Deep Learning - COSC2779

Neural Network Model Interpretation

Dr. Ruwan Tennakoon

RMIT UNIVERSITY

October 4, 2021

Outline

Deep Learning Models

- So far we explored how to develop models for image, text, voice, . . . , analysis.
- Usually involve:
  1. Determine your goals
  2. Default Baseline Model: Baseline Model architecture, cost functions, optimization.
  3. Setup the diagnostic instrumentation
  4. Make incremental changes
- Performance of the model is evaluated on an **independent and identically distributed** test set.

Deep Learning Models

- So far we explored how to develop models for image, text, voice, . . . , analysis.
- Usually involve:
  1. Determine your goals
  2. Default Baseline Model: Baseline Model architecture, cost functions, optimization.
  3. Setup the diagnostic instrumentation
  4. Make incremental changes
- Performance of the model is evaluated on an **independent and identically distributed** test set.

- Many see Deep Networks an **black boxes**.
- Important questions that need to be answered are:
  - **How can I trust your model?**
  - **How does your model really make its decisions?**
- Performance on iid test set alone cannot answer these questions.

How does your model really make its decisions?

Urban Legend: "Once upon a time, the US Army wanted to use neural networks to automatically detect camouflaged enemy tanks. The researchers trained a neural net on 50 photos of camouflaged tanks in trees, and 50 photos of trees without tanks. Using standard techniques for supervised learning... Without further training the neural network classified all remaining (test) photos correctly. It turned out that in the researchers' dataset, photos of camouflaged tanks had been taken on cloudy days, while photos of plain forest had been taken on sunny days. The neural network had learned to distinguish cloudy days from sunny days, instead of distinguishing camouflaged tanks from empty forest."

How does your model really make its decisions?

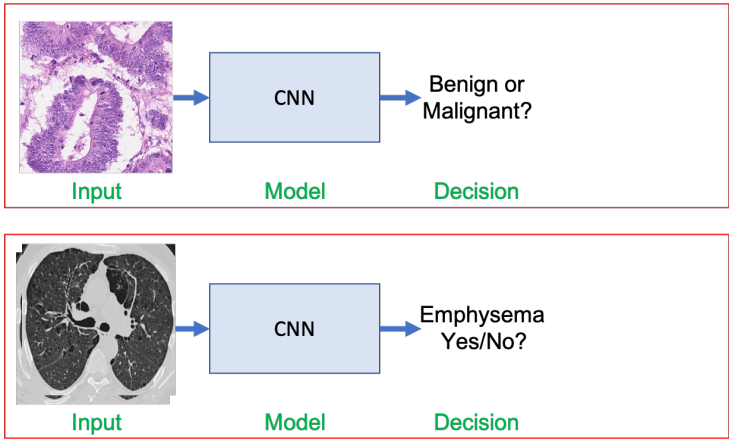
Urban Legend: "Once upon a time, the US Army wanted to use neural networks to automatically detect camouflaged enemy tanks. The researchers trained a neural net on 50 photos of camouflaged tanks in trees, and 50 photos of trees without tanks. Using standard techniques for supervised learning... Without further training the neural network classified all remaining (test) photos correctly. It turned out that in the researchers' dataset, photos of camouflaged tanks had been taken on cloudy days, while photos of plain forest had been taken on sunny days. The neural network had learned to distinguish cloudy days from sunny days, instead of distinguishing camouflaged tanks from empty forest."

What we expected: A neural network which detects camouflaged army tanks.
What we got: A neural network which distinguishes between cloudy and sunny days

Good story. It's very likely, though not certain, that this didn't actually happen.

