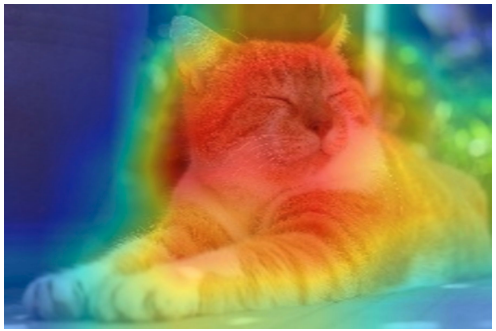
The Many Flavours of CAM

May 9 - 2022

Felipe Torres Figueroa



Centrale Marseille

LIS
LABORATOIRE D'INFORMATIQUE & SYSTÈMES
UMR 7020

Interpretability Recap: Why?

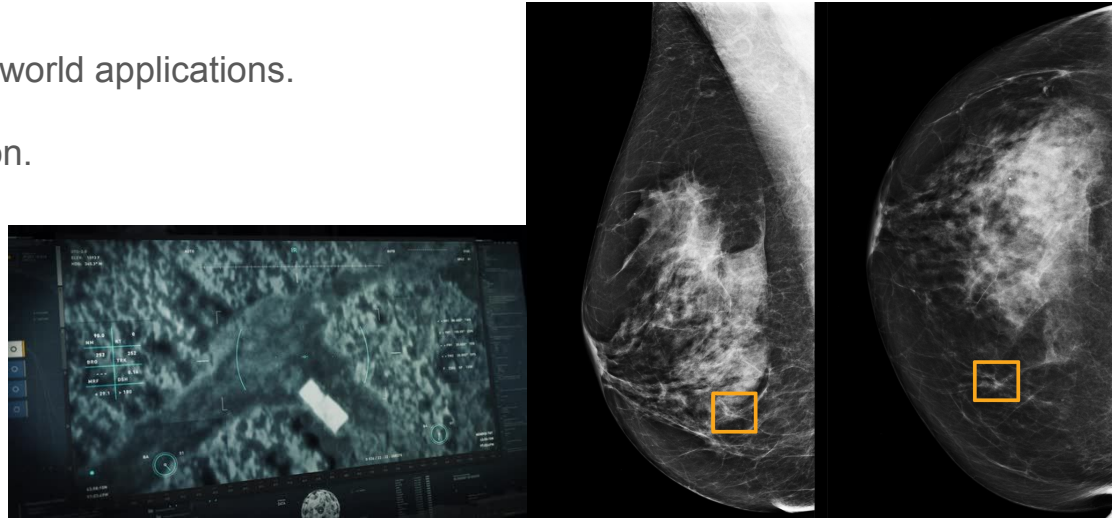
Definitions on Interpretability:

- Mechanism to understand our models? (features, parameters, training schemes)
- Process in which we uncover/understand the hidden structure of data.
- Right of an explanation of a prediction for a model and a given input.

Lipton, 2018

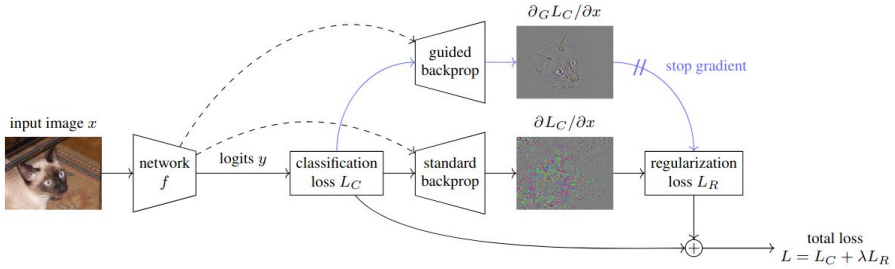
Why is Interpretability Desired:

- Science constrained bubble vs real world applications.
- Accountability and Responsibility.
- Right of an explanation for a decision.



