Computer Vision Explainability

A Machine Learning Engineer's Overview

Navaneeth Kamballur Kottayil
AltaML

2021 embedded VISION summit
VIRTUAL | MAY 25-28

AltaML

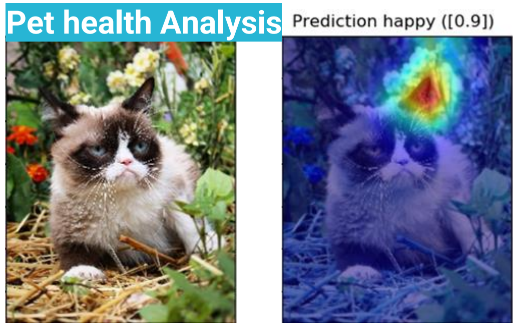
Outline

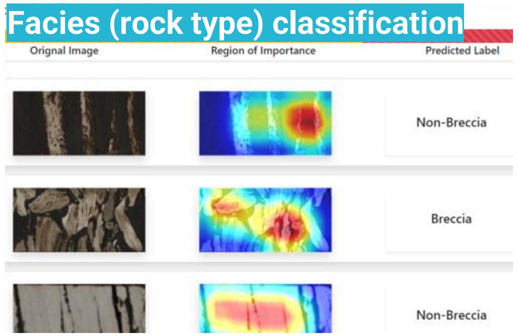
- ➢ Company introduction
- ➢ Deep learning and trust and why explainability is needed
- ➢ Categories of techniques in explainability
  - Basic idea + explanation of a representative method
- ➢ Case studies

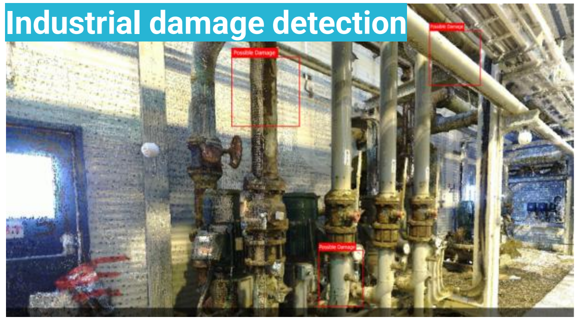
AltaML and Computer Vision

- **AltaML** is a Canadian applied Machine learning company that works with industry partners to augment their capabilities with AI&ML.

- **AltaML** has had great success in generating value for its partners with use of computer vision-based ML systems.



