Explainable AI summary

I. Papastratis, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 1.0.1

Artificial Intelligence &
Information Analysis Lab

Explainable AI

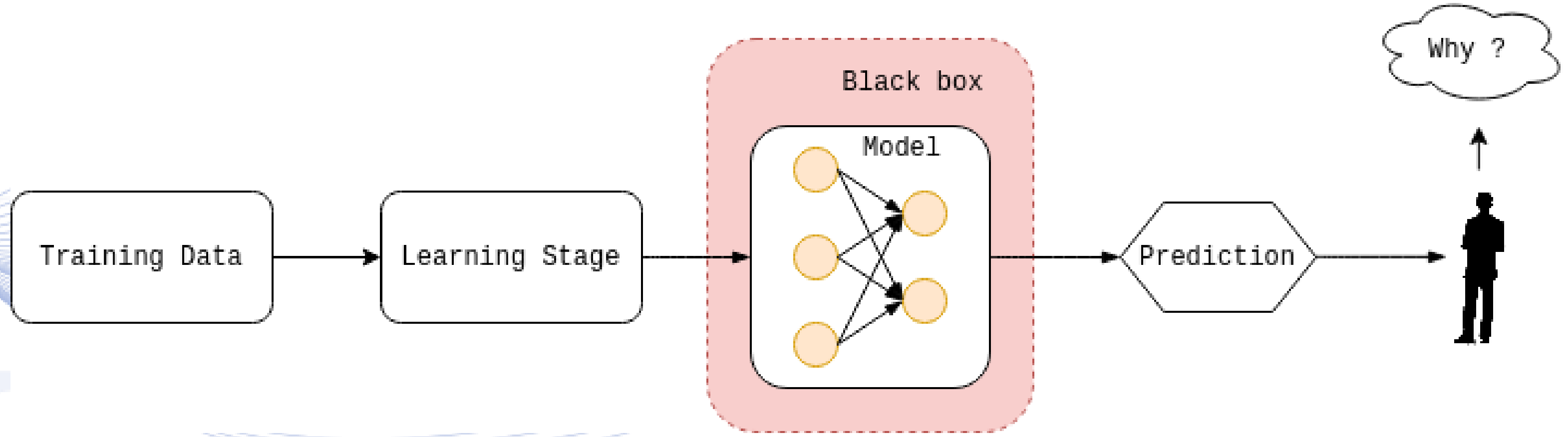
- Introduction to Explainable Artificial Intelligence
- Interpretability
- Types of Interpretability
  - Visual explanations
    - Image-based
    - Plot visualizations
  - Textual explanations
  - Numerical-Mathematical explanations
- Applications
- Frameworks

Introduction to Explainable AI

Machine and Deep Learning have surpassed humans in many tasks (image and speech recognition, recommendation systems, medical diagnosis)

Drawbacks:

- Machine learning (ML) architectures are usually considered as blackbox models



Introduction to Explainable AI

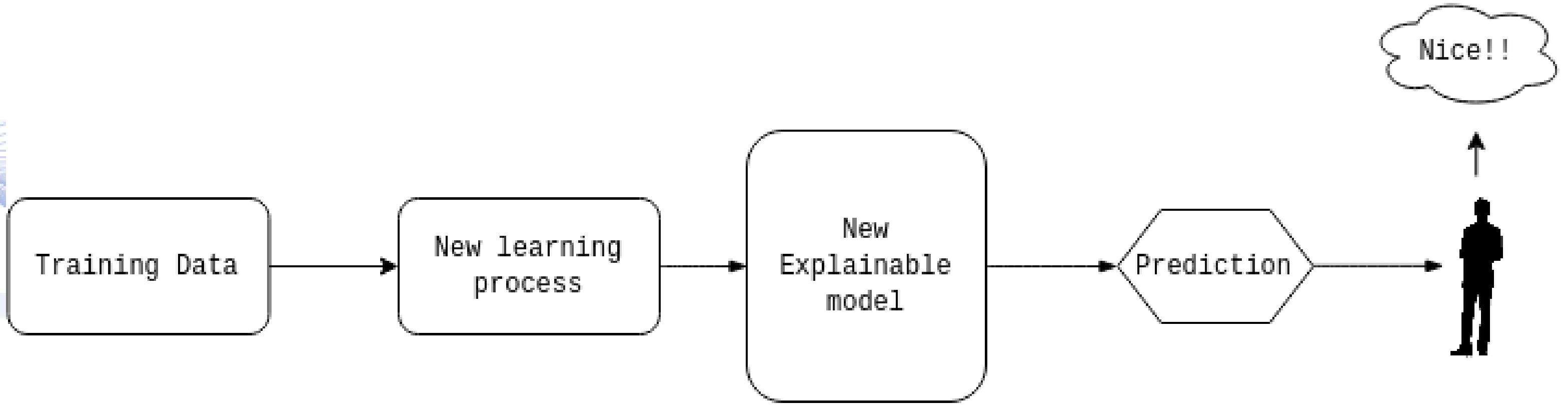
- Deep learning (DL) have achieved outstanding performance but they don't justify their reliability
- Failure
    - An error in any moment of a self-driving car can lead to a fatal crash.
    - in the medical area, human lives may be dependent on these decisions

Introduction to Explainable AI

Issues

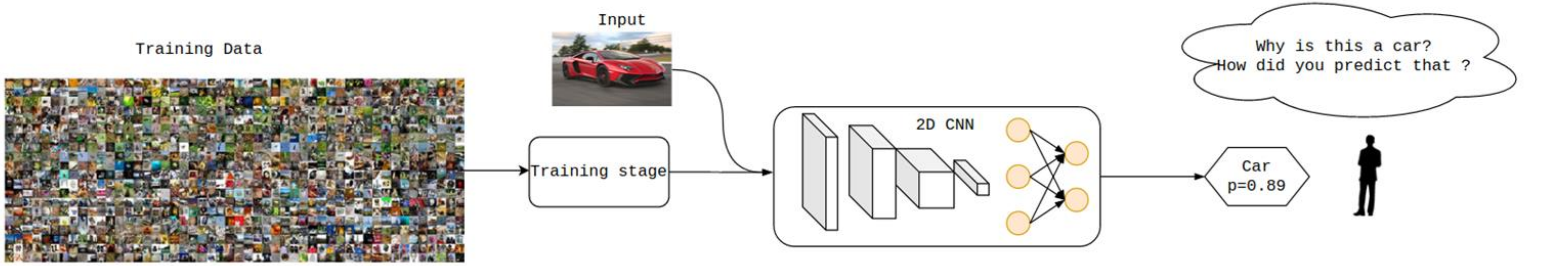
- responsibility for bad AI decisions
- explain errors of AI decisions
- improvement of AI models

Solution: Explainable AI models easily understandable by humans

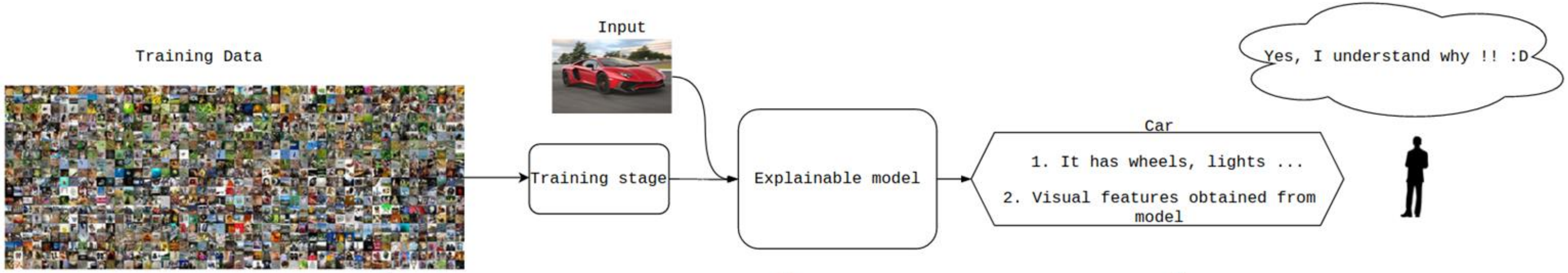


Artificial Intelligence & Information Analysis Lab

Introduction to Explainable AI



S



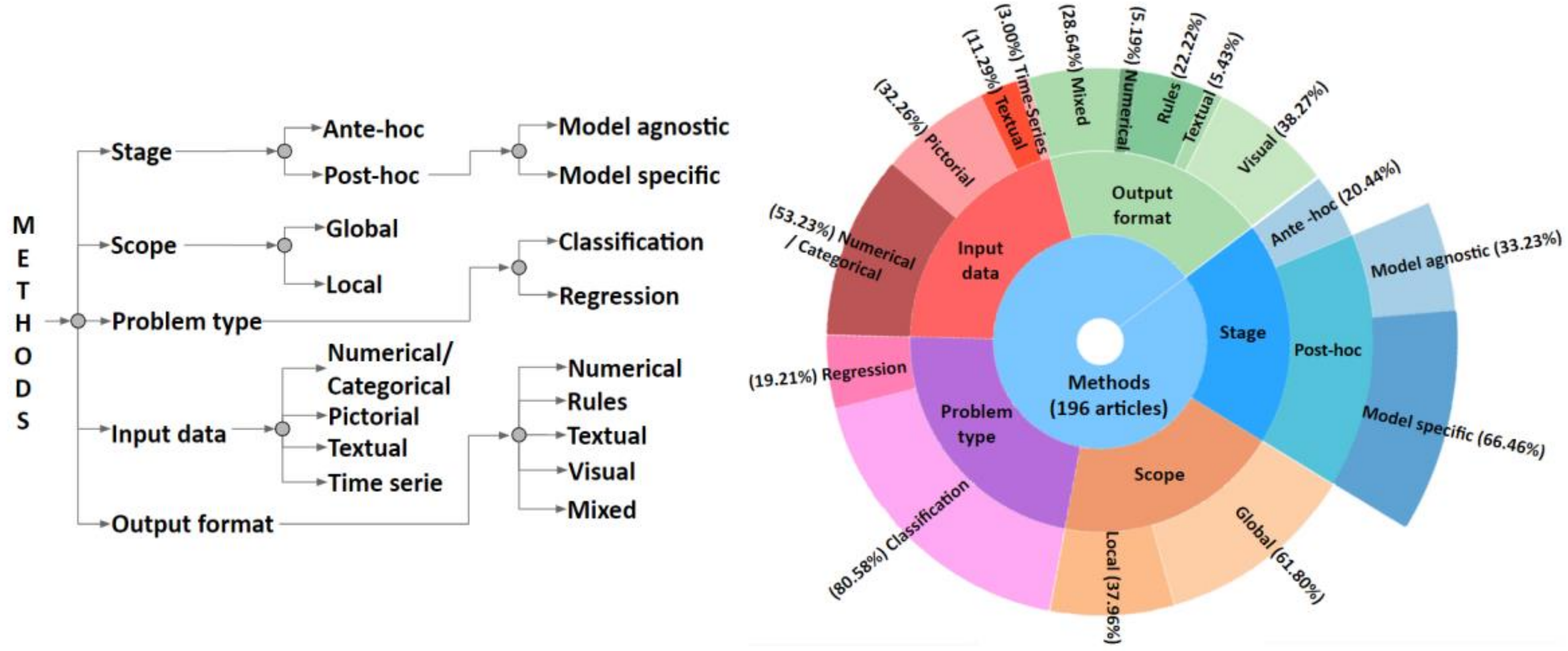
Interpretability

How easily we can understand the cause of an algorithm's decision or action
The categorization of Interpretability methods is based on how interpretable information is provided.

Categories:

- Visual Interpretability ("obviously" interpretable information, easily perceived from human eye
- Textual Explanations (Given in form of text)
- Mathematical-Numerical Explanations

Types of Interpretability Methods



[Vilone2020] Categorization of explainability methods

Visual Explanations

- Visual explainable methods produce pictures or maps in order to provide information about the model’s decision
  - Most common: **Saliency** methods explain results of model by producing outputs to show which components are responsible.
  - These values take the form of output probabilities or images like heatmaps.
- Plot visualization methods produce scatter plots to explain decisions or visualize the data

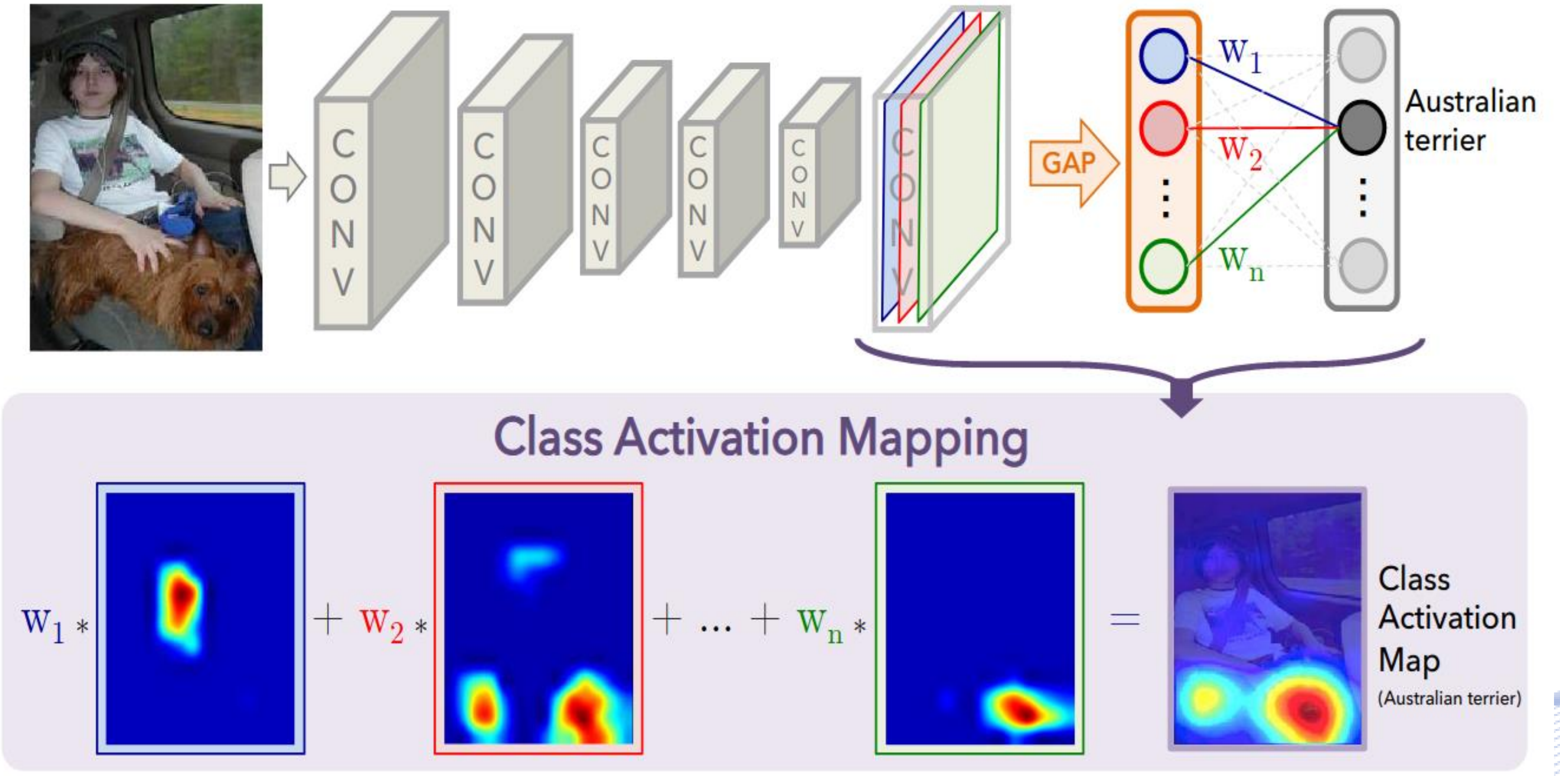
Visual Explanations

- Methods
  - CAM (Class Activation Maps)
  - Grad-CAM: Gradient based CAM
  - LRP (Layer-wise Relevance Propagation)
  - Peak Response Map (PRM)
  - CLass-Enhanced Attentive Response (CLEAR)
  - DeConvNet
  - DeepResolve
  - SCOUTER

Class Activation Maps (CAM)

- Generate activation maps for a decision
- Global average pooling layer is after convolutional layers
- The output features of the convolutional neural network are passed through a fully-connected layer that makes the prediction
- CAM indicates the region on the image that correspond to the prediction result

Class Activation Maps(CAM)



[Zhou2016] Overview of CAM

Class Activation Maps(CAM)