Interpretability Recap: Transparency - Posthoc

Transparency



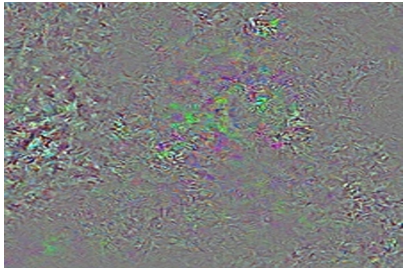
On the model

Algorithm 1: Interpretable gradient loss

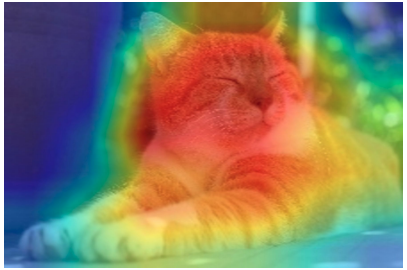
Input: network f, parameters θ
Input: input images $X = \{x_i\}_{i=1}^n$
Input: target labels $T = \{t_i\}_{i=1}^n$
Output: loss L
$L_C \leftarrow \frac{1}{n}\sum_i \mathrm{CE}(f(x_i;\theta), t_i)$ ▷ class. loss (7)
foreach $i \in \{1, \ldots, n\}$ **do**
  $\delta_G x_i \leftarrow \mathrm{DETATCH}(\partial_G L_C/\partial x_i)$ ▷ guided grad
  $\delta x_i \leftarrow \partial L_C/\partial x_i$ ▷ standard grad
$L_R \leftarrow \frac{1}{n}\sum_i E(\delta x_i, \delta_G x_i)$ ▷ reg. loss (9)
$L \leftarrow L_C + \lambda L_R$ ▷ total loss (10)