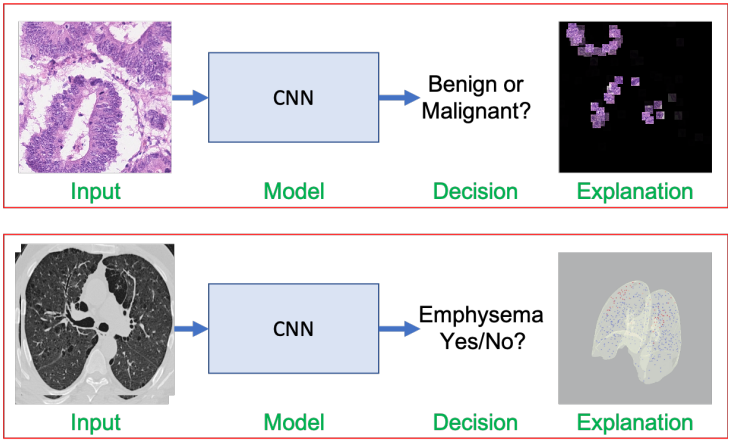
Yudkowsky, E., 2008. Artificial intelligence as a positive and negative factor in global risk.

Why did the model make a particulate decision?



Why does the model say input is malignent (or has emphysema)?

Why did the model make a particulate decision?



Tennakoon, R., et. al. "Classification of Volumetric Images Using Multi-Instance Learning and Extreme Value Theorem".IEEE Transactions on Medical Imaging, 2019.

If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily (Doshi-Velez and Kim 2017):

- **Trust**: It is easier for humans to trust a system that explains its decisions compared to a black box.
- **Causality**: Check that only causal relationships are picked up.
- **Reliability or Robustness**: Ensuring that small changes in the input do not lead to large changes in the prediction.
- **Fairness**: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g.racial, gender) bias.
- **Privacy**: Ensuring that sensitive information in the data is protected.

Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

Methods and models that make the behaviour and predictions of **machine learning** systems understandable to humans.

- Intrinsic or post hoc.
  - **Intrinsic**: Using a machine learning model which are interpretable in nature (linear models or tree-based models).
  - **Post hoc**: Using a black box model (ensemble methods or neural networks) and applying interpretability methods after.

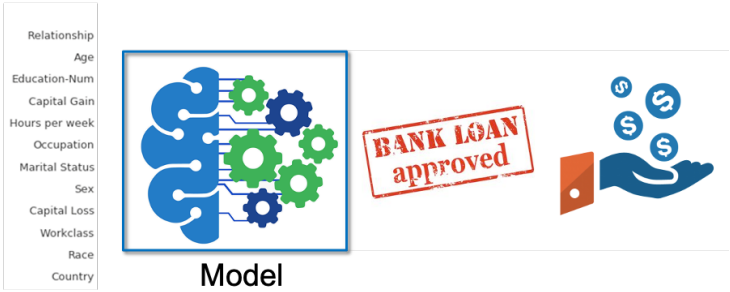
Methods and models that make the behaviour and predictions of **machine learning** systems understandable to humans.

- Intrinsic or post hoc.
  - **Intrinsic**: Using a machine learning model which are interpretable in nature (linear models or tree-based models).
  - **Post hoc**: Using a black box model (ensemble methods or neural networks) and applying interpretability methods after.
- Model-specific or model-agnostic
  - **Model-specific**: specific to intrinsic model interpretation methods which depend purely on the capabilities and features on a per-model basis.
  - **Model-agnostic**: Mostly post hoc methods that can be used on any machine learning model. These methods usually operate on feature input and output pairs.

Methods and models that make the behaviour and predictions of **machine learning** systems understandable to humans.

- Intrinsic or post hoc.
  - **Intrinsic**: Using a machine learning model which are interpretable in nature (linear models or tree-based models).
  - **Post hoc**: Using a black box model (ensemble methods or neural networks) and applying interpretability methods after.
- Model-specific or model-agnostic
  - **Model-specific**: specific to intrinsic model interpretation methods which depend purely on the capabilities and features on a per-model basis.
  - **Model-agnostic**: Mostly post hoc methods that can be used on any machine learning model. These methods usually operate on feature input and output pairs.
- Local or global
  - **Local**: Explains a single prediction
  - **Global**: Explains the behaviour of the the entire model

Local Vs Global



- **Local**: Why did (not) my loan get approved?
- **Global**: Is The model biased to a particular race or gender?

Interpretation methods for neural networks

- **Feature Visualization**: What features has the neural network learned?
- **Feature Attribution** : How did each input contribute to a particular prediction?
- **Concepts**: Which more abstract concepts has the neural network learned?
- **Model Distillation**: How can we explain a neural network with a simpler model?