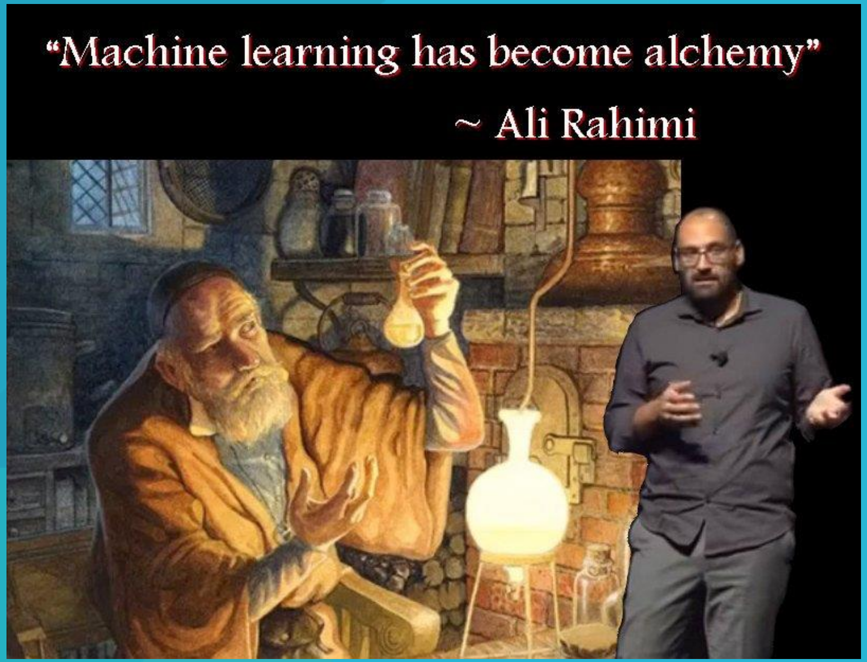




AltaML

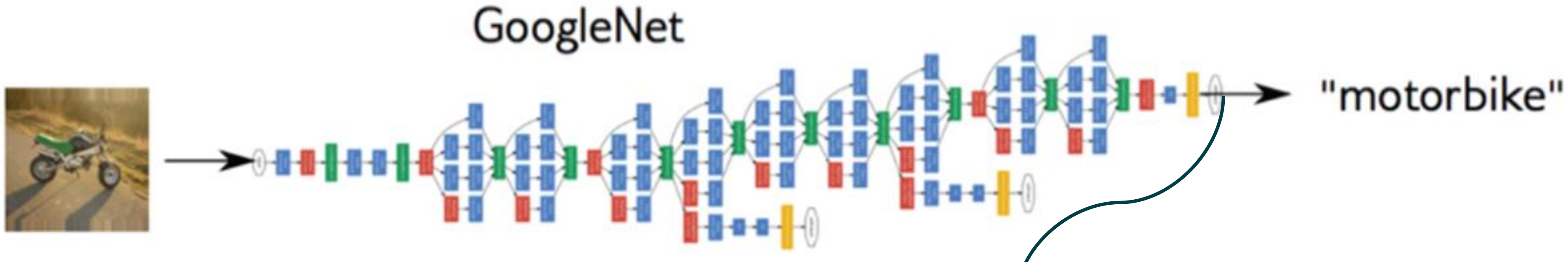
Deep Learning and Trust

“Deep learning is a black box”



AltaML

Why Black Boxes



The picture is classified as a motorcycle on the basis of the following set of pixels

Common questions:

- **Clients** : “Why does it make this prediction?”
- **ML Dev/Data Scientist** : “Why does this work?”; “Is my algorithm looking at the right things?”

This is a serious problem even if performance is high.

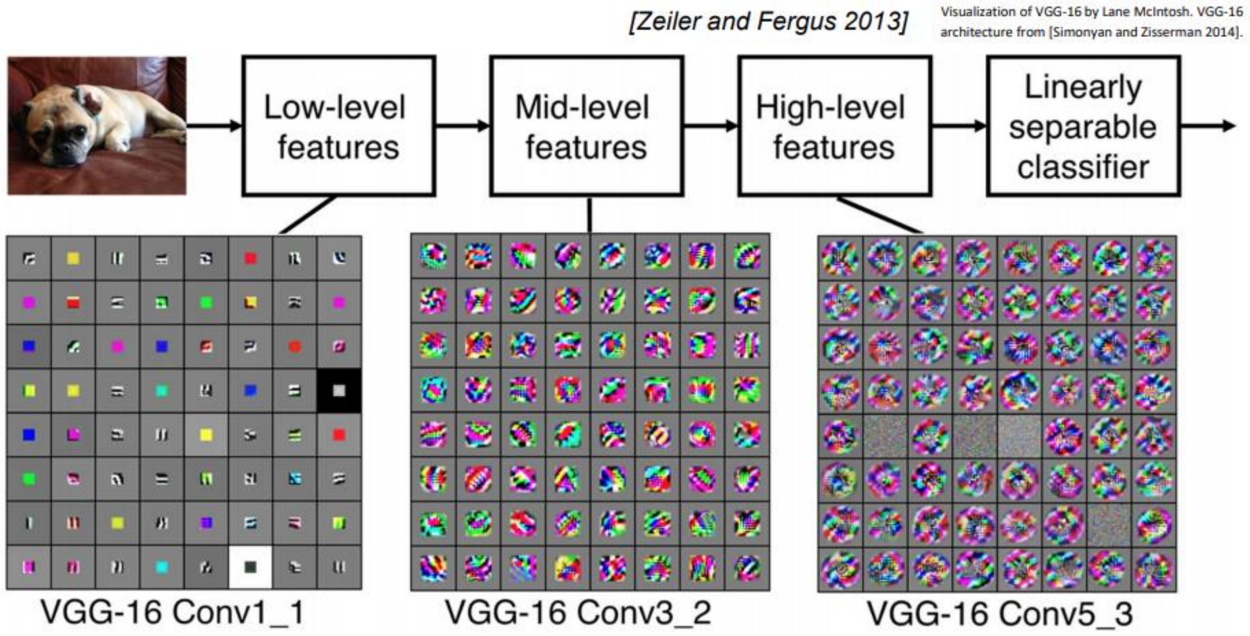
Image reference: https://medium.com/@RaghavPrabhu/cnn-architectures-lenet-alexnet-vgg-googlenet-and-resnet-7c81c017b848

Opening up Black Boxes

We can visualize features detected at each of the layers.