On the training

Post-Hoc Explanations

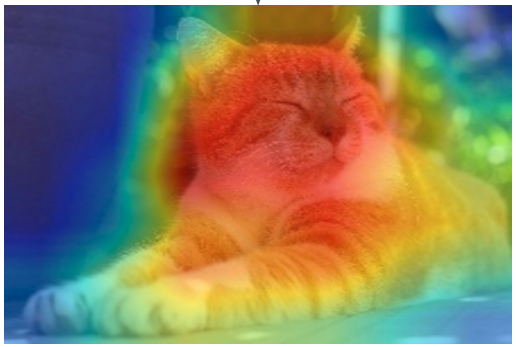


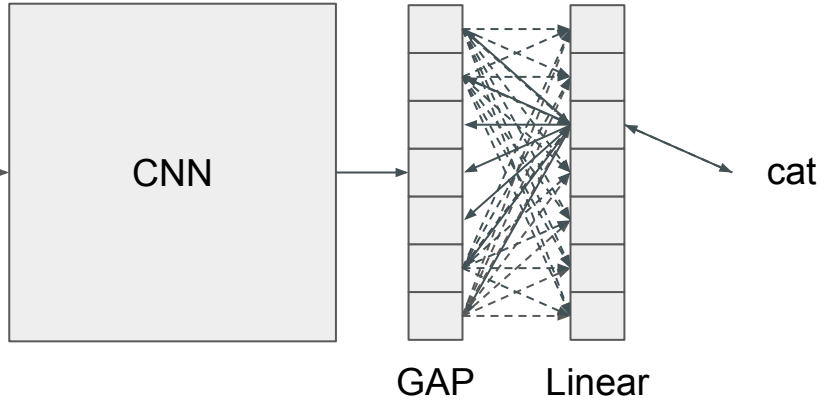
Gradient



Activations

Class Activation Maps



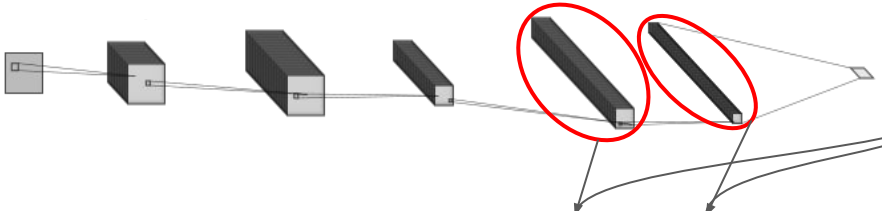


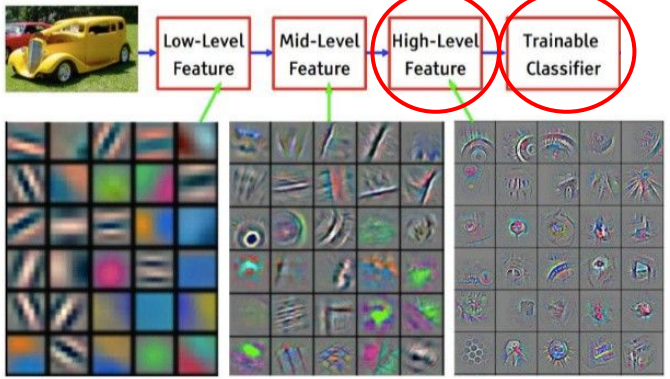
$$a_c^k = \sum_k y^c \frac{1}{Z} \sum_i \sum_j A_{i,j}^k$$

General Equation

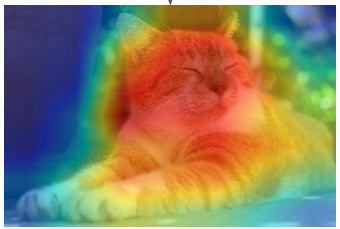
(Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016)

Why Class Activation Maps?

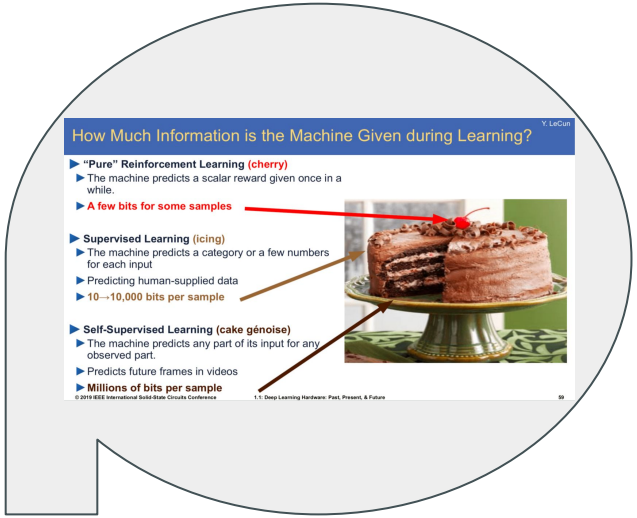




Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]



The Many Flavours of CAM



The CAM family of methods can be seen as a cake too:

$$a_c^k = \sum_k y^c \frac{1}{Z} \sum_i \sum_j A_{i,j}^k$$

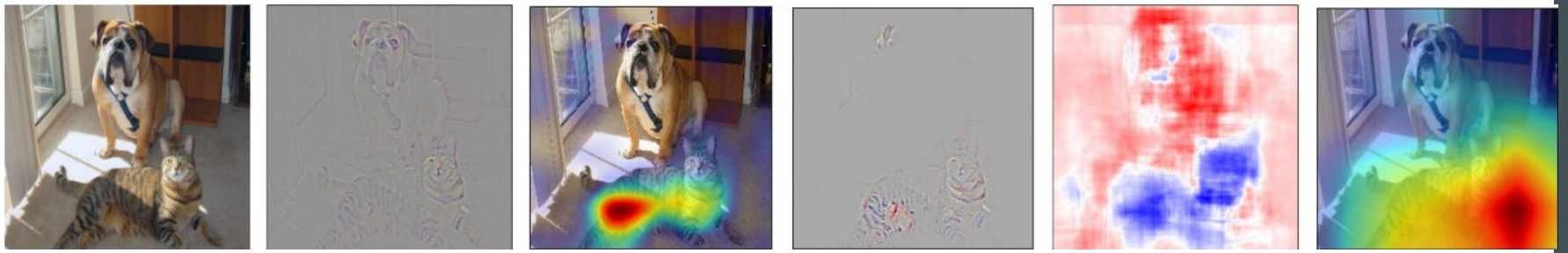
The taste of a cake can be determined by the flavour of the frosting you add to it.

