ANTHILL INSIDE 2018

Going beyond what and asking why

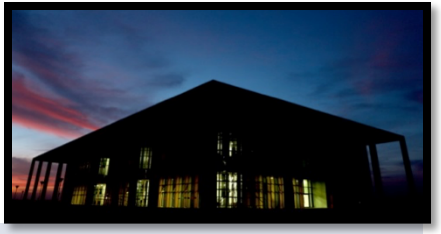
Explainability in Machine/Deep Learning

25 Jul 2018

Vineeth N Balasubramanian
Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

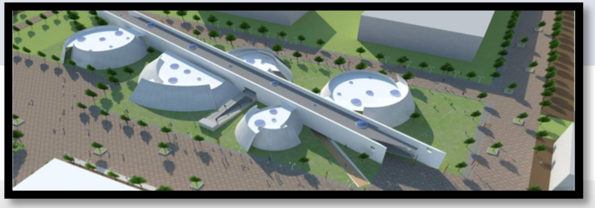
आई आई टी हैदराबाद
IIT Hyderabad

About IIT-H



- Started in Aug 2008
- 14 Departments covering all major engineering, sciences and humanities
- BTech, MTech, MDes, MPhil, MSc, PhD degrees offered
- ~ 2200 students (~ 50:50 undergrad: grad)
- Effective Oct 2015, functioning from the permanent campus

Next phase of buildings coming up



CSE @ IIT-H

- 20 faculty covering most areas of CS
- Opening/Closing JEE ranks this year: 450/770 (improving each year)
- Several projects with Govt, academia and industry
- Several student and faculty awards
  - S N Bose/Viterbi Fellowships, Best Paper Awards, GSoCs, etc
- Come visit us!

Our Group's Research

Algorithmic

- Non-convex optimization for DL*
- Explainable ML§
- Deep generative models⌘
- Deep graph representations

Applied

- Recognition of Expressions, Poses, Gestures, Actions
- Vision on UAVs/Drones
- Computer Vision for Agriculture
- *etc*

* On Noise and Optimality in Neural Networks, **ICML 2018 Workshops**

⌘Adversarial Data Programming, **CVPR 2018**

§ Grad-CAM++: Generalized Gradient-based Visual Explanations for Convolutional Networks, **WACV 2018**

⌘Attentive Semantic Video Generation using Captions, **ICCV 2017**, **ACM MM 2017**

Today

Outline

- Explainability in ML: An Overview
- Visual Interpretability of CNNs
- Looking Forward: Directions

Note

- Semi-technical Talk
- Intermediate-level
- Basic background in deep learning assumed
- Focus on Computer Vision

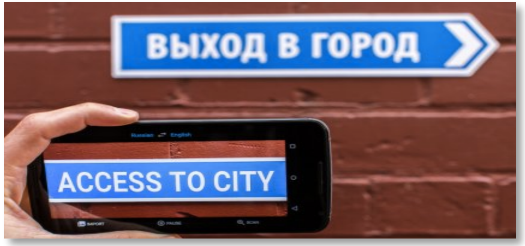
Machine Learning Successes

- Science (Astronomy, neuroscience, medical imaging, bio-informatics)
- Environment (Energy, climate, weather, resources)
- Retail (Intelligent stock control, demographic store placement)
- Manufacturing (Intelligent control, automated monitoring, detection methods)
- Security (Intelligent smoke alarms, fraud detection)
- Marketing (Promotions, ...)
- Management (Scheduling, timetabling)
- Finance (Credit scoring, risk analysis...)
- Web data (Information retrieval, information extraction, ...)

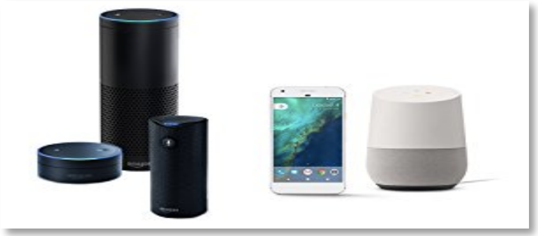
The Deep Learning Revolution

Explosive Growth in Recent Years

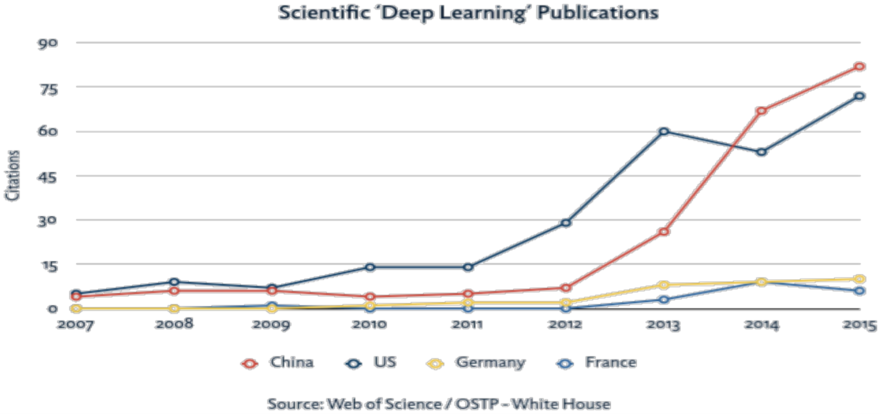
Vision

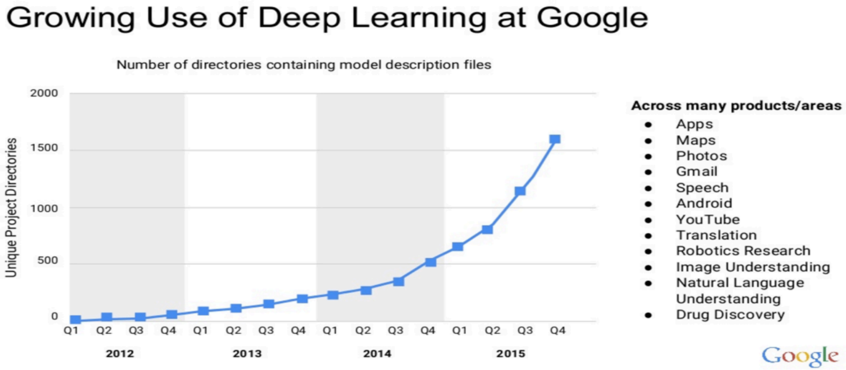


Text



Speech





Characterizing Today's ML Applications

Input → **ML Model** → Output

What is the product relevant to the user? What is the sentiment of this tweet? What are the objects in this image?

What is X?

- Cost of a bad decision is low
  - E.g. Bad recommendation => Bad movie => 500 Rs + 3 hours loss
- Accuracy is all-important
  - "Why" does not matter, as long as revenue is optimized
- Highly one-dimensional
  - Only one (or two more) simple mathematical metric(s) matter(s)

Where ML is yet to fulfill its promise

- Complex real-world systems
  - Risk-sensitive systems
    - E.g. Medical diagnosis, Financial modeling/prediction
  - Safety-critical systems
    - E.g. Cockpit decision support

Characterizing these applications

- Cost of a bad decision can be very high
- Accuracy is not the only objective
- Need for a multi-dimensional perspective

What then do we need in ML?

- Human-understandable rationale in decision-making
- Trust/confidence in a system
- Compliance with ethical principles
- Enhanced control and robustness
- Openness of discovery and scientific research

General
Data
Protection
Regulation

"By 2018, half of business ethics violations will occur through improper use of big data analytics." (Gartner)

https://www.gartner.com/newsroom/id/3144217

Explainability in ML

What does it mean?