[Zhou2016] Examples of CAM

Grad-CAM: Gradient-based CAM

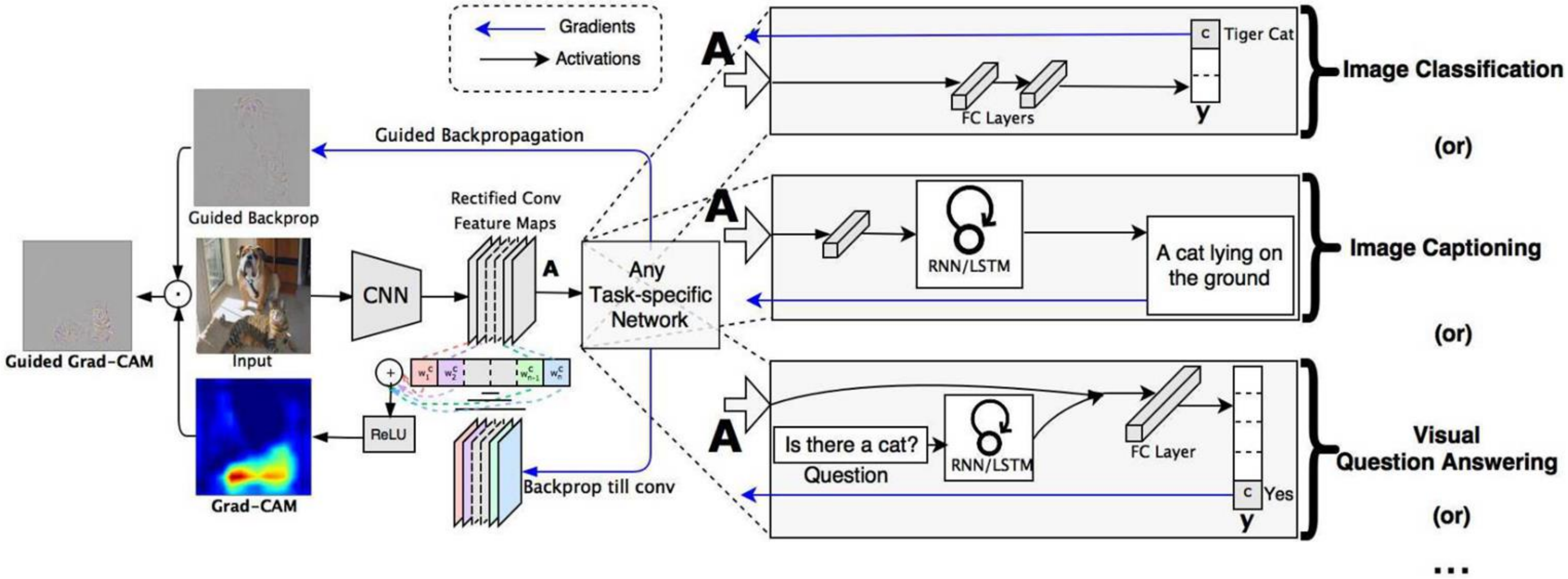
Gradient-weighted Class Activation Maps(Grad-CAM)

- an extended version of CAM by computing the gradients with respect to the target that flow to the final convolutional layer
- produces a map, which highlights the most useful pixels for classification

Processing steps

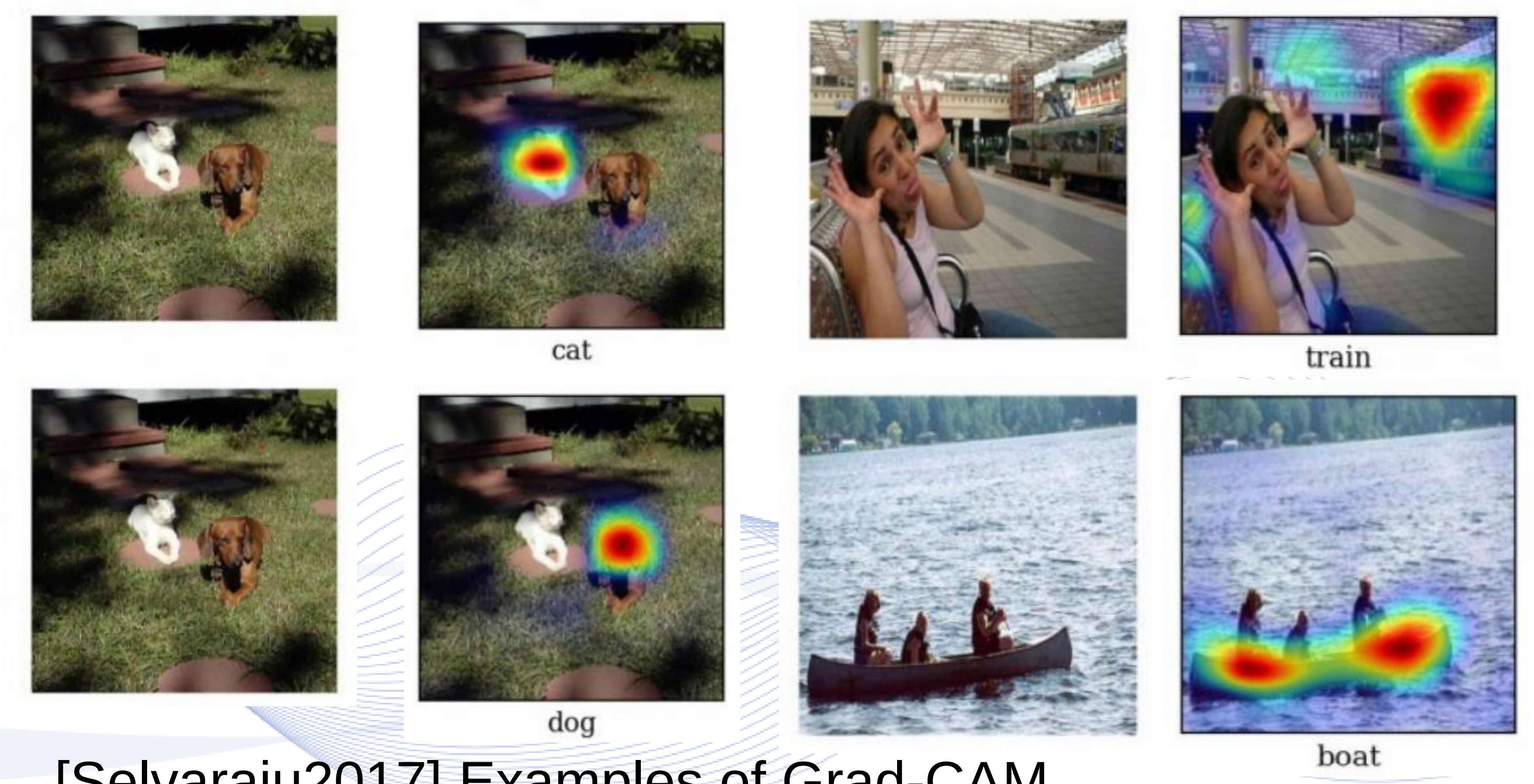
- Forward pass of the input image to produce the prediction
- Gradients of the target class
- The gradients of the target are back-propagated to last convolutional layer
- Find the important locations of the image

Grad-CAM: Gradient-based CAM



[Selvaraju2017] Overview of Grad-CAM

Grad-CAM: Gradient-based CAM



[Selvaraju2017] Examples of Grad-CAM

Layer-wise Relevance Propagation (LRP)

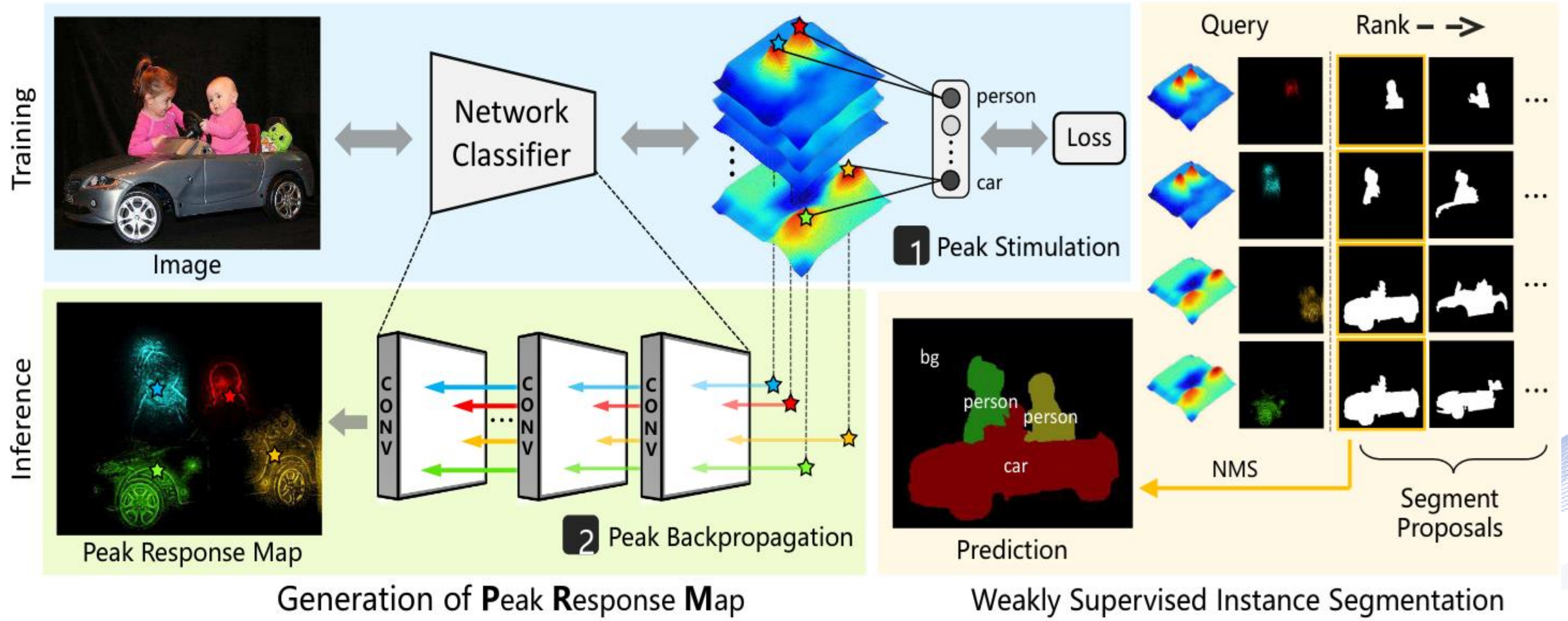
- Decomposition of the classification decision of a DNN model in using the input
- The classification layer is decomposed into several layers
- Backward pass to produce the pixel-wise conributions to the final output from last to first layers
- We denote as $w(i,j)$ the weight between of the connection from neuron i to neuron j

Layer-wise Relevance Propagation (LRP)

[Samek2016] LRP calculation for the input image

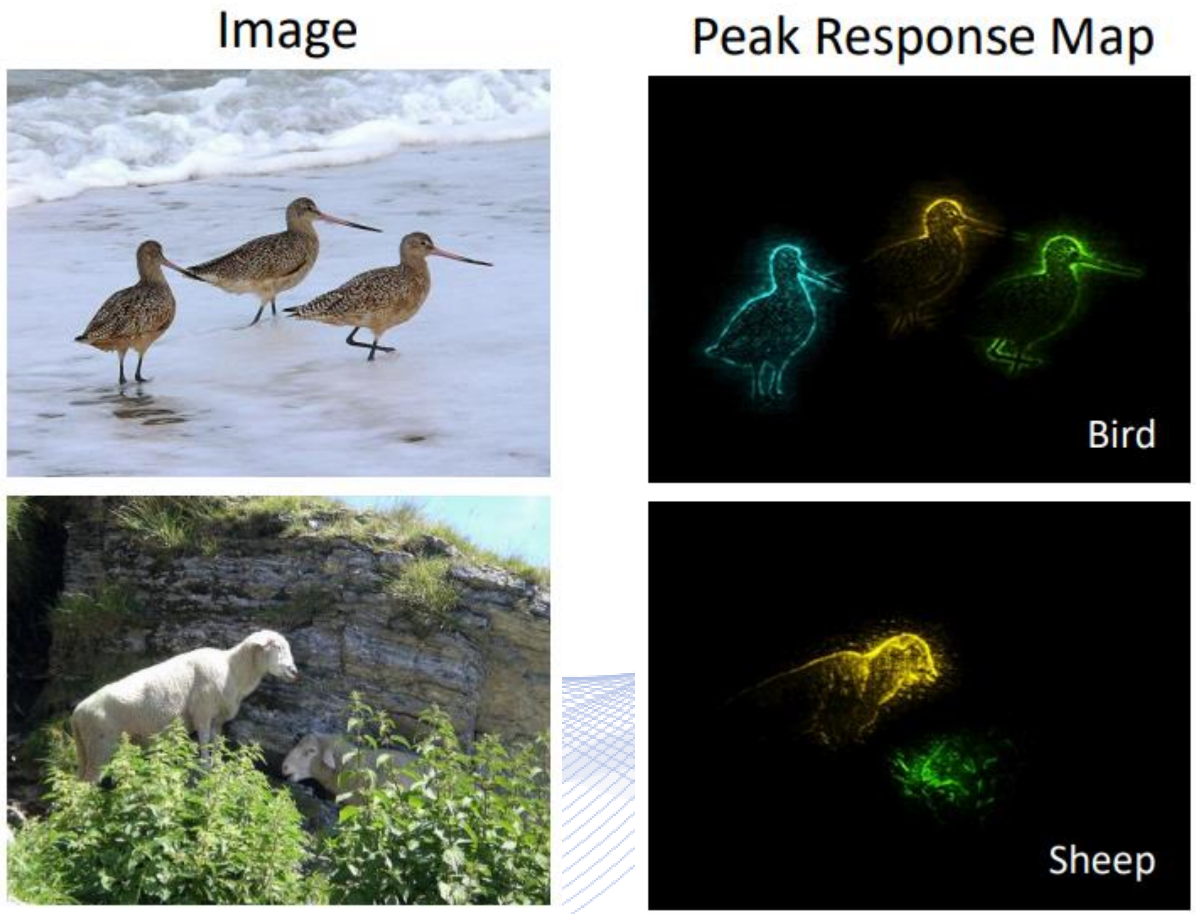
Artificial Intelligence & Information Analysis Lab

Peak Response Map (PRM)



[Zhou2016] Overview of PRM

Peak Response Map (PRM)



[Zhou2016] Examples of PRM

CLass-Enhanced Attentive Response (CLEAR)

- Visualizes the decisions of image classification applications with attention maps produced by back-propagating the activations of the last layer
  - After forward pass using deconvolutions we obtain the deconvolved output

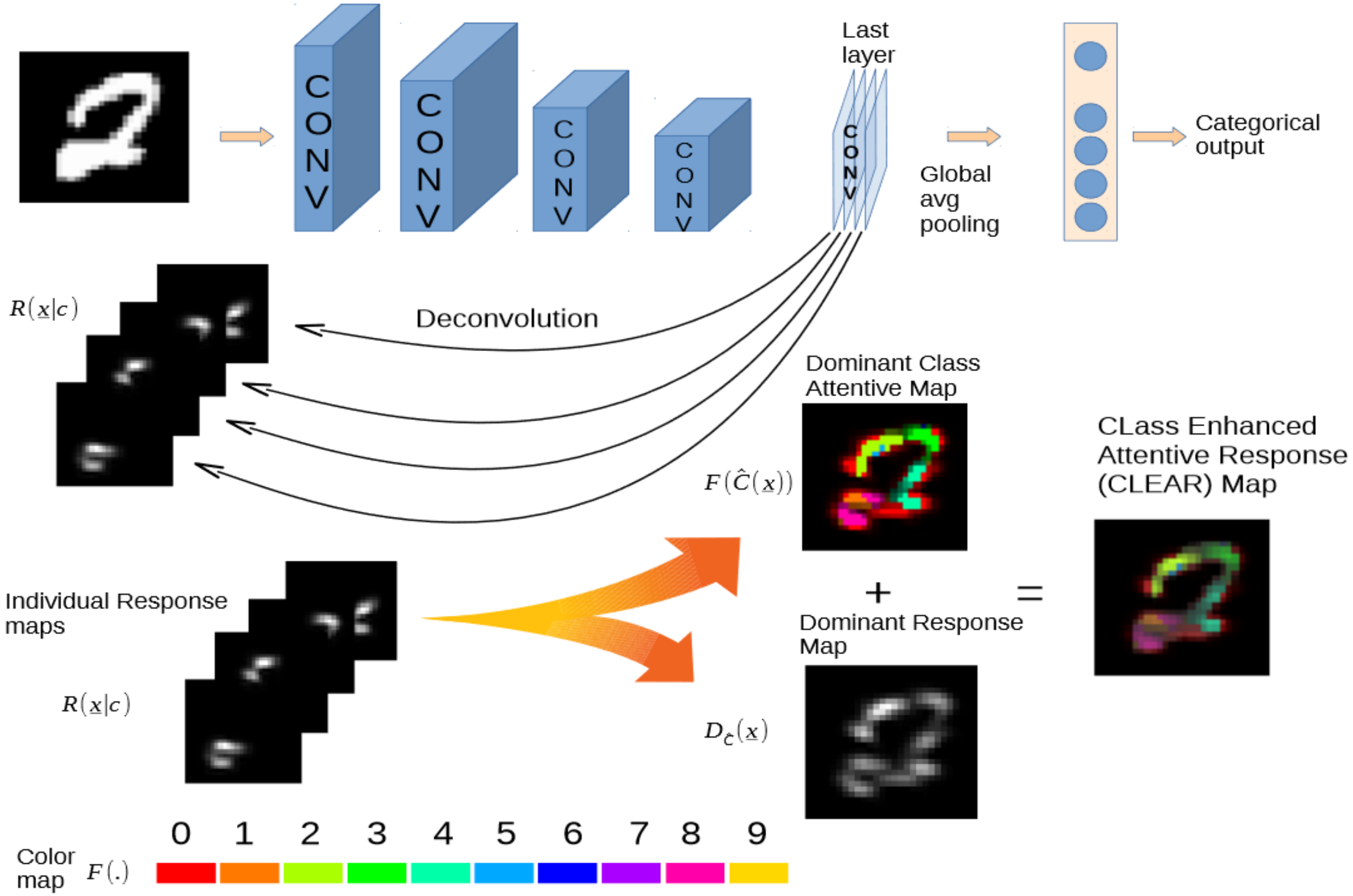
$$\mathbf{h}(l) = \sum_{k=1}^{K} z(k,l) * w(k,l)$$

k: kernel index $\mathbf{z}(l)$: feature maps of layer l, $\mathbf{w}(l)$: kernel weights, K kernels

- Final response of layer l is the product: $\mathbf{R}(l) = \mathbf{h}(1)\mathbf{h}(2) \dots \mathbf{h}(l)$
- Compute individual attention maps $\mathbf{R}(\mathbf{x}', c)$ of class c and back-projected input $\mathbf{x}'$ from all L layers of the deconvolutional network as :

$$\mathbf{R}(\mathbf{x}', c) = \mathbf{h}(1)\mathbf{h}(2) \dots \mathbf{h}(L)$$

CLass-Enhanced Attentive Response (CLEAR)

