Explanation techniques for neural networks

An overview with practical examples

DTU Compute
Department of Applied Mathematics and Computer Science

Different approaches

- We assume
  - Complicated non-linear task
  - Can **not** be solved by an intuitively explainable model
- Complete understanding is not possible
- Three basic approaches shown
  - Backpropagation approach
  - Local approximation
  - Network representation

Based on backpropagation

- Derivative of output in regards to input
- Optimized in most libraries
- Simple implementation
- Many variants
  - SmoothGrad [11]
  - CAM [2]
  - GradCAM [13]
  - LRP [12]
  - ...

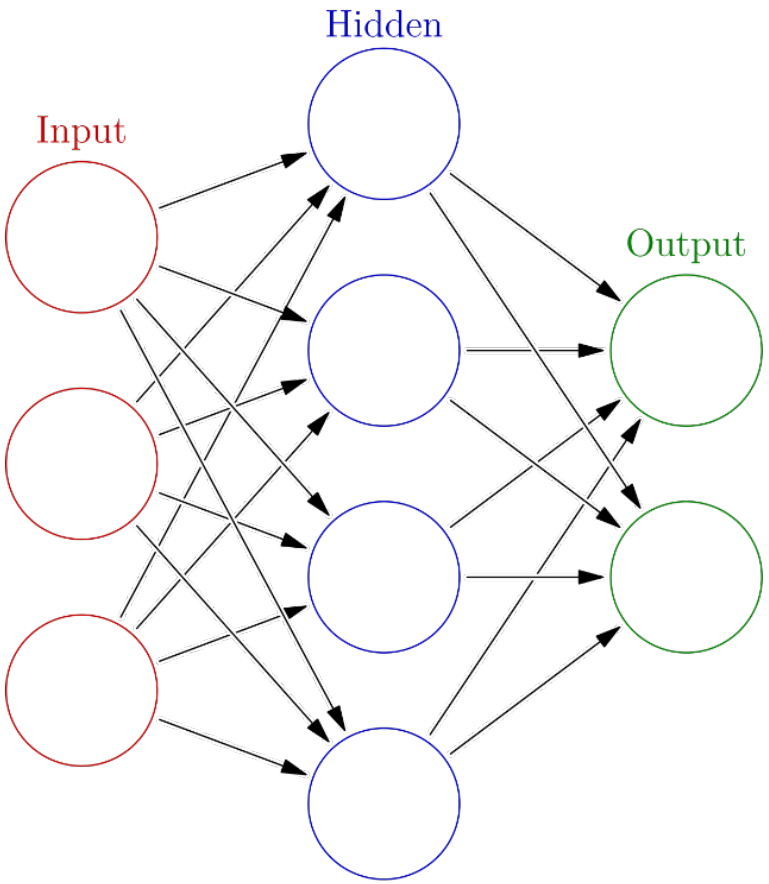


Image from https://en.wikipedia.org/wiki/Artificial_neural_network

Saliency

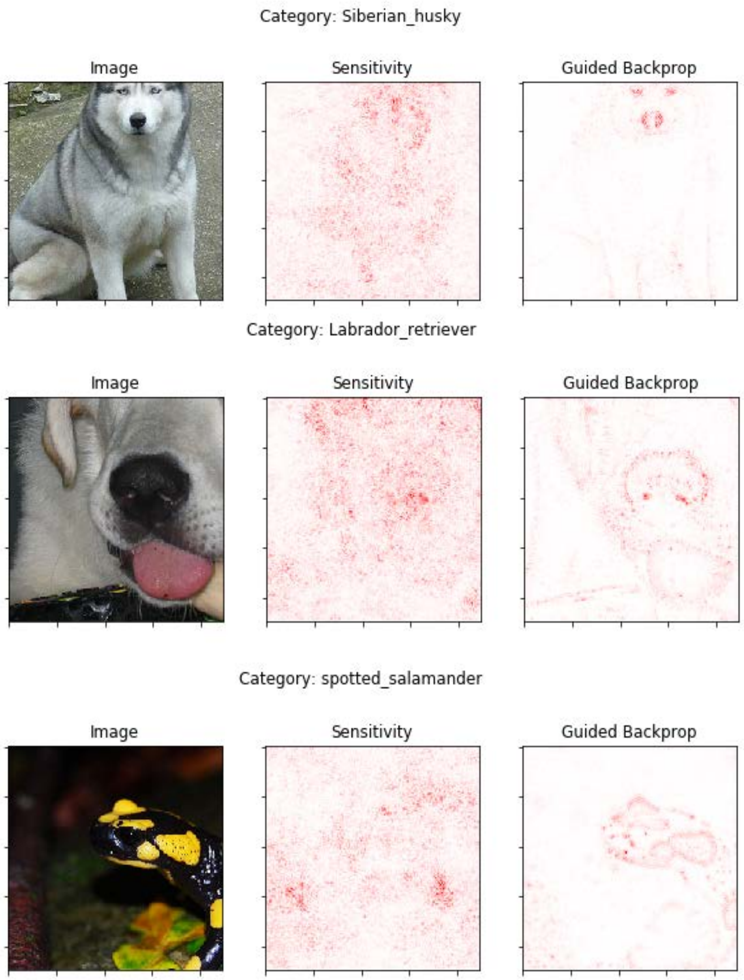
- Basic variant [1]