- *Keywords:* Explainability, Trust, Interpretability

Interpretability	Explainability
Comprehending what a model did or might have done	Summarizing the reasons for neural network behavior, gaining trust of users, producing insights or causes of decisions

Gilpin et al, Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning, arXiv, Jun 2018

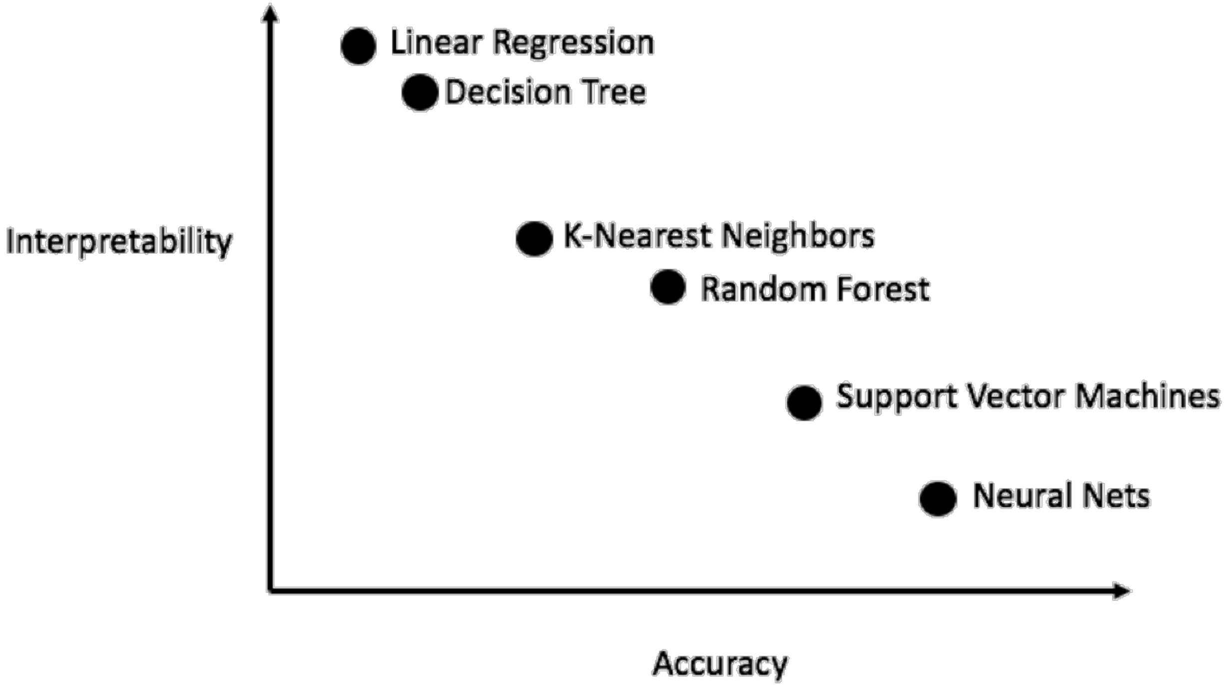
Today's ML Models

Accuracy vs Interpretability Tradeoff

67%

of the businesses leaders taking part in PwC's 2017 Global CEO Survey believe that AI and automation will impact negatively on stakeholder trust levels in their industry in the next five years.

Source: PwC 20th Annual CEO Survey, 2017

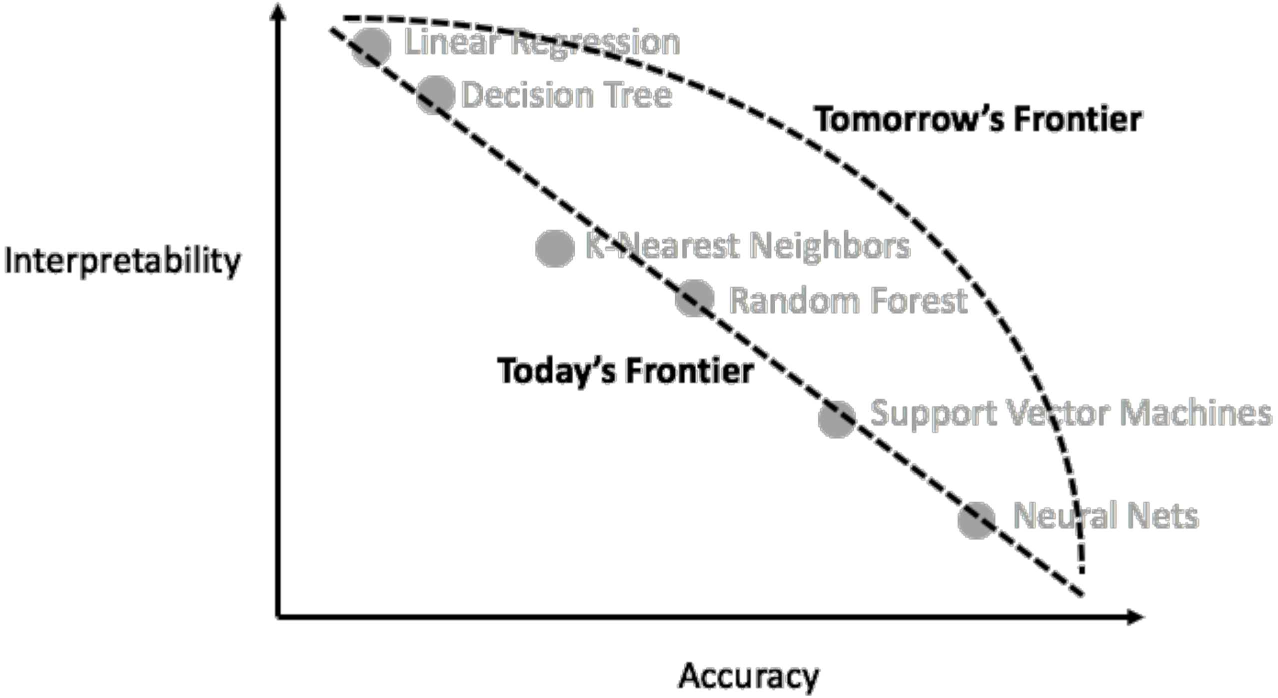


Today's ML Models

67%

of the businesses leaders taking part in PwC's 2017 Global CEO Survey believe that AI and automation will impact negatively on stakeholder trust levels in their industry in the next five years.

Source: PwC 20th Annual CEO Survey, 2017



Explainability in ML: What has been done?

Processing/Input-output Analysis	Explanation-Producing	Representation Analysis
• Linear Proxy Methods • Decision Trees • Saliency Maps • Automatic Rule Extraction	• Scripted Conversations • Attention-based • Disentangling Representations	• Role of Layers • Role of Neurons • Role of Vectors

Gilpin et al, Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning, arXiv, Jun 2018

Explainability in ML: What has been done?

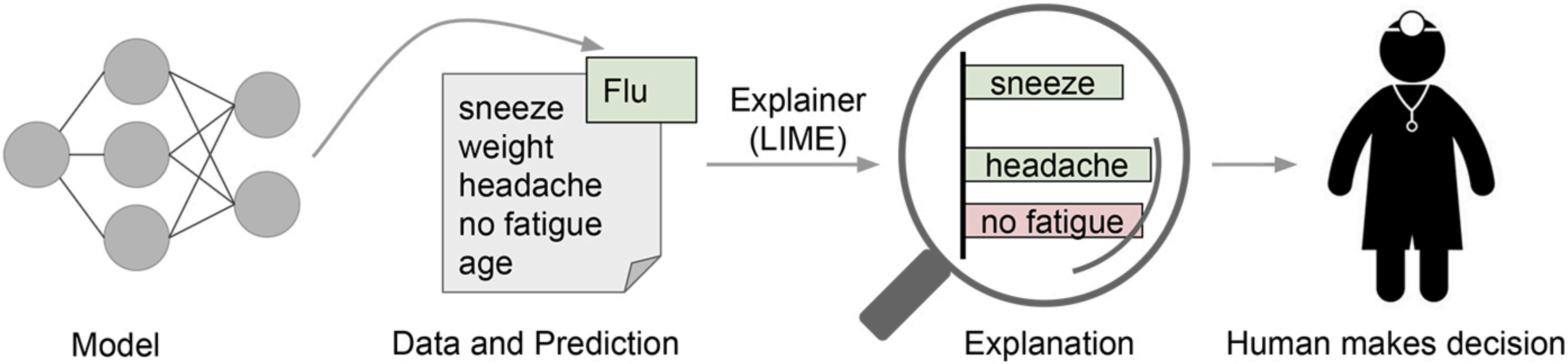
LIME, KDD 2016

Processing/Input-output Analysis	Explanation-Producing	Representation Analysis
• Linear Proxy Methods • Decision Trees • Saliency Maps • Automatic Rule Extraction	• Scripted Conversations • Attention-based • Disentangling Representations	• Role of Layers • Role of Neurons • Role of Vectors