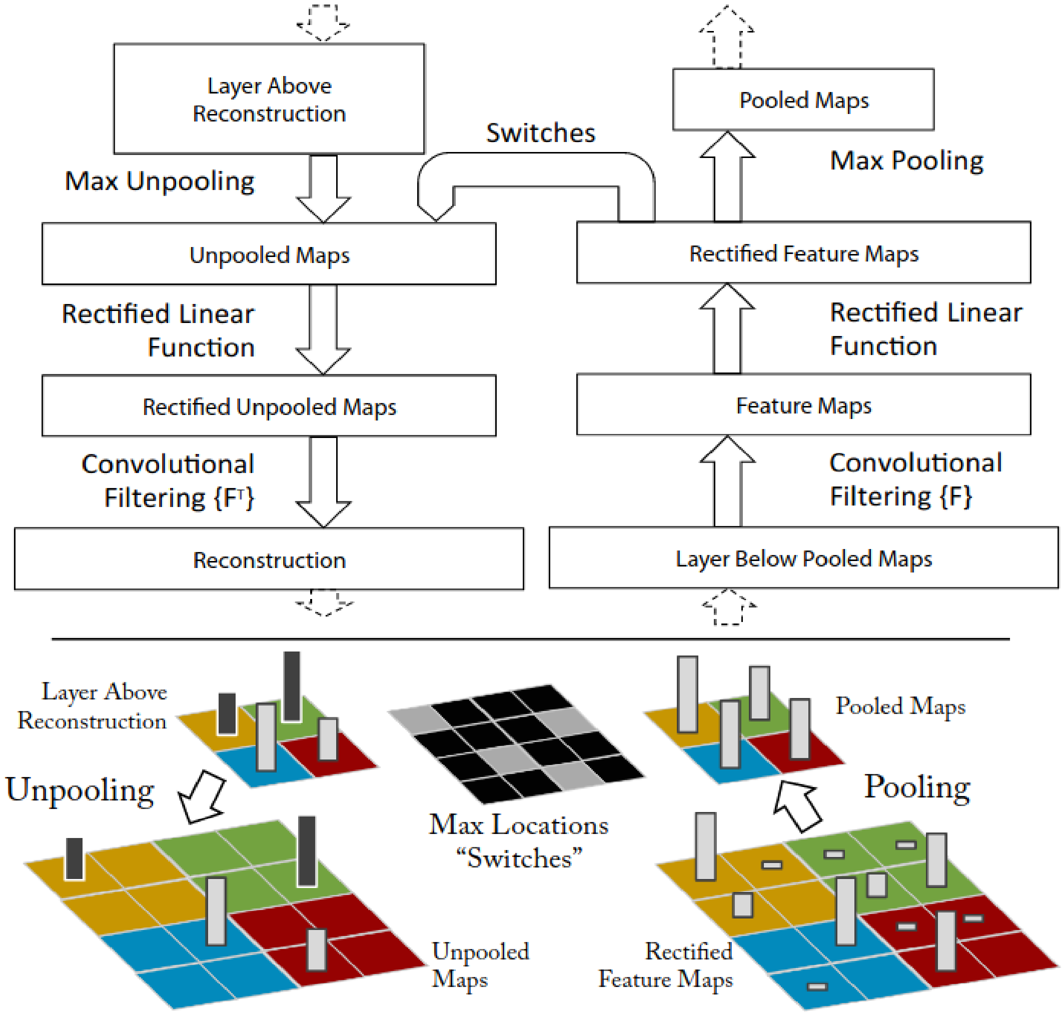


[Kumar2017] CLEAR method

DeConvNet

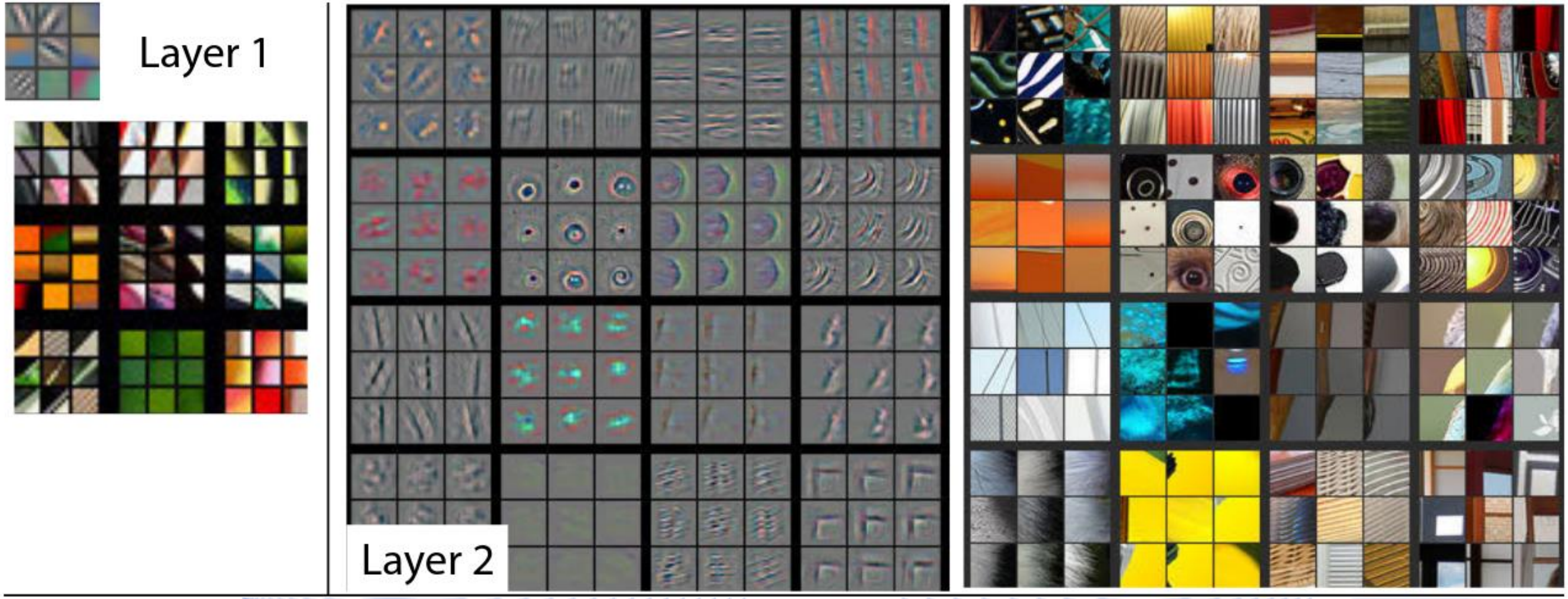
- A deconvnet is actually a convnet model with the same convolutional layers but in reverse since it maps pixels to features
- Deconvet is used here to generate the input images from features
- Following operations are adopted:
- Trsnspose Convolution (deconvolution):
  - Use transposed learned filters of cnn to reconstruct the input of the layer from the output
- Unpooling
  - record maximum values locations during maxpooling and use the locations to reconstruct input during deconvolution

DeConvNet



[Zeiler2014] Reconstruction of the convnet features up to the pixel space

DeConvNet

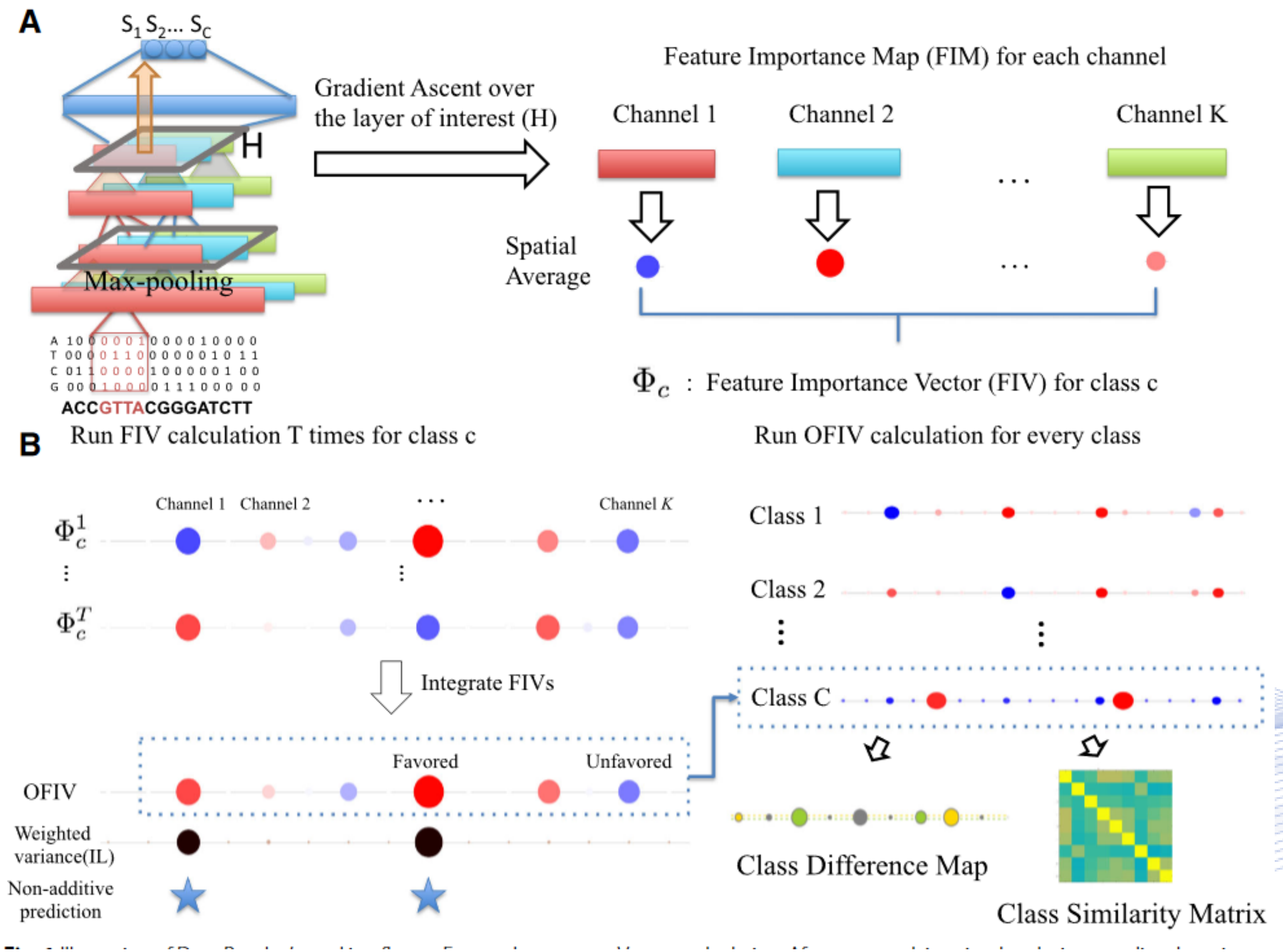


[Zeiler2014] Visualization of intermediate layers

DeepResolve

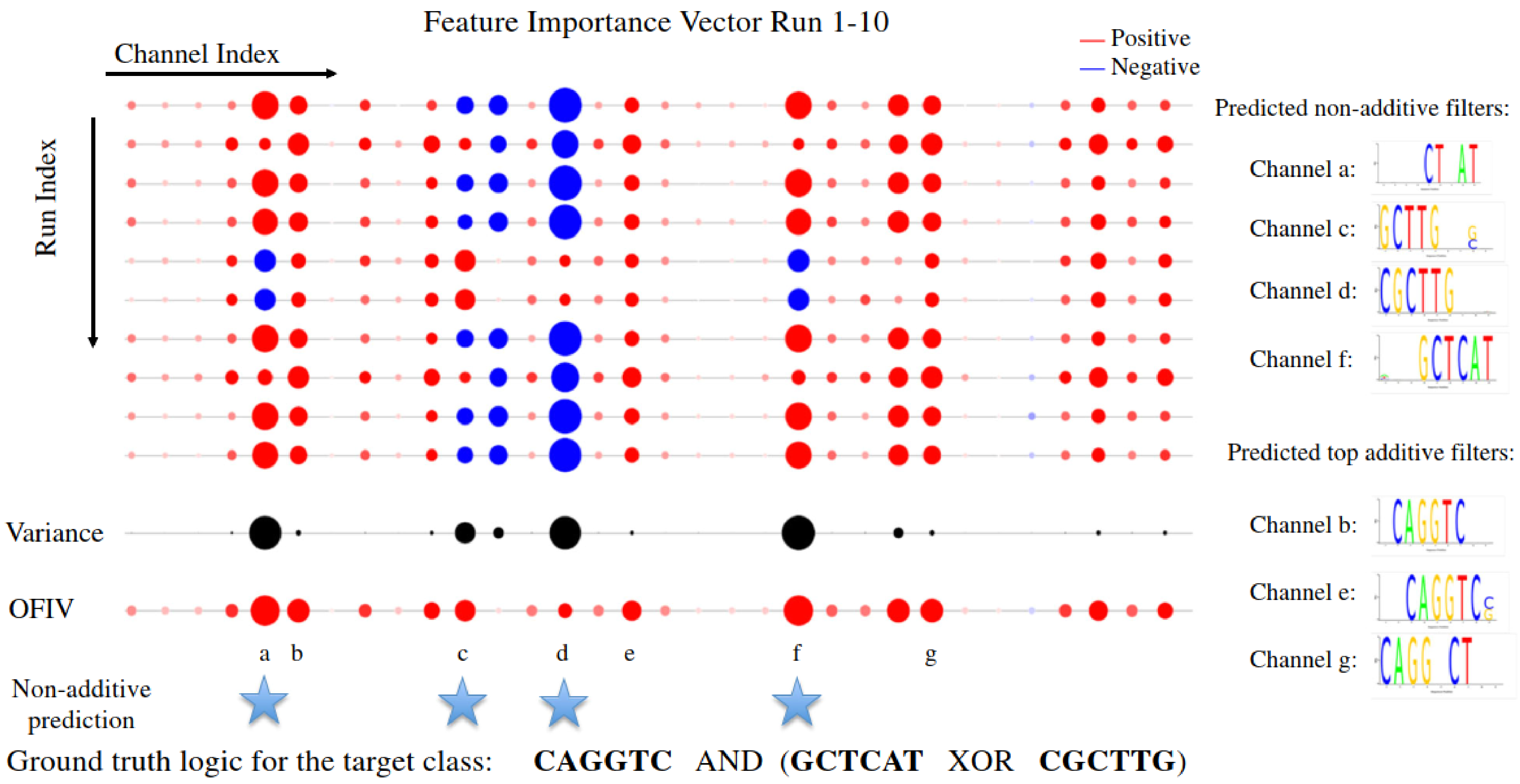
- Generate intermediate layer heatmaps to show how the network combines features for classification
- Compute optimized feature map $\mathbf{H_c} = argmax_{\mathbf{H}} S_c(\mathbf{H}) - \lambda||\mathbf{H}||_2^2$, S_c is the score of class c obtained from the last layer, λ tunable hyperarameter and $\mathbf{H} \in \mathbf{R}^{K \times W}$
- Global average of Feature Importance Maps (FIM) $\mathbf{H_c}$ to obtain feature importance vector (FIV) $\boldsymbol{\Phi}_\mathbf{c} = \left(\varphi_c^1, \ldots \varphi_c^k\right)$ where $\varphi_c^k = \frac{1}{W}\sum_{i=1}^{W}\left(H^k(i)\right)_c$
- This procedure is ran T times with different initial parameters to get several estimations of $\mathbf{H_c^t}$ and $\boldsymbol{\Phi}_\mathbf{c}^{\boldsymbol{t}}$

DeepResolve



[Liu2019] Working flow of DeepResolve

DeepResolve

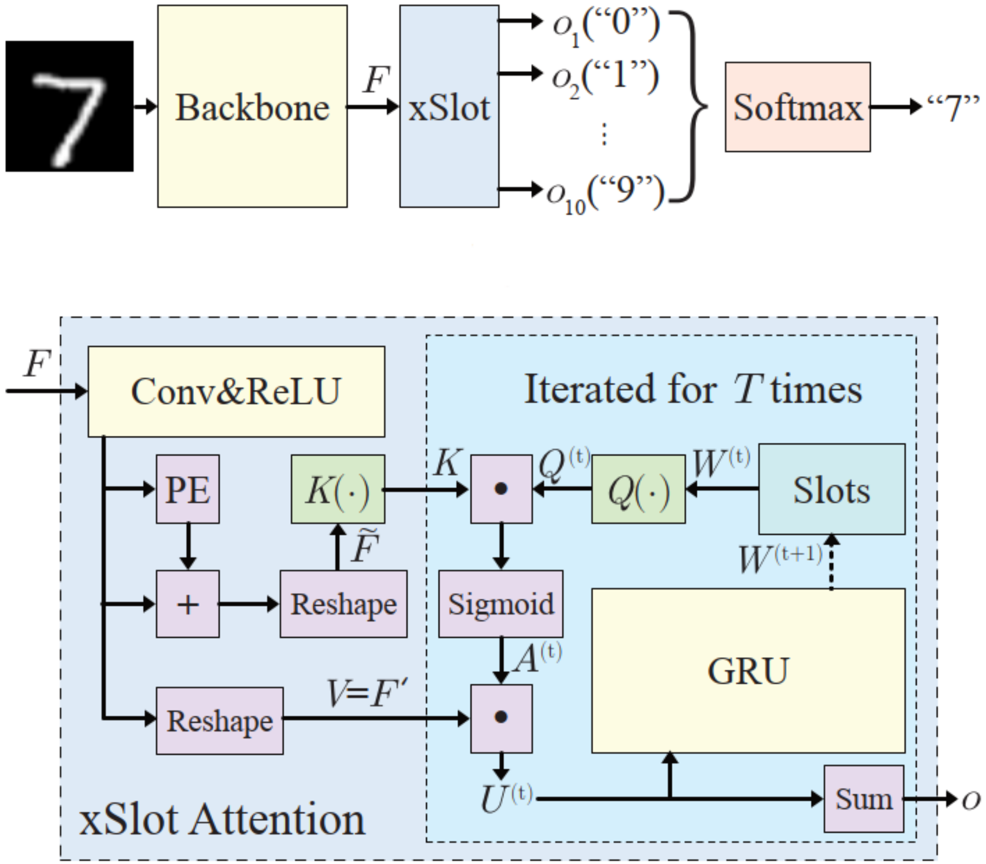


[Liu2019] Generation of importance vectors

SCOUTER

- Explain the reasons an image is classified or not to a specific class
- The cnn's classifier (fully connected layer) is replaced with a slot attention model
- Each slot produces a confidence score for each class
- The cnn's features $\mathbf{F}$ are passed through a convolutional layer and a position embedding to model the spatial information
- Then, a self-attention mechanism is utilized to compute the weighted sum of features as: $\mathbf{A}^{(t)} = \sigma(Q(\mathbf{W}^{(t)})K(\mathbf{F}))$
- σ: sigmoid function Q, K: fully-connected networks $\mathbf{W}^{(t)}$: slot weights