Outline

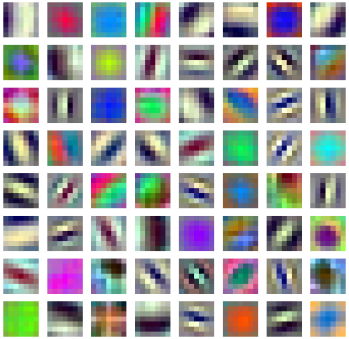
What are the intermediate layers in a neural network looking for?

German Shepherd

Image of the dog: Unsplash

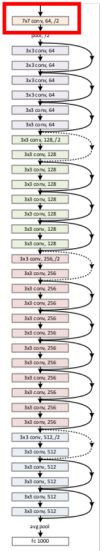
Global Interpretability: Explains what the network has learned?

First Layer: Visualize Filters



ImageNet trained weights for ResNet50V2.

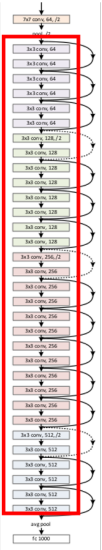
Plot each weight filter of the first convolution layer.
ResNet50 First Conv layer 64 filters, each 7x7 kernel with 3 channels.
Similar patterns seen in many CNNs.



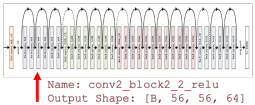
Middle Layers: Visualize Filters

From ConvNetJS CIFAR-10 demo

We can visualize filters in middle layers, but not that interesting.



Middle Layers: Visualize Outputs

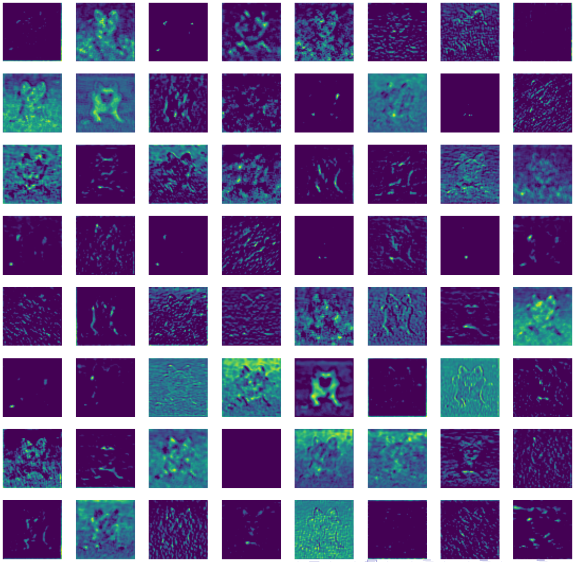


From ConvNetJS CIFAR-10 demo

We can visualize Outputs of middle layers.

- Forward propagate for the image of interest.
- Extract the output at any layer and plot each channel separately.

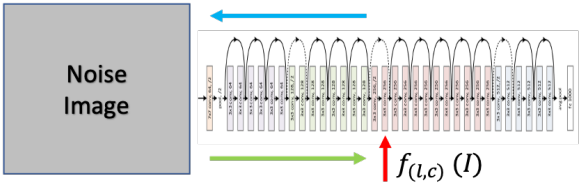
Observing a particular filter for many inputs will provide an insight into what it has learned.



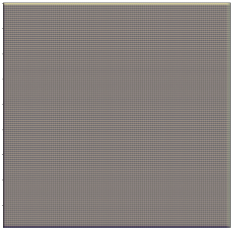
Generate a synthetic image that maximally activated a given neuron.

$$I_{syn} = \arg\max_{I} f_{(l,c)}(I) + R(I)$$

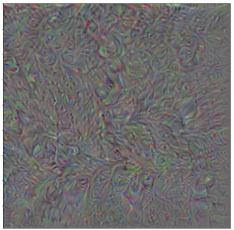
- $f_{(l,c)}(I)$: Activation of layer l channel c for input image I.
- $R(I)$: Some regularization function on image I.



- Initialize image to random noise.
- Repeat
  1. Forward image to calculate current score.
  2. Backprop to get gradient of score with respect to image pixels.
  3. Update the image pixel values based on gradients.