LRP, PlosOne 2015; CAM, CVPR 2016, Grad-CAM, ICCV 2017; DeepLIFT, ICML 2017

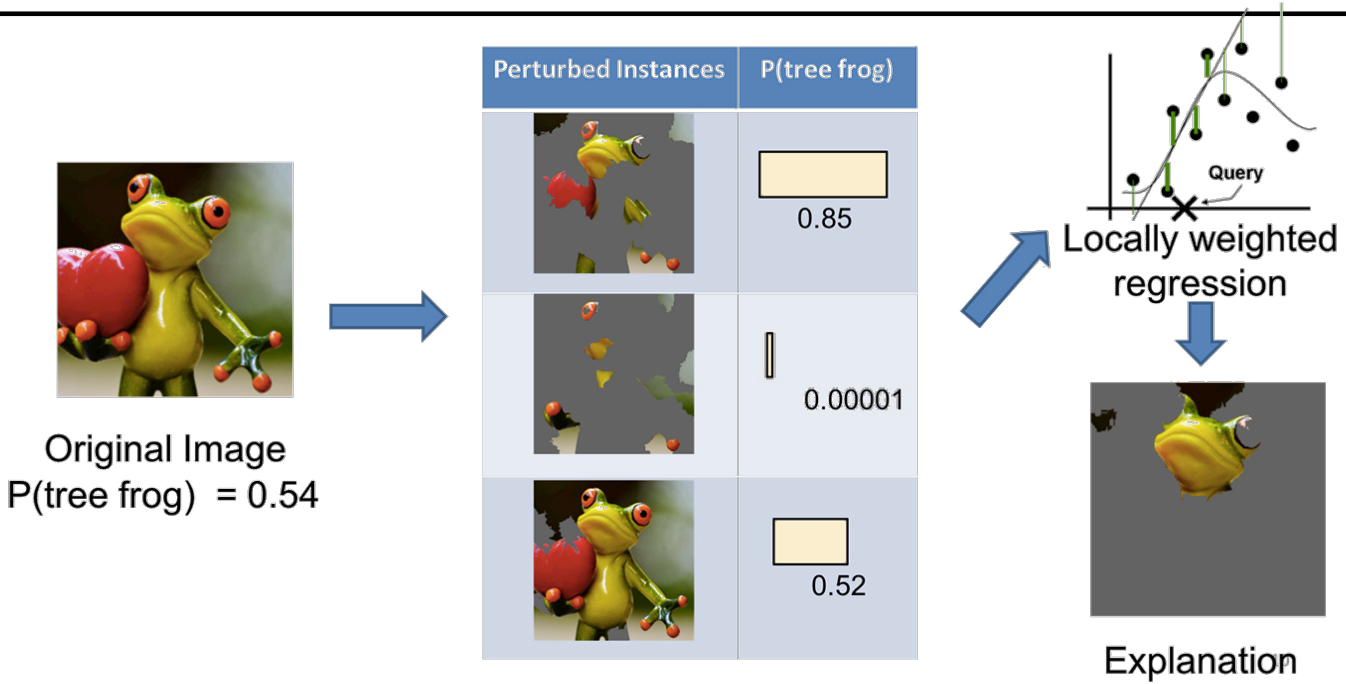
Gilpin et al, Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning, arXiv, Jun 2018

LIME: Local Interpretable Model-Agnostic Explanations



Ribiero et al, Why Should I Trust You? Explaining the Predictions of Any Classifier, KDD 2016
Image Credit: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

LIME: Local Interpretable Model-Agnostic Explanations



Popularly used

Code: https://github.com/marcotcr/lime

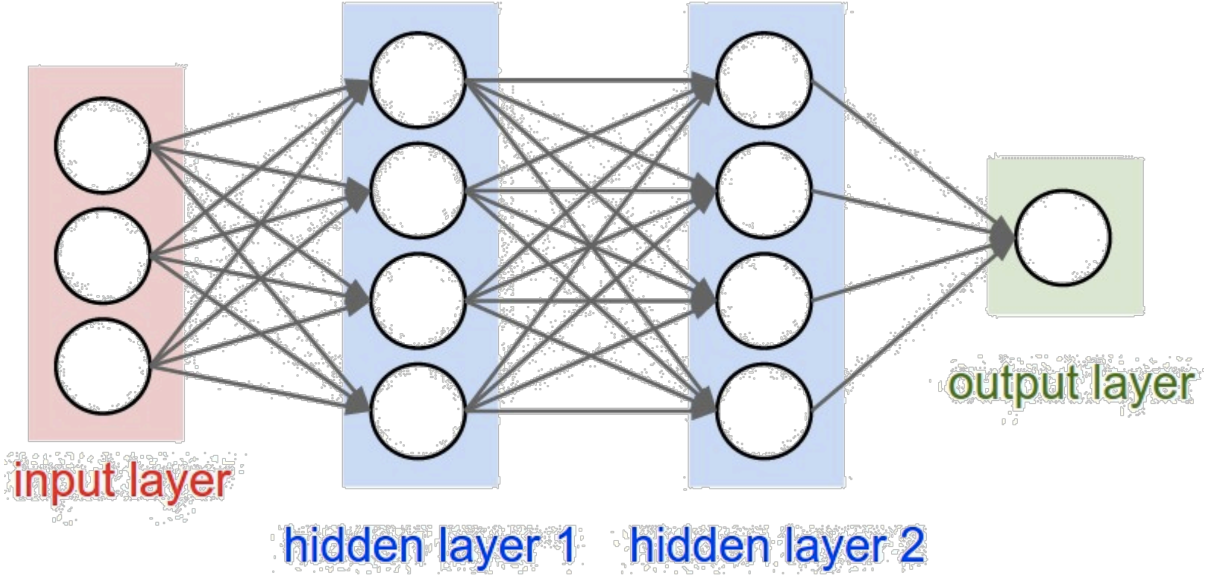
Ribiero et al, Why Should I Trust You? Explaining the Predictions of Any Classifier, KDD 2016
Image Credit: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

Today

Outline

- Explainability in ML: An Overview
- Visual Interpretability of CNNs
- Looking Forward: Directions

Basic Neural Networks

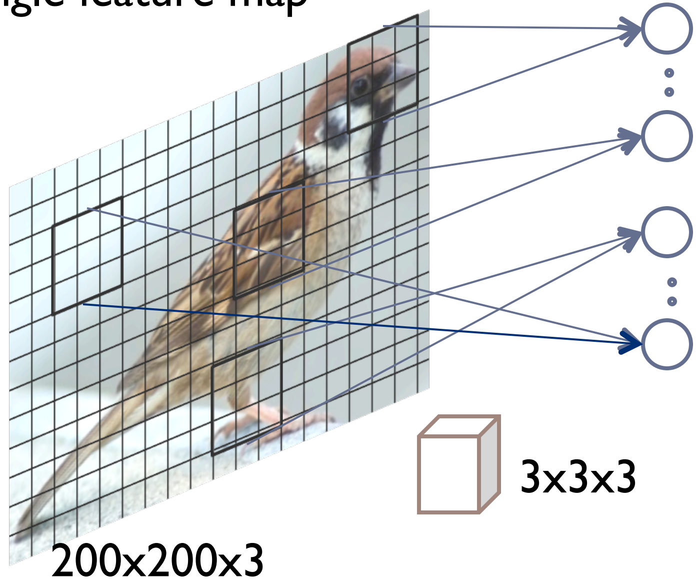


Trained using Backpropagation + Gradient Descent, given a loss function for a particular application