- Initial layer filters detect Gabor like edges !

Deeper layer filters convey no meaningful information.



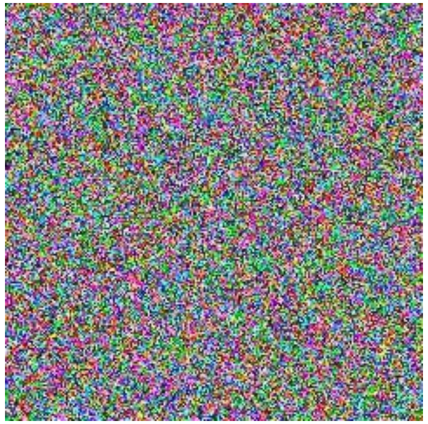
Results cannot be explained with visualizations of filter coefficients or outputs.

Image reference:Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

AltaML

Filter Visualization Techniques

Basic idea: Synthesize inputs that can maximize a specific neuron activation.



- Input a random noise image as input. Say x, to trained CNN.
- Perform a fwd pass of the image.
- Assuming that the filter that one wants to visualize is of index **i,** such that activation of the specific layer of interest is $a_i(x)$
- The visualization of the fitler is obtained by adding the backpropagated gradients from $a_i(x)$ back to the image x (usually with a scale factor to control the amount by which update is done)

$$x = x + \alpha . \frac{\partial a_i(x)}{\partial x}$$

AltaML

Explainability

Provide either, ***a set of pixels*** or a ***heat map*** showing the **pixels that were important** for a classification decision.

Input image

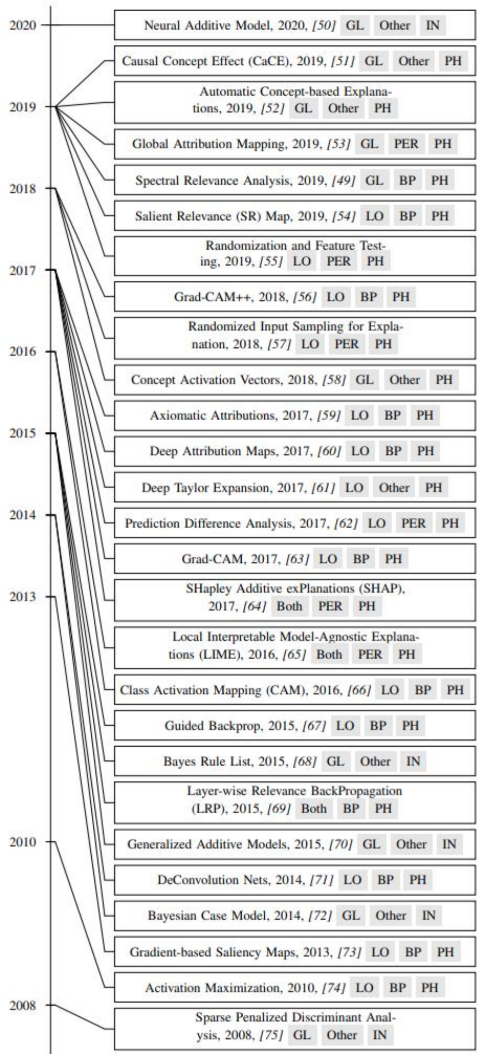
Explainability output

Explainability

Active field of research with *very large number* of research papers, tools and techniques

Broad categories of research

- **Perturbation based methods**
- **Backpropagation based methods**
- **Activation based methods**



Abbreviation	Definition
ACE	Automatic Concept-based Explanations
AI	Artificial Intelligence
API	Application Programming Interface
BAM	Benchmarking Attribution Methods
BRL	Bayesian Rule List
CaCE	Causal Concept Effect
CAM	Class Activation Mapping
CAV	Concept Activation Vectors
CNN	Convolutional Neural Network
DeConvNet	Deconvolution Neural Network
DL	Deep Learning
DNN	Deep Neural Network
EG	Expected Gradients
FMRI	Functional Magnetic Resonance Imaging
GAM	Generalized Additive Models
IG	Integrated Gradients
IRT	Interpretability Randomization Test
LIME	Local Interpretable Model-Agnostic Explanations
LRP	Layer-wise Relevance BackPropagation
ML	Machine Learning
NAM	Neural Additive Models
OSFT	One-Shot Feature Test
ReLU	Rectified Linear Unit
RISE	Randomized Input Sampling for Explanation
RNN	Recurrent Neural Network
SCS	System Causability Scale
SHAP	SHapley Additive exPlanations
SPDA	Sparse Penalized Discriminant Analysis
SpRAy	Spectral Relevance Analysis
SR	Salient Relevance
TCAV	Testing with Concept Activation Vectors
t-SNE	t-Stochastic Neighbor Embedding
VAE	Variational Auto Encoders
XAI	Explainable Artificial Intelligence

Image reference: Das, Arun, and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey." arXiv preprint arXiv:2006.11371 (2020).