Available at: https://t.co/BZZgcTW4jO

Grad-CAM

- A generalization of CAM
  - Now the weighting coefficient is obtained from the gradients flowing backwards from the classification layer. (Rumelhart, Hinton, & Williams, 1986) (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014)
  - Some networks don't have a simple classifier: i.e. VGG, thus having a CAM representation is not easy to achieve.

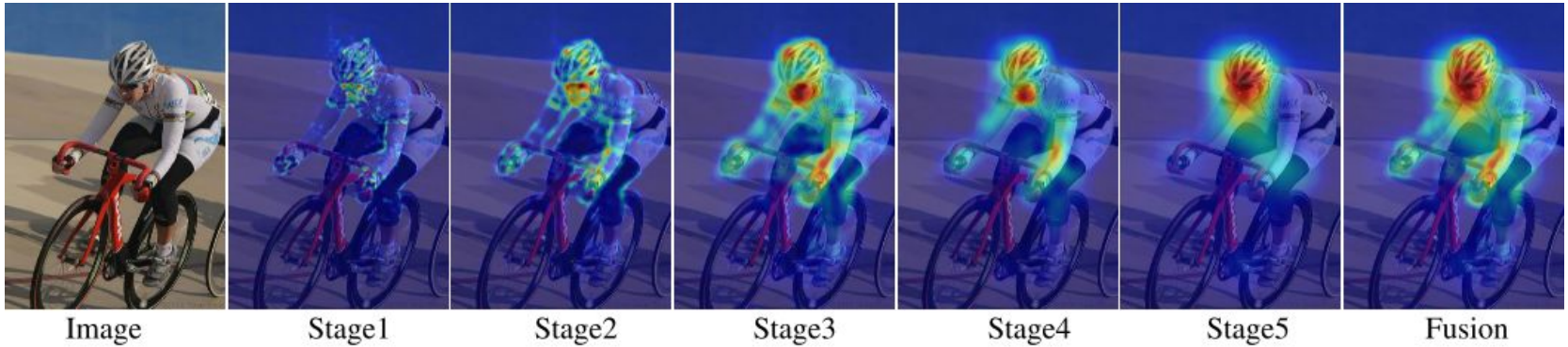
$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\delta y^c}{\delta A_{ij}^k}$$



Selvaraju et al., 2016

Layer-CAM

- Answer to the question of Grad-CAM answers on non semantic layers
  - Now we don’t take into consideration the last convolution before the classifier-
  - One convolution per layer can be taken into consideration.
  - A representation of the pyramidal structure of the network is built.

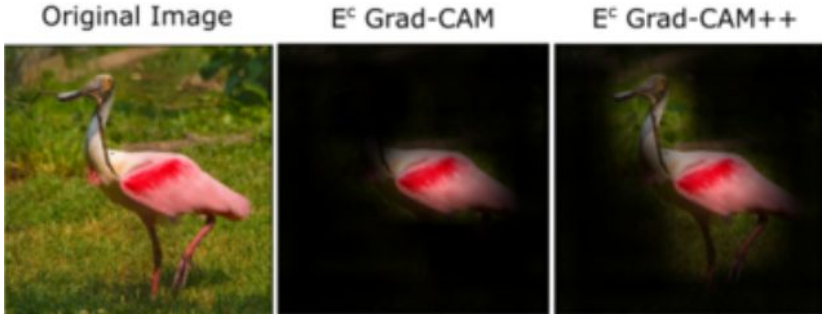


Jiang, Zhang, Hou, Cheng, & Wei, 2021

Grad-CAM++

- Use of a combination of the positive partial derivatives of the last convolutional layer's feature maps w.r.t. an specific class score as weights.
  - Improved localization and sharper activation maps.
  - Better robustness towards more objects on the image