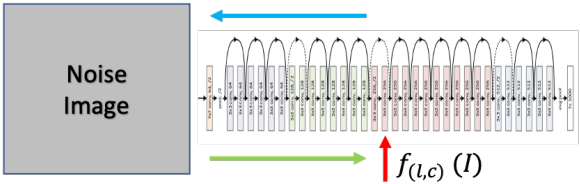


Layer 1: ResNet50



Layer 5: ResNet50

From ResNet50 trained on ImageNet.



- Initialize image to random noise.
- Repeat
  1. Forward image to calculate current score.
  2. Backprop to get gradient of score with respect to image pixels.
  3. Update the image pixel values based on gradients.

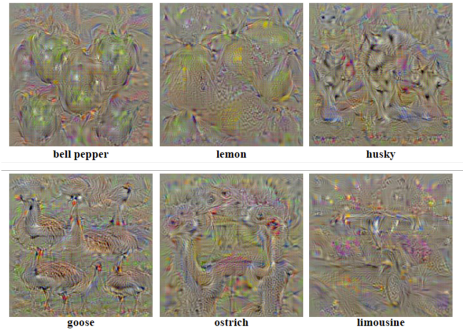
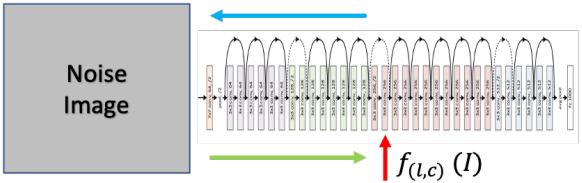


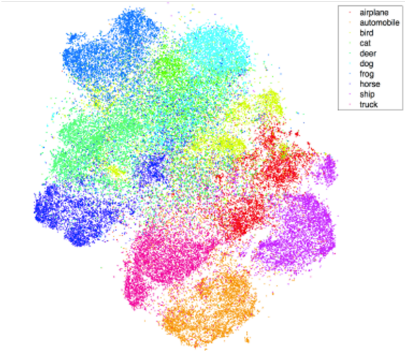
Image: Simonyan, K., Vedaldi, A. and Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Better regularization lead to better visualizations.



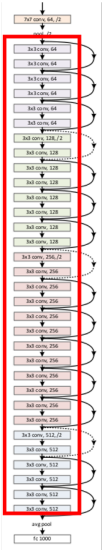
- Initialize image to random noise.
- Repeat
  1. Forward image to calculate current score.
  2. Backprop to get gradient of score with respect to image pixels.
  3. Update the image pixel values based on gradients.

Fully Connected Layers: Visualize Filters



tSNE visualization of CIFAR10 data of final fully connected layer.

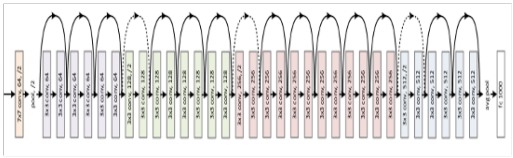
We can visualize filters in middle layers, but not that interesting



- Pros:
  - Feature visualizations give unique insight into the working of neural networks.
  - Through feature visualization, we have learned that neural networks first learn simple edge and texture detectors and more abstract part and object detectors in higher layers.
  - Communicate in a non-technical way how neural networks work.
- Cons:
  - Many feature visualization images are not interpretable at all.
  - There are too many units to look at.
  - Illusion of Interpretability?

Outline

How did each input contribute to a particular prediction?

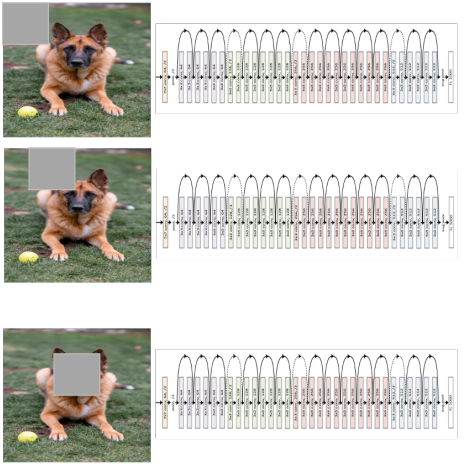


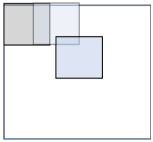
German Shepherd

Image of the dog: Unsplash

Local interpretability: Explains the rational for single input.