Perturbation Based Methods

Basic idea: learn the behavior by perturbing the input and see how the predictions change.

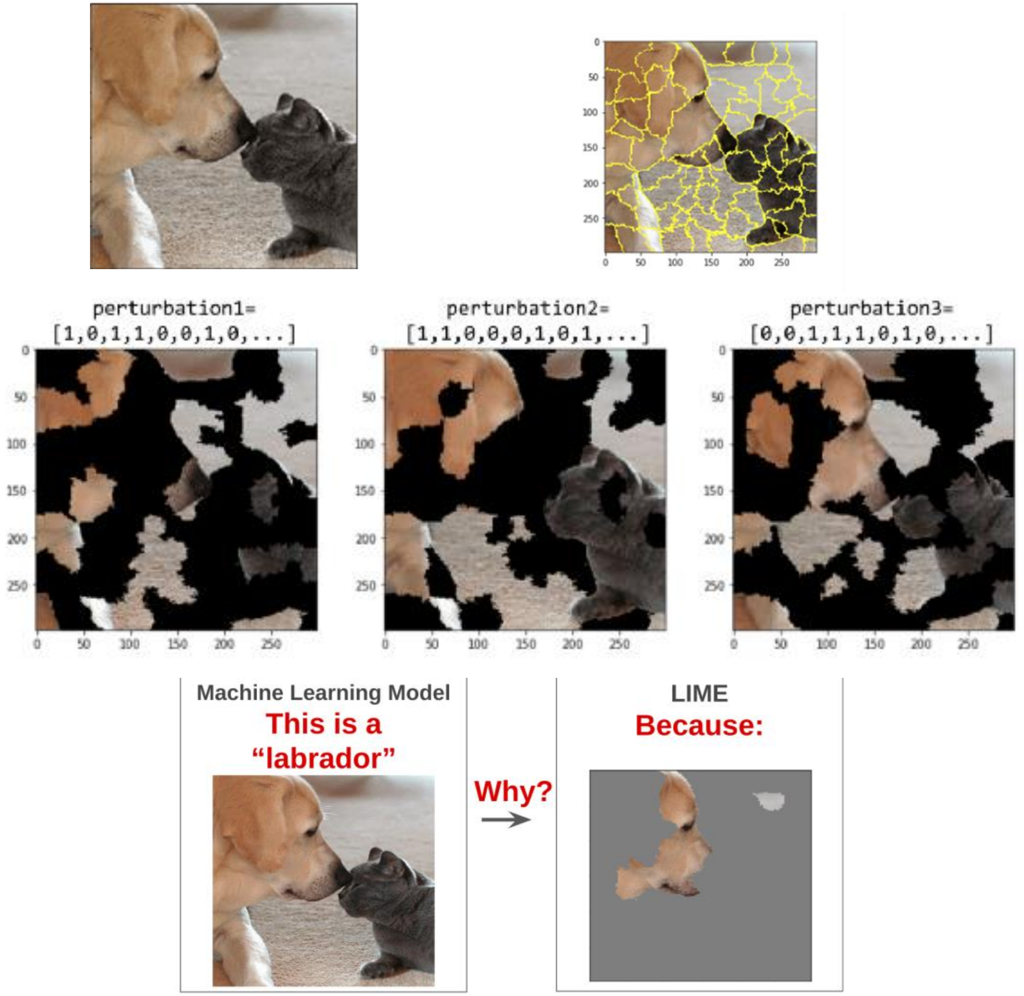


Image reference: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

LIME (Local Interpretable Model-Agnostic Explanations)

- Choose the ML model and a reference class to be explained
- Generate perturbations all over the image space (*an approximation of this is done by dividing image to superpixels and randomly turn off superpixels)*
- Predict the output Y, *for each* perturbed image, using the ML model
- Find the contribution of each of the superpixel by, training the following **Linear Ridge Regression** on the generated output:

  $$\mathbf{E}(Y) = \beta_0 + \sum \beta_j X_j$$

- *The β coefficients are regarded as LIME explanation.* The superpixels with the largest weight is the explanation (in terms of pixels)