$$\frac{\partial y_c}{\partial x_{i,j}} = \left(\frac{\partial y_c}{\partial l_{-1}}\right) \dots \left(\frac{\partial l_1}{\partial x_{i,j}}\right)$$

- Variant: Guided Backprop [9]

$$\frac{\partial' y_c}{\partial' x_{i,j}} = \mathrm{ReLU}\left(\frac{\partial y_c}{\partial l_{-1}}\right) \dots \mathrm{ReLU}\left(\frac{\partial l_1}{\partial x_{i,j}}\right)$$

- Easy to implement
- Very noisy



Sensitivity heatmaps obtained with kerasvis library from pretrained VGG16

Gradient-weighted Class Activation Mapping (Grad-CAM)

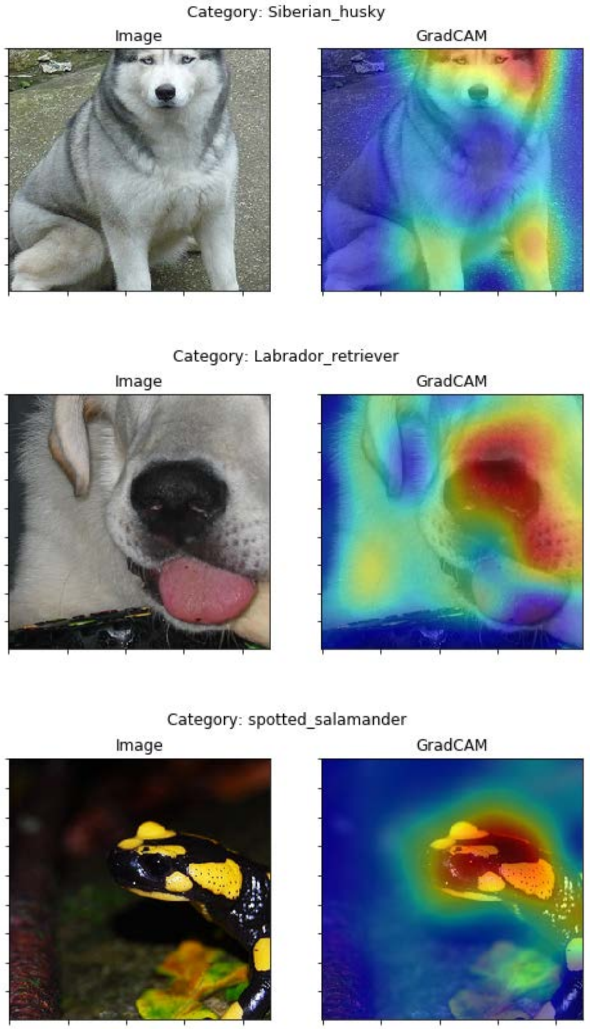
- Combining CAM[2] and gradients
- Requires CNN structure
- Coarse localization due to upfiltering
- CAM [2]:

$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

- Grad-CAM [13]:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

$$M_{GradC}^c(x,y) = ReLU\left(\sum_k a_k^c f_k(x,y)\right)$$



GradCAM heatmaps obtained with kerasvis library from pretrained VGG16

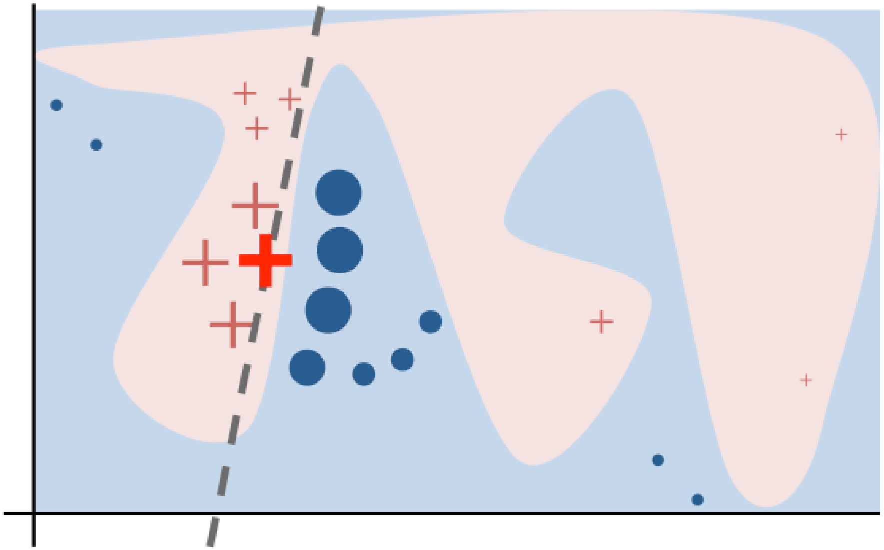
Overview

	Fidelity	Understandability	Sufficiency	Low construction overhead	Efficiency
Backprop	+	-	O	+	+
Local					
High-level					

Desiderate taken from [14]

Local approximation with interpretable model – LIME

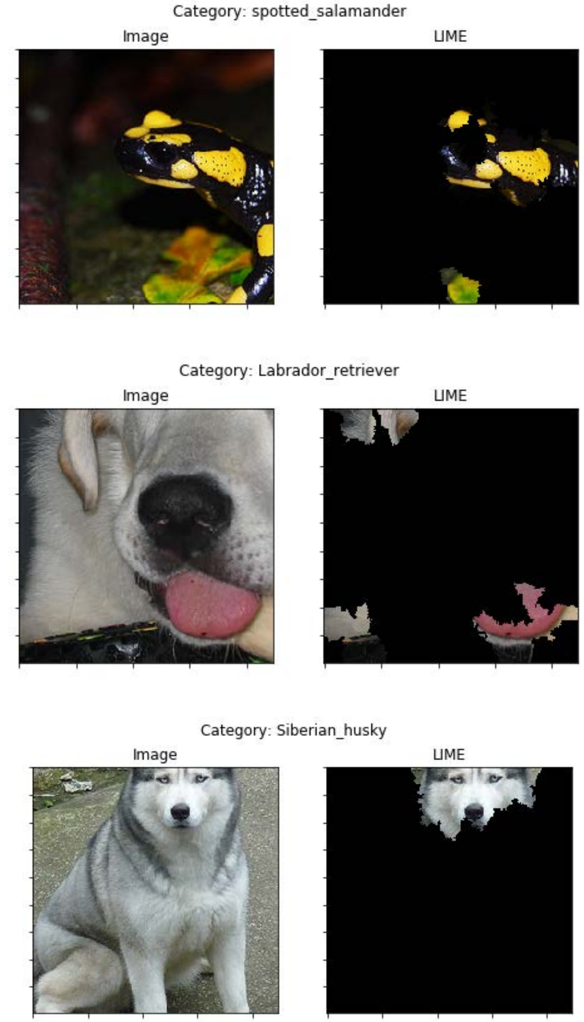
- Intuition:
  - Sample around **x**
  - Weigh samples according to distance
  - Train linear classifier
  - Obtain explanation
- Low-dimensional representation necessary
  - For images: segment into super-pixels
  - For text: bag of words



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

Local approximation with interpretable model – LIME

- Intuition:
  - Sample around **x**
  - Weigh samples according to distance
  - Train linear classifier
  - Obtain explanation
- Low-dimensional representation necessary
  - For images: segment into super-pixels
  - For text: bag of words



Images obtained with LIME library from pretrained VGG16

Overview

	Fidelity	Understandability	Sufficiency	Low construction overhead	Efficiency
Backprop	+	-	O	+	+
Local	O	+	O	+	-
High-level					