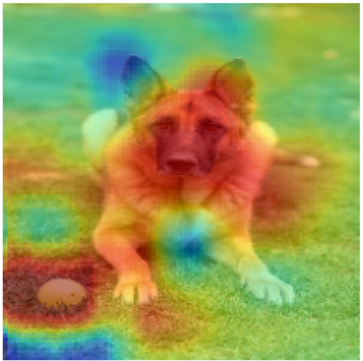
Which pixels matter: Mask part of the image before it is fed to the CNN. See how much the prediction probability for a class change.





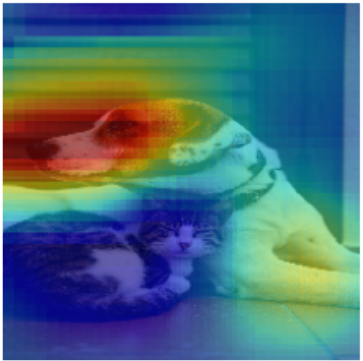
Which pixels matter: Mask part of the image before it is fed to the CNN. See how much the prediction probability for a class change.

Class: German Shepard

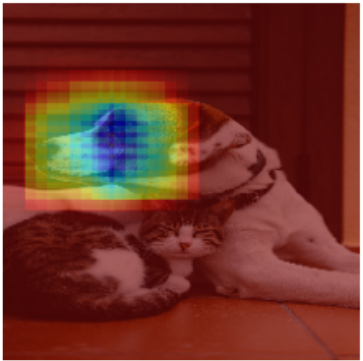


Class: Tennis Ball

Which pixels matter: Mask part of the image before it is fed to the CNN. See how much the prediction probability for a class change.

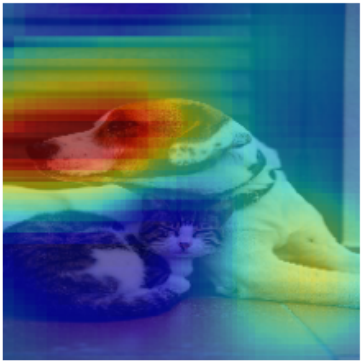


Class: Walker hound

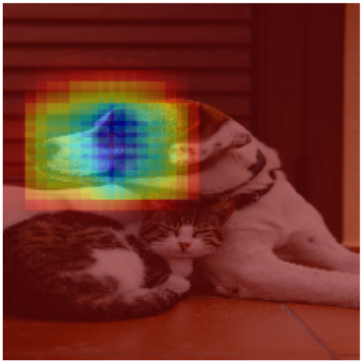


Class: Tabby Cat

Which pixels matter: Mask part of the image before it is fed to the CNN. See how much the prediction probability for a class change.

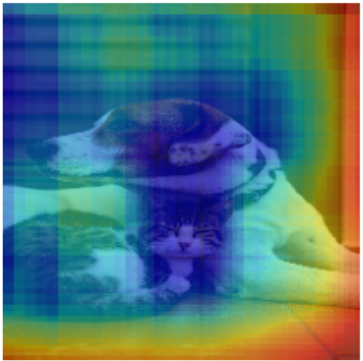


Class: Walker hound

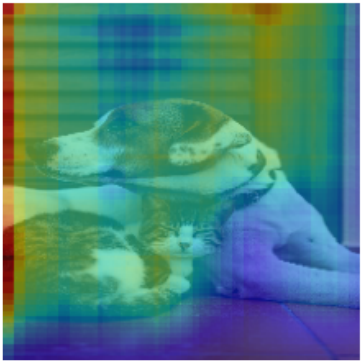


Class: Tabby Cat

What if the network is not correctly trained: Lets check Saliancy map for randomly initialized weights.