SCOUTER



[Li2020] SCOUTER overview

SCOUTER

[Li2020] Examples of positive and negative decisions

Visual explanation of deep neural networks by interpretation

- Identification of relevant features to the predictions of the network F
- Forward-propagation of N training images to obtain m–dimensional features

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_N) \in \mathrm{R}^{N \times m}$$

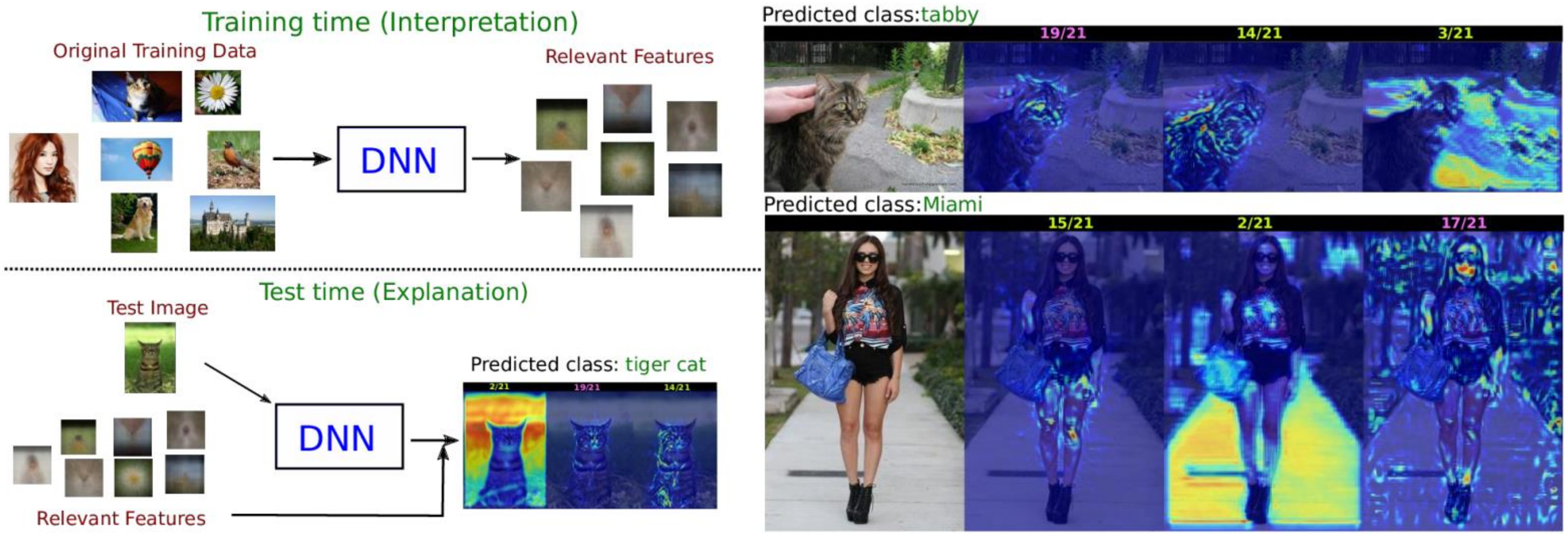
- Target labels $\mathbf{L} = (\mathbf{l}_1, \mathbf{l}_2, \ldots \mathbf{l}_N) \in \mathrm{R}^{C \times N}$
- Predict a linear combination of activations $\mathbf{X}$ for each class using

$$\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_N) \in \mathrm{R}^{C \times m}$$

by solving the optimization problem :

$$\mathbf{W}^* = argmin_{\mathbf{W}} \left|\left|\mathbf{X}^{\mathbf{T}}\mathbf{W} - \mathbf{L}^{\mathbf{T}}\right|\right|_F^2$$

Visual explanation of deep neural networks by interpretation

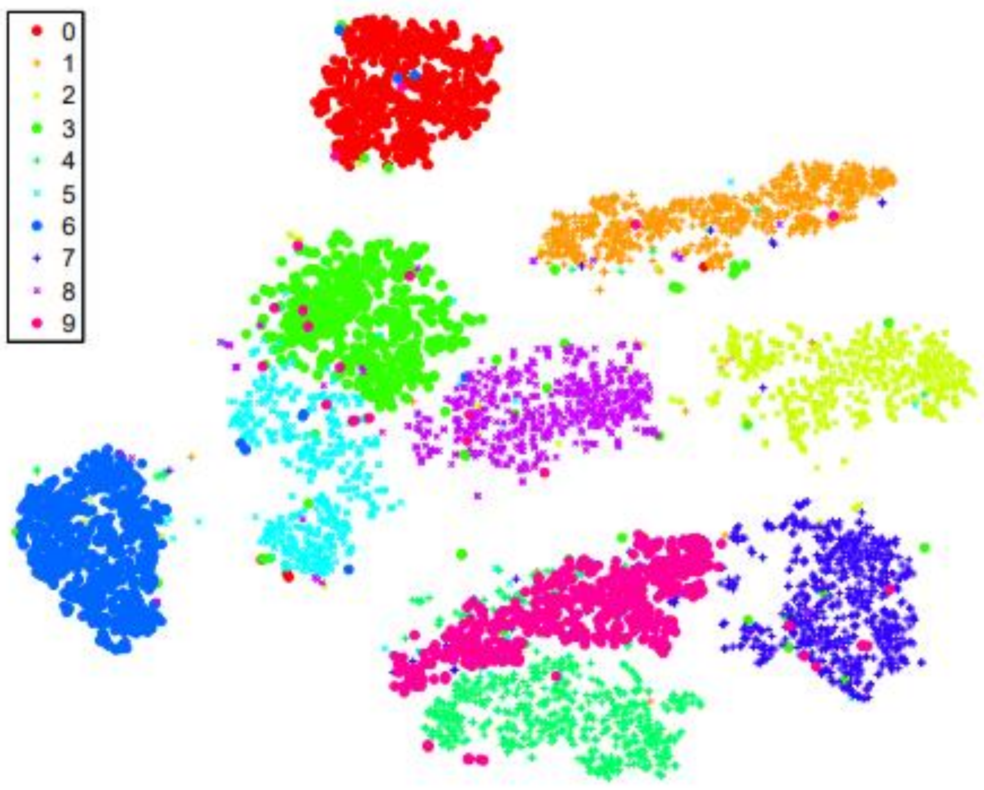


[Oramas2019] Overview of training and testing pipeline

Plot visualizations

- Produce scatter plots to explain decisions or visualize the data
- Methods
  - t-SNE
  - Understanding deep features using PCA
  - Visualization of hidden layers
  - TreeView

t-distributed stochastic neighbor embeddings (t-SNE)

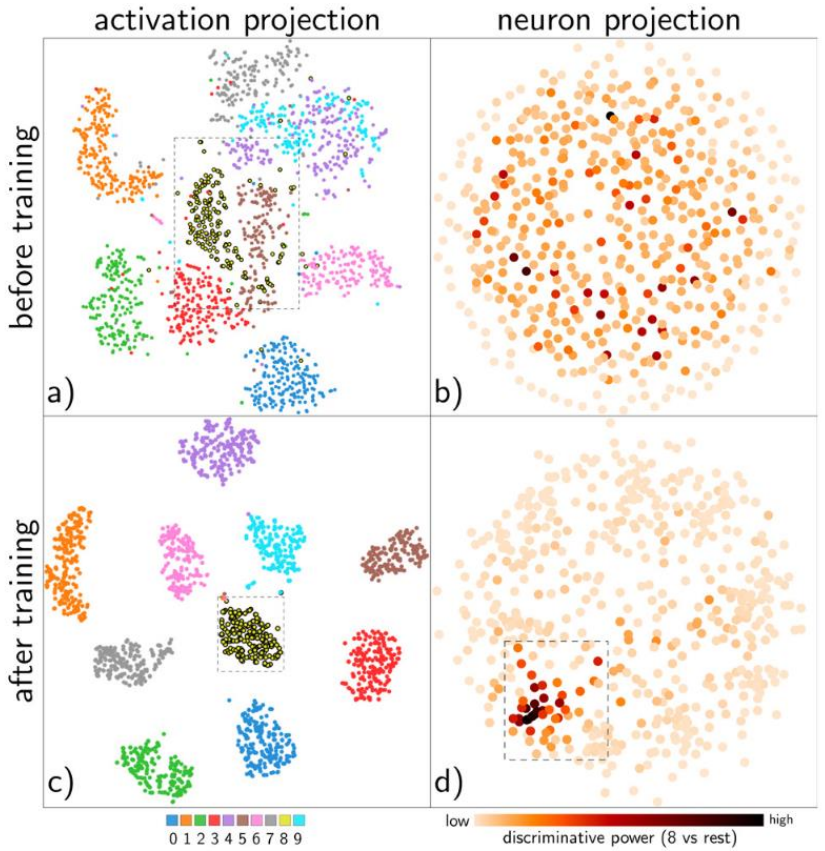


[Maaten2008] Visualization of MNIST

Visualization of hidden layers

- 2D scatter-plot the projections of the hidden neurons' activation coloured according to the class
- Dimensionality reduction and visualization of:
  - observations
  - neuron's relationships
- Clustering of activations in groups to explain predictions

Visualization of hidden layers



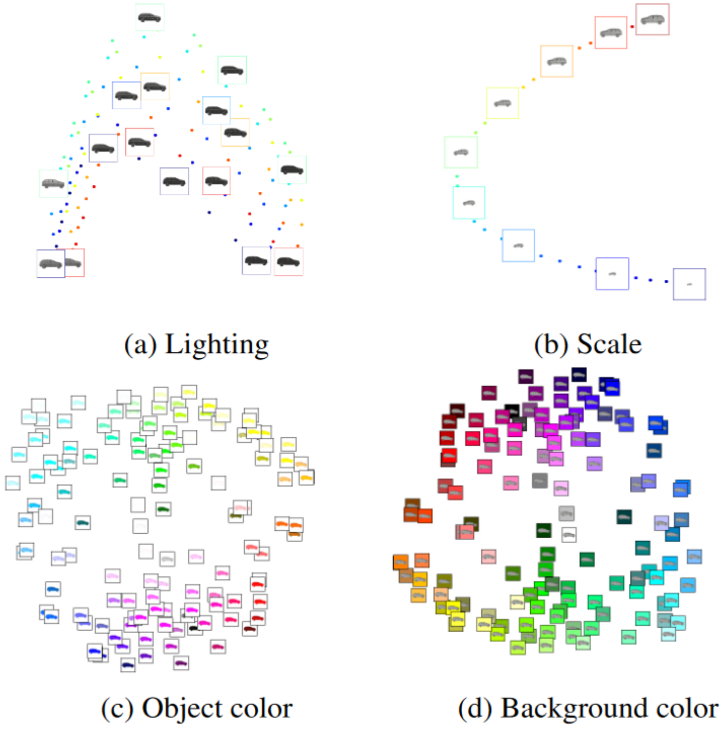
[Rauber2016]T-SNE projection of neurons and classes on MNIST test set

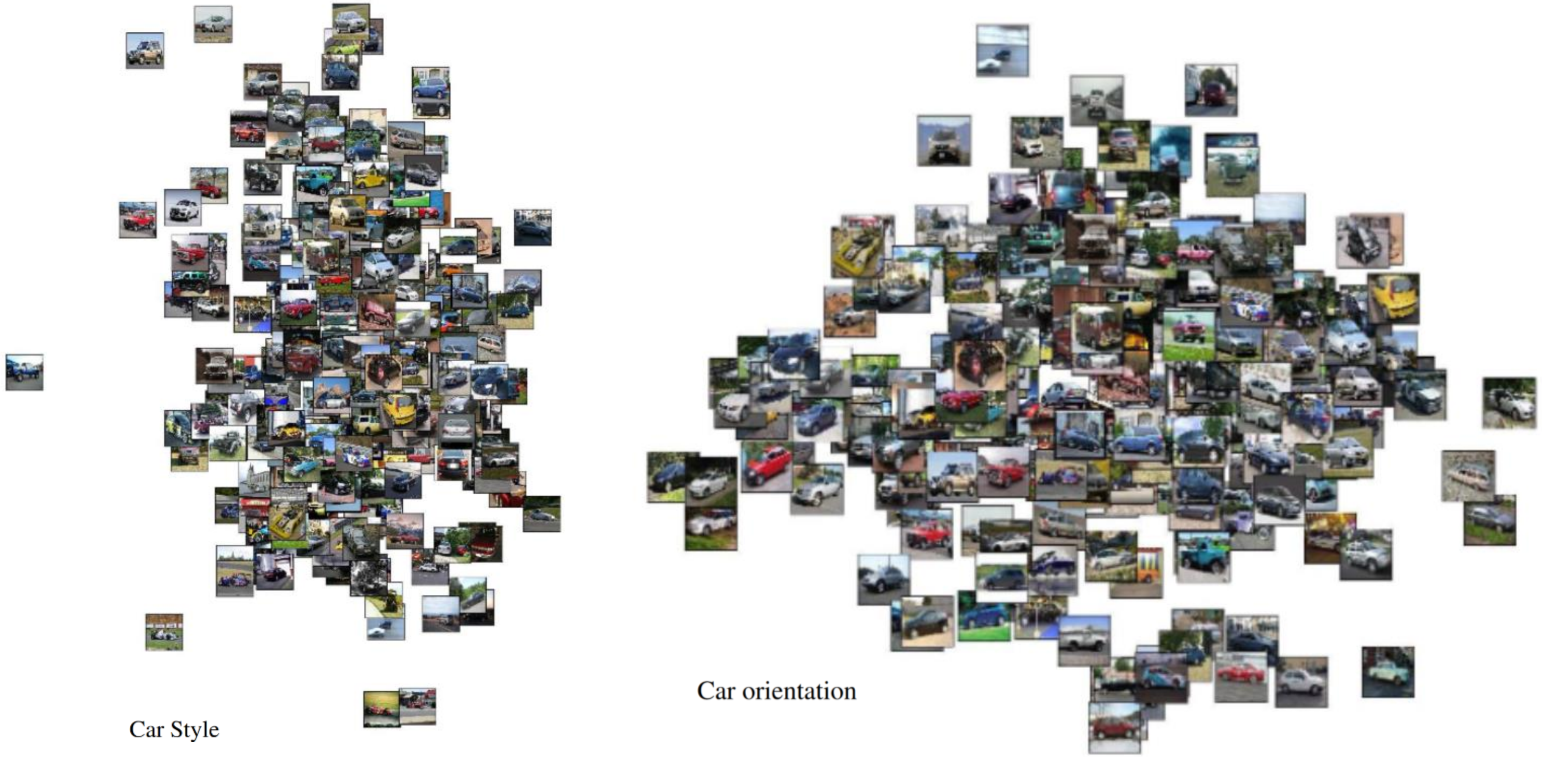
Understanding deep features using PCA

- analyzes CNN feature responses of layers with decomposition into a linear combination of principal components using knowledge from the input scene
- Given
  - an image r_θ from a set $\mathbf{\Omega}$ with Θ images indexed from $\theta \in [1, \Theta]$
  - features $\hat{\mathbf{F}}^L(r_\theta)$ of the L_{th} layer of CNN
- Calculate centered features $\mathbf{F}^L(r_\theta) = \hat{\mathbf{F}}^L(r_\theta) - \frac{1}{\Theta}\sum_{t=1}^{\Theta} \hat{\mathbf{F}}^L(r_t)$
- r_t: t image from set $\mathbf{\Omega}$
- Compute eigenvectors of the covariance matrix:

$$\frac{1}{\Theta}\sum_{\theta=1}^{\Theta} \mathbf{F}^L(r_\theta)\mathbf{F}^L(r_\theta)^{\mathbf{T}}$$