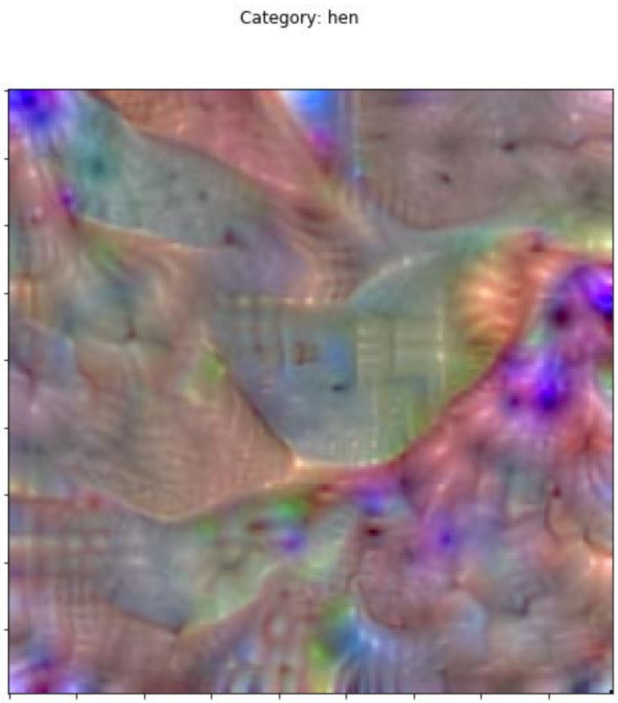
Desiderate taken from [14]

Higher-level

- Network level explanations
- Requires domain knowledge
- Interesting for risk and fairness analysis
- Two approaches presented
  - Analyzing specific network parts
  - Analyzing specific aspects

Probing the network

- “Understanding Neural Networks Through Deep Visualization” [5]
  - Idea: iteratively optimize activation of neurons with backpropagation
  - Regularize to encourage realistism
  - For output or intermediate layers
- Alternatives
  - Bau, David, et al. "Network Dissection: Quantifying Interpretability of Deep Visual Representations." *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017.
  - Alain, Guillaume, and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes." (2016).



Obtained with kerasvis
from pretrained VGG16

Testing with Concept Activation Vectors CAV

- Recently proposed by Kim et al [8]
- Requires high domain knowledge
- Idea:
  - assemble dataset {P,N}
  - Train linear classifier on representation of given layer
  - Obtain weights
- Useful for
  - Evaluating given input
  - Identifying a known bias in dataset and model

Kim, Been, et al.
"TCAV: Relative concept importance testing with Linear Concept Activation Vectors." (2018).

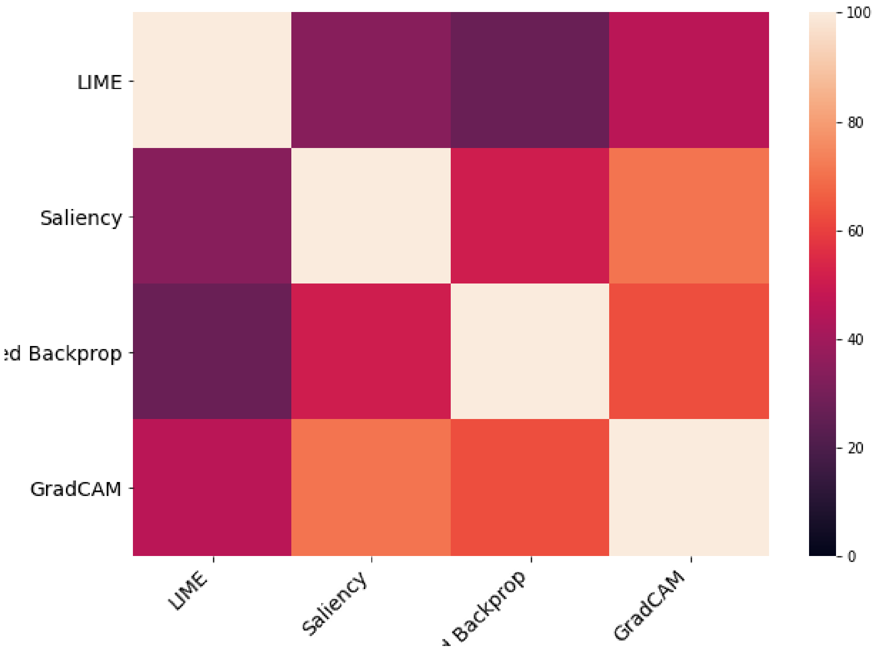
Overview

	Fidelity	Understandability	Sufficiency	Low construction overhead	Efficiency
Backprop	+	-	O	+	+
Local	O	+	O	+	-
High-level	+	+	-	-	O

Desiderate taken from [14]

How much disagreement is there?