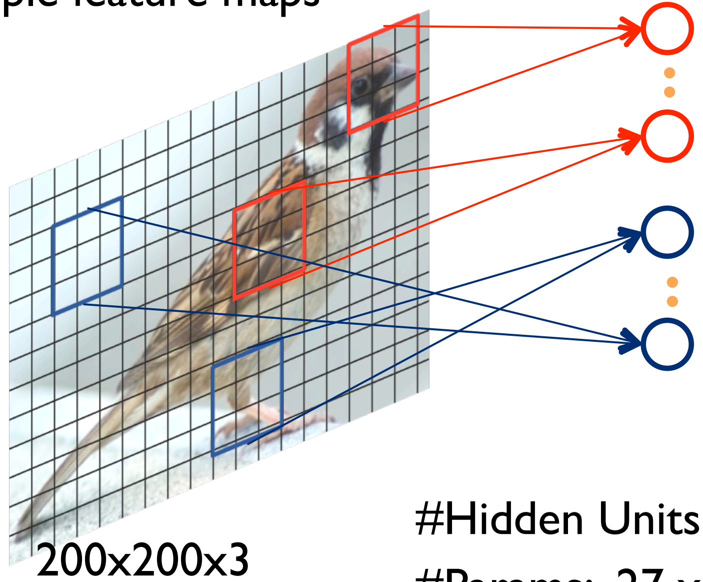
Convolutional Neural Networks

A Brief Review

Convolutional layer with single feature map



Convolutional layer with multiple feature maps



- Sharing of parameters
- Preserving locality of pixel dependencies

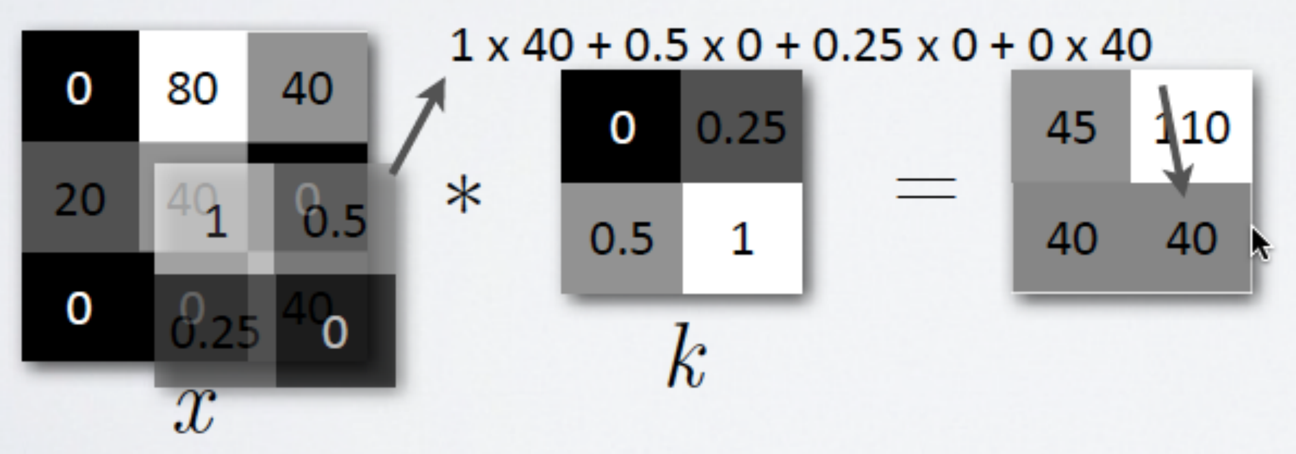
#Hidden Units: 120,000

#Params: 27 x #Feature Maps

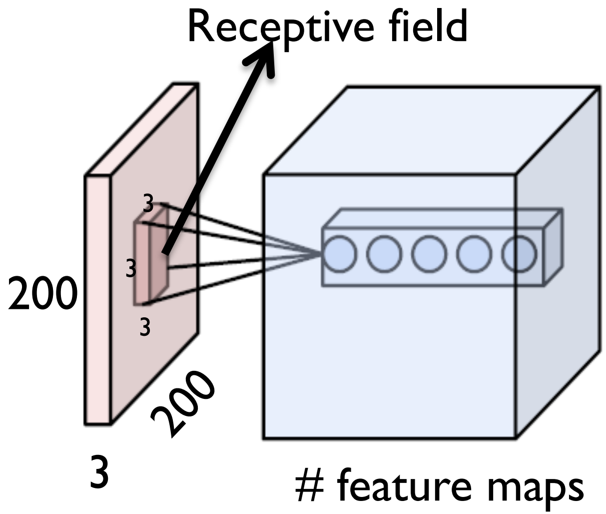
Convolution

- The convolution of an image *x* with a kernel *k* is computed as:

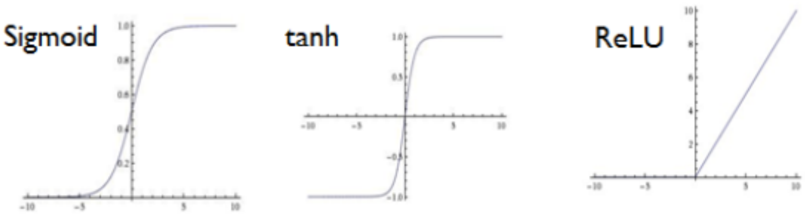
$$(x * k)_{ij} = \sum_{pq} x_{i+p,j+q} \, k_{r\text{-}p,r\text{-}q}$$



Convolutional Layer



Activation Functions

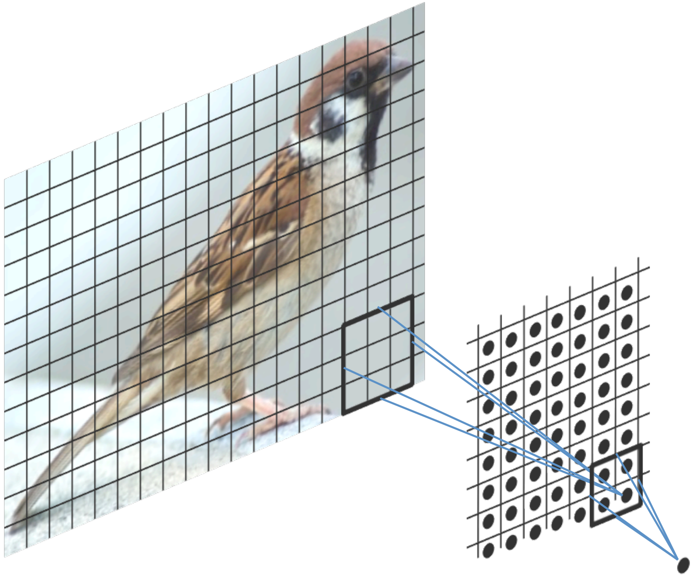


$$y = \frac{1}{1+e^{-x}}$$

$$y = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$y = max(0, x)$$

Pooling Layer



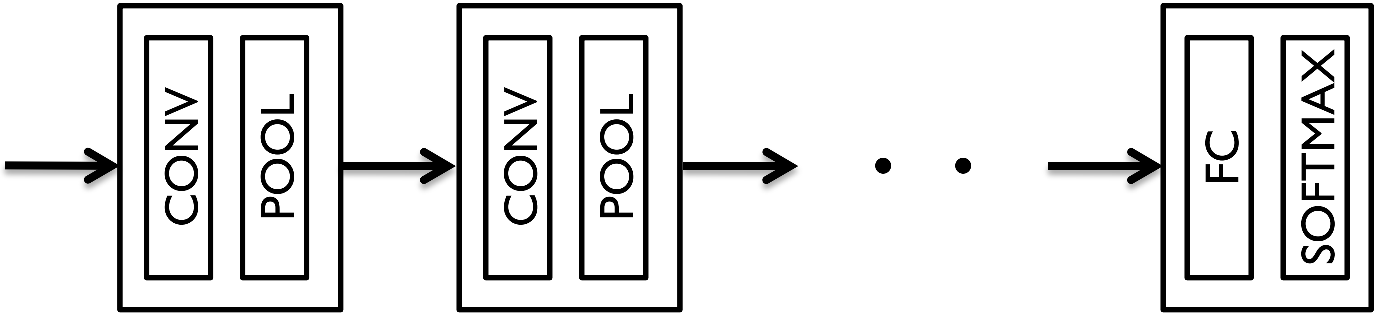
2	8	9	4
3	6	5	7
3	1	6	4
2	5	7	3

Max pooling →

8	9
5	7

- Role of an aggregator
- Invariance to image transformation and increases compactness to representation
- *<u>Pooling types:</u>* Max, Average, L2, etc