Understanding deep features using PCA





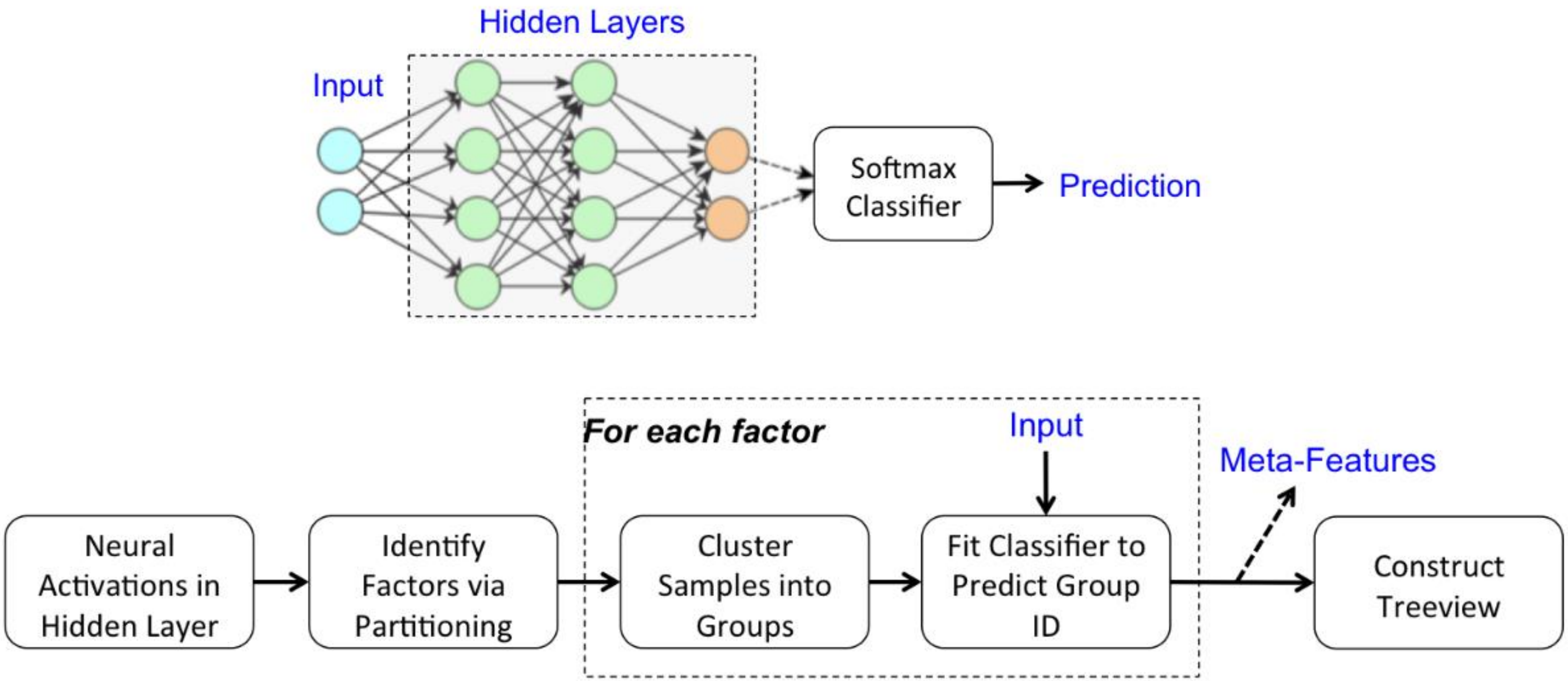
[Aubry2015] PCA embeddings based on different factors

TreeView

- Hierarchical decomposition of the feature subspaces
- Transformation $T_1: \mathbf{X} \rightarrow \mathbf{Y}$ input from space: $\mathbf{X}$ into new space of features: $\mathbf{Y}$,
- Transformation $T_2: \mathbf{Y} \rightarrow \mathbf{Z}$ classify features $\mathbf{Y}$ to label space $\mathbf{Z}$

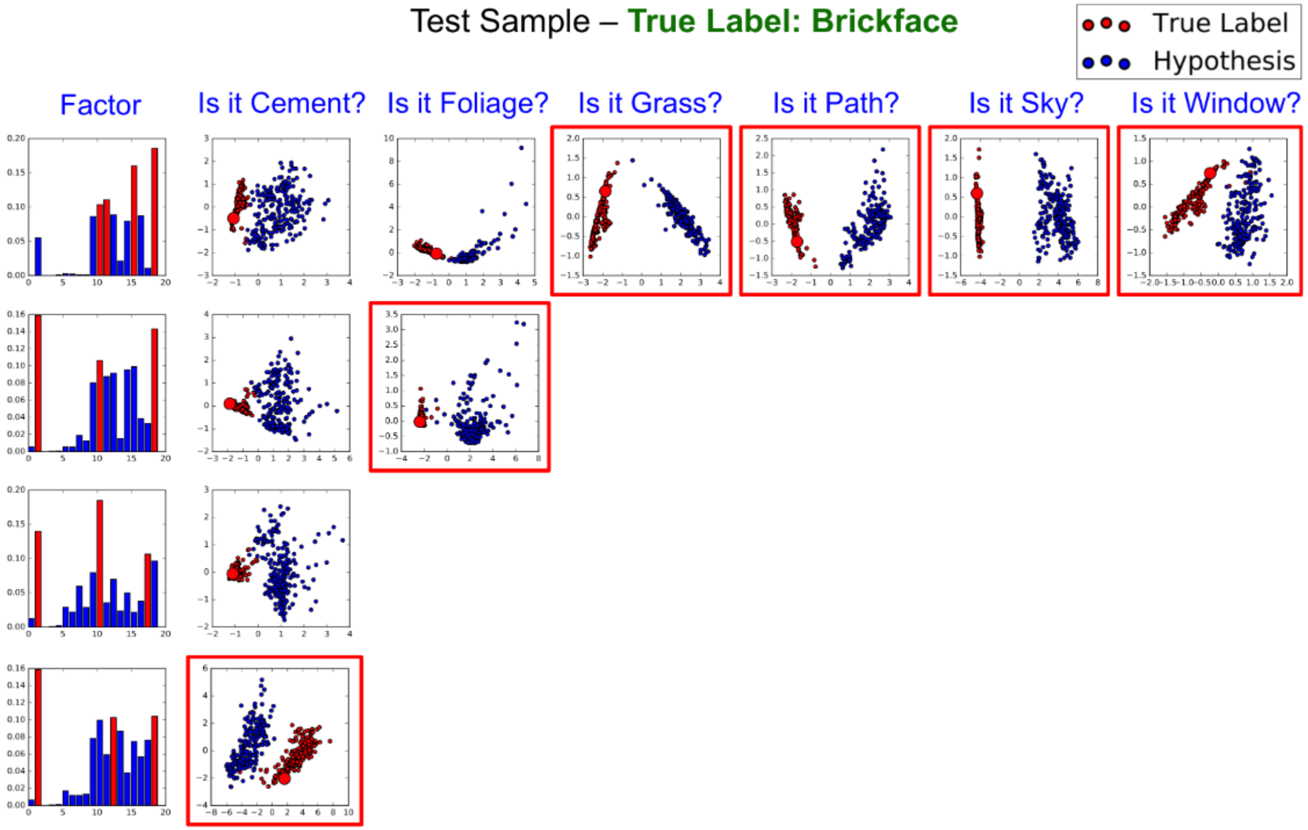
- Partition space of features: $\mathbf{Y}$ into K subspaces with similar activations of the hidden layers
- Each cluster i describes a specific factor S_i
- A new K-dimensional vector from cluster labels is constructed and used for visualization

TreeView



[Thiagarajan2016] TreeView method

TreeView

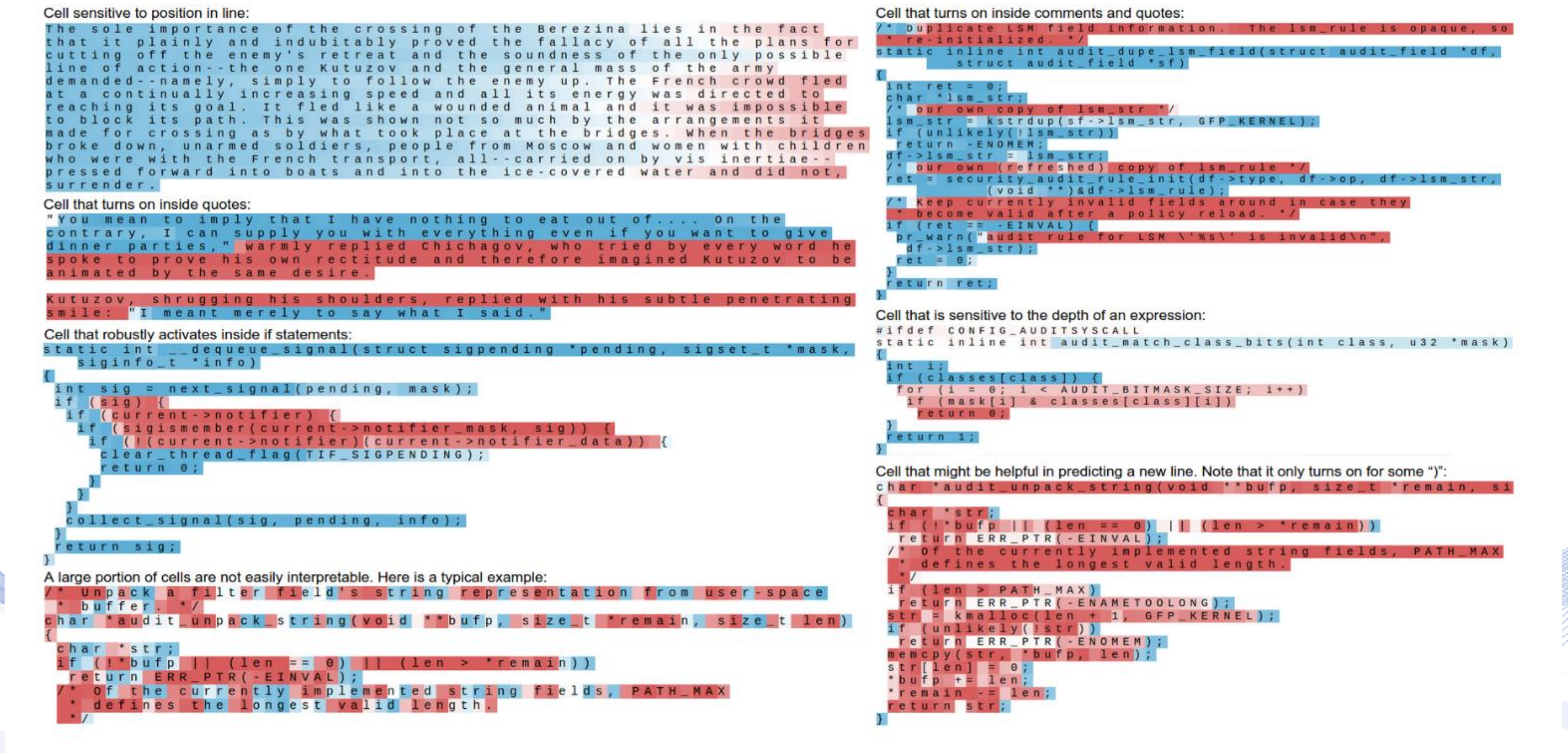


[Thiagarajan2016] Visualization of a correctly classified example for each factor

Textual explanations

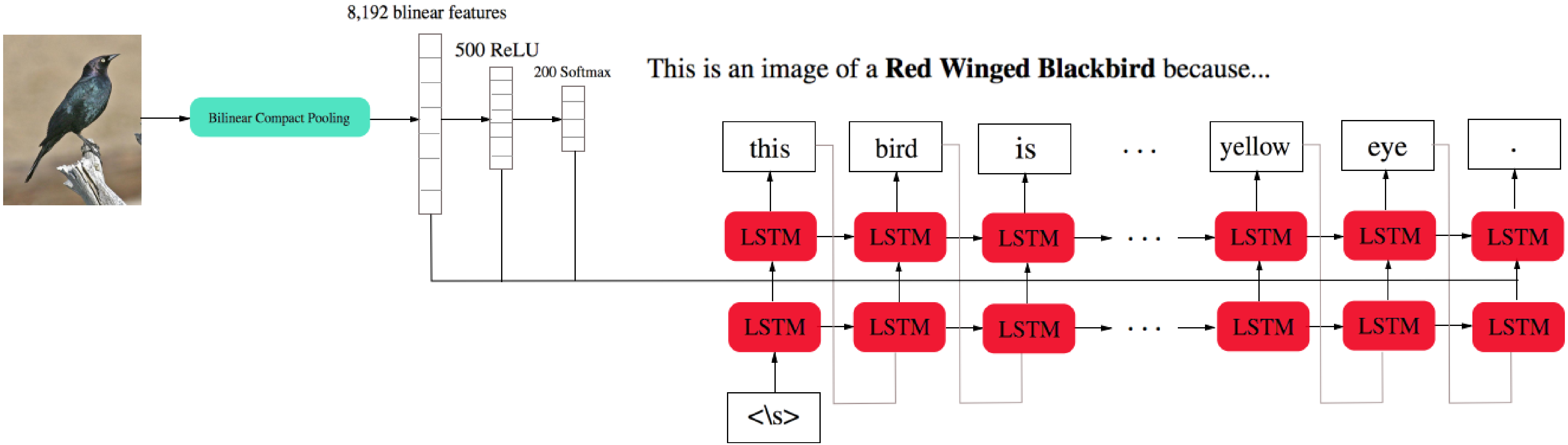
- Produce natural language-text to explain the decisions of the algorithm
- Find semantic words that provide qualitative explanations
- Methods
  - Cell Activation Value
  - InterpNET
  - Hierarchical Question and Image Co-Attention for Visual Question Answering
  - Visual Dialog
  - Explain Deep Neural Networks with Semantic Information

Cell Activation Value



[Karpathy2015] activations of LSTM. Text color corresponds to activation value

InterpNET



[Barratt2017] Overview of Interpnet

InterpNET

This is an image of a **Tree Sparrow** because...

InterpNET(0)	this bird has a brown crown, brown primaries, and a brown belly.
InterpNET(1)	this bird has wings that are brown and has a white belly.
InterpNET(2)	this bird has wings that are brown and has a white belly.
InterpNET(3)	this bird has a brown crown, brown primaries, and a brown belly.
Captioning	this bird has a brown crown, brown primaries, and a brown belly.

This is an image of a **Philadelphia Vireo** because...

InterpNET(0)	this bird has wings that are grey and has a yellow belly.
InterpNET(1)	this bird has wings that are grey and has a yellow belly.
InterpNET(2)	this bird has wings that are brown and has a yellow belly.
InterpNET(3)	this bird has a yellow belly and breast with a short pointy bill.
Captioning	this bird has a yellow belly and breast with a gray crown and white wingbars.

This is an image of an **Ivory Gull** because...

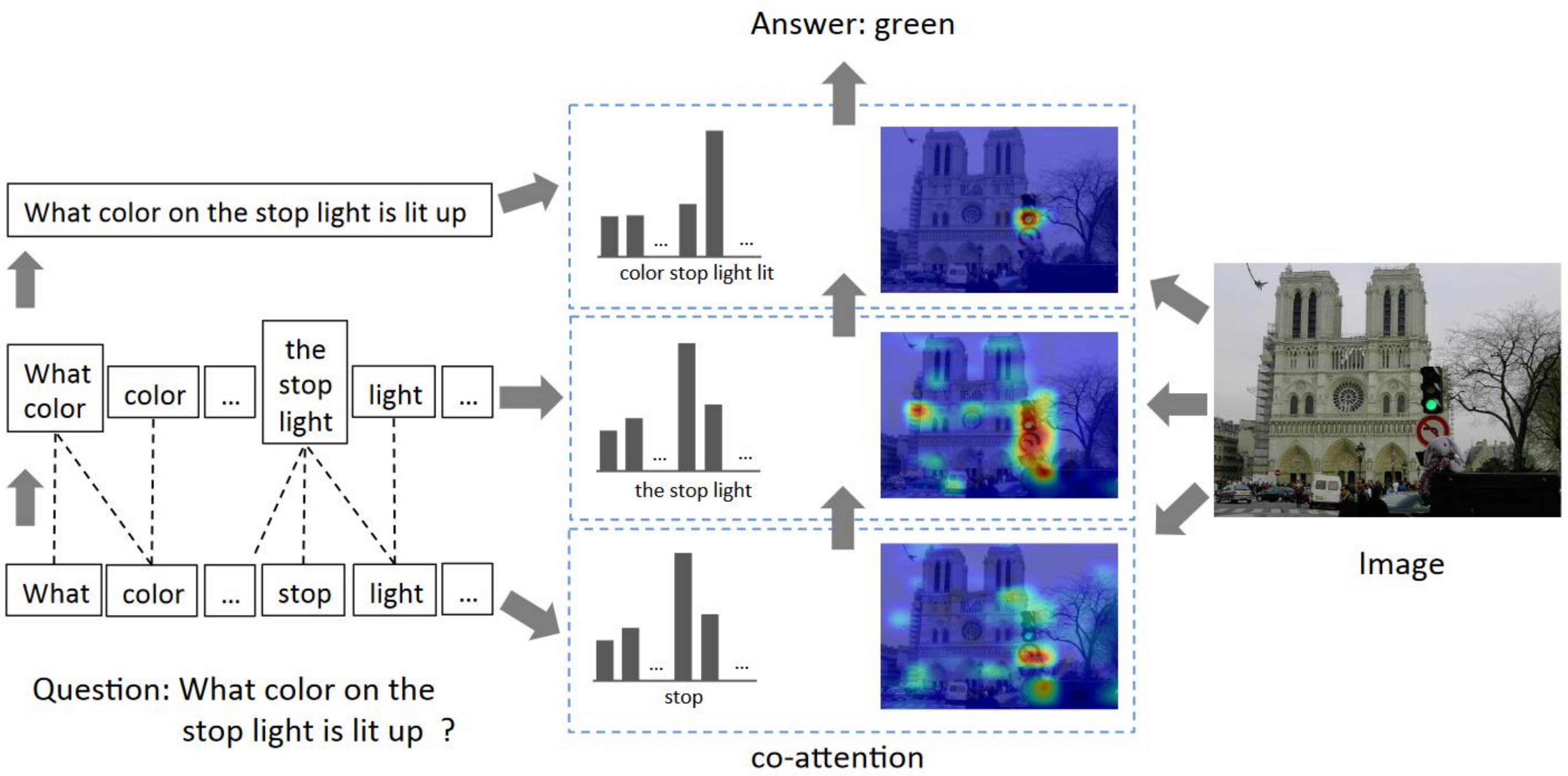
InterpNET(0)	this bird has wings that are white and has a yellow bill.
InterpNET(1)	this bird has wings that are black and has a white belly.
InterpNET(2)	this bird has wings that are grey and has a white belly.
InterpNET(3)	this bird has wings that are grey and has a white belly.
Captioning	this bird has a white belly and breast with a black wing and long hooked bill.

This is an image of a **Scott Oriole** because...

InterpNET(0)	this bird has a yellow belly and breast with a black superciliary and white wingbars.
InterpNET(1)	this bird has a black crown, black primaries, and a yellow belly.
InterpNET(2)	this bird has wings that are black and has a yellow belly.
InterpNET(3)	this bird has a black crown, a black bill, and a black breast.
Captioning	this bird has a black crown, black primaries, and a white belly.