Image reference: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Backpropagation Based Methods

Basic idea: Trace the signals from classification output back to the input

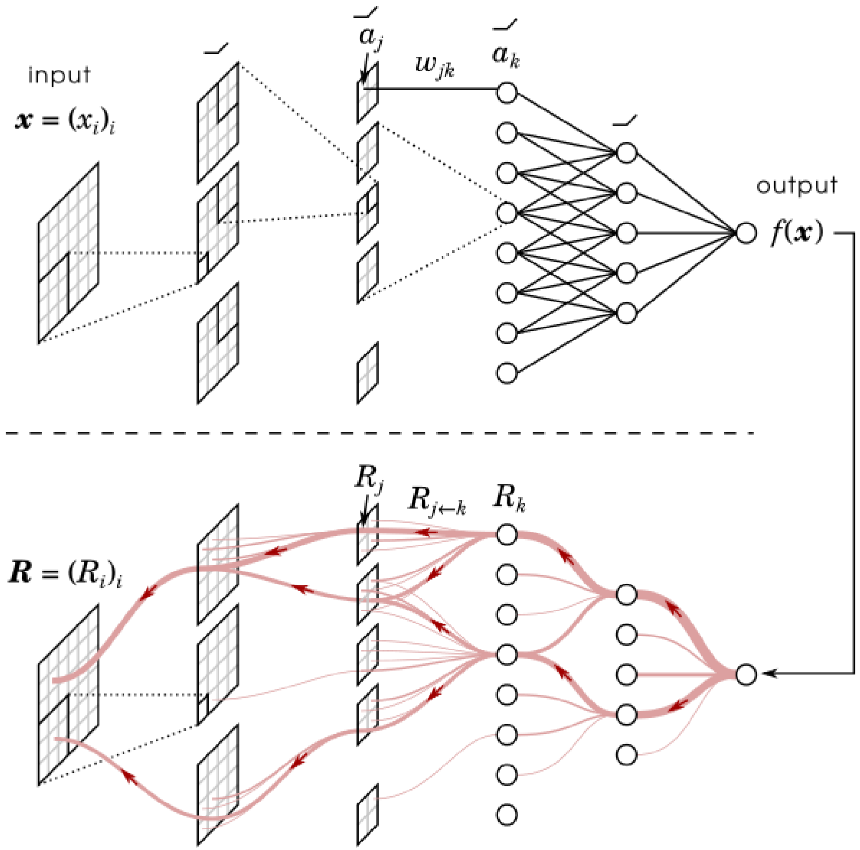


Image reference:http://www.heatmapping.org/

Paper reference: Binder, Alexander, et al. "Layer-wise relevance propagation for neural networks with local renormalization layers." International Conference on Artificial Neural Networks. Springer, Cham, 2016.

LRP (Layer-wise relevance propagation)

- Choose the ML model and a reference class to be explained
- Start with output neuron of class c (its probability will be considered as relevance R at output layer) and trace the result to its previous layer with formula

$$R_i(l) = \sum_j \frac{Z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)} \ where \ z_{ij} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

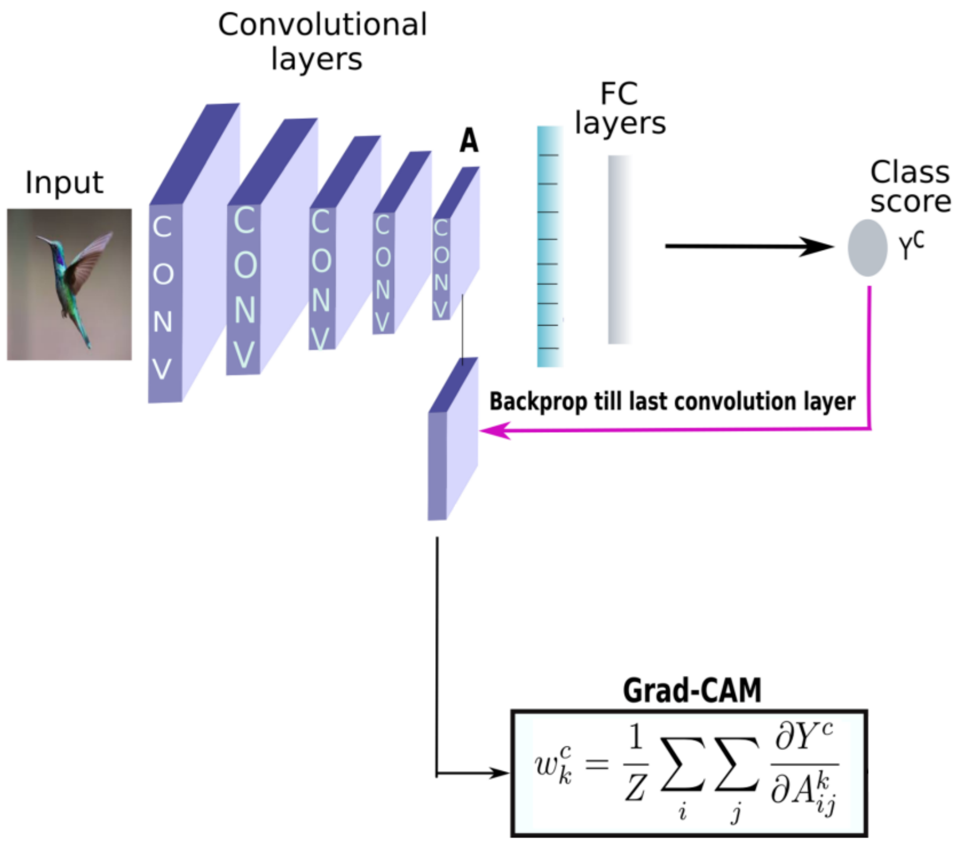
  for neuron i, at layer *l. In the notation used,* $x_j^{(l)}$ *is output of neuron j at layer l*

- This is continued till the input image

Image reference:http://www.heatmapping.org/
Paper reference: Binder, Alexander, et al. "Layer-wise relevance propagation for neural networks with local renormalization layers." International Conference on Artificial Neural Networks. Springer, Cham, 2016.

Activation Based Methods

Basic idea: Express classification results in terms of strength of feature maps*



* Note that these are feature maps, ie, outputs of filters, NOT filters themselves

Image reference: Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

AltaML

Activation Based Methods

GradCAM

- Choose the ML model and a reference class c to be explained
- Choose a set of k, CNN layer outputs, ie, feature maps A^k
- Compute scale factor for the feature maps as

$$\alpha_k^c = GAP(\frac{\partial y^c}{\partial A_{ij}^k})$$

where GAP is average operation on 2D

- GradCAM is computed as

$$L_{Grad-CAM} = ReLU(\sum_k \alpha_k^c A_k)$$

for all of the k chosen feature maps

* Note that these are feature maps, ie, outputs of filters, NOT filters themselves

Image reference: Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

Comparisons

	Perturbation based	Backpropagation based	Activation based
Advantage	• Model agnostic • Easy to implement • No modification to model	• Quick to compute Fine-grained interpretation • No modification to model	• Easy to interpret • Reasonably fast
Disadvantage	• Time consuming to run	• Need access to model weights and architecture • Sometimes will be hard to interpret	• Only for CNN • Different explanations based on selected feature maps • Minimal modification of model (in some implementations)

Courtesy: https://yozey.github.io/files/CNNInterpretability-1-41.pdf

AltaML

Applications in Use Cases

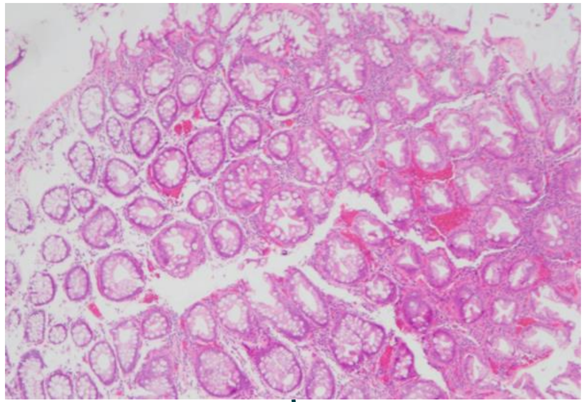
Debugging and Building Trust

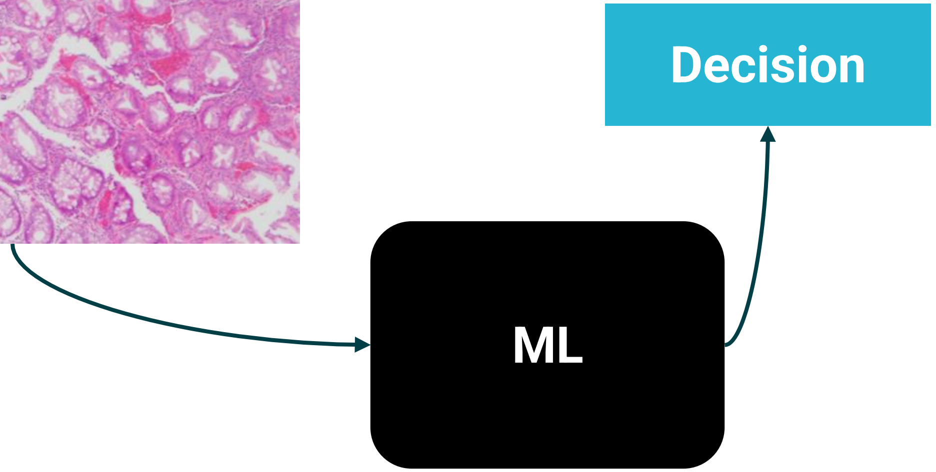
Use case:

Classify high resolution microscopy slides into normal or abnormal

ML: Image classification problem

Input	Images of biopsy slides
Labels	Per tissue label; [`Normal`/`Abnormal`]





Used *image segmentation* to give the ML model only images of tissue and not the rest of the slide. ***Decisions on an entire tissue sample.*** No further information provided by the model.

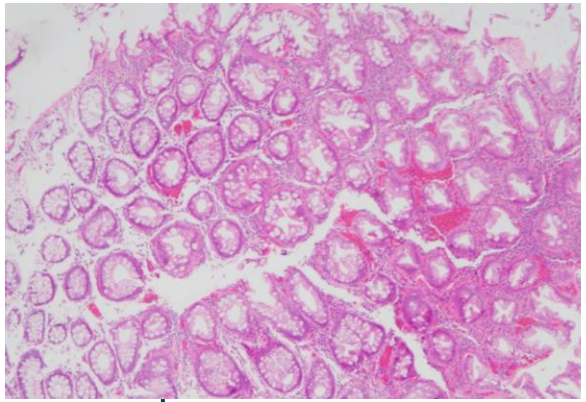
Simulated images are used to respect the privacy agreements

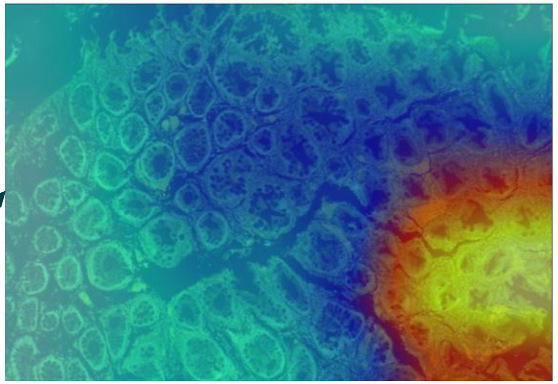
Debugging and Building Trust

Explanations significantly helped in building trust in the proposed model for ML team and doctors.

Doctors identified that the heat map pointed to the damaged cells of interest to the doctors.

Input	Images of biopsy slides
Labels	Per tissue label; [Normal/Abnormal]





ML + Explanation

Decision

Localized heat map of the cells that contribute to the decision (*not directly available in labels*).

Simulated images are used to respect the privacy agreements

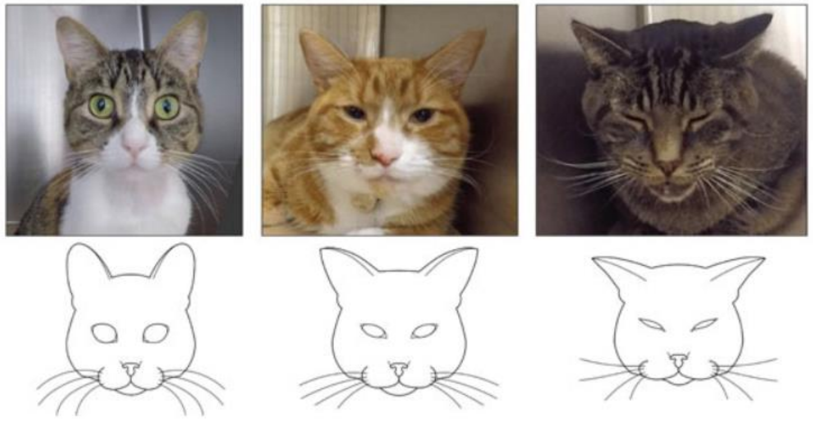
AltaML

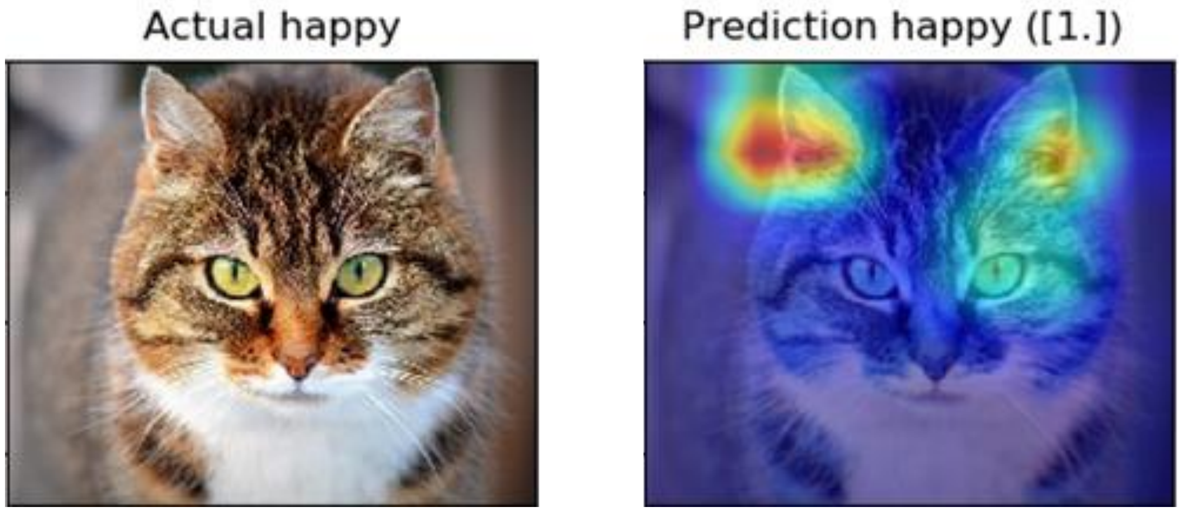
Additional Insights Generated

Use case:

CV model for assessment of pet health

Input	Images from hospital
Labels	Per image label; [**happy**/**not happy**]

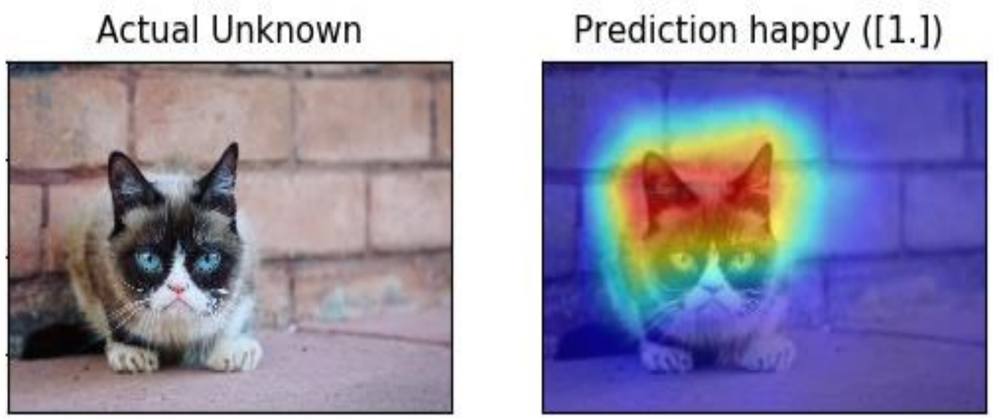




Model predictions 'rediscovered' the idea of ***Cat Grimace Scale***.

Conclusions

- Explainability techniques are vital in computer vision-based use cases to explain the decisions of deep learning-based models.
- Implementation of these techniques is not expensive.
- Most basic techniques can give a lot of useful insights.



Fun fact: We found out that grumpy cat is in fact, *not grumpy at all*!

Resources

Perturbation based explainability: *Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.*

***Backpropagation based* explainability**: *Binder, Alexander, et al. "Layer-wise relevance propagation for neural networks with local renormalization layers." International Conference on Artificial Neural Networks. Springer, Cham, 2016.*

Activation map explainability: Chattopadhay, Aditya, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839-847. IEEE, 2018.

Visualizing Convolutional Feature maps: Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

Explainability in CV survey paper: Das, Arun, and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey." arXiv preprint arXiv:2006.11371 (2020).

Thank You