- Expectation: explanations are correlated across similar networks
- Compute agreement between heatmaps
  - Between methods
  - Between networks
- LIME is highly changeable compared to other methods

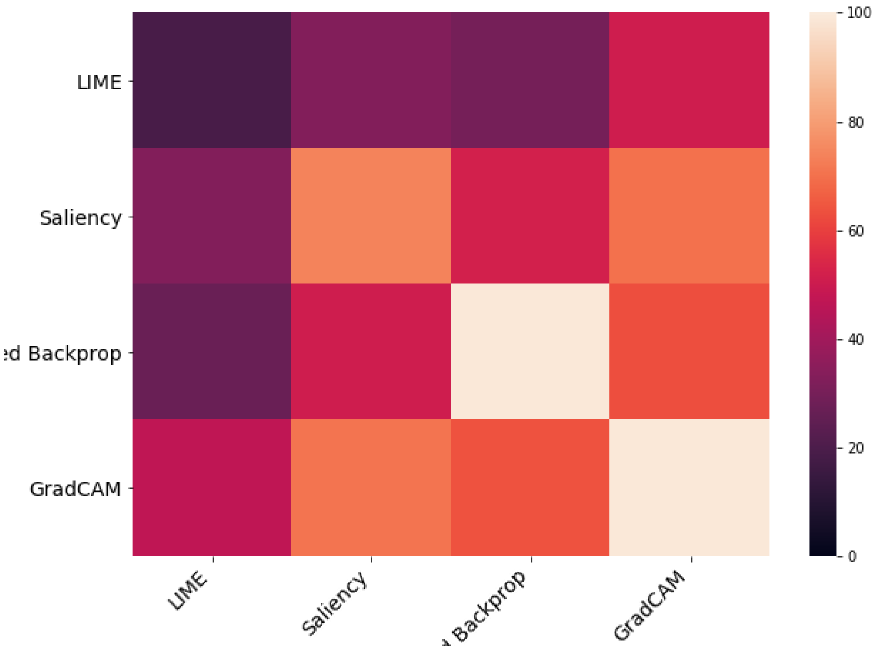


Agreement between explanation methods for VGG16

Pretrained VGG16 and VGG19 from keras,

How much disagreement is there?

- Expectation: explanations are correlated across similar networks
- Compute agreement between heatmaps
  - Between methods
  - Between networks
- LIME is highly changeable compared to other methods



Agreement between explanation methods between VGG16 and VGG19

Pretrained VGG16 and VGG19 from keras,

Take-away

- We have to make a trade-off when obtaining explanations from NNs
- Different approaches have different pros and contras
- Task in question needs to be considered

	Fidelity	Understandability	Sufficiency	Low construction overhead	Efficiency
Backprop	+	-	o	+	+
Local	o	+	o	+	-
High-level	+	+	-	-	o

References

1. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).
2. Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
3. Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu. "Interpretable Convolutional Neural Networks.”
4. 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. ACM, 2016.
5. Yosinski, Jason, et al. "Understanding neural networks through deep visualization." In ICML Workshop on Deep Learning.
6. Bau, David, et al. "Network dissection: Quantifying interpretability of deep visual representations." arXiv preprint arXiv:1704.05796 (2017).
7. Alain, Guillaume, and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes." arXiv preprint arXiv:1610.01644 (2016).
8. Kim, Been, et al. "TCAV: Relative concept importance testing with Linear Concept Activation Vectors." (2018).
9. Springenberg, Jost Tobias, et al. "Striving for simplicity: The all convolutional net." arXiv preprint arXiv:1412.6806 (2014).
10. 1Zhang Q, Wu YN, Zhu S-C. Interpretable Convolutional Neural Networks. https://arxiv.org/pdf/1710.00935.pdf. Accessed February 20, 2018.
11. Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." arXiv preprint arXiv:1706.03825 (2017).
12. Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.
13. Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." ICCV. 2017.
14. Swartout, William R., and Johanna D. Moore. "Explanation in second generation expert systems." Second generation expert systems. Springer, Berlin, Heidelberg, 1993. 543-585.