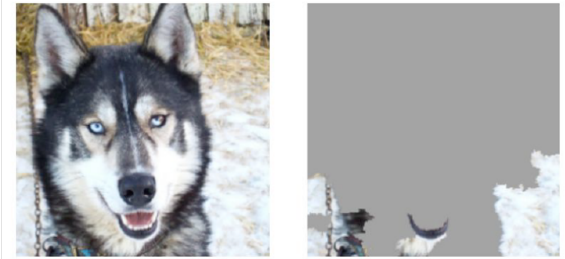
Class: Walker hound



Class: Tabby Cat

There are better ways to occlude images than rectangle patches.

Such methods can also find biases and explain why the model is making mistakes.



(a) Husky classified as wolf (b) Explanation

Image: Ribeiro et al. "Why should I trust you?" Explain the predictions of any classifier. 2016.

ResNet50 trained on imagenet does not make such mistakes as it pays attention to face.

Which pixels matter: Compute gradient of class score with respect to image pixels. Take absolute value and max over RGB channels.

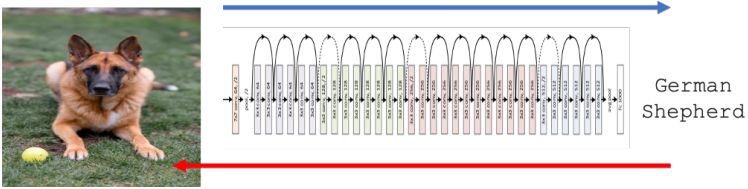
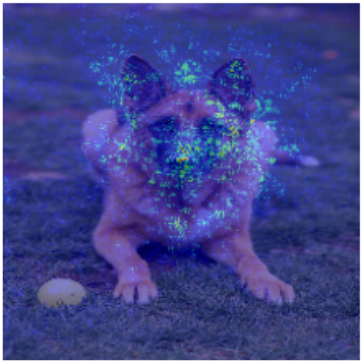
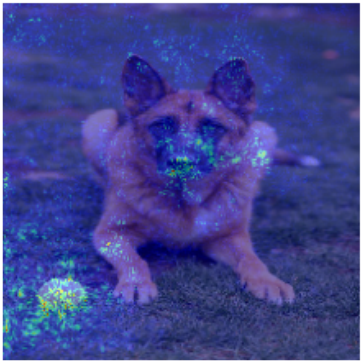


Image of the dog: Unsplash

Which pixels matter: Compute gradient of class score with respect to image pixels. Take absolute value and max over RGB channels.



Class: German Shepard



Class: Tennis Ball

Which pixels matter: Compute gradient of class score with respect to image pixels. Take absolute value and max over RGB channels.

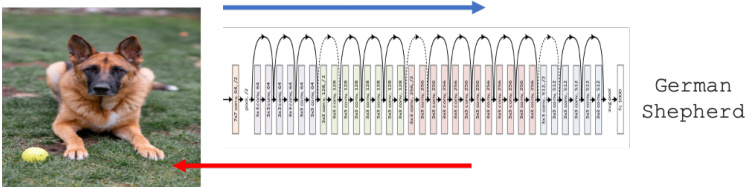


Image of the dog: Unsplash

This technique can also be used to check what pixels in an image maximally activate a given neuron.

Modern CNN architectures (e.g. ResNet, GoogleNet) has the following general structure:

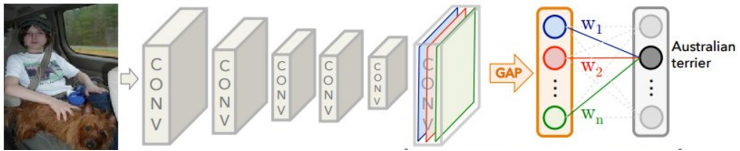


Image: Zhou et. al. CVPR 2016

- The feature elements going into the last linear classifier can be thought of as an 'attribute' in the image. E.g. element 1 maybe for human faces, element 2 for dog faces, ...
- The 'attributes' location information should be in the activation going into the GAP.
- In linear classifier feature importance is related to weights.