$$a_k^c = \sum_i \sum_j w_{ij}^{kc} \circ ReLU\left(\frac{\delta Y^c}{\delta A_{ij}^k}\right) \longrightarrow w_{ij}^{kc} = \frac{\frac{\delta^2 Y^c}{(\delta A_{ij})^2}}{2\frac{\delta^2 y^c}{(\delta A_{ij}^k)^2} \sum_a \sum_b A_{ab}^k \left[\frac{\delta^3 Y^c}{(\delta A_{ij}^k)^3}\right]}$$



Chattopadhyay, Sarkar, Howlader, & Balasubramanian, 2017

Grad-CAM++: Interpretable Metrics

Faithfulness in image Recognition

- Average Drop

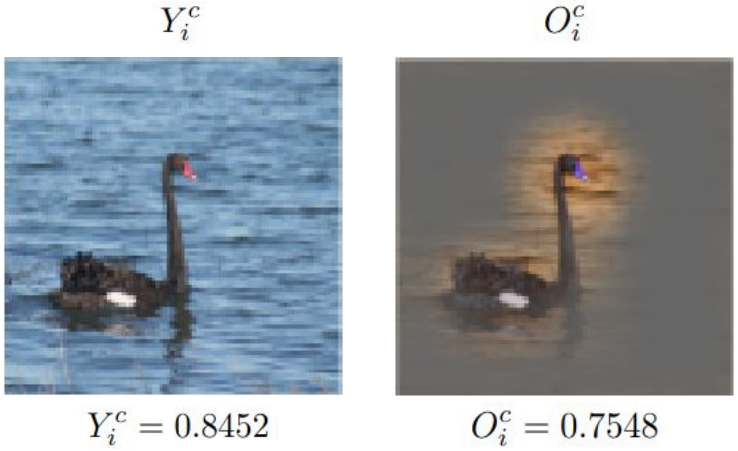
$$AD(\%) = \sum_{i}^{N} \frac{max(0, Y_i^c - O_i^c)}{Y_i^c} * 100$$

- Increase in Confidence

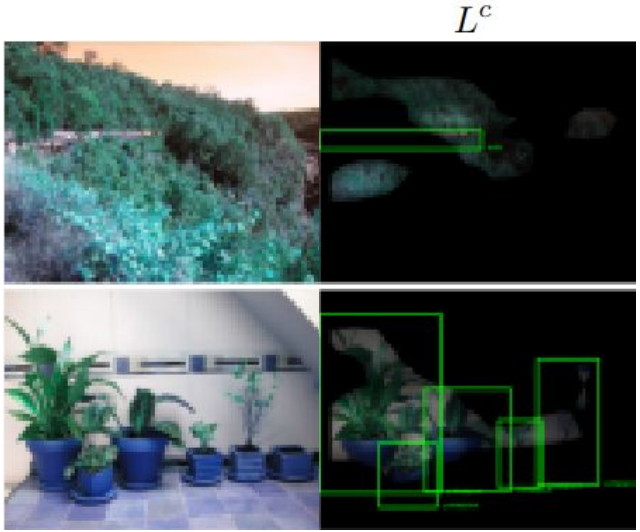
$$IC(\%) = \frac{1}{N} \sum_{i}^{N} sign(Y_i^c < O_i^c) * 100$$

