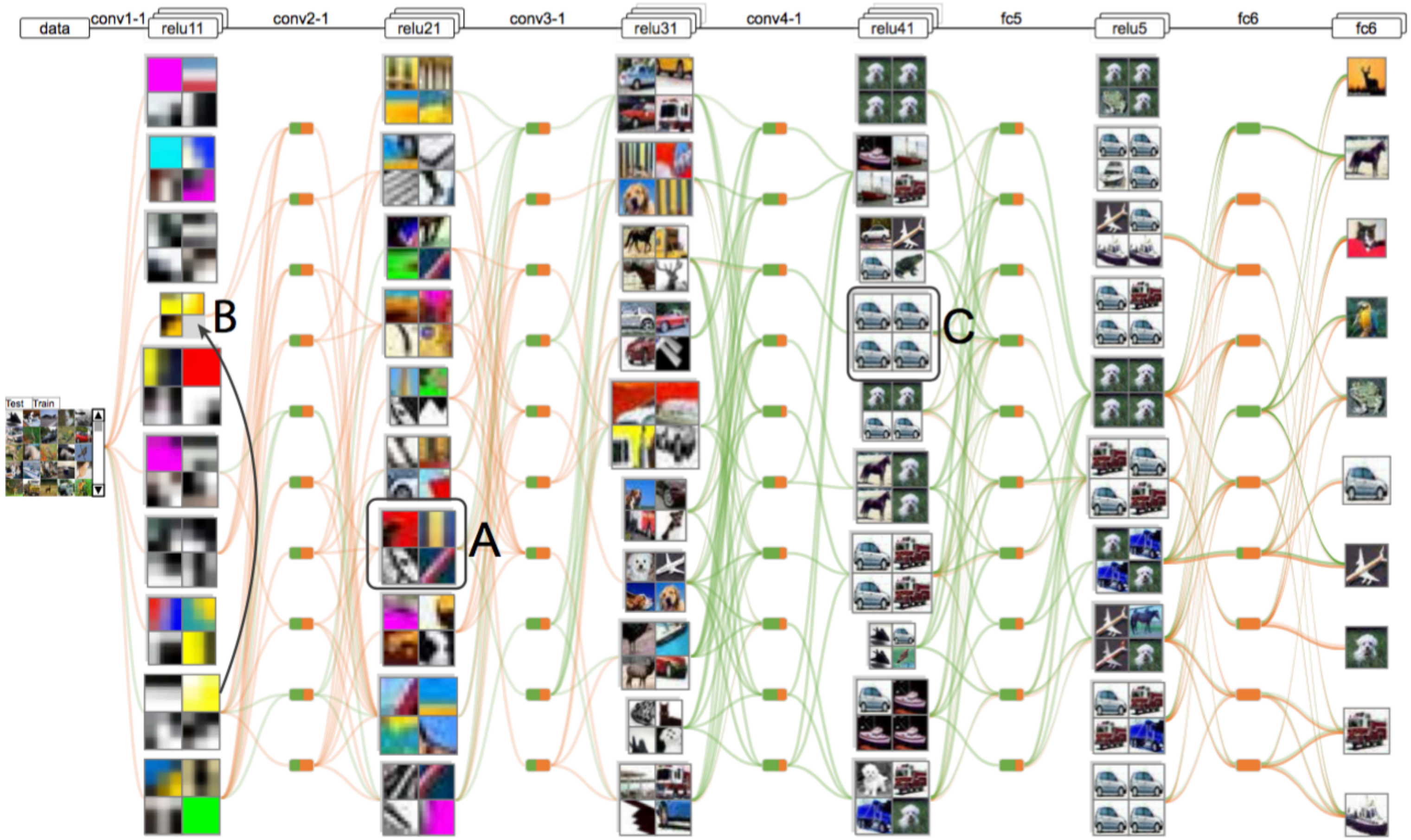
- ~~Architecture: What is the classifier? How to compute?~~
- **Training:** How the model gradually improves? **How to diagnose?**
- Evaluation: What has the model learned from the data?
- Comparison: Which classifier should I choose?

Vis for Operation

- Deploy: How to establish users' trust?
- Operation: How to identify possible failure?

Visualization for Model Development

Training: Why the training fails? Analyzing CNN snapshots



#Global, #Model-aware (f, θ)

CNNVis. Liu et al. 2016

HKUST

Visualization for Model Development

Training: Why the training fails? Analyzing snapshots

Setting:

4 conv layer

2 fully connected layer

RELU activation

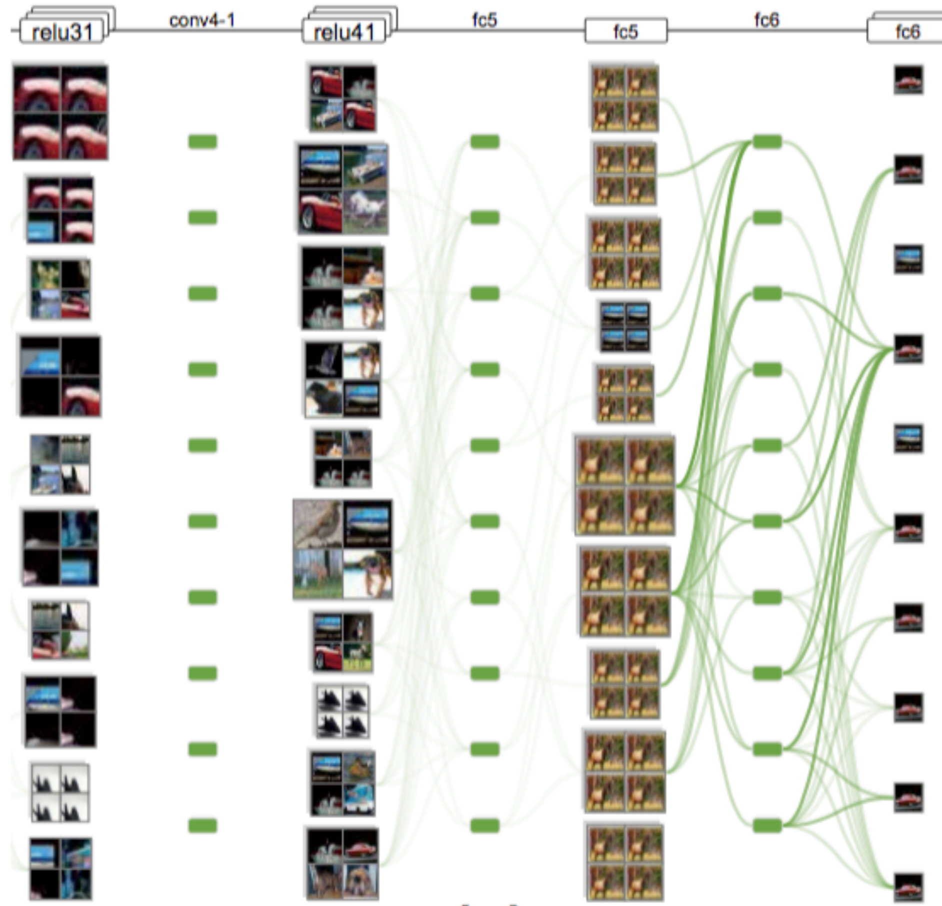
Identity output: $f(x) = x$

Hinge loss: $l(\hat{y}, y) = \max(0, 1 - \hat{y}y)$

$\hat{y}$: output, y: label, ± 1

Cifar-10 dataset

Loss stuck at around 2.0

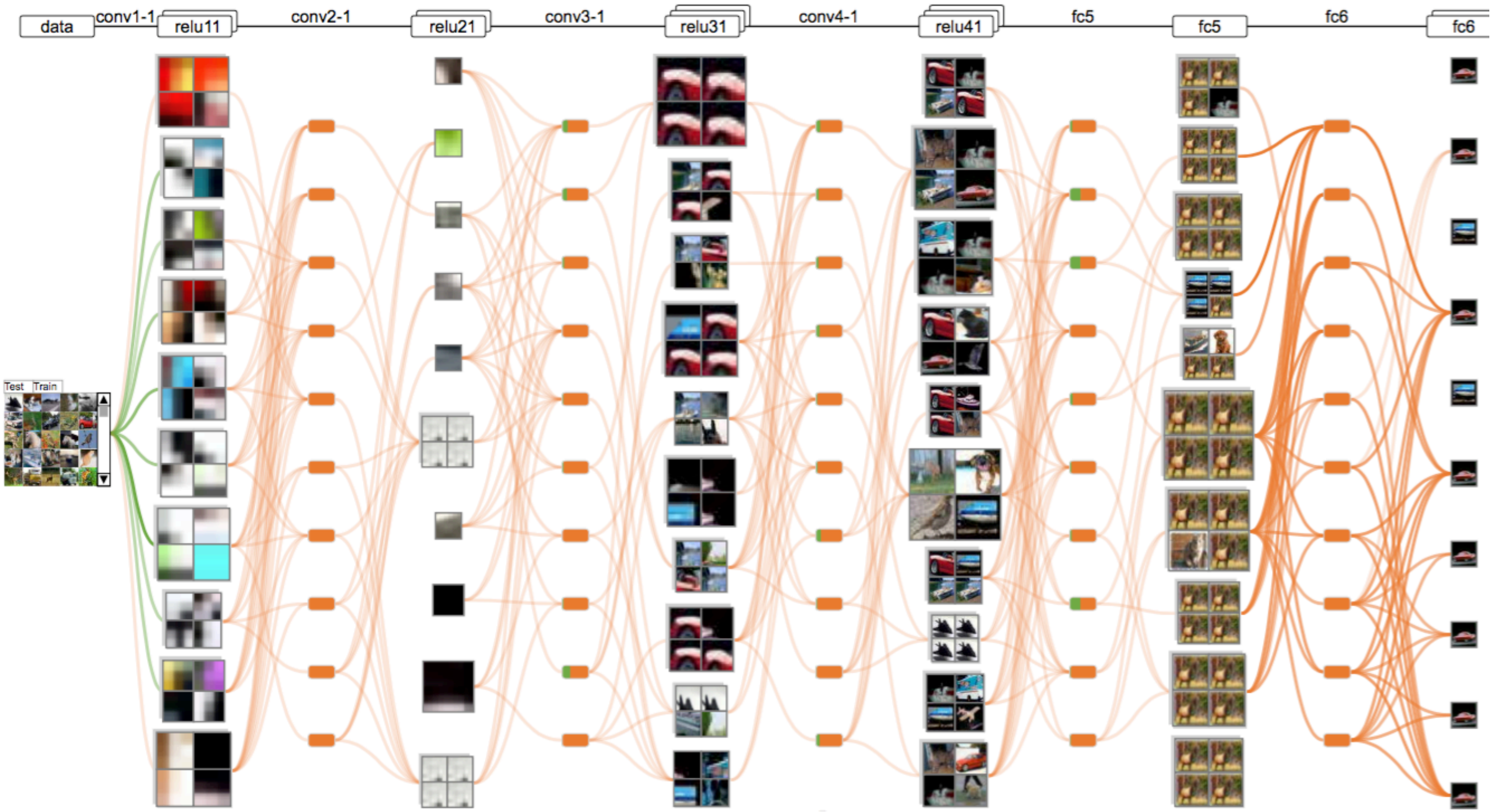


Color: $\Delta\theta$--> 0

CNNVis. Liu et al. 2017

Visualization for Model Development

Training: Why the training fails? Analyzing snapshots

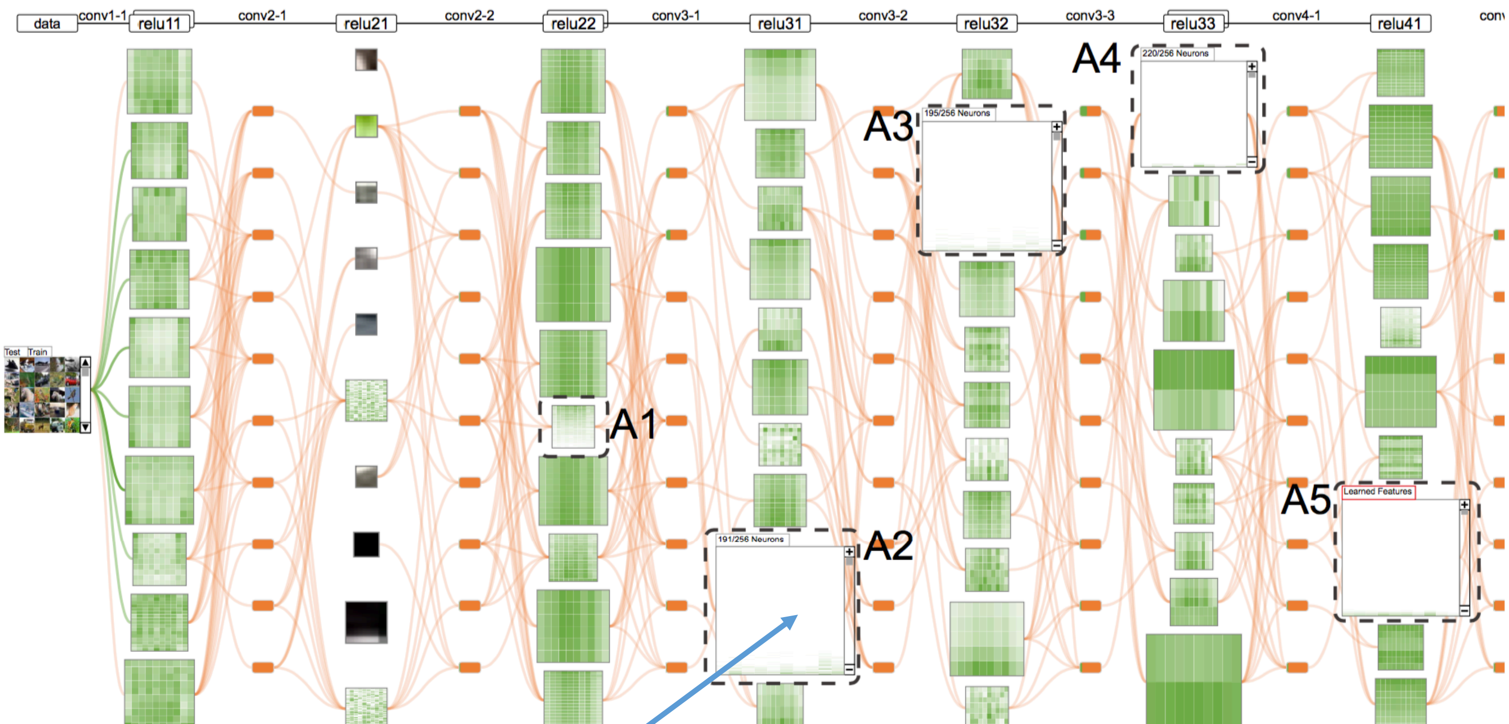


Color: θ (all negative)

CNNVis. Liu et al. 2017

Visualization for Model Development

Training: Why the training fails? Analyzing snapshots



Activation Ratio --> 0

Color: θ (all negative)

Explain: Negative weights
$\Rightarrow$ Negative outputs
$\Rightarrow$ Zero activations (RELU)

Solution:
Add batch-norm to
force non-negative

CNNVis. Liu et al. 2017

What are the problems?

Vis for Exploratory Data Analysis

- ~~What does my dataset look like? Any mislabels?~~

Vis for Model Development

- ~~Architecture: What is the classifier? How to compute?~~
- ~~Training: How the model gradually improves? How to diagnose?~~
- **Evaluation: What has the model learned from the data?**
- Comparison: Which classifier should I choose?

Vis for Operation

- Deploy: How to establish users' trust?
- Operation: How to identify possible failure?

Visualization for Model Development

Evaluation: Do CNN learn class hierarchy?

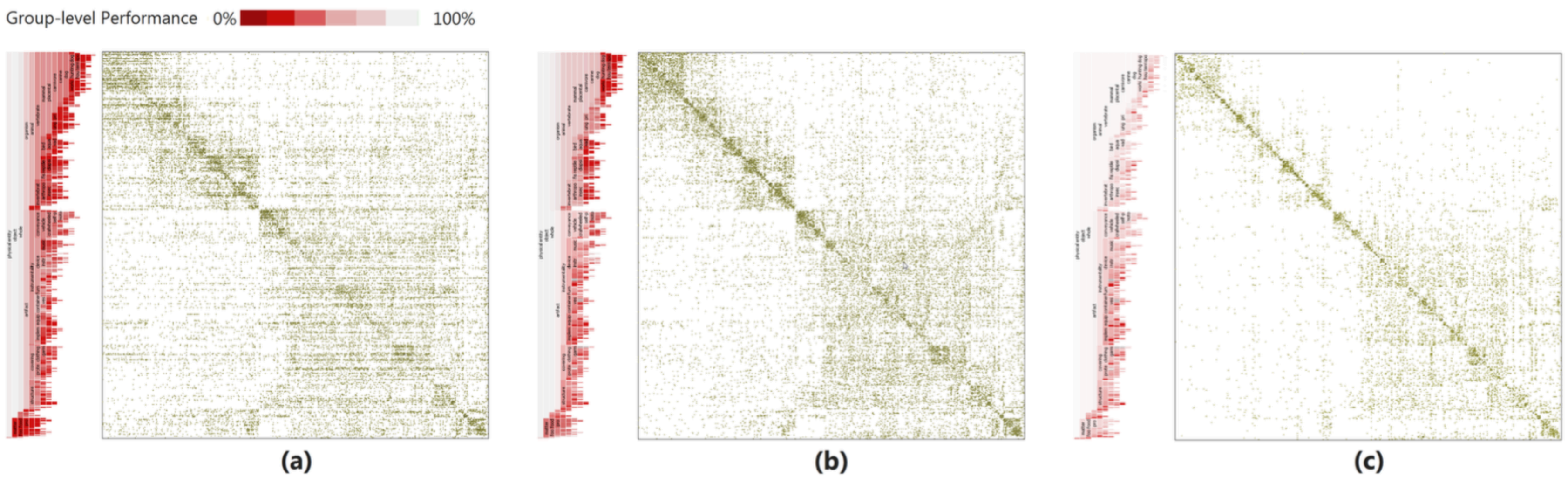
Confusion matrix of the classification results of the ImageNet using GoogleNet

#Global, #Model-unaware

Blocks. Alsallakh et al. 2017

Visualization for Model Development

Evaluation: Do CNN learn class hierarchy?

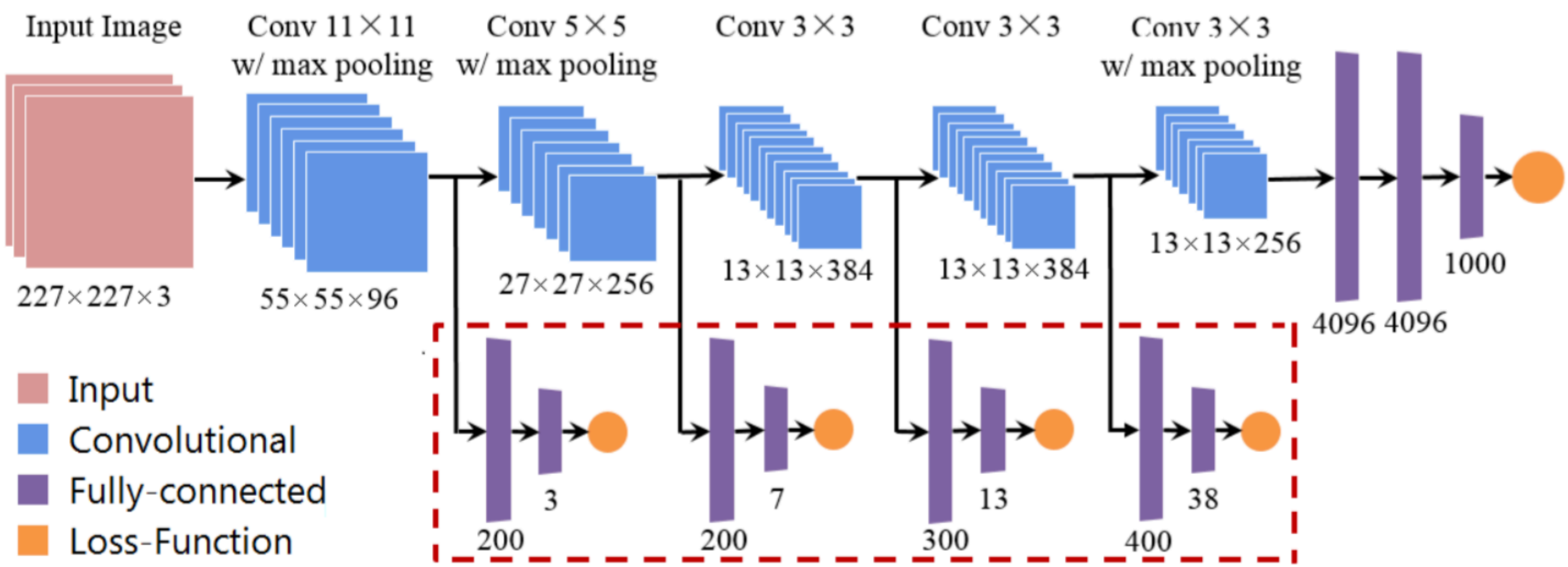


The confusion matrix after the first epoch (a), the second epoch (b), and the final epoch (c) during the training of AlexNet.
The network starts to distinguish high-level groups already after the first epoch.

Blocks. Alsallakh et al. 2017

Visualization for Model Development

Evaluation: Do CNN learn class hierarchy?



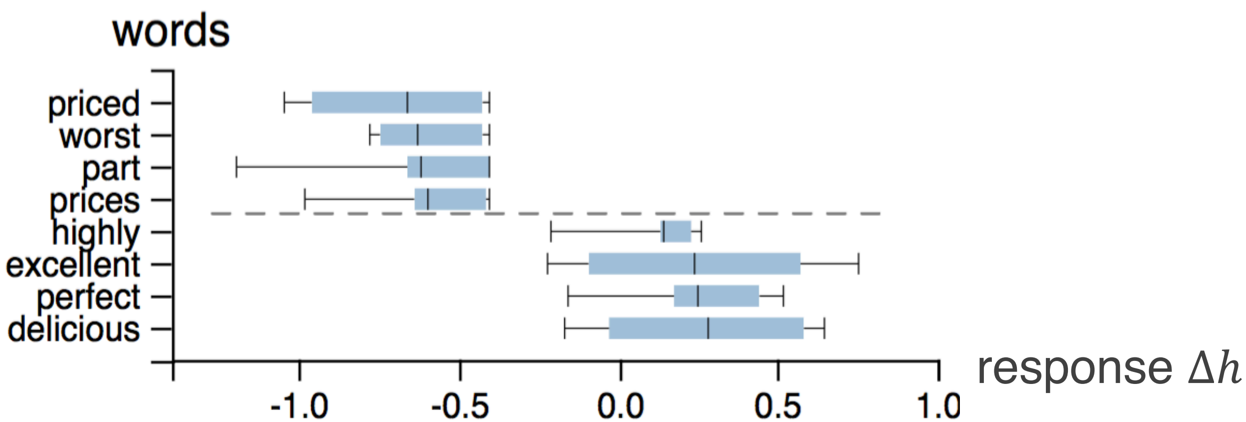
Explicitly add hierarchy loss between layers.

Architecture	Top-1 error	Top-5 error
Standard AlexNet	42.6%	19.6%
Hierarchy-Aware AlexNet	**34.33%**	**13.02%**

Blocks. Alsallakh et al. 2017

Visualization for Model Development

Evaluation: What has an RNN learned from the data?



Unit: #36

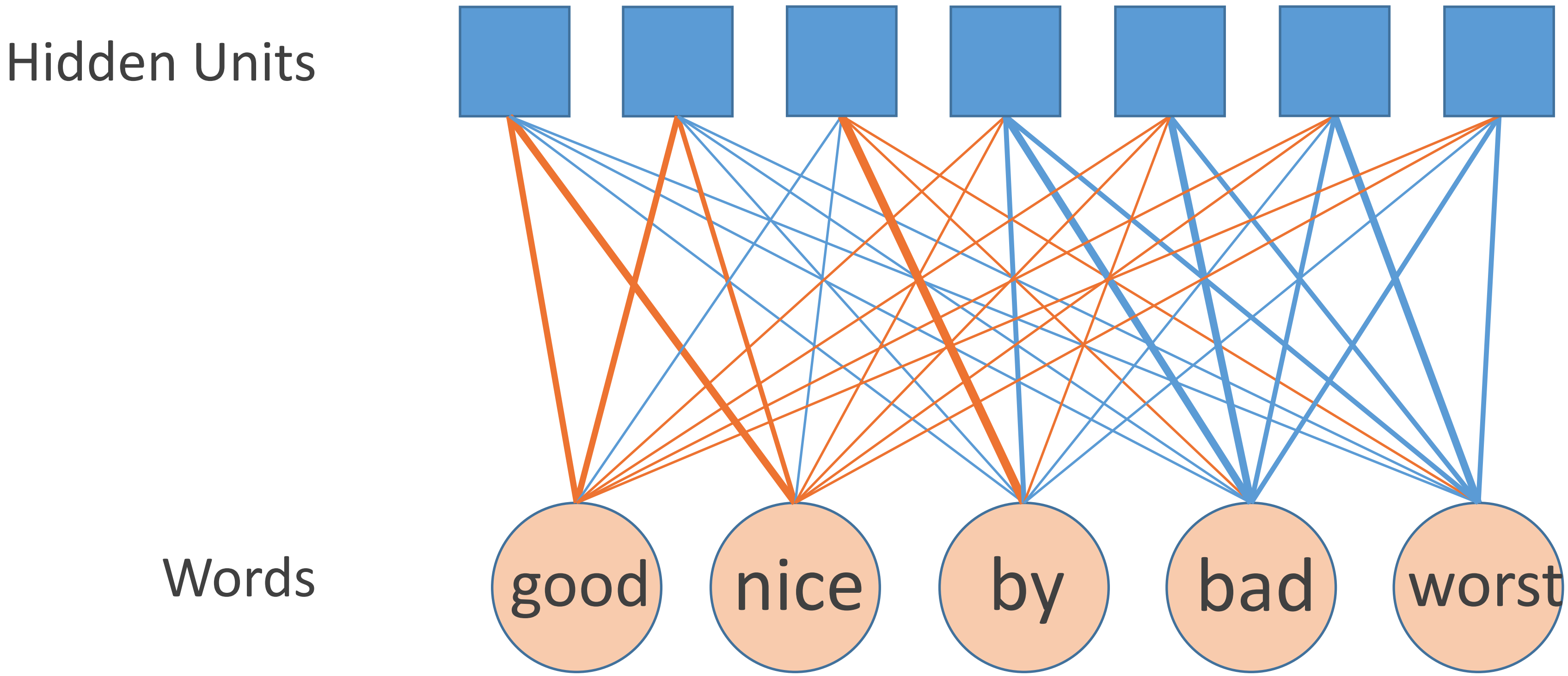
Top 4 positive/negative salient words of unit 36 in
an RNN (GRU) trained on Yelp review data.

600 units in h ! Investigate one at a time is too difficult!

RNNVis: Ming et al. 2017

Visualization for Model Development

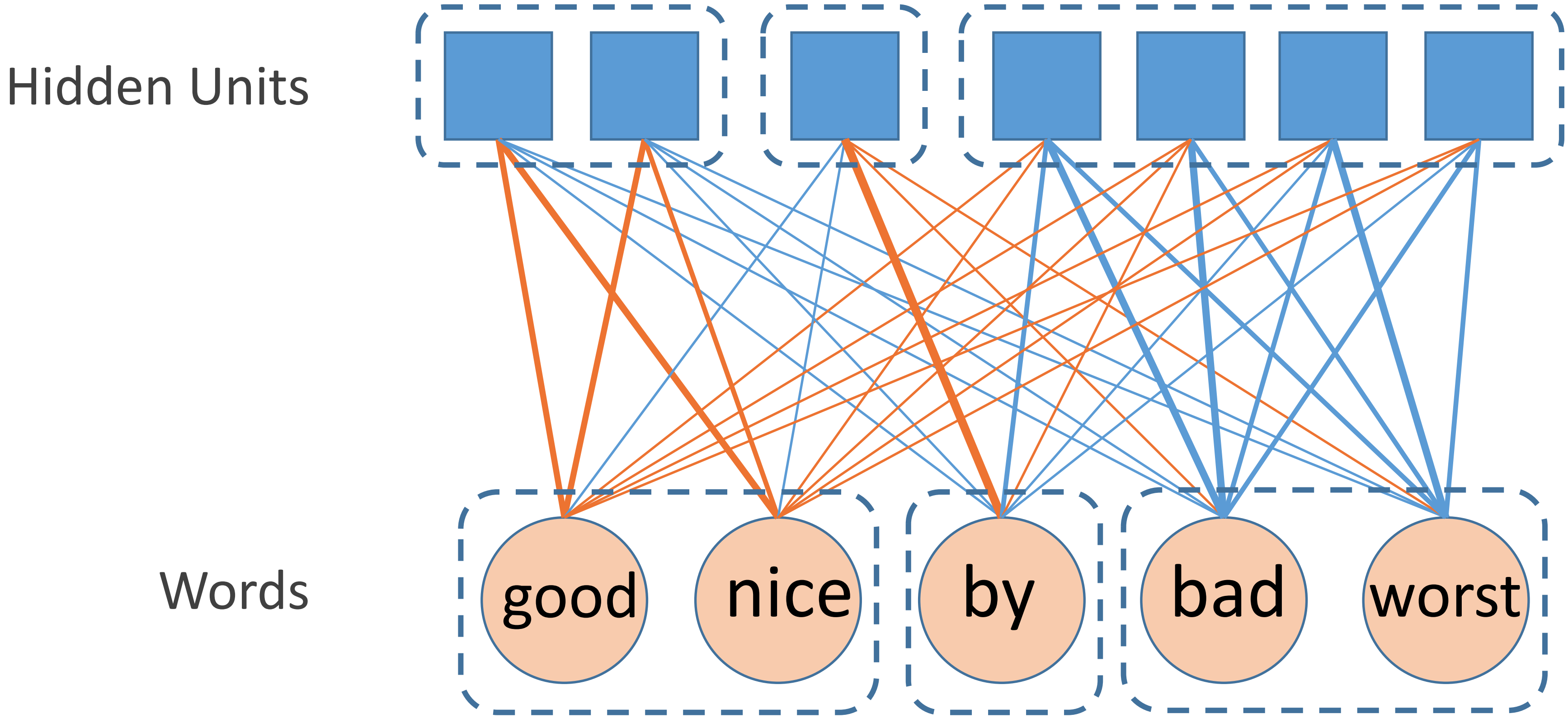
Evaluation: What has an RNN learned from the data?



RNNVis: Ming et al. 2017

Visualization for Model Development

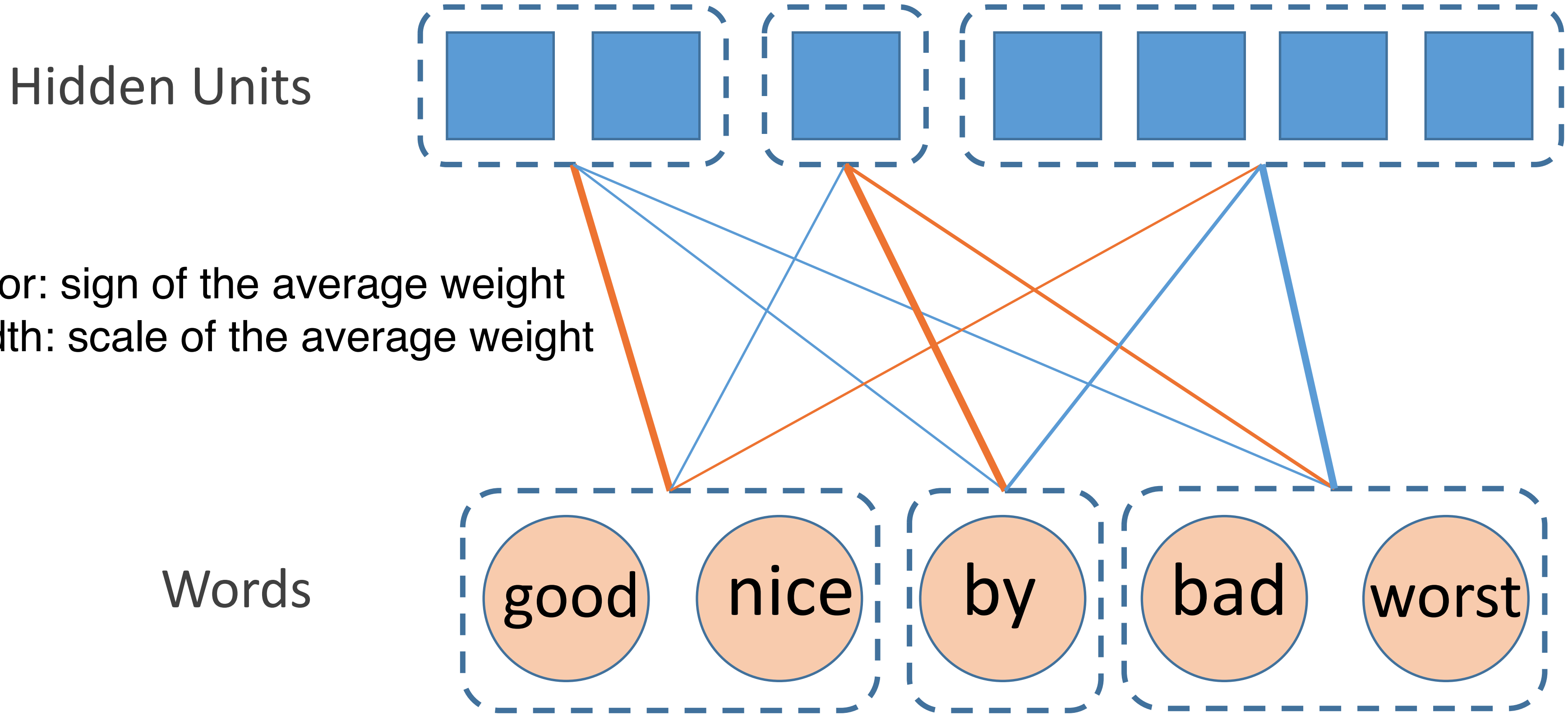
Evaluation: What has an RNN learned from the data?



RNNVis: Ming et al. 2017

Visualization for Model Development

Evaluation: What has an RNN learned from the data?



Color: sign of the average weight
Width: scale of the average weight

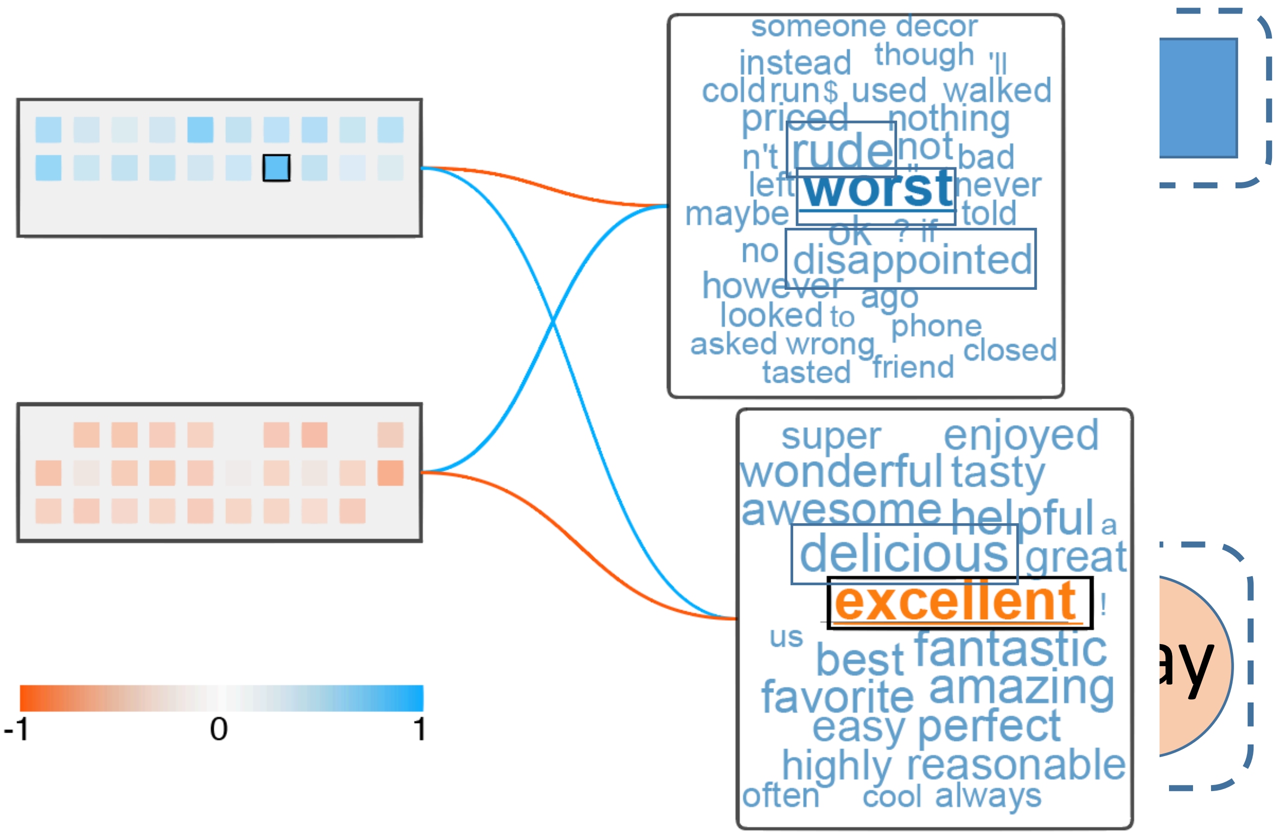
RNNVis: Ming et al. 2017

Visualization for Model Development

Evaluation: What has an RNN learned from the data?

Hidden Units

Words



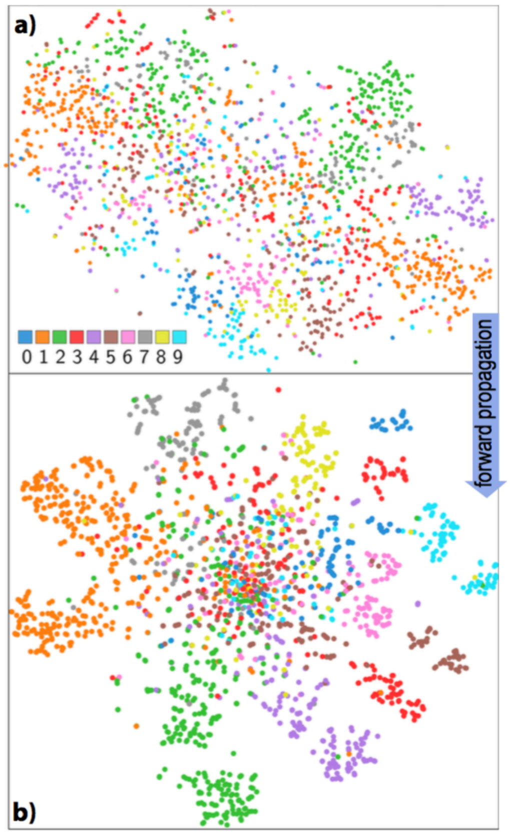
Hidden Units Clusters
(Memory Chips)

Words Clusters
(Word Clouds)

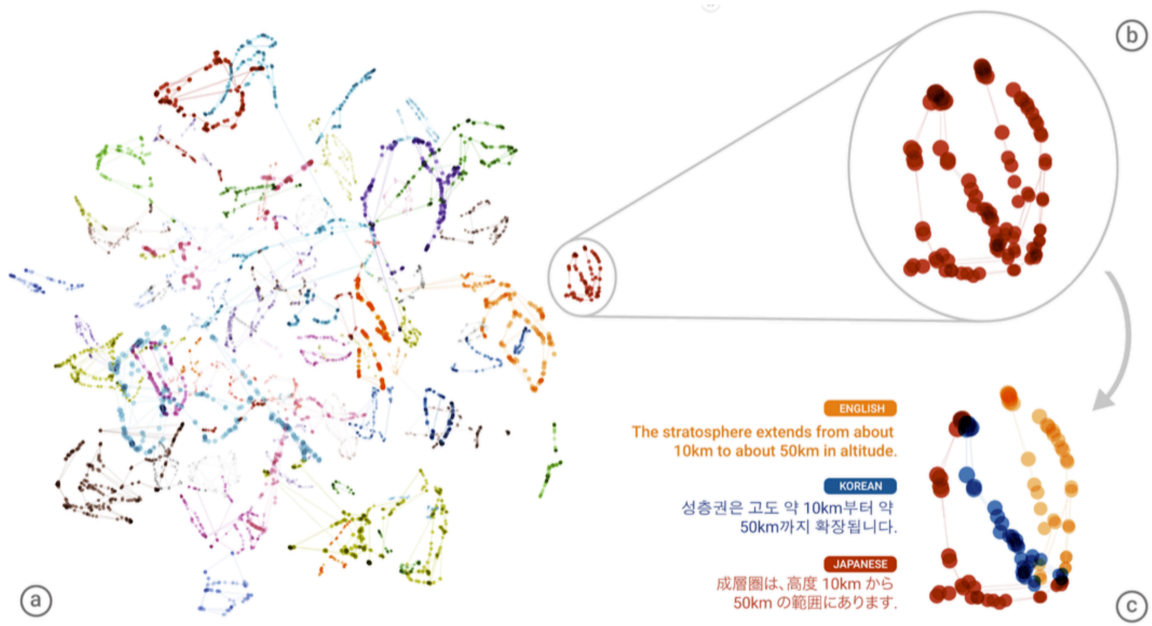
RNNVis: Ming et al. 2017

Visualization for Model Development

Understanding - Others (Embedding Projection)



Embedding projection
SVHN test set.
Rauber et al. 2017

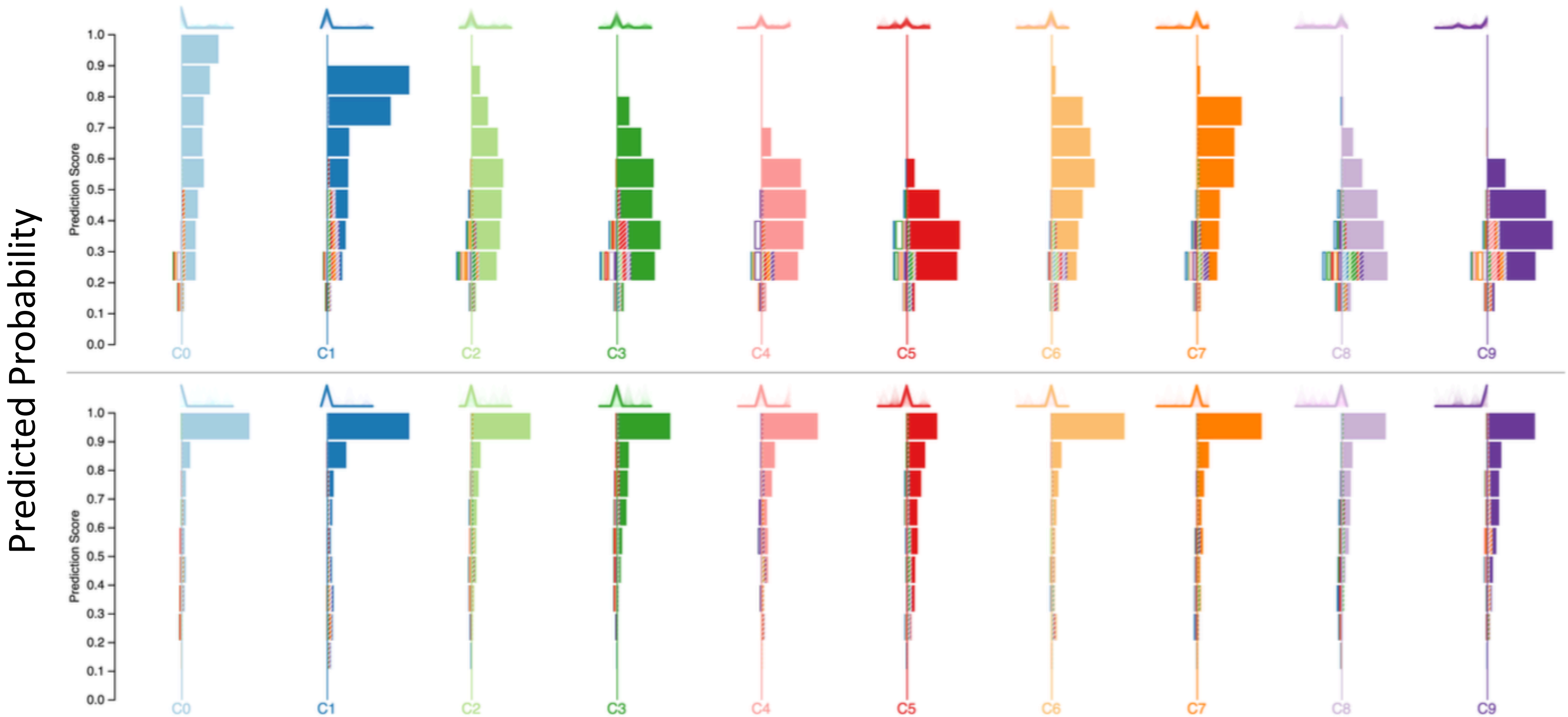


Multilingual translation model
t-SNE projection
Each node is a word
Johnson et al. 2016

#Global, #Model-unaware (f)

Visualization for Model Development

Assessment & Comparison



Histograms of predicted probability of instances of each class. Top: RF. Bottom: SVM. Acc: 0.87 (solid: TP, dashed-left: FP, dashed-right: FN)

Squares (Microsoft). Ren et al. 2017

Others: ModelTracker. Amershi et al. 2015

#Global, #Model-unaware (summarizing y)

Visualization for Model Development

Comments

Scalability

- Most only tested for small datasets like MNIST

How to evaluate understanding?

- Most use expert reviews

Is it possible to qualitatively evaluate fairness (non-discrimination) and robustness of classifiers?

What are the problems?

Vis for Exploratory Data Analysis

- ~~What does my dataset look like? Any mislabels?~~

Vis for Model Development

- ~~Architecture: What is the classifier? How to compute?~~
- ~~Training: How the model gradually improves? How to diagnose?~~
- ~~Evaluation: What has the model learned from the data?~~
- Comparison: Which classifier should I choose?

Vis for Operation

- Deploy: How to establish users' trust?
- Operation: How to identify possible failure?

Visualization for Operation

Deploy: How to establish users' trust?

- If users don't trust the model, they will not use it! (Lieberman 1998)
- Trust is based on experience.
- Interaction boost trust. (Stumpf 2007)

Operation: How to cope with possible failure?

- Human taking over in case of failure
- Identify failure for safety-critical applications
- Better user experience

Few studies in this part

Conclusion

Theory

- Rigorous theory (cognition+CS) of explainability and explanation
- Proper evaluation of explainability and the quality an explanation
- How to model the bias and variance of human

Application

- Real-world applications for end-users
- Design guidelines
- Human learn from AI?