Modern CNN architectures (e.g. ResNet, GoogleNet) has the following general structure:

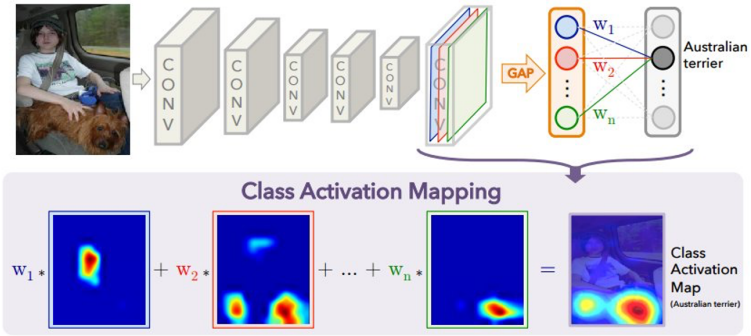


Image: Zhou et. al. CVPR 2016

An Improvement that combines gradient based saliency and Class Activation Maps:

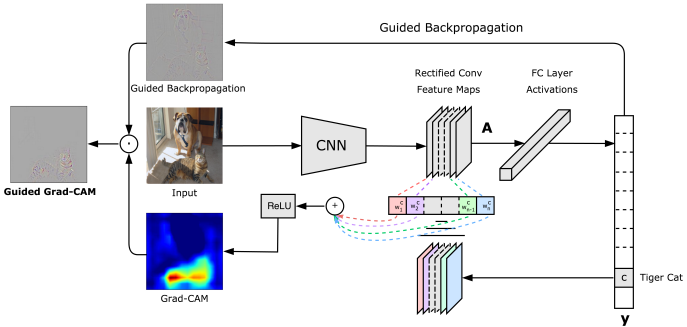


Image: Salvaraju, R et. al. ICCV 2017

Other Interesting Feature attribution methods

- **LIME**: Ribeiro et. al. "Why Should I Trust You?" Explaining the Predictions of Any Classifier, 2016.
- **SHAP, Deep SHAP**: Lundberg et. al. A unified approach to interpreting model predictions, NIPS, 2017.

Outline

Abstract Concepts Learned by Neural Network

- Imagine you build a ML model to switch traffic light that can detect fire-engines using cameras and give priority.
- If the model is trained using data from Melbourne, Can it be used in Canberra?

Fire Engine Melbourne

Abstract Concepts Learned by Neural Network

- Imagine you build a ML model to switch traffic light that can detect fire-engines using cameras and give priority.
- If the model is trained using data from Melbourne, Can it be used in Canberra?

Fire Engine Melbourne

Fire Engine Canberra

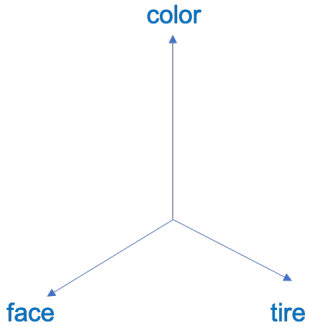
Abstract Concepts Learned by Neural Network

- Imagine you build a ML model to switch traffic light that can detect fire-engines using cameras and give priority.
- If the model is trained using data from Melbourne, Can it be used in Canberra?
- Does you fire engine detection model rely too heavily on Color?

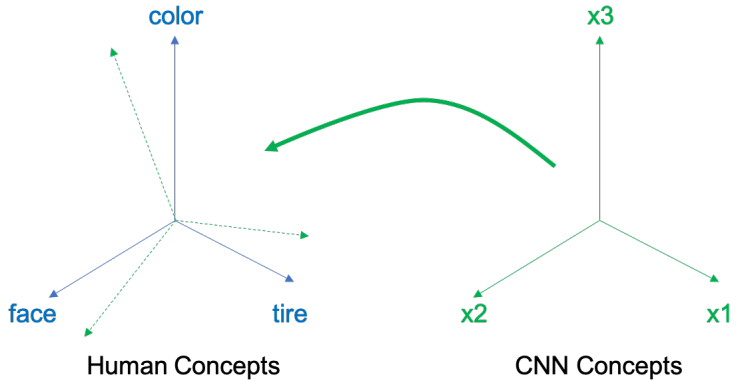
Fire Engine Melbourne

Fire Engine Canberra

Provide an interpretation of a neural net's internal state in terms of human-friendly concepts.



Provide an interpretation of a neural net's internal state in terms of human-friendly concepts.