- Win(%)

$$Win(\%) = \frac{1}{N} \sum_{i}^{N} sign(drop(GC++)_i < drop(GC)_i) * 100$$



Grad-CAM++: Interpretable Metrics



Harnessing Explanations for Object Loc

- Loc at threshold

$$IoU(\delta) = \frac{Area(internal)}{Area(bbox) + Area(external)}$$

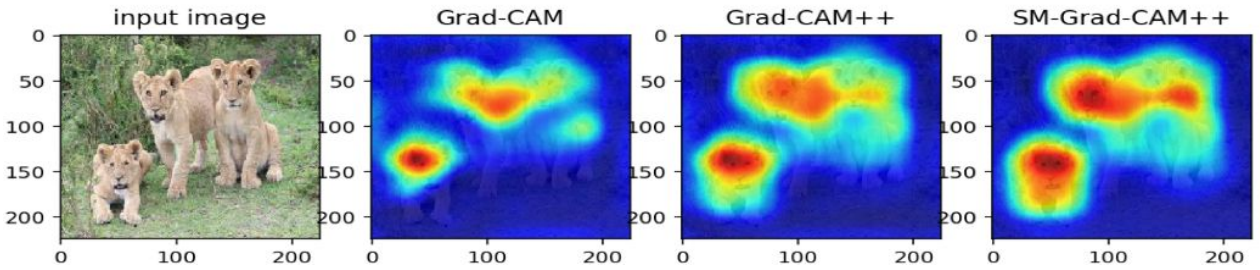
- Proportion

$$Proportion(\%) = \frac{\sum L^c_{(i,j)\in bbox}}{\sum L^c_{(i,j)\in bbox} + \sum L^c_{(i,j)\notin bbox}}$$

Smooth Grad-CAM++

- Coming from purely gradient based methods.
- Introducing noise to denoise gradients:
  - Activation functions like ReLU don't operate properly on small neighborhoods.
  - A way of improving gradient computation to get better weighting coefficients.

$$\hat{M_c}(x) = \frac{1}{n}\sum_{1}^{n} M_c(x + N(0, \sigma^2))$$



Smilkov, Thorat, Kim, Viégas, & Wattenberg, 2017

Score-CAM

What if we remove the need of gradients and focus on increasing the model's confidence?

$$C(A_l^k) = F(x \circ s(Up(A_l^k))) - F(x_b)$$

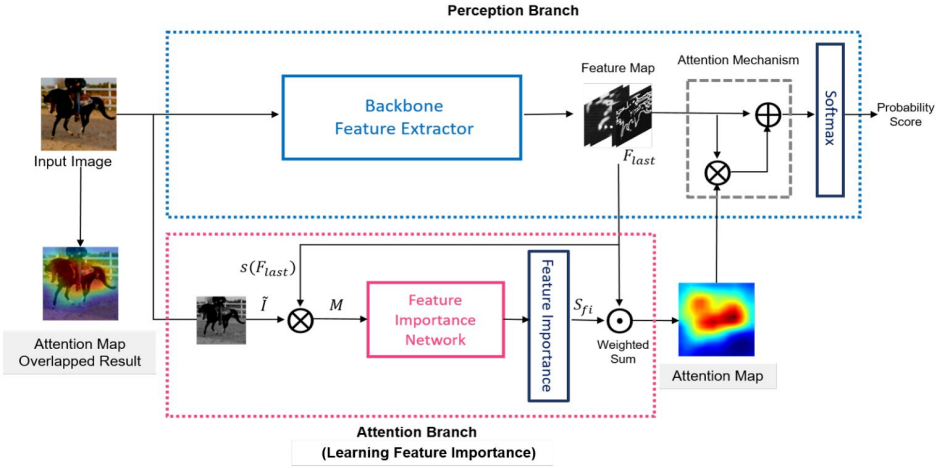
Wang, Du, Yang, & Zhang, 2019

CAM constraints.

- CAM takes on an specific convolutional layer to get insight into the network, many more layers could be looked upon. (Layer CAM)
- Activation Map size on semantic level matters.
- We want insight on the network’s decision process, not solely on the groundtruth responses.