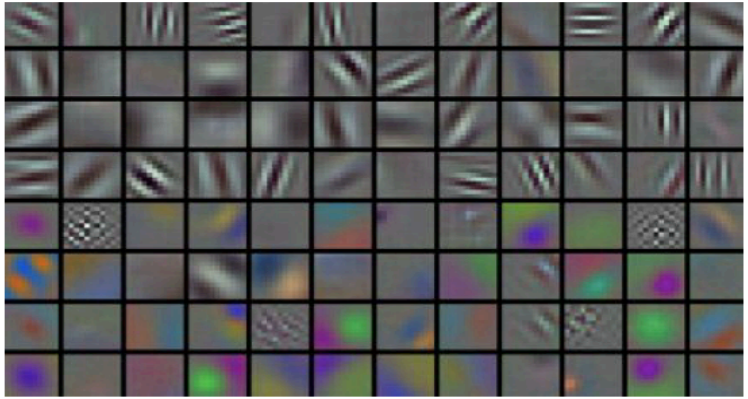
(Vanilla) CNN

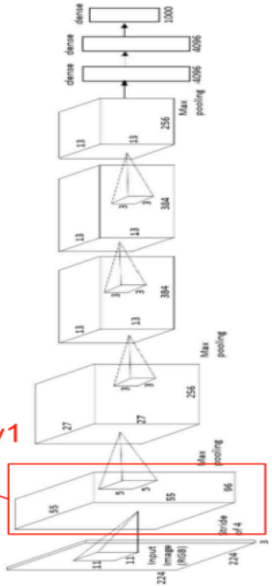


- CONV: Convolutional Layer
- POOL: Pooling Layer
- FC: Fully Connected Layer
- SOFTMAX: Classification Layer

Understanding CNNs

Visualize the filters/kernels (raw weights)





only interpretable on the first layer :(

Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Winter 2016

Interpreting CNNs

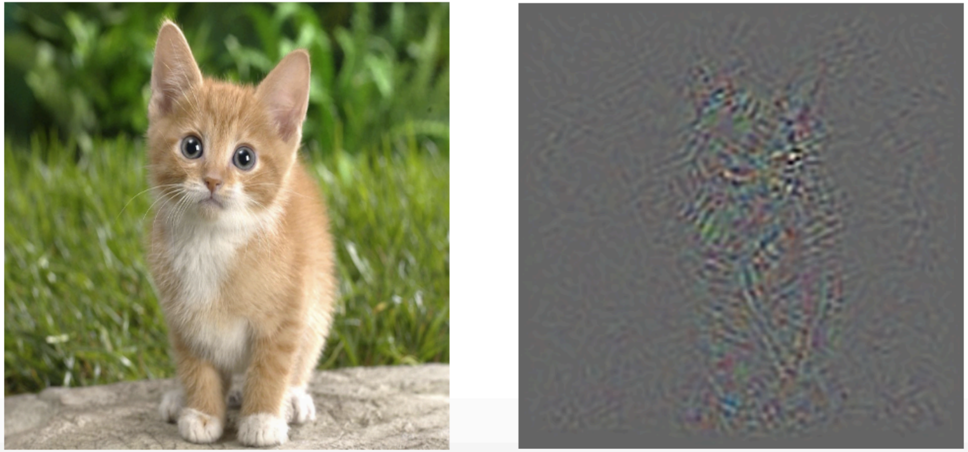
Backprop Methods

Input Image → CNN → Output Class

f(x)

You can compute gradient of loss w.r.t. x, instead of weights w

Input image | Backpropagation



Interpreting CNNs

DeconvNet and Guided Backpropagation

Backward pass of ReLU

$$\frac{\partial L}{\partial h^l} = [[h^l > 0]] \frac{\partial L}{\partial h^{l+1}}$$

Backward pass: backpropagation

-2	0	-1
6	0	0
0	-1	3

$\Leftarrow$

-2	3	-1
6	-3	1
2	-1	3

$\frac{\partial L}{\partial h^{l+1}}$

$$\frac{\partial L}{\partial h^l} = [[h^{l+1} > 0]] \frac{\partial L}{\partial h^{l+1}}$$

Backward pass: "deconvnet"

0	3	0
6	0	1
2	0	3

$\Leftarrow$

-2	3	-1
6	-3	1
2	-1	3

$\frac{\partial L}{\partial h^{l+1}}$

$$\frac{\partial L}{\partial h^l} = [[(h^l > 0) \& \& (h^{l+1} > 0)]] \frac{\partial L}{\partial h^{l+1}}$$