[Barratt2017] Examples of InterpNet text generations

Hierarchical Question-Image Attention for Visual Question Answering

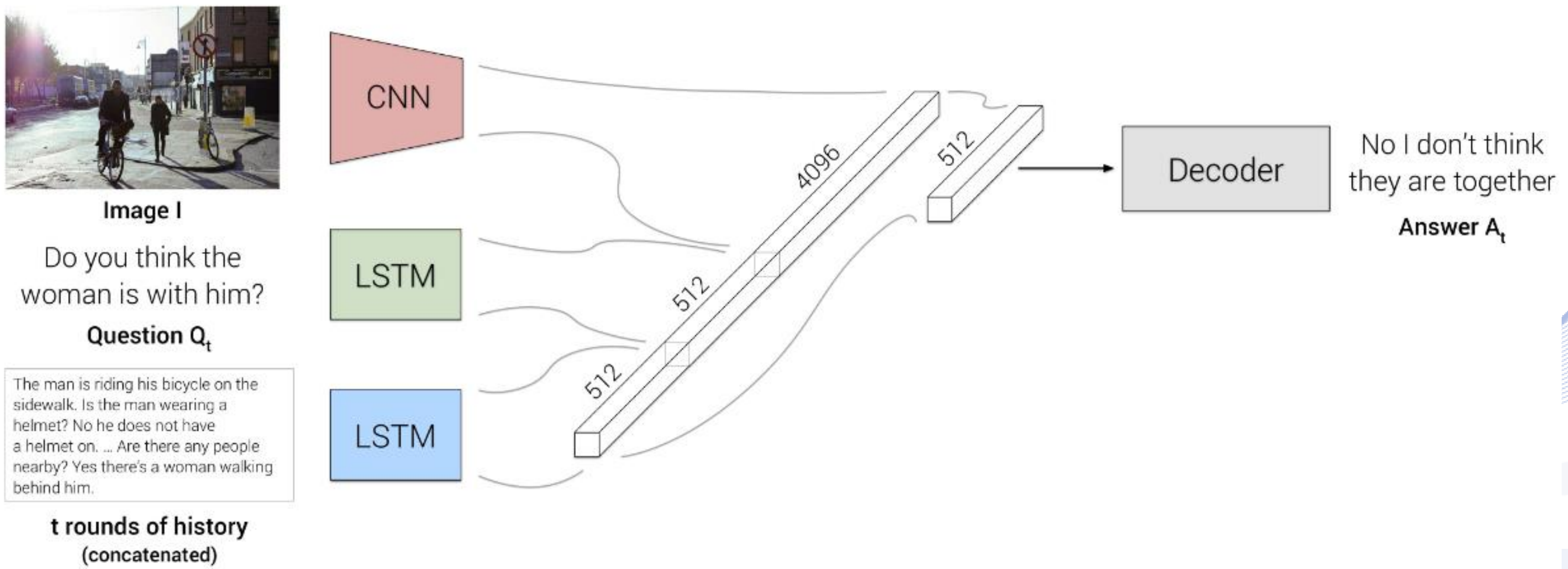


[Lu2016] Types of attention

Hierarchical Question-Image Attention for Visual Question Answering

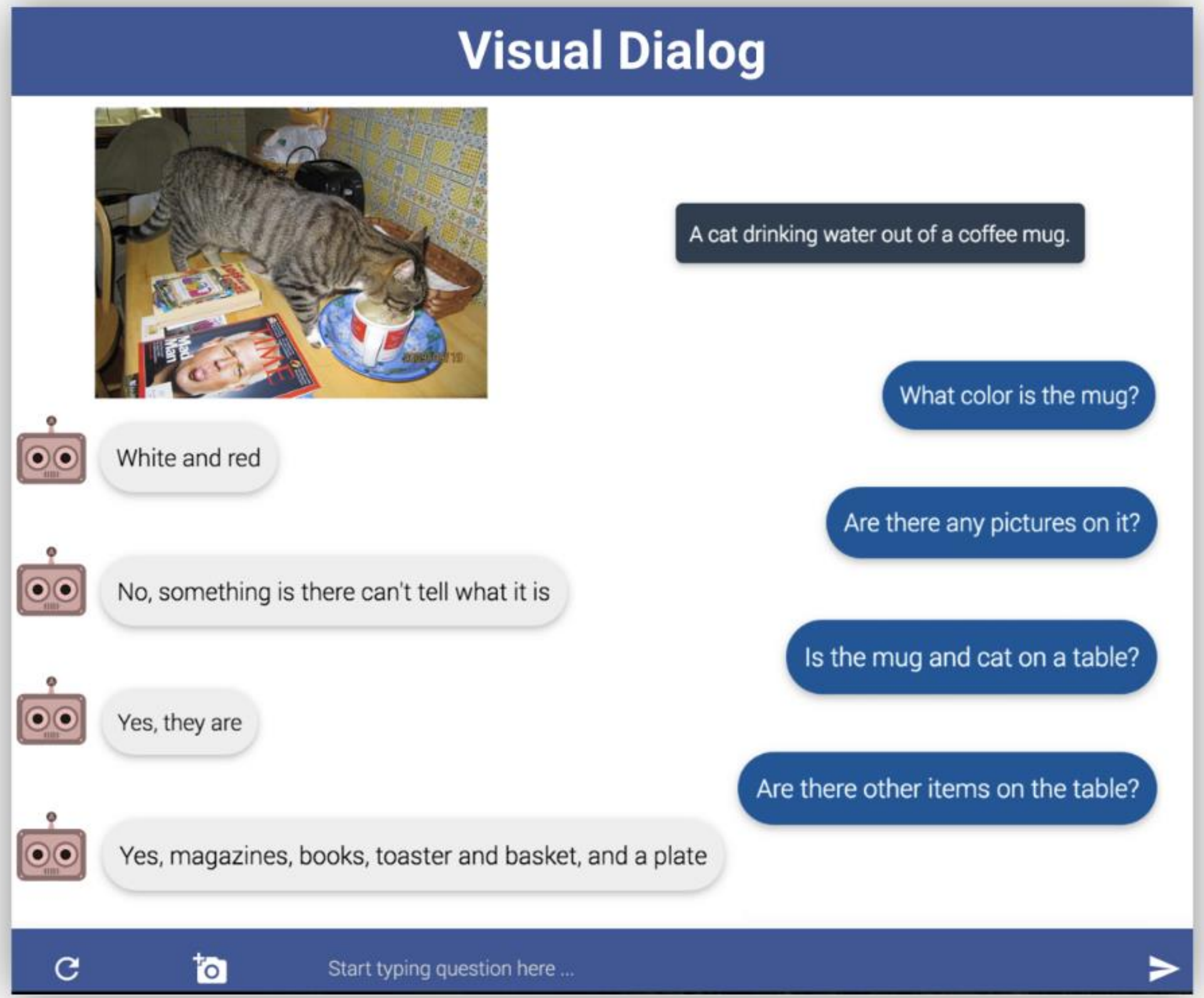
[Lu2016] Example of attention maps

Visual dialog



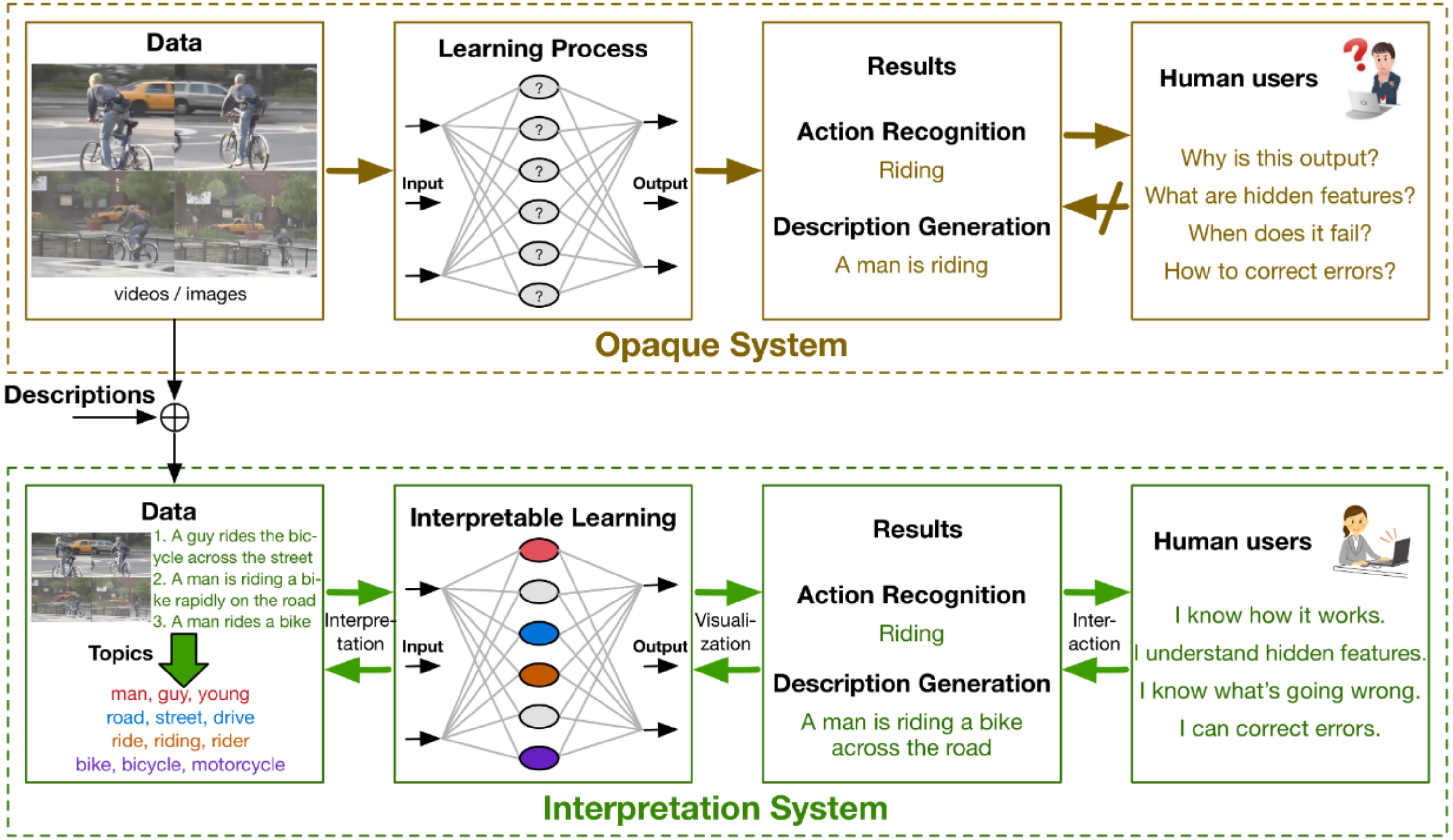
[Das2018] Visual dialog system to predict the answer

Visual dialog



[Das2018] Dialog of AI and human

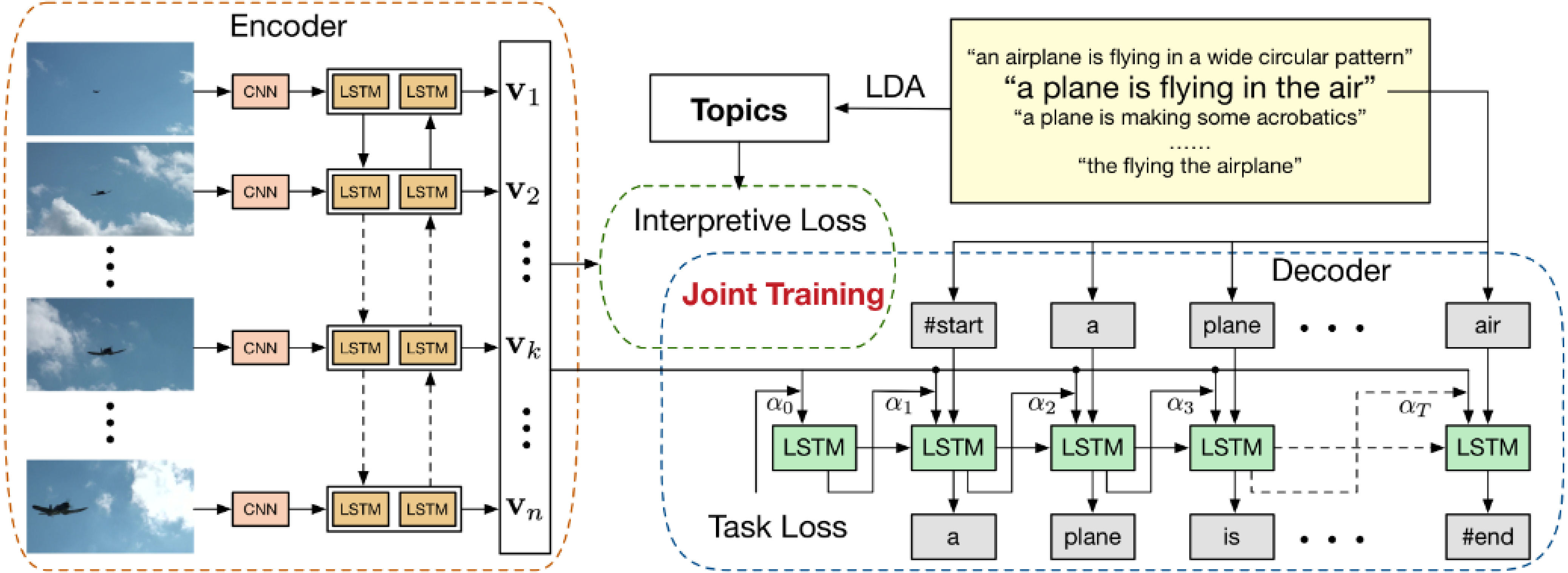
Semantic explanation of Deep Neural Networks



[Dong2017] Overview of the method

Artificial Intelligence & Information Analysis Lab

Semantic explanation of Deep Neural Networks

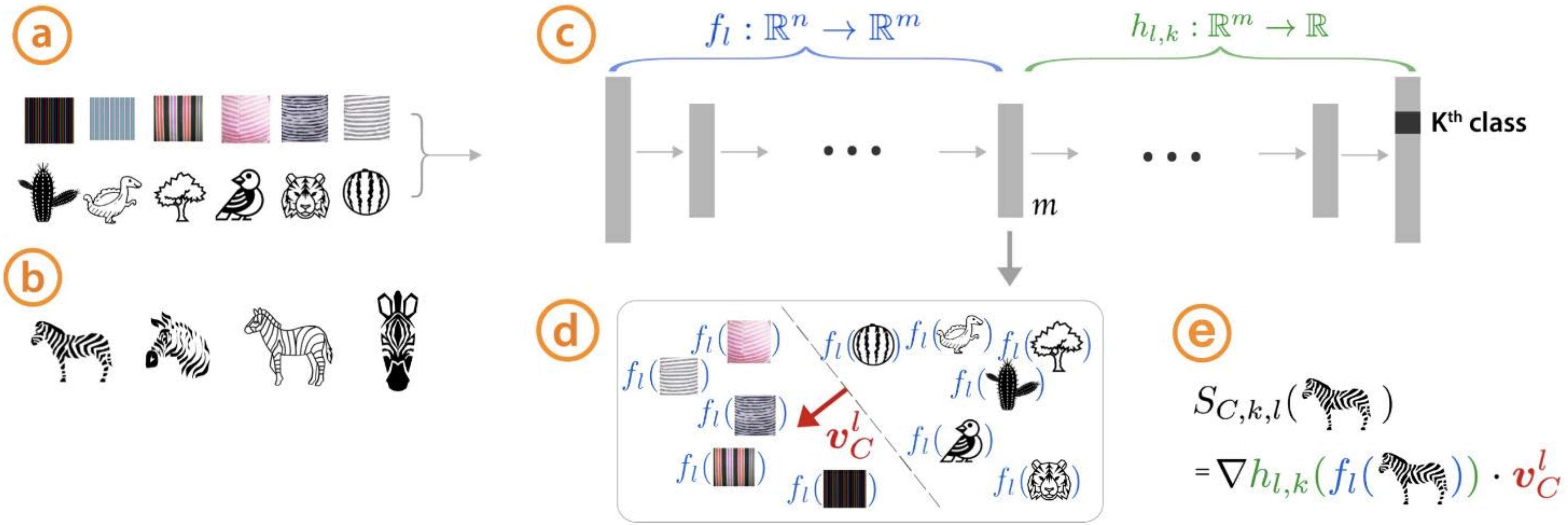


[Dong2017] Video captioning encoder's training process along with human descriptions decoder

Numerical Explanations

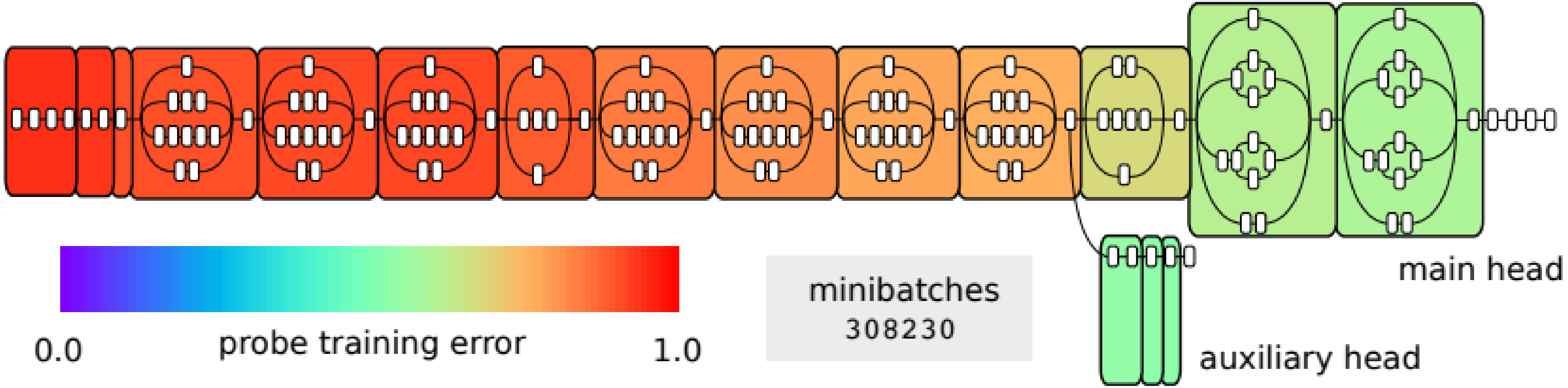
- Provide numerical outputs to interpet models
- Train classifiers that explain the model
  - Concept Activation Vectors
  - Linear classifier probes
  - LIME

