Visual Analytics for Explainable Deep Learning

By Jaegul Choo and Shixia Liu

Presented by:
Ameya Samak,
Bhavishyya Muppalaneni,
Neeraj Bandi

Introduction

What is explainable deep learning?

- It is an emerging field of research that seeks to make the decision-making process of deep learning models more transparent and interpretable.
- Goal is to increase trust and transparency in deep learning models, making it easier for users to understand and trust the predictions and decisions made by these models.
- Particularly important in the fields such as precision medicine, law enforcement and financial investment.

Deep learning models are often considered as "black boxes" because their internal processes and decision-making mechanisms are not easily understandable to humans.

Why explainable deep learning?

- Gender and racial biases learnt by artificial intelligence programs recently emerged as a serious issue.

- In April 2016, the European Union legislated the human right to request an explanation regarding machine-generated decisions.

- For instance, Defense Advanced Research Projects Agency (DARPA) of the United States launched a large initiative called Explainable Artificial Intelligence (XAI).

3 main research directions:

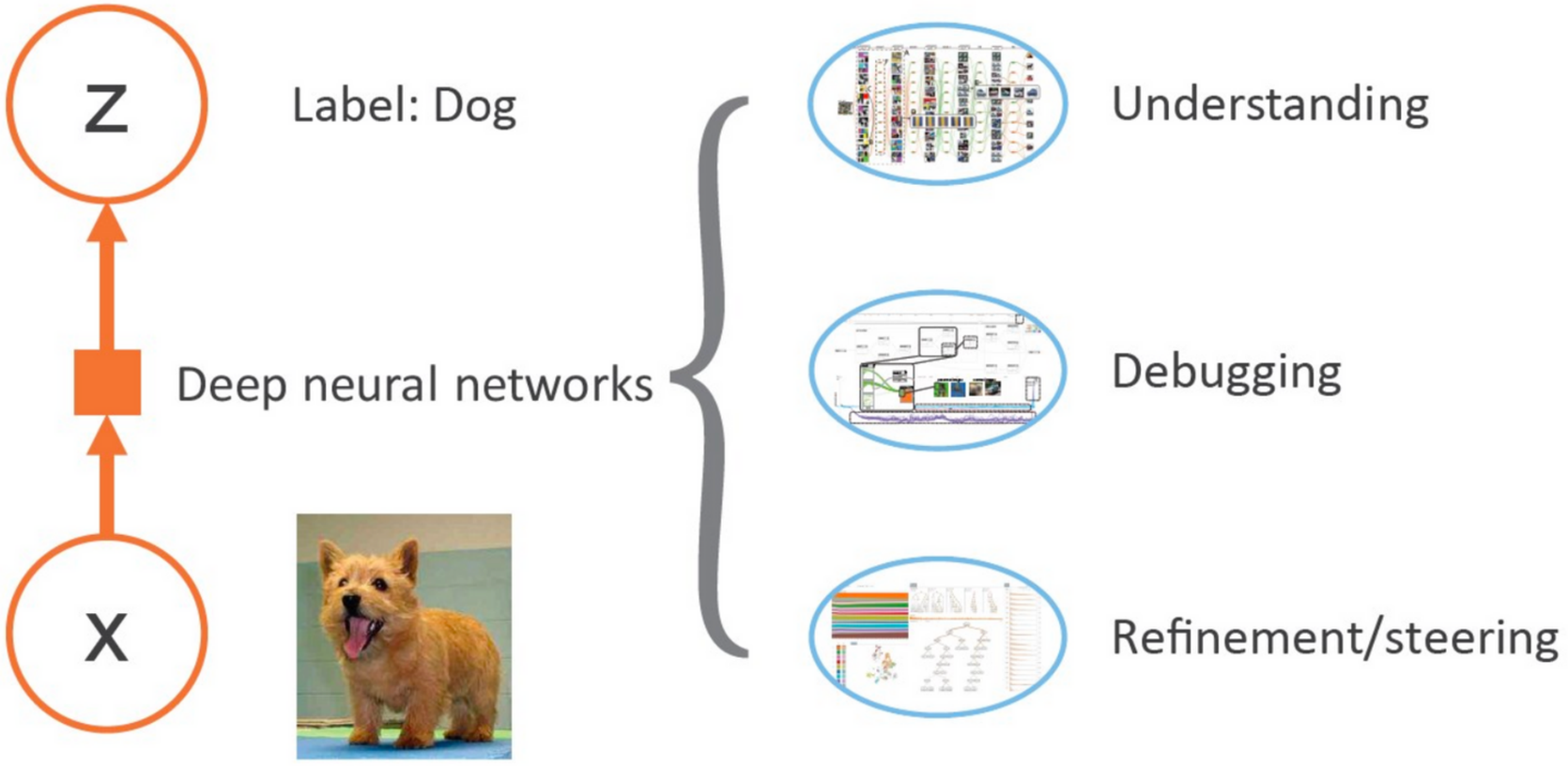


Fig. Overview of explainable deep learning

Visual Analytics in Deep Learning | Interrogative Survey Overview

§4 WHY

Why would one want to use visualization in deep learning?

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

§6 WHAT

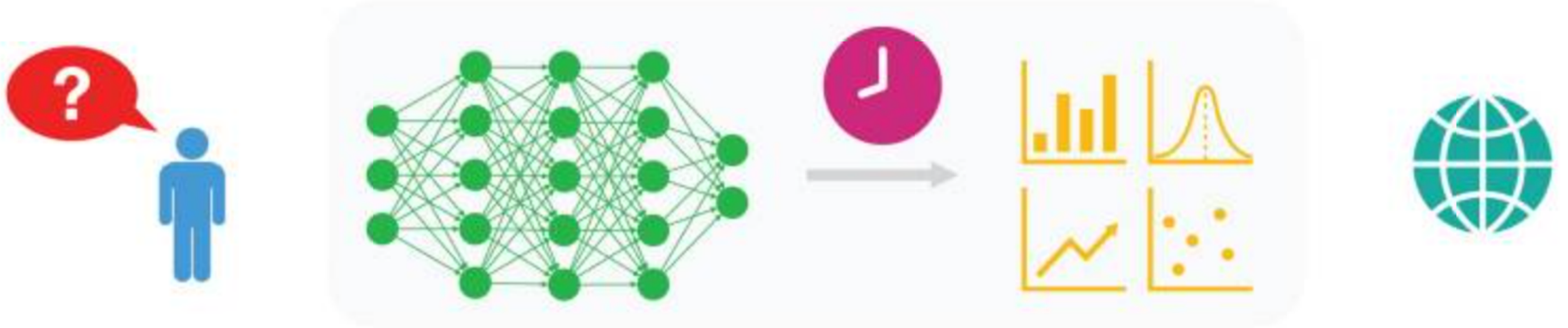
What data, features, and relationships in deep learning can be visualized?

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

§8 WHEN

When in the deep learning process is visualization used?

During Training
After Training



§5 WHO

Who would use and benefit from visualizing deep learning?

Model Developers & Builders
Model Users
Non-experts

§7 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

§9 WHERE

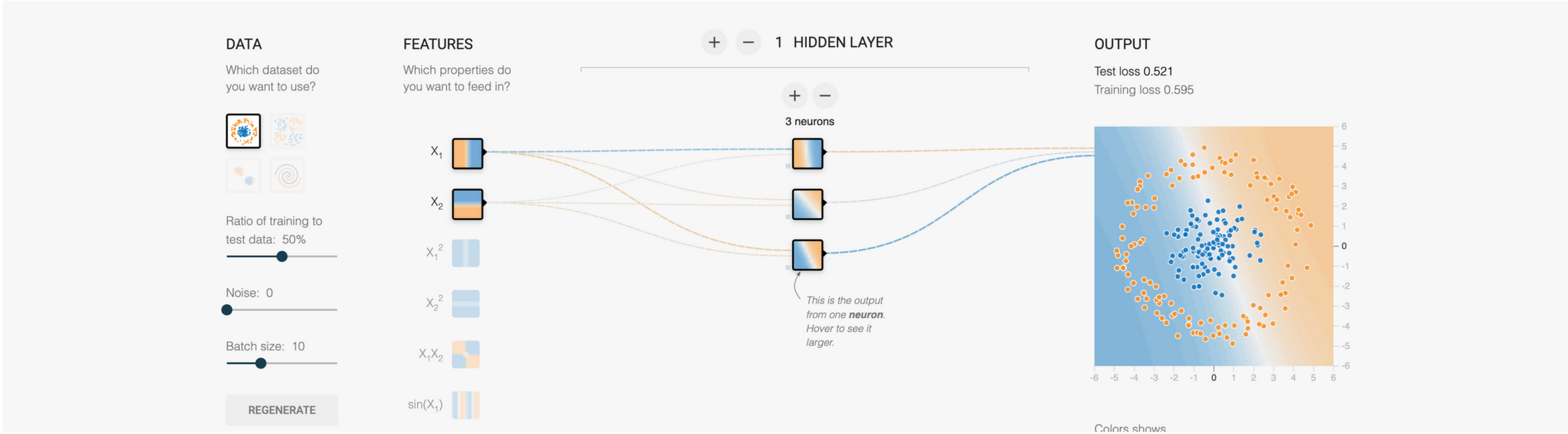
Where has deep learning visualization been used?

Application Domains & Models
A Vibrant Research Community

INTUITIVE UNDERSTANDING WITH INTERACTIVE VISUALIZATION

- Tensorflow Playground is an interactive and educational tool that allows users to experiment with neural network configurations and visualize the activation of nodes using two-dimensional data sets.

- ConvNetJS is a web-based deep learning library with rich visualization modules that make deep learning accessible via a web browser, while DeepVis toolbox can dynamically visualize activation maps of filters in user-selected layers of CNNs in real-time webcam videos.

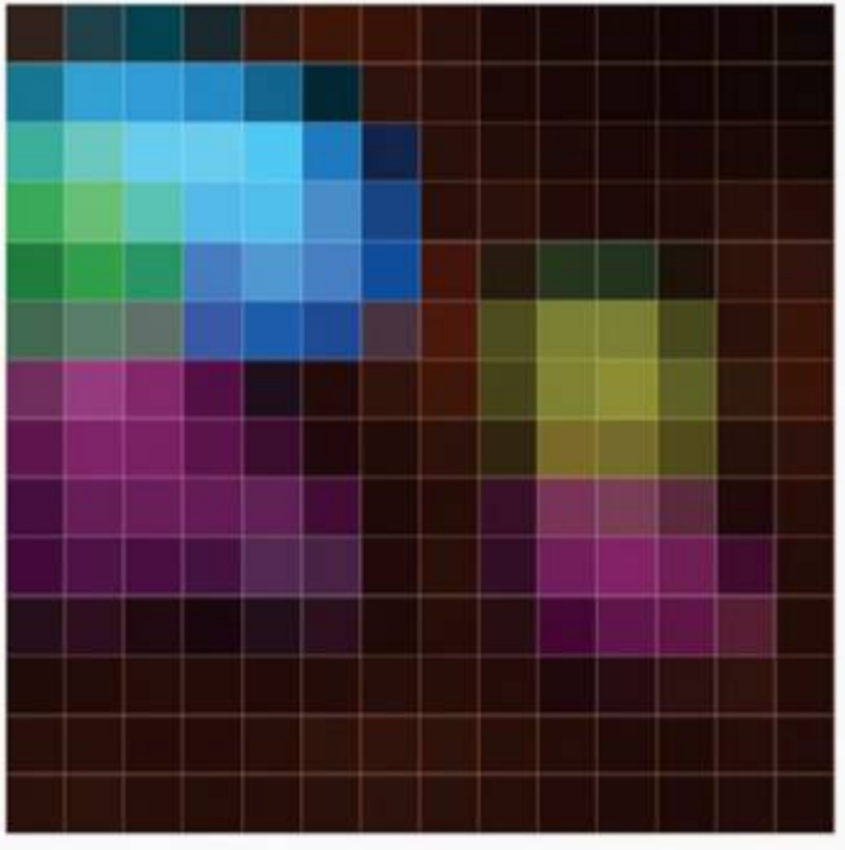


By using non-negative matrix factorization we can reduce the large number of neurons to a small set of groups that concisely summarize the story of the network.

REPRODUCE IN A NOTEBOOK

INPUT IMAGE

ACTIVATIONS of neuron groups



NEURON GROUPS based on matrix factorization of mixed4d layer

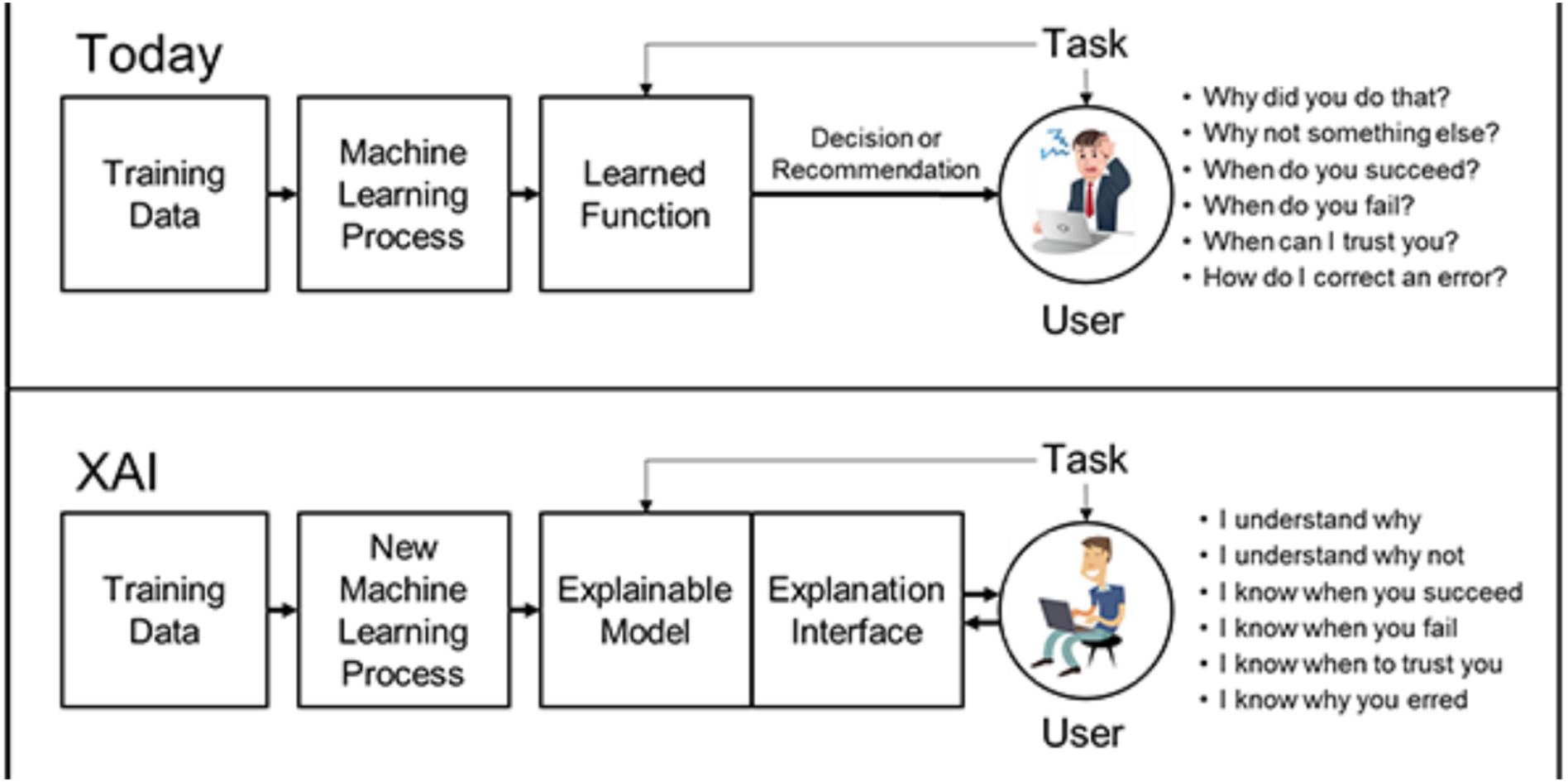
6 groups

color key

feature visualization of each group

hover to isolate →

The XAI program aims to tackle challenge problems in two areas:
(1) developing machine learning systems for classifying events of interest in multimedia data, and
(2) constructing decision policies for autonomous systems to perform various simulated missions.



MODEL DEBUGGING THROUGH VISUALIZATION TOOLKITS

- Deep learning libraries typically come with rudimentary visualization toolkits, which can assist users in debugging their models and optimizing their performance.
- The new Embedding Projector module in TensorFlow enables 2D/3D embedding views using principal component analysis and t-distributed stochastic neighbor embedding, unveiling relationships between multi-dimensional representations of data points in a layer.
- Visdom is a user-friendly, web-based interactive visualization toolkit that seamlessly integrates with deep learning libraries like PyTorch.
- Deeplearning4j UI is a visual user interface that provides basic visualization components to help users monitor the training process. It provides user interface to visualize in your browser (in real time) the current network status and progress of training. The UI is typically used to help with tuning neural networks - i.e., the selection of hyperparameters (such as learning rate) to obtain good performance for a network.

DATA

5 tensors found
Word2Vec 10K

Label by
word

Color by
No color map

Edit by
word

Tag selection as

Load | Publish | Download | Label

Sphereize data

Checkpoint: Demo datasets

Metadata: oss_data/word2vec_10000_200d_labels.tsv

UMAP | T-SNE | PCA | CUSTOM

X
Component #1

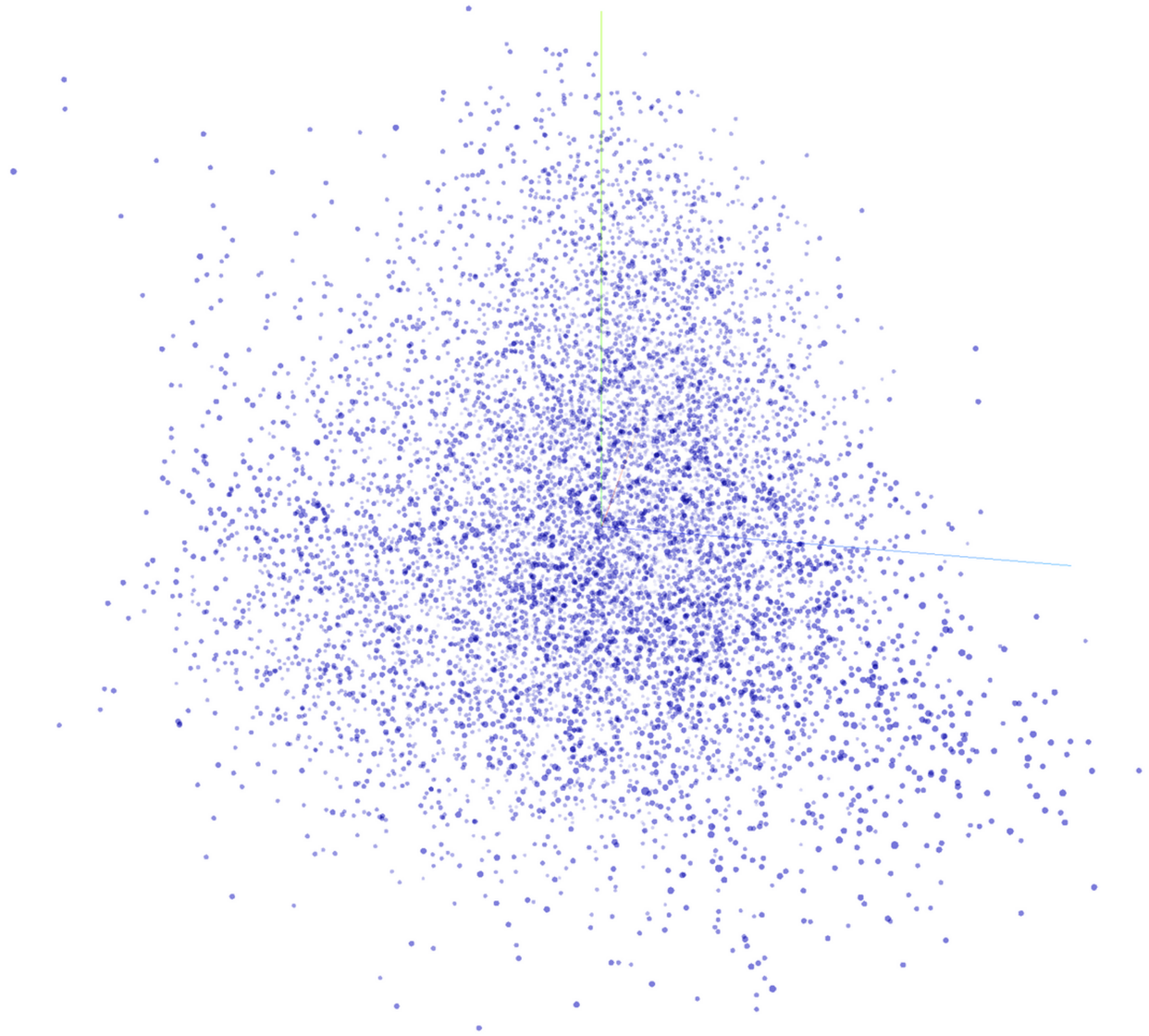
Y
Component #2

Z
Component #3

PCA is approximate.

Total variance described: 8.5%.

Points: 10000 | Dimension: 200



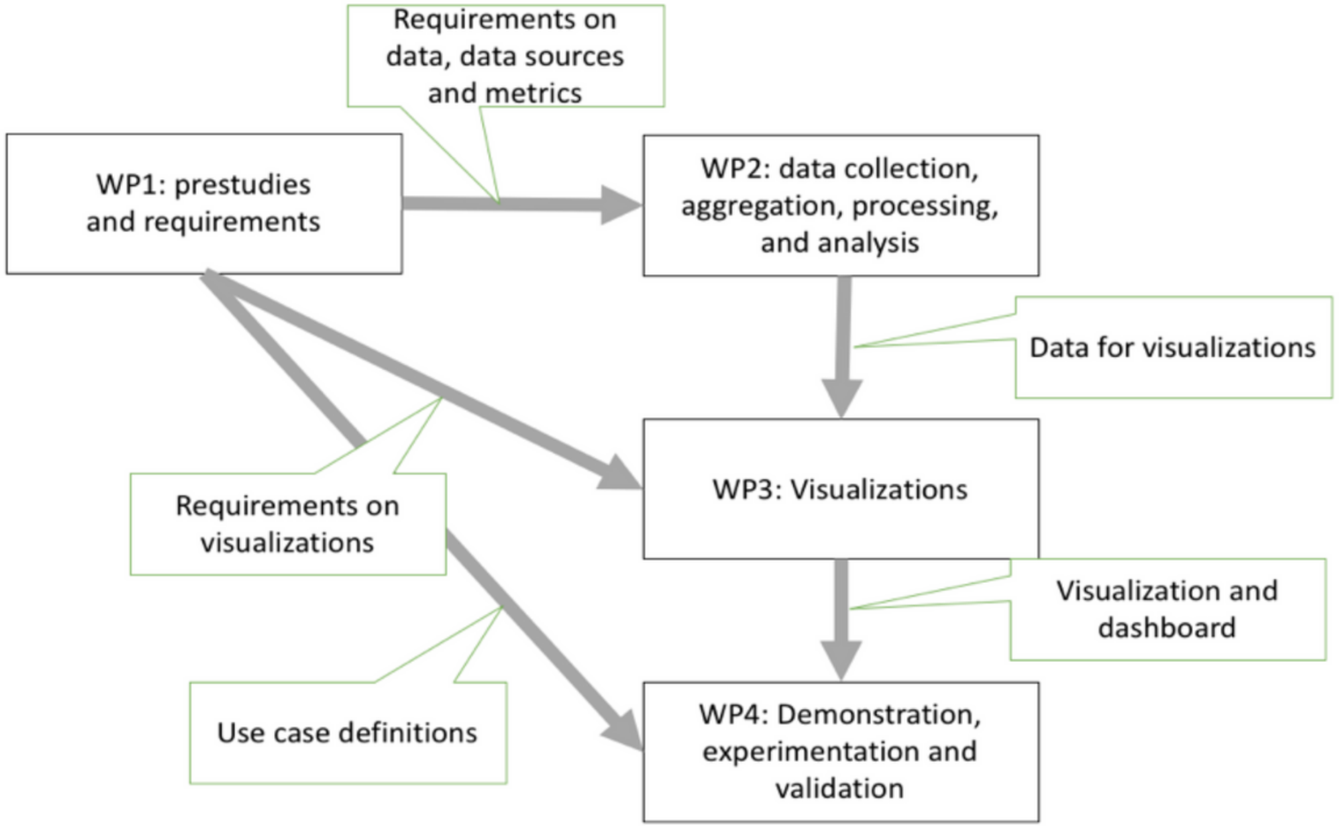
Show All Data | Isolate selection | Clear selection

Search .*

by
word

BOOKMARKS (0)

Workpackages



Technical workpackages

WP1: Prestudies and requirements

WP2: Data collection, aggregation, processing and analysis

WP3: Visualizations

WP4: Demonstration, experimentation and validation

Supportive workpackages

WP5: Exploitation and dissemination

WP6: Project management

COMPUTATIONAL METHODS FOR INTERPRETATION AND EXPLANATION

- Interpreting a deep learning model usually involves determining the feature importance score, such as identifying which part of an input feature in a data item is responsible for a prediction result in the output layer or the high activation of an internal node/layer.

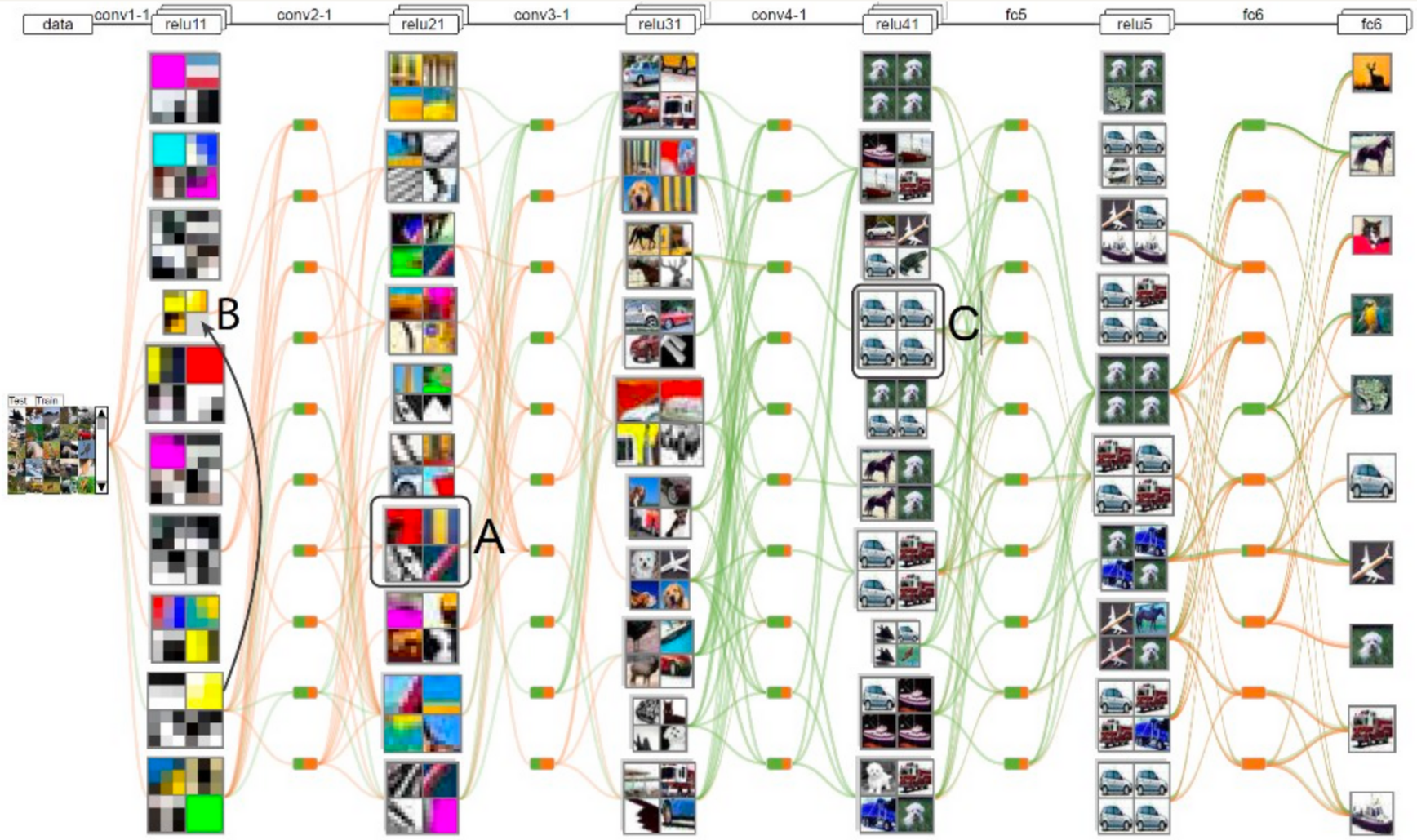
- One such computational method is both perturbation experiments and saliency map-based methods have demonstrated their effectiveness in identifying the part of an input image that has the most impact on a model's final prediction.

- The LIME (Local Interpretable Model-agnostic Explanations)technique, which is a new method, creates an approximately locally linear model from a complex model, enabling linear combination weights to be interpreted as the feature importance score.

- The training data items most responsible for a particular prediction output have been identified in an efficient manner by adopting a classical technique called influence functions.

VISUAL ANALYTICS FOR IN-DEPTH UNDERSTANDING AND MODEL REFINEMENT

- CNNVis is an example of a visual analytics system that helps with the understanding and diagnosis of CNNs by exploring the relationship between the depth and width of neural networks, learned representations, and the performance of model classification through a DAG(Directed Acylic Graph) layout.



- ActiVis is a tool that offers a visual exploratory analysis of a deep learning model through various coordinated views, including a matrix view and an embedding view.

- To help understand the semantic significance of cells in language modeling applications, LSTMVis displays the activation patterns of individual cells over time steps or sequences as line graphs.

- RNNV is a visualization tool that is used to analyze Recurrent Neural Networks (RNNs) by clustering hidden-state nodes that have similar activation patterns. RNNVis then creates a grid-style heatmap of the clustered nodes along with their most strongly associated keywords.

- ReVACNN is a visual analytics system designed for Convolutional Neural Networks (CNNs) that offers real-time model steering capabilities during the training process. It allows users to adjust and optimize the CNN model while it is training, which can improve its performance. This is achieved through interactive visualization tools that enable users to explore and manipulate the CNN's architecture and data inputs in real-time.

- One of the key features of DeepEyes is its ability to allow users to interactively steer and monitor deep learning models in real-time. This can help users to identify and address issues with their models as they arise, which can improve their overall performance.

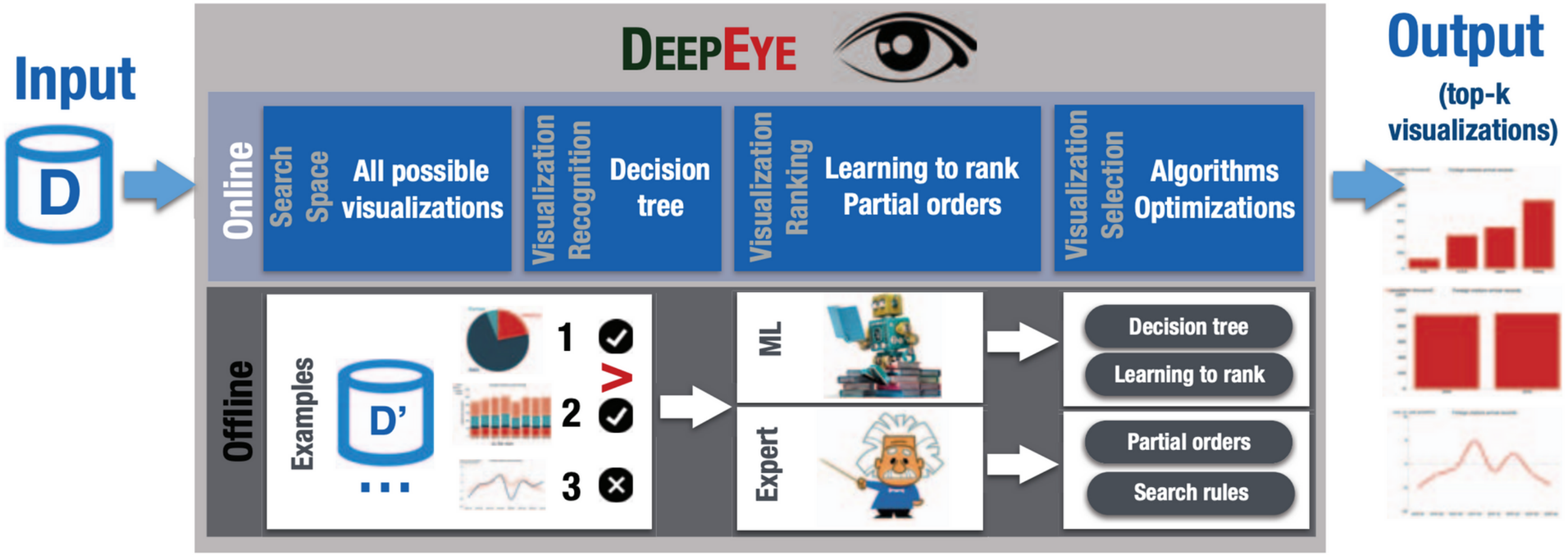


Figure 4. Overview of DEEPEYE

Research Gaps and Opportunities

Opp 1. Injecting external human knowledge

- Currently, most deep learning models constitute data-driven methods, whereas knowledge- driven perspectives have received comparatively little attention.

- An open research opportunity is to combine human expert knowledge and deep learning techniques through interactive visualization.

- The topics include:

1. domain knowledge representation and interpretation
2. expert knowledge propagation
3. knowledge-based visual explanation

- Visual analytics could be utilized to intuitively verify that a model correctly follows human-injected knowledge and rules.

Example: Medical diagnosis

- Let's say we are developing a deep learning model to predict whether a patient with certain symptoms is likely to have a particular disease.

Domain knowledge representation and interpretation

- Incorporating expert knowledge through knowledge graph representation can guide the decision-making process.
- Example: "fever" and "cough" might be linked to "pneumonia" and "flu"

Expert knowledge propagation

- A set of expert rules to dictate how certain symptoms should be interpreted in the context of different diseases.
- For instance, a patient with a fever and cough is more likely to have pneumonia than the flu.

Knowledge-based visual explanation

- Interactive visualizations are used to help experts understand how the model is making its predictions.

Opp 2. Progressive visual analytics of deep learning

- Most of the existing explainable deep learning approaches mainly focus on understanding and analyzing model predictions or the training process offline after the model training is complete.

- As the training of many deep learning models is time-consuming, progressive visual analytics techniques are needed to incorporate experts into the analysis loop.

- To this end, deep learning models are expected to produce semantically meaningful partial results during the training process.

- Experts can then leverage interactive visualizations to explore these partial results, examine newly incoming results, and perform new rounds of exploratory analysis without having to wait for the entire training process to be completed.

Example: Object Detection

- Let's say we are training a deep learning model to detect objects in images, such as cars, pedestrians, and traffic signs.

- During the training process, we could use progressive visual analytics techniques to monitor the model's performance on a subset of the training data, displaying semantically meaningful partial results such as the number of objects detected and their locations in the image

- Experts could then use interactive visualizations to explore the data and perform new rounds of analysis, such as examining the types of images where the model is struggling and adjusting the training parameters accordingly.

Opp 3. User-driven generative models

- Conventional deep learning techniques limit human interaction by offering a single response for each data point.

- User-driven generative models enable users to interact with the model and affect the output

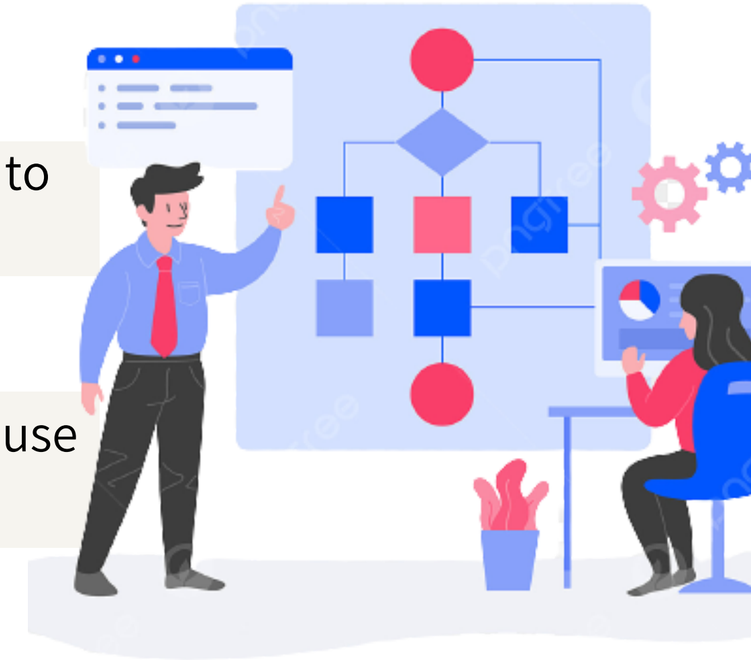
- They have the potential to enable more creativity and expression in deep learning applications, making them more engaging and personalized.

- The interactive nature of generative models can be used in a variety of fields, including language translation and video game design.

- Deep learning models that are simple to include user input are currently being developed by researchers.

User-driven generative models demonstrate the power of human-machine collaboration

Opp 4. Improving the robustness of deep learning for secure artificial intelligence



- Deep learning models used for predictions and judgments are vulnerable to adversarial perturbations in real-world applications.

- Adversarial perturbations are small adjustments to input data that can cause deep learning models to make incorrect predictions.

- Subtle changes can cause the model to make mistakes, which attackers can exploit by generating adversarial examples to cause incorrect predictions.

- Incorporating human knowledge into deep learning model training can improve its robustness and security against adversarial perturbations by recognizing patterns and features less likely to be affected.

Example: Self-Driving car

- Self-driving cars rely heavily on deep learning models for driving decisions
- Deep learning models used in self-driving cars can be vulnerable to adversarial attacks
- Adversarial attacks can cause a stop sign to be misidentified as a yield sign
- An attacker could use an undetectable sticker to cause a self-driving car to fail to stop at an intersection
- Adversarial attacks can put passengers and other drivers at risk

Opp 5. Reducing the size of the required training set

- Deep learning models require many parameters and thousands of training examples to perform accurately. However, obtaining such large amounts of data is often impractical.

- Researchers are thus seeking to reduce the training set size by leveraging prior knowledge from previous models and human expertise.

- Suppose, You're training a deep learning model to predict heart disease risk based on a patient's medical history and lifestyle factors, but limited labeled data makes it hard to predict accurately for patients with less common medical conditions or risk factors.

- Incorporating external human knowledge such as medical expertise or academic research on risk factors improves the model's accuracy and generalization to diverse patient histories without requiring many labeled examples.

- This approach is valuable for real-world applications where labeled data collection is not feasible.

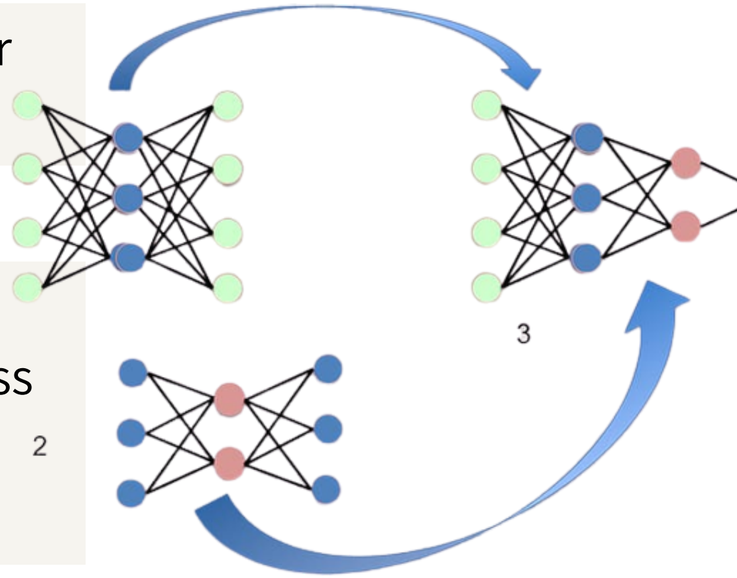
Cont.

- This can be achieved through One-shot or zero-shot learning techniques which enable the model to identify new objects with minimal user input.

- One-shot learning trains the model to recognize new medical conditions or risk factors with a single or few labeled examples

- Zero-shot learning trains the model to recognize new medical conditions or risk factors without labeled examples.

- Both techniques are useful in scenarios where it may be difficult or impractical to obtain large amounts of training data for every possible class or scenario.



Opp 6. Visual analytics for advanced deep learning architectures

- Deep learning architectures are complex and harder to control than simpler versions

- Advanced deep learning programs are difficult to interpret and interact with, making it challenging to understand their internal processes, decision-making, and control or modification.

- Researchers can develop visualization techniques to help understand and control deep learning programs, such as visualizing their internal workings.

- Researchers can also create interactive tools that let users change and manage the behavior of the program.

Conclusion

- The study emphasizes the necessity for tools and methods to comprehend and work with deep learning models as they become more crucial in day-to-day activities.

- The authors explore initiatives taken by academia and business to address this issue .

- They also proposed potential research directions for the future, such as

1. incorporating human knowledge and data-driven learning approaches
2. creating advanced visual analytics systems
3. exploring user-driven generative models

Thank you!