Backward pass: *guided backpropagation*

0	0	0
6	0	0
0	0	3

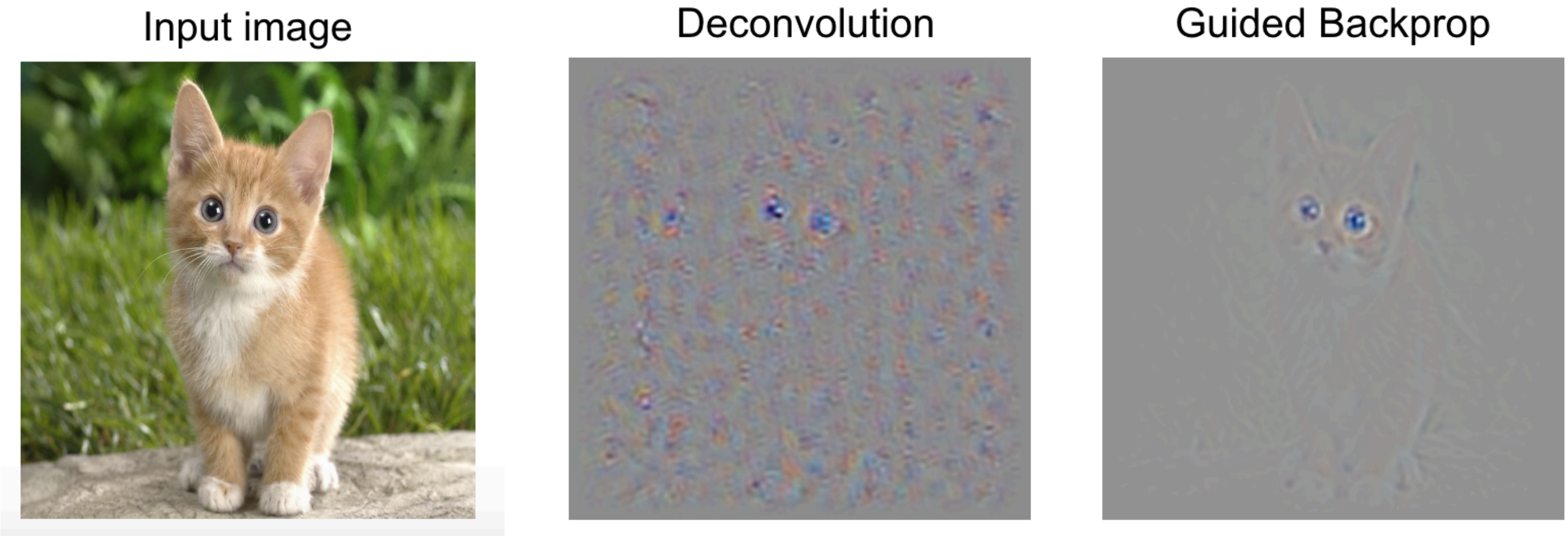
$\Leftarrow$

-2	3	-1
6	-3	1
2	-1	3

$\frac{\partial L}{\partial h^{l+1}}$

Interpreting CNNs

DeconvNet and Guided Backpropagation



Interpreting CNNs

Limitations of Guided Backpropagation

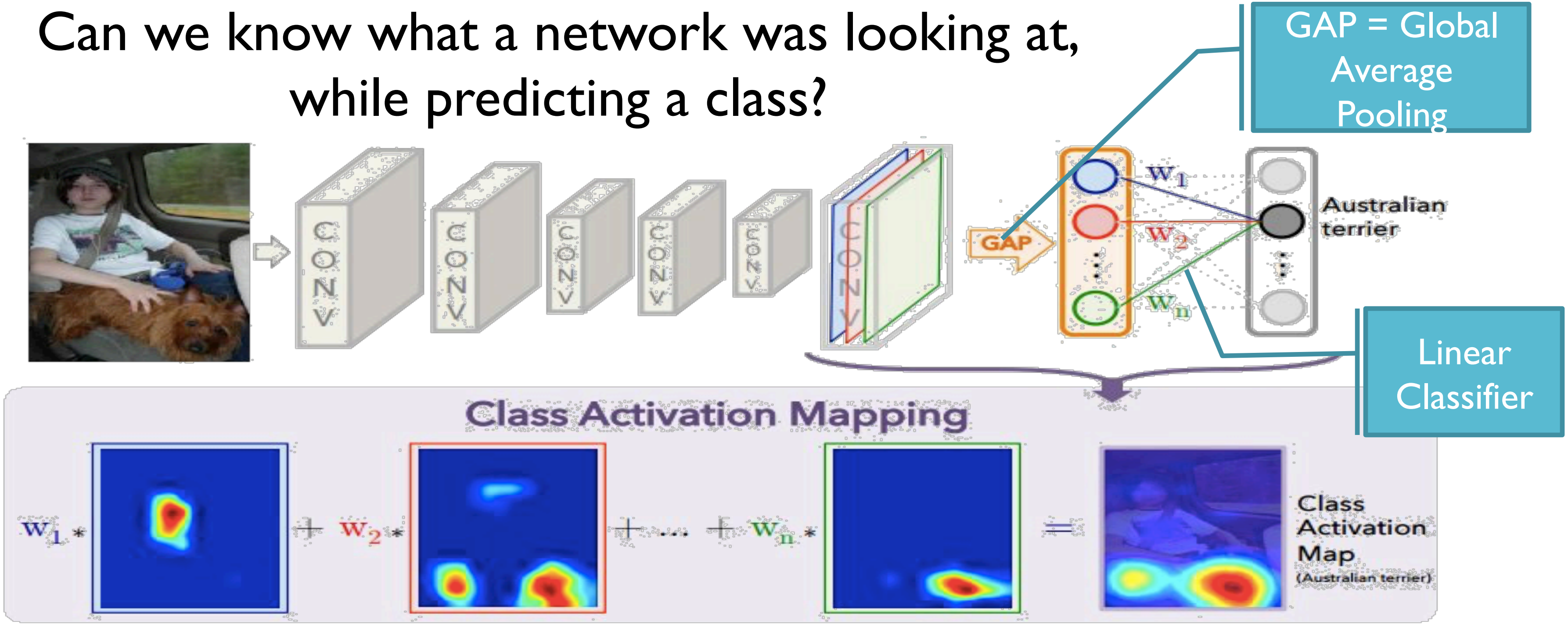
GB for “Cat”

GB for “Dog”

No class specificity

CAM: Class Activation Maps

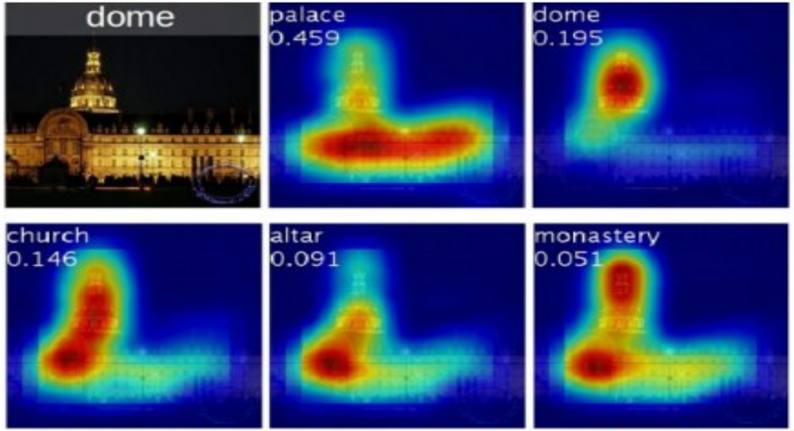
Can we know what a network was looking at, while predicting a class?



Zhou et al, Learning Deep Features for Discriminative Localization, CVPR 2016

CAM: Class Activation Maps

Sample Results



Class activation maps of top 5 predictions

Class activation maps for one object class

Zhou et al, Learning Deep Features for Discriminative Localization, CVPR 2016

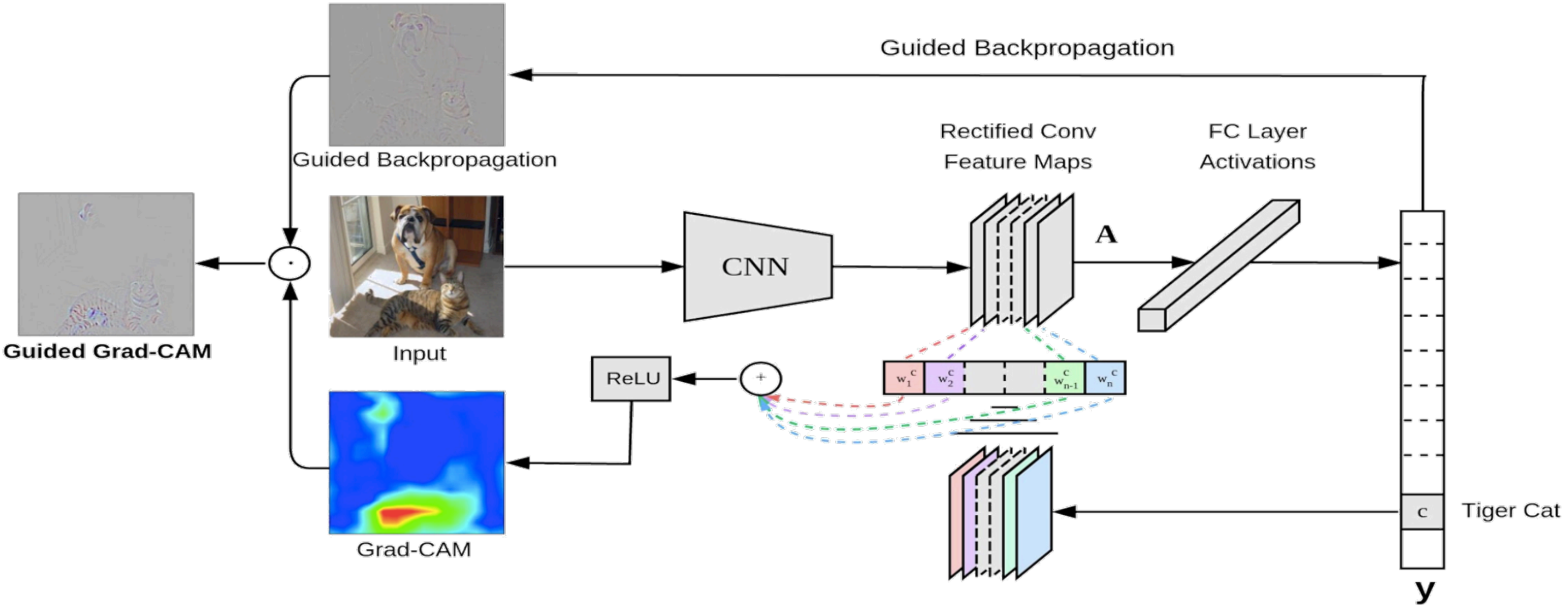
Grad-CAM

CAM requires re-training the network. How can we avoid this?

Note $$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$$ Retraining not required!

Selvaraju et al, Grad-CAM: Why did you say that? ICCV 2017

Grad-CAM



Selvaraju et al, Grad-CAM: Why did you say that? ICCV 2017

Grad-CAM

Sample Results

Selvaraju et al, Grad-CAM: Why did you say that? ICCV 2017

Grad-CAM

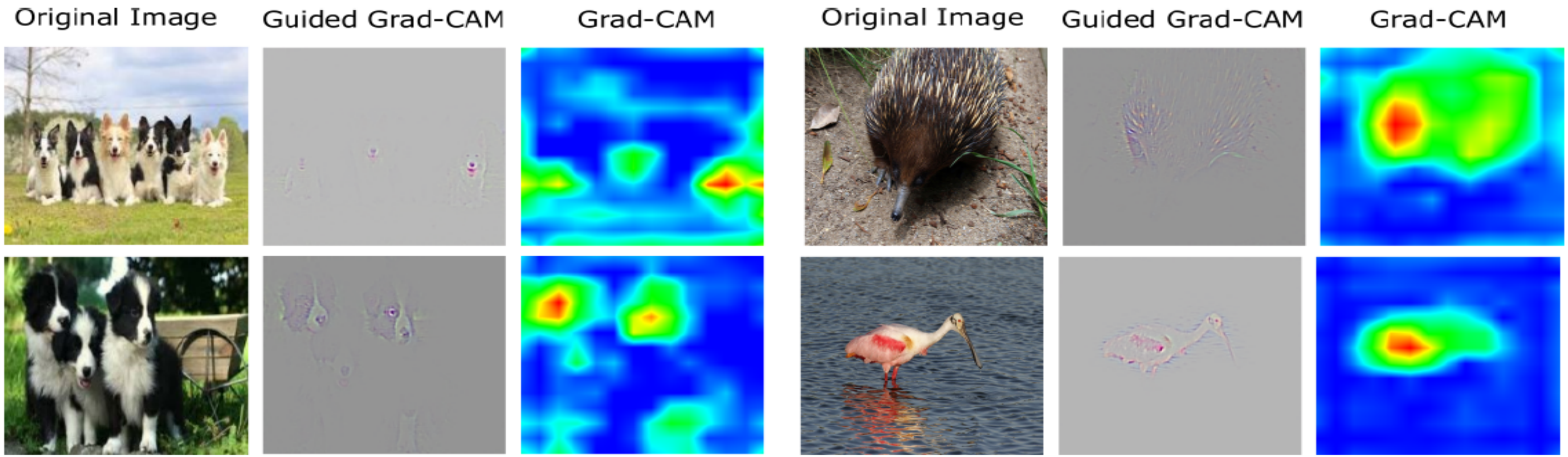
Sample Results for Image Captioning Models