Concept Activation Vectors (CAV)



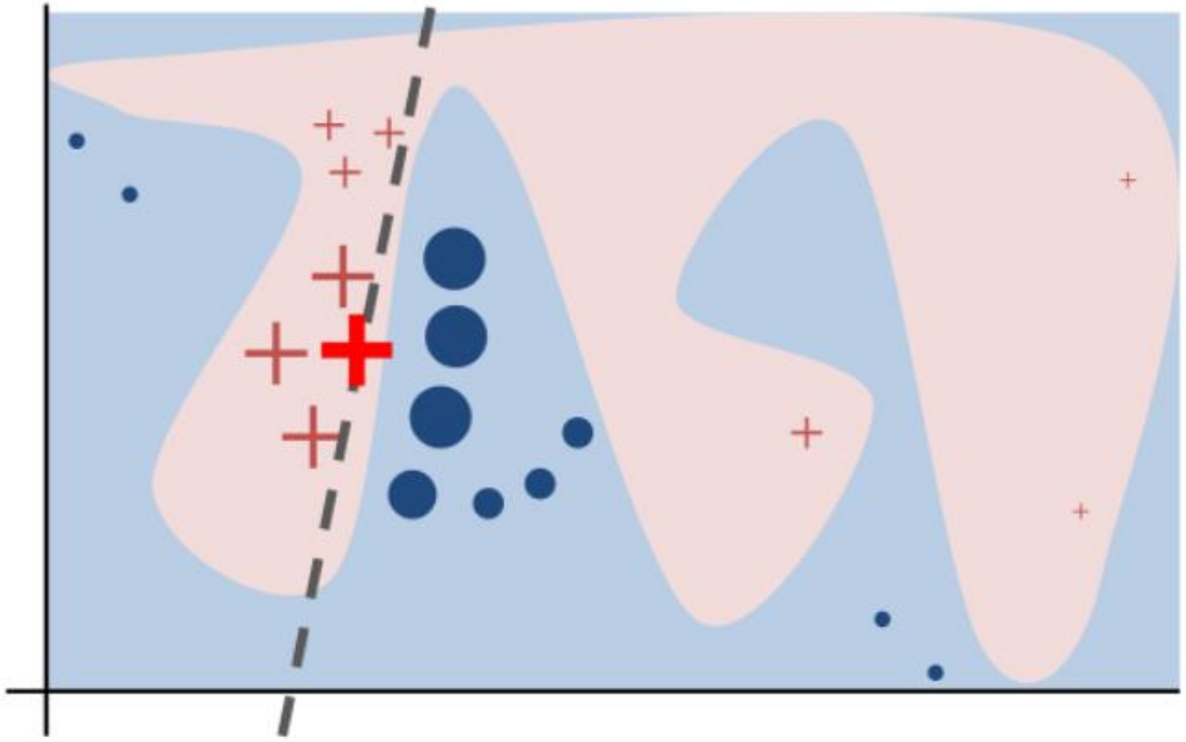
[Kim2018] Overview of CAVs learning process

Linear classifier probes



[Alain2016] Prediction error of each layer using probe

LIME (Local Interpretable Model-agnostic Explanations)



[Ribeiro2016] The data represented with the red cross is explained locally using the dashed line.

Applications

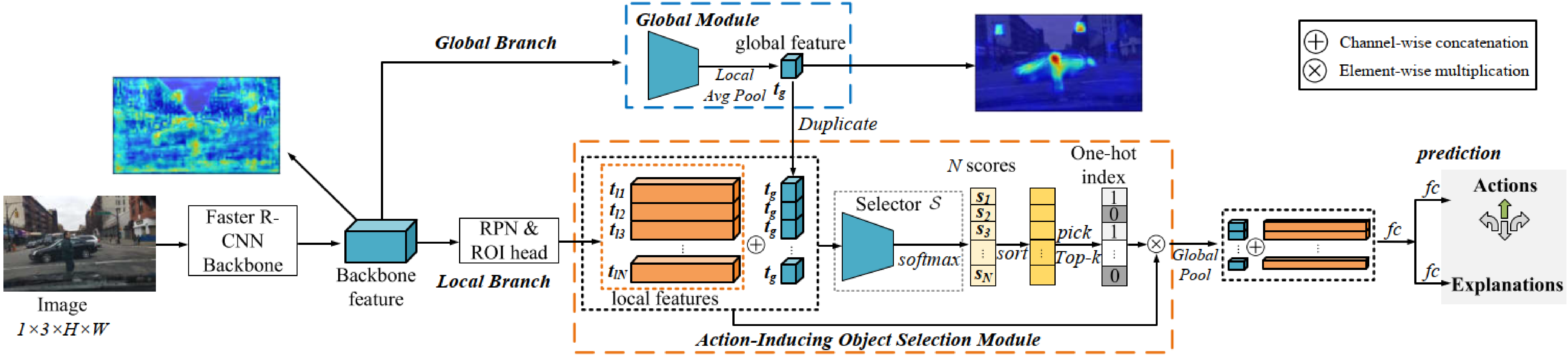
of explainable methods in important tasks

- Autonomous Driving
  - Advisable Learning for Self-driving Vehicles by Internalizing Observation-to-Action Rules
  - Explaining Autonomous Driving by Learning End-to-End Visual Attention
- Medical Applications
  - COVID detection

Explainable Object-induced Action Decision for Autonomous Vehicles

- Explainable autonomous driving architecture
- Global module to generate scene context
- Local module to predict actions and explanations
- Concatenation of the two modules to improve decision accuracy
- Predict next action such as (stop, turn left, go forward) and the explanations e.g. Stop, the traffic light is red , an obstacle is in front stop the car)

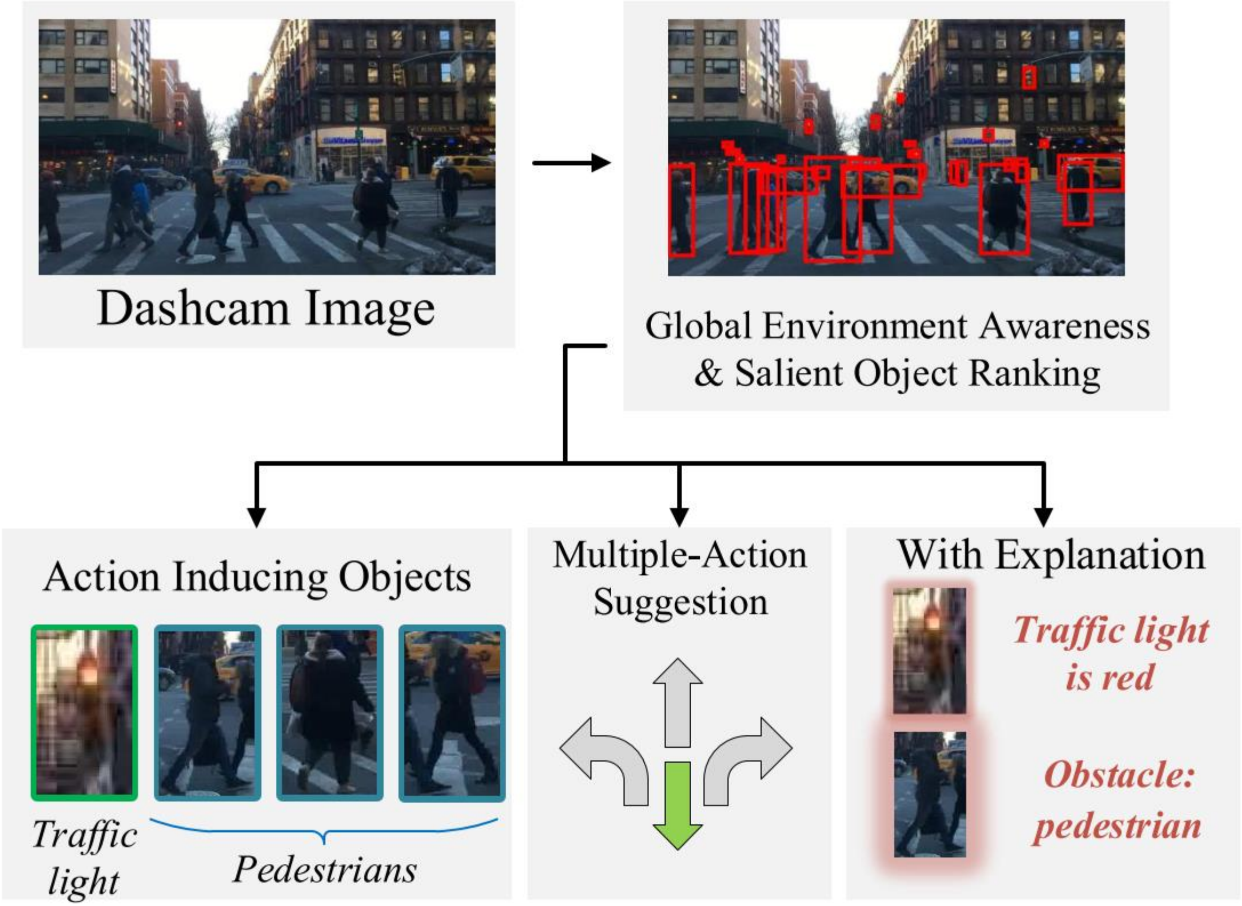
Explainable Object-induced Action Decision for Autonomous Vehicles



[Yu2020] Overview of the proposed explainable self-driving method

Artificial Intelligence & Information Analysis Lab

Explainable Object-induced Action Decision for Autonomous Vehicles



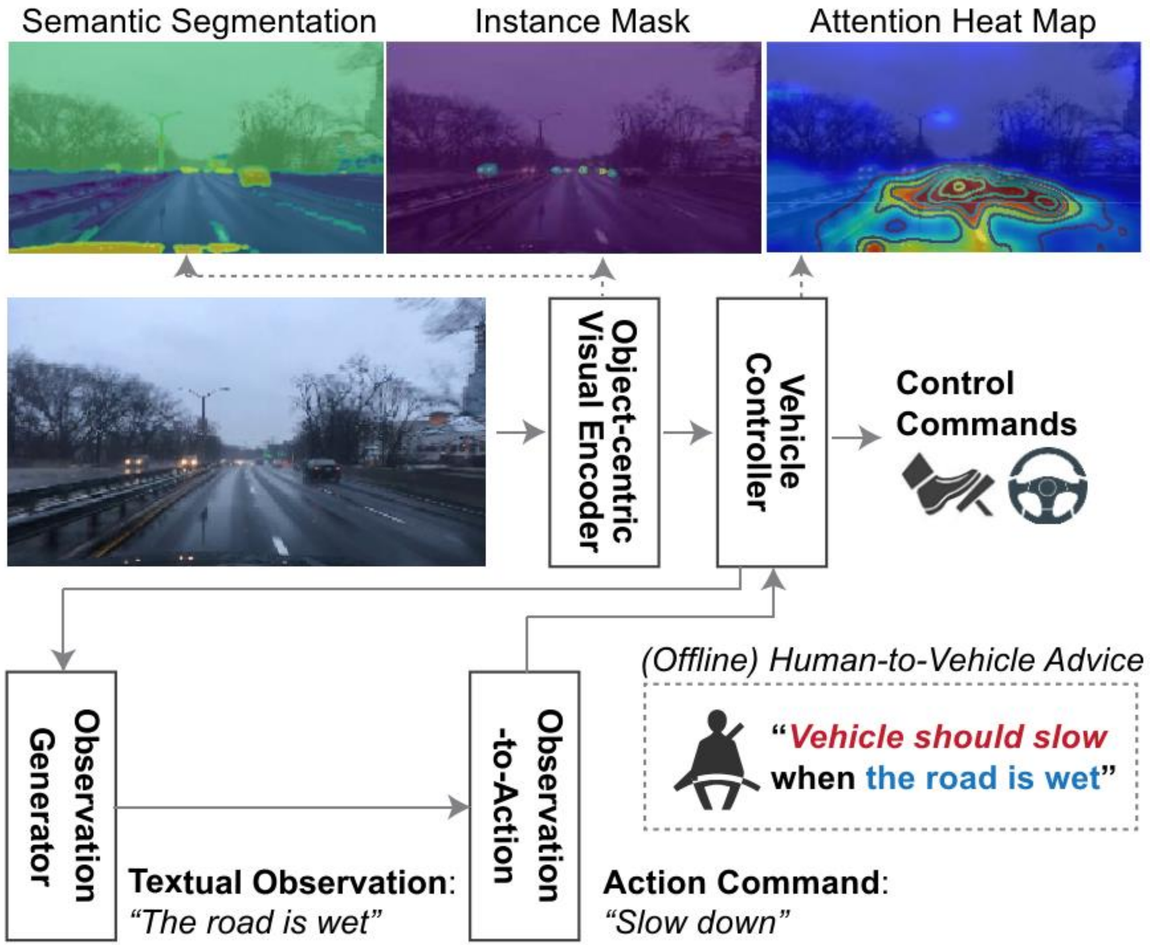
[Yu2020] Examples of action and explanations of the system

Artificial Intelligence & Information Analysis Lab

Advisable Learning for Self-driving Vehicles by Internalizing Observation-to-Action Rules

- Explainable self-driving system
- Human advice integrated on the architecture
- Visual (attention maps) and textual explanations (sentences)
- Main modules
  1. Object segmentation encoder
  2. Vehicle controller
  3. Observation generator
  4. Action generator

Advisable Learning for Self-driving Vehicles by Internalizing Observation-to-Action Rules

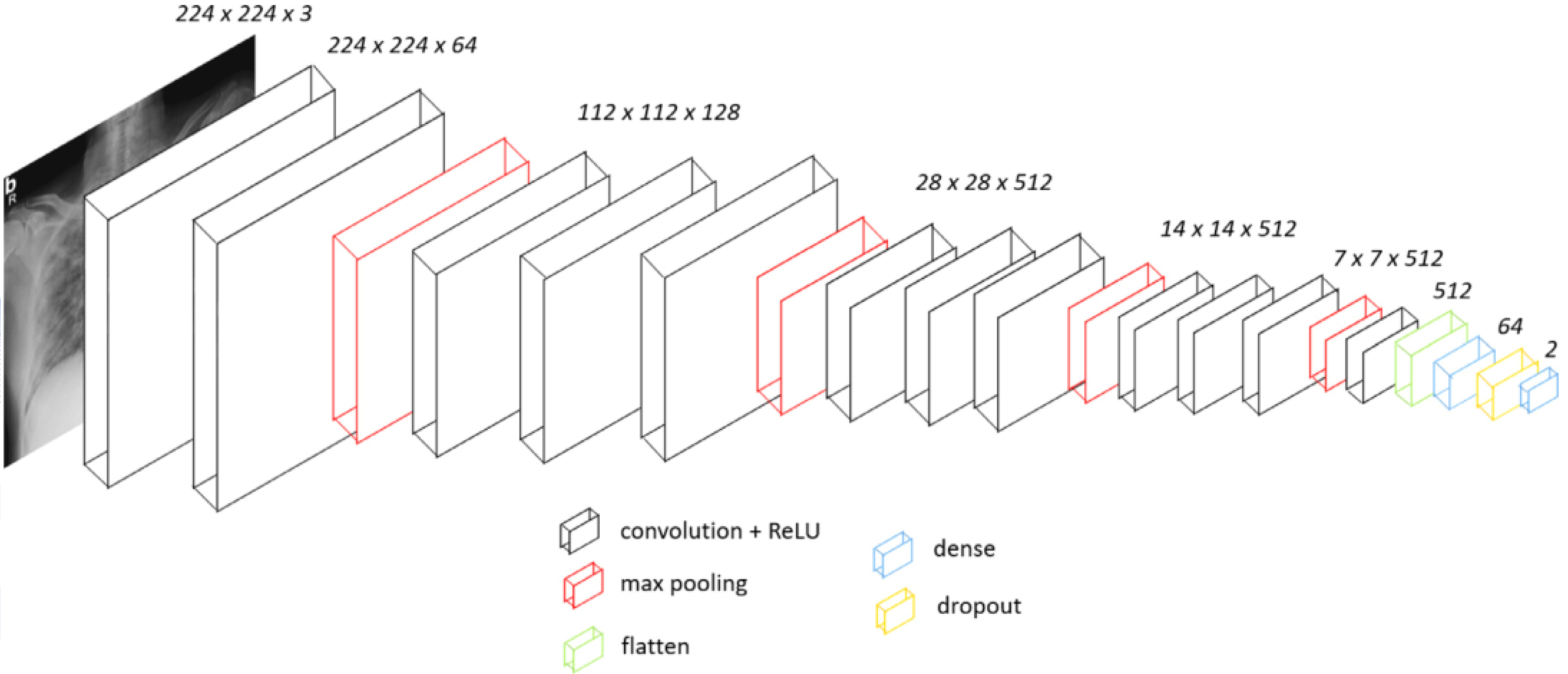


[Kim2020] Workflow of the self-driving system

Explainable Pulmonary Disease and COVID-19 Detection from X-rays

- Deep convolutional network VGG-16 is used to distinguish between healthy lungs, pneumonia or covid-19
- Grad-CAM provides interpets decisions by visualizing feature maps
- Localize areas responsible for the detection of pneumonia or covid-19