Provide an interpretation of a neural net's internal state in terms of human-friendly concepts.

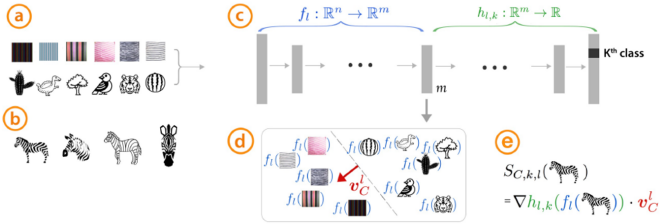


Image: Kim, B. et. al . Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML 2018.

[Step a] A concept is represented by a set of images:

- Some example image patches of the concept (e.g. stripes, red color).
- Some random image patches that does not belong to the concept.

Provide an interpretation of a neural net's internal state in terms of human-friendly concepts.

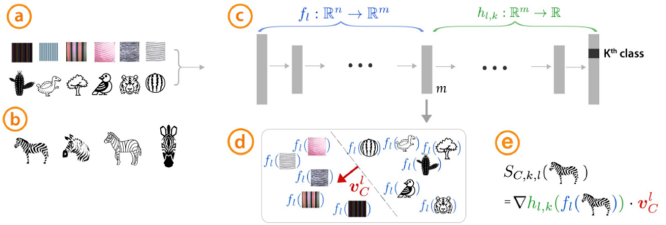


Image: Kim, B. et. al . Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML 2018.

[Step b] Select a set of images from the category we are interested in exploring (e.g. Zebra, Fire truck).

Provide an interpretation of a neural net's internal state in terms of human-friendly concepts.

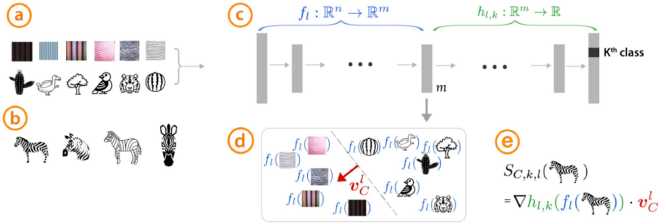


Image: Kim, B. et. al . Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML 2018.

[Step c] Forward propagate examples set (step a) through the CNN to get intermediate feature activation at a pre defined layer.

Provide an interpretation of a neural net's internal state in terms of human-friendly concepts.

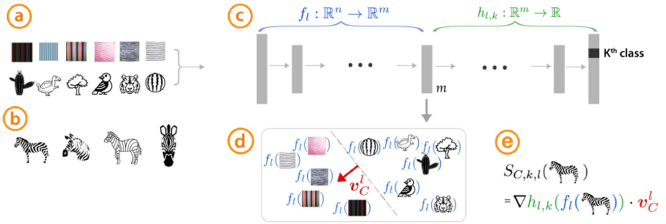


Image: Kim, B. et. al . Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML 2018.

[Step d] Do logistic regression on activations (from step c) to get the vector perpendicular to decision boundary v_C^l.

Provide an interpretation of a neural net's internal state in terms of human-friendly concepts.

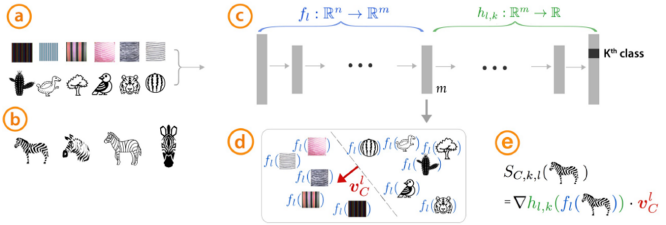


Image: Kim, B. et. al . Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML 2018.

[Step e] Compute the gradinets of class (e.g. Zebra, fire engine) with respect to hidden feature activation for for images from the class (step b).
Compute the similarity between that vector and v_C^l.
if the similarity is high the concept is important.

Summary

Basic model interpretation methods applicable to NN:

- **Feature Visualization**
- **Feature Attribution**
- **Concepts**
- Model Distillation

Next week: Revision.

Lab: Model Interpretation.
Tutorial: Bring your own questions & Assignment 2 Discussion.