A group of people flying kites on a beach

A man is sitting at a table with a pizza

Selvaraju et al, Grad-CAM: Why did you say that? ICCV 2017

Limitations of Grad CAM



Struggles when there are multiple occurrences of the same class; Localization sometimes incomplete

Grad-CAM++: Our Recent Work

Generalized version of Grad-CAM

- The weights w_k^c capture importance of particular activation map A^k.
- <u>Idea:</u> For a particular feature map A^k , only the positive gradients of a class score Y^c w.r.t. each spatial location (i,j) contribute towards its importance for that class *c*.
- Can we use this to impose a structure on the weights w_k^c ?

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} . relu(\frac{\partial Y^c}{\partial A_{ij}^k})$$

Chattopadhyay, Balasubramanian, et al, Grad-CAM++, WACV 2018

Grad CAM++

Obtaining Gradient weights

- Impose weights determined by:

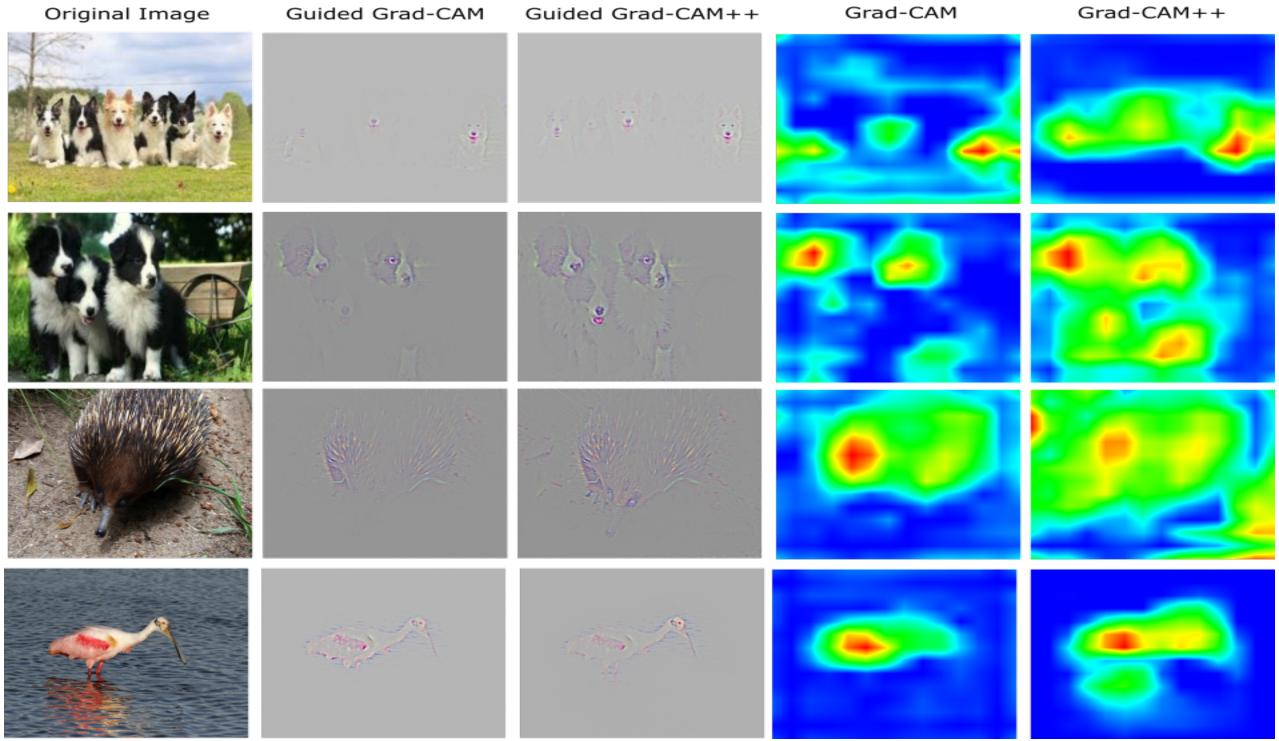
$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \{\frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3}\}}$$

- The class-discriminative saliency maps (reminiscent to Grad-CAM) are then calculated as:

$$L_{ij}^c = relu(\sum_k w_k^c . A_{ij}^k)$$

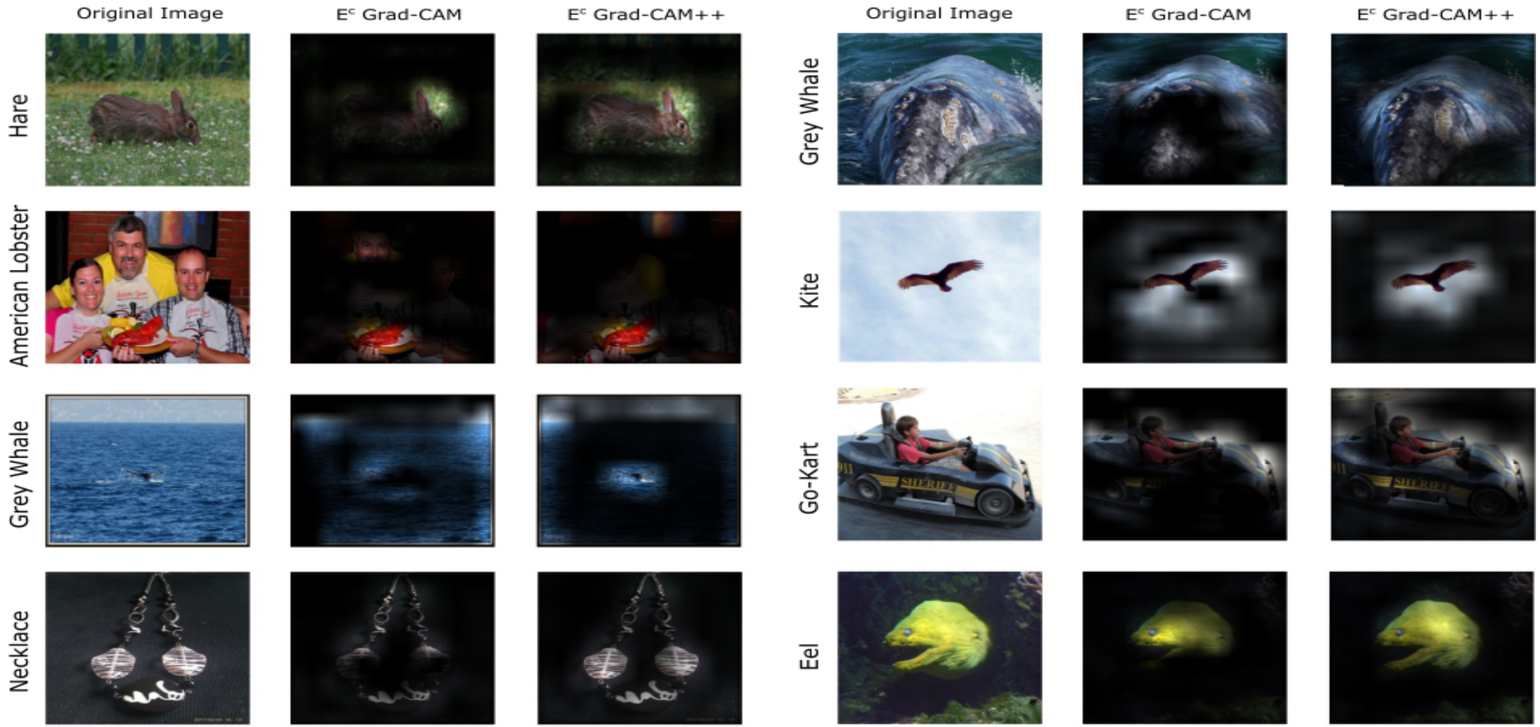
Grad-CAM++ Results

Visual Examples

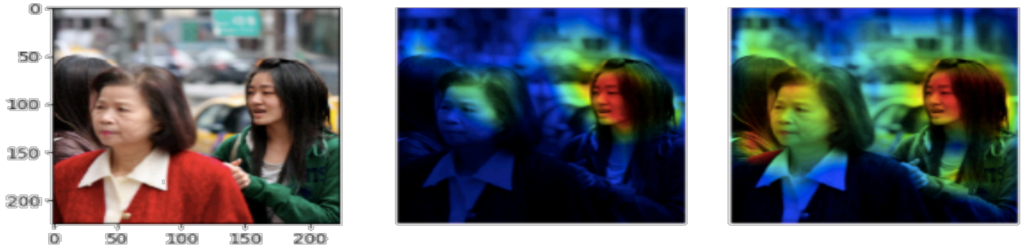


More Visual Examples

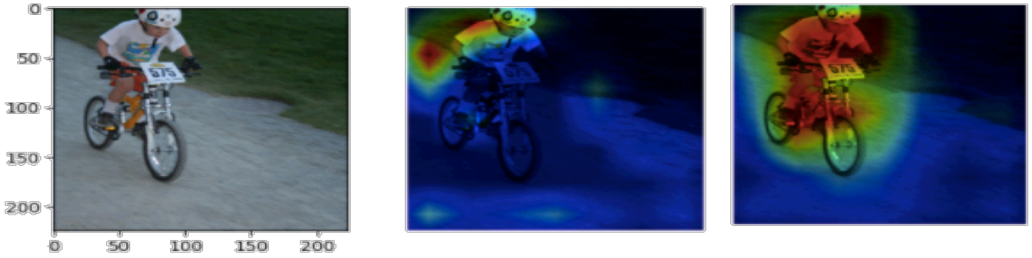
Visual Examples for the class localization problem for AlexNet



Grad-CAM++: More Details



two girls focused on their faces on a sunny day .



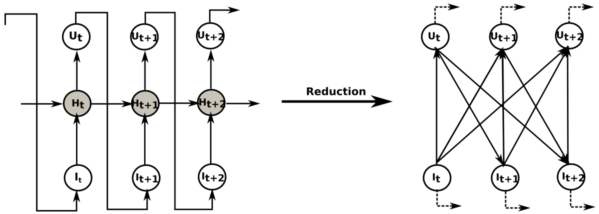
a motocross bike race four little kids are riding a bike race .

arXiv:
https://arxiv.org/abs/1710.11063

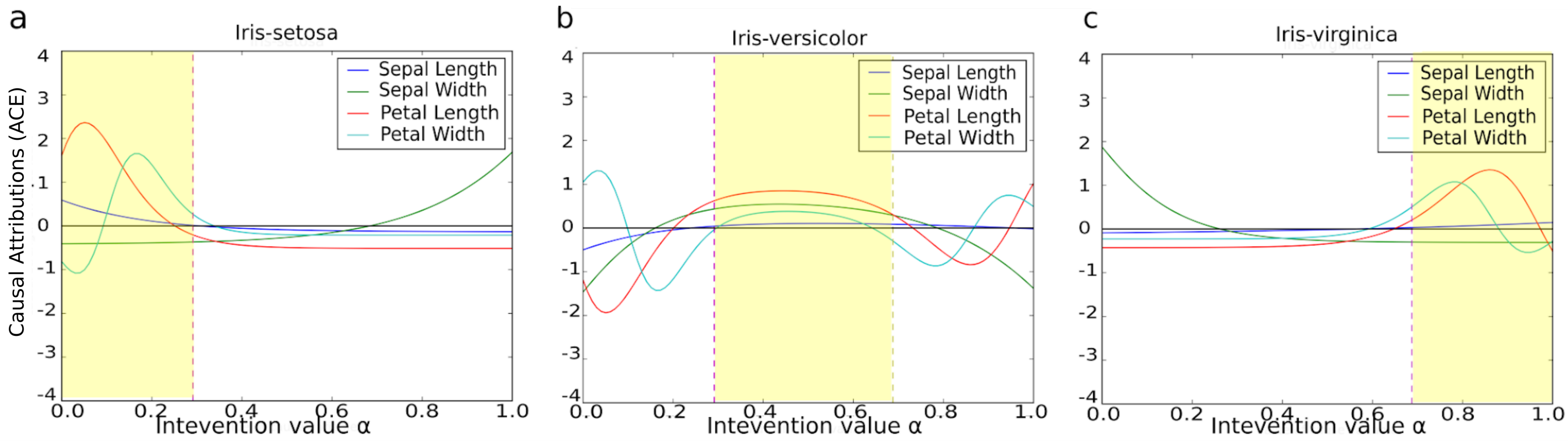
Github:
https://github.com/adityac94/Grad_CAM_plus_plus

More of our Recent Work

Causal Attribution in Neural Networks



Neural Networks as Structural Causal Models



Today

Outline

- Explainability in ML: An Overview
- Visual Interpretability of CNNs
- Looking Forward: Directions

Explainability in ML: What has been done?

LIME, KDD 2016

Processing/Input-output Analysis	Explanation-Producing	Representation Analysis
• Linear Proxy Methods • Decision Trees • Saliency Maps • Automatic Rule Extraction	• Scripted Conversations • Attention-based • Disentangling Representations	• Role of Layers • Role of Neurons • Role of Vectors

LRP, PlosOne 2015; CAM, CVPR 2016, Grad-CAM, ICCV 2017; DeepLIFT, ICML 2017

Gilpin et al, Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning, arXiv, Jun 2018

Looking Forward

Open Questions and Research Directions

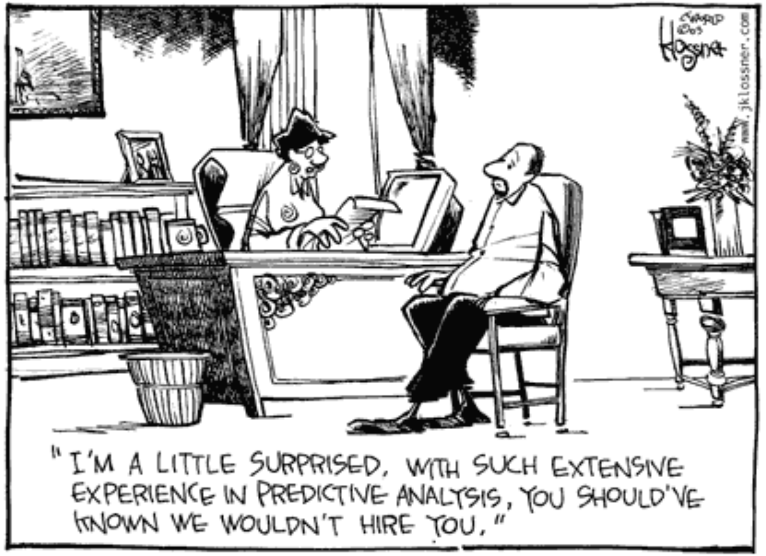
- Is there a universal formalization for explainable ML?
- How to balance the accuracy/performance vs interpretability tradeoff?
  - Is interpretability always required?
- What kind of data and what class of problems are more amenable for explainable systems?
- How to evaluate explainable systems?
- Who owns the explanation? Model or explanation methodology?

References and Resources

- Guidotti et al, A Survey Of Methods For Explaining Black Box Models, Jun 2018 [arXiv]
- Gilpin et al, Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning, Jun 2018 [arXiv]
- Lipton, The Mythos of Model Interpretability, Mar 2017 [arXiv]
- Velez and Kim, Towards A Rigorous Science of Interpretable Machine Learning, Mar 2017 [arXiv]
- Abdul et al, Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda, CHI 2018 [ACM link]
- Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Jul 2018 (Online Book)

Thank you!

Questions?



If only the model could explain itself…

vineethnb@iith.ac.in

Department of Computer Science and Engineering, IIT-Hyderabad

http://www.iith.ac.in/~vineethnb