Explainable Pulmonary Disease and COVID-19 Detection from X-rays



[Brunese2020] CNN architecture

Explainable Pulmonary Disease and COVID-19 Detection from X-rays

[Brunese2020] Response maps of input chest x-ray

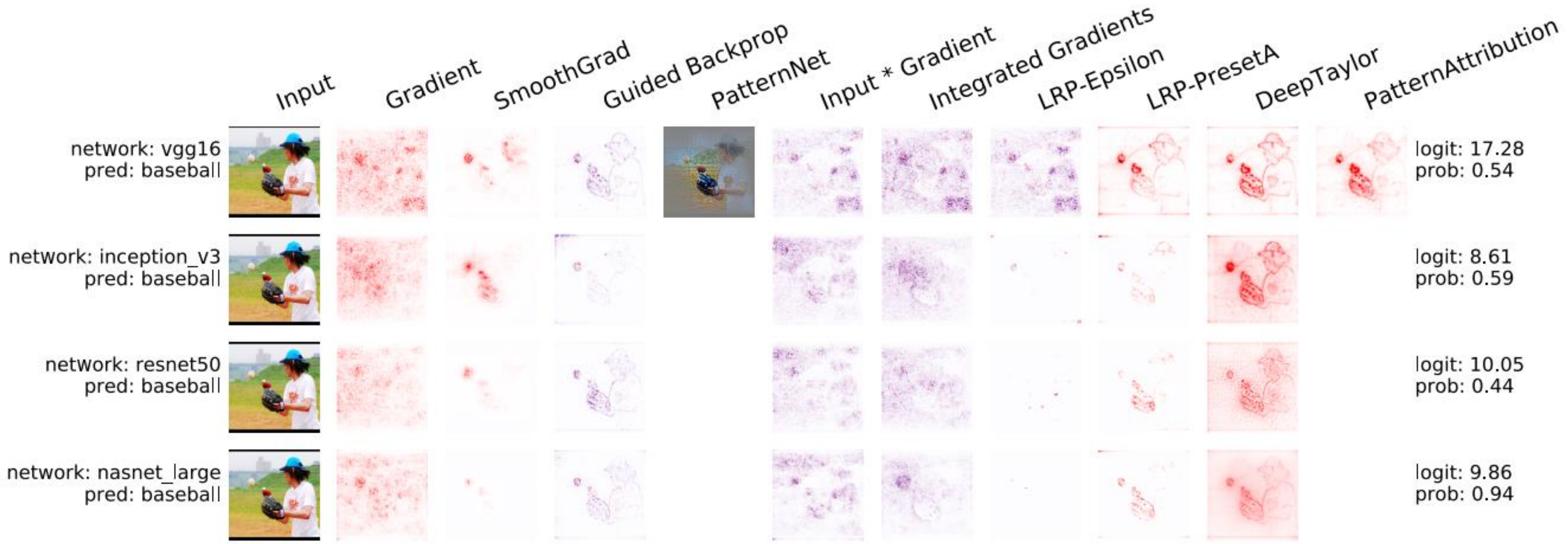
Explainable AI frameworks

with implemented explainable methods can be used for interpetation

- iNNvestigate Neural Networks
- ExplAIner
- InterpretML

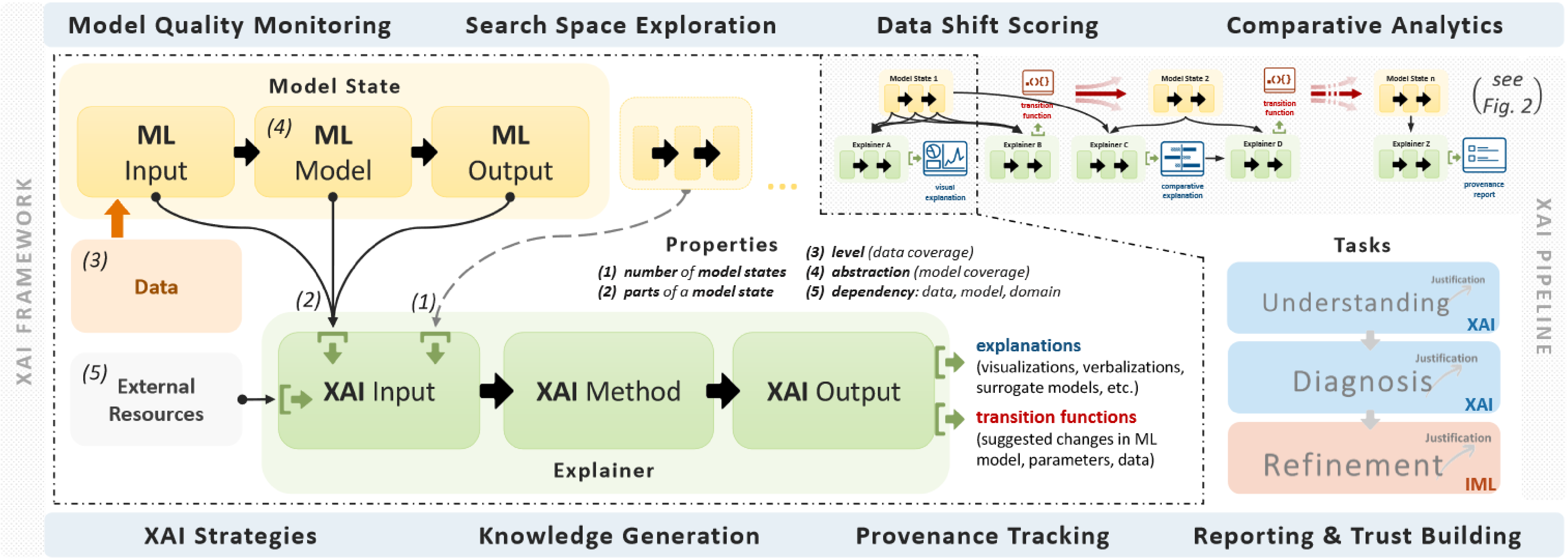
iNNvestigate Neural Networks

```
import innvestigate
model = create_a_keras_model()
analyzer = innvestigate.create_analyzer(''analyzer_name'', model)
analyzer.fit(X_train) # if needed
analysis = analyzer.analyze(X_test)
```



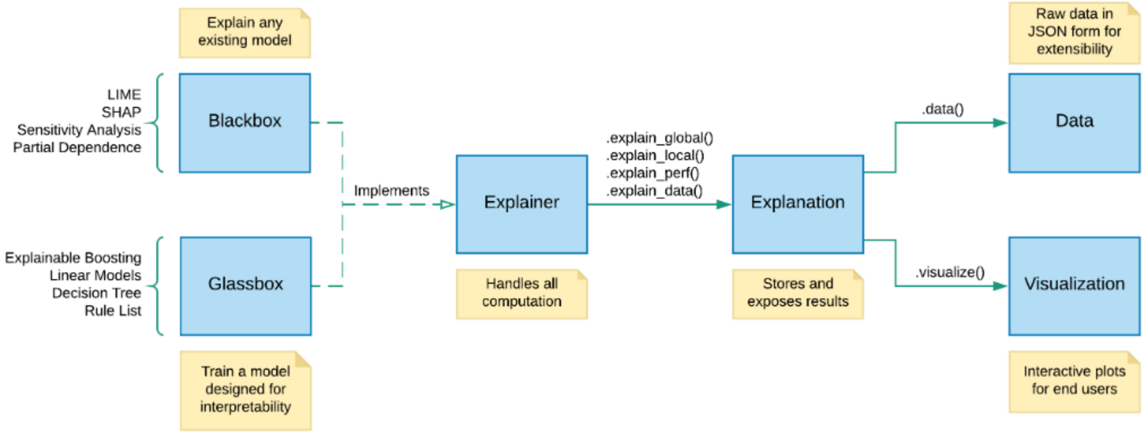
[Alber2019] Usage Example of iNNvestigate Neural Networks

ExplAIner



[Spinner2019] General approach of explAIner

InterpretML



Glassbox

```python
from interpret import show
from interpret.glassbox import LogisticRegression


clf = LogisticRegression()
clf.fit(X, y)

global_exp = clf.explain_global()
local_exp = clf.explain_local(X, y)

show([global_exp, local_exp])
```

Blackbox

```python
from interpret import show
from interpret.blackbox import PartialDependence
from sklearn.neural_network import MLPClassifier

blackbox = MLPClassifier()
blackbox.fit(X, y)

pdp = PartialDependence(blackbox.predict_proba, X)
global_exp = pdp.explain_global()

show(global_exp)
```

[Nori2019] Example of InterpetML usage

References

[Vilone2020] Vilone, Giulia, and Luca Longo. "Explainable Artificial Intelligence: a Systematic Review." *arXiv preprint arXiv:2006.00093* (2020).
[Tjoa2020] Tjoa, Erico, and Cuntai Guan. "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI." *IEEE Transactions on Neural Networks and Learning Systems* (2020).
[Adadi2018] Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6 (2018): 52138-52160.
[Zhou2016] Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921-2929. 2016.
[Selvaraju2017] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In *Proceedings of the IEEE international conference on computer vision*, pp. 618-626. 2017.
[Samek2016] Samek, Wojciech, Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, and Klaus-Robert Müller. "Interpreting the predictions of complex ml models by layer-wise relevance propagation." *arXiv preprint arXiv:1611.08191* (2016).
[Ribeiro2016] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144. 2016.

References

[Zhou2016] Zhou, Yanzhao, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. "Weakly supervised instance segmentation using class peak response." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3791-3800. 2018.
[Zeiler2014] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In *European conference on computer vision*, pp. 818-833. Springer, Cham, 2014.
[Dosovitskiy2015] Dosovitskiy, Alexey, and Thomas Brox. "Inverting convolutional networks with convolutional networks." *arXiv preprint arXiv:1506.02753* 4 (2015).
[Kumar2017] Kumar, Devinder, Alexander Wong, and Graham W. Taylor. "Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 36-44. 2017.
[Liu2019] Liu G, Zeng H, Gifford DK. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. BMC bioinformatics. 2019 Dec;20(1):1-4.
[Aubry2015] Aubry, Mathieu, and Bryan C. Russell. "Understanding deep features with computer-generated imagery." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2875-2883. 2015.
[Rauber2016] Rauber, Paulo E., Samuel G. Fadel, Alexandre X. Falcao, and Alexandru C. Telea. "Visualizing the hidden activity of artificial neural networks." *IEEE transactions on visualization and computer graphics* 23, no. 1 (2016): 101-110.

References

[Thiagarajan2016] Thiagarajan, Jayaraman J., Bhavya Kailkhura, Prasanna Sattigeri, and Karthikeyan Natesan Ramamurthy. "Treeview: Peeking into deep neural networks via feature-space partitioning." *arXiv preprint arXiv:1611.07429* (2016).

[Maaten2008] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9, no. Nov (2008): 2579-2605.

[Karpathy2015] Karpathy, Andrej, Justin Johnson, and Li Fei-Fei. "Visualizing and understanding recurrent networks." *arXiv preprint arXiv:1506.02078* (2015).

[Barratt2017] Barratt, Shane. "Interpnet: Neural introspection for interpretable deep learning." *arXiv preprint arXiv:1710.09511* (2017).

[Lu2016] Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical question-image co-attention for visual question answering." In *Advances in neural information processing systems*, pp. 289-297. 2016.

[Dong2017] Dong, Yinpeng, Hang Su, Jun Zhu, and Bo Zhang. "Improving interpretability of deep neural networks with semantic information." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4306-4314. 2017

[Das2018] Das, A., S. Kottur, K. Gupta, A. Singh, D. Yadav, S. Lee, J. Moura, D. Parikh, and D. Batra. "Visual Dialog." IEEE transactions on pattern analysis and machine intelligence (2018).

References

[Kim2018] Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." In *International conference on machine learning*, pp. 2668-2677. PMLR, 2018.

[Alain2016] Alain, Guillaume, and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes." *arXiv preprint arXiv:1610.01644* (2016).

[Li2020] Li, Liangzhi, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. "SCOUTER: Slot attention-based classifier for explainable image recognition." arXiv preprint arXiv:2009.06138 (2020).

[Oramas2019] Oramas Mogrovejo, J. A., Wang, K., & Tuytelaars, T. (2019). Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In https://iclr. cc/Conferences/2019/AcceptedPapersInitial. openReview.

[Kim2020] Kim, Jinkyu, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. "Advisable Learning for Self-Driving Vehicles by Internalizing Observation-to-Action Rules." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9661-9670. 2020.

[Yu2020] Xu, Yiran, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. "Explainable Object-induced Action Decision for Autonomous Vehicles." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9523-9532. 2020.

References

[Brunese2020] Brunese, Luca, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays." Computer Methods and Programs in Biomedicine 196 (2020): 105608.

[Alber2019] Alber, Maximilian, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. "iNNvestigate neural networks!." J. Mach. Learn. Res. 20, no. 93 (2019): 1-8.

[Spinner2019] Spinner, Thilo, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. "explAIner: A visual analytics framework for interactive and explainable machine learning." IEEE transactions on visualization and computer graphics 26, no. 1 (2019): 1064-1074.

[Nori2019] Nori, Harsha, Samuel Jenkins, Paul Koch, and Rich Caruana. "InterpretML: A Unified Framework for Machine Learning Interpretability." arXiv preprint arXiv:1909.09223 (2019).

Bibliography

[Vilone2020] Vilone, Giulia, and Luca Longo. "Explainable Artificial Intelligence: a Systematic Review." *arXiv preprint arXiv:2006.00093* (2020).
[Tjoa2020] Tjoa, Erico, and Cuntai Guan. "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI." *IEEE Transactions on Neural Networks and Learning Systems* (2020).
[Adadi2018] Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6 (2018): 52138-52160.
[Zhou2016] Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921-2929. 2016.
[Selvaraju2017] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In *Proceedings of the IEEE international conference on computer vision*, pp. 618-626. 2017.
[Samek2016] Samek, Wojciech, Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, and Klaus-Robert Müller. "Interpreting the predictions of complex ml models by layer-wise relevance propagation." *arXiv preprint arXiv:1611.08191* (2016).
[Ribeiro2016] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144. 2016.

Bibliography

[Zhou2016] Zhou, Yanzhao, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. "Weakly supervised instance segmentation using class peak response." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3791-3800. 2018.

[Zeiler2014] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In *European conference on computer vision*, pp. 818-833. Springer, Cham, 2014.

[Dosovitskiy2015] Dosovitskiy, Alexey, and Thomas Brox. "Inverting convolutional networks with convolutional networks." *arXiv preprint arXiv:1506.02753* 4 (2015).

[Kumar2017] Kumar, Devinder, Alexander Wong, and Graham W. Taylor. "Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 36-44. 2017.

[Liu2019] Liu G, Zeng H, Gifford DK. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. BMC bioinformatics. 2019 Dec;20(1):1-4.

[Aubry2015] Aubry, Mathieu, and Bryan C. Russell. "Understanding deep features with computer-generated imagery." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2875-2883. 2015.

[Rauber2016] Rauber, Paulo E., Samuel G. Fadel, Alexandre X. Falcao, and Alexandru C. Telea. "Visualizing the hidden activity of artificial neural networks." *IEEE transactions on visualization and computer graphics* 23, no. 1 (2016): 101-110.

Bibliography

[Thiagarajan2016] Thiagarajan, Jayaraman J., Bhavya Kailkhura, Prasanna Sattigeri, and Karthikeyan Natesan Ramamurthy. "Treeview: Peeking into deep neural networks via feature-space partitioning." *arXiv preprint arXiv:1611.07429* (2016).

[Maaten2008] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9, no. Nov (2008): 2579-2605.

[Karpathy2015] Karpathy, Andrej, Justin Johnson, and Li Fei-Fei. "Visualizing and understanding recurrent networks." *arXiv preprint arXiv:1506.02078* (2015).

[Barratt2017] Barratt, Shane. "Interpnet: Neural introspection for interpretable deep learning." *arXiv preprint arXiv:1710.09511* (2017).

[Lu2016] Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical question-image co-attention for visual question answering." In *Advances in neural information processing systems*, pp. 289-297. 2016.

[Dong2017] Dong, Yinpeng, Hang Su, Jun Zhu, and Bo Zhang. "Improving interpretability of deep neural networks with semantic information." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4306-4314. 2017

[Das2018] Das, A., S. Kottur, K. Gupta, A. Singh, D. Yadav, S. Lee, J. Moura, D. Parikh, and D. Batra. "Visual Dialog." IEEE transactions on pattern analysis and machine intelligence (2018).

Bibliography

[Kim2018] Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." In *International conference on machine learning*, pp. 2668-2677. PMLR, 2018.

[Alain2016] Alain, Guillaume, and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes." *arXiv preprint arXiv:1610.01644* (2016).

[Li2020] Li, Liangzhi, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. "SCOUTER: Slot attention-based classifier for explainable image recognition." arXiv preprint arXiv:2009.06138 (2020).

[Oramas2019] Oramas Mogrovejo, J. A., Wang, K., & Tuytelaars, T. (2019). Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In https://iclr. cc/Conferences/2019/AcceptedPapersInitial. openReview.

[Kim2020] Kim, Jinkyu, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. "Advisable Learning for Self-Driving Vehicles by Internalizing Observation-to-Action Rules." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9661-9670. 2020.

[Yu2020] Xu, Yiran, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. "Explainable Object-induced Action Decision for Autonomous Vehicles." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9523-9532. 2020.

Bibliography

[Brunese2020] Brunese, Luca, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays." Computer Methods and Programs in Biomedicine 196 (2020): 105608.

[Alber2019] Alber, Maximilian, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. "iNNvestigate neural networks!." J. Mach. Learn. Res. 20, no. 93 (2019): 1-8.

[Spinner2019] Spinner, Thilo, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. "explAIner: A visual analytics framework for interactive and explainable machine learning." IEEE transactions on visualization and computer graphics 26, no. 1 (2019): 1064-1074.

[Nori2019] Nori, Harsha, Samuel Jenkins, Paul Koch, and Rich Caruana. "InterpretML: A Unified Framework for Machine Learning Interpretability." arXiv preprint arXiv:1909.09223 (2019).

Bibliography

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[PIT2017] I. Pitas, “Digital video processing and analysis ” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television ” , Createspace/Amazon, 2013.
[NIK2000] N. Nikolaidis and I. Pitas, 3D Image Processing Algorithms, J. Wiley, 2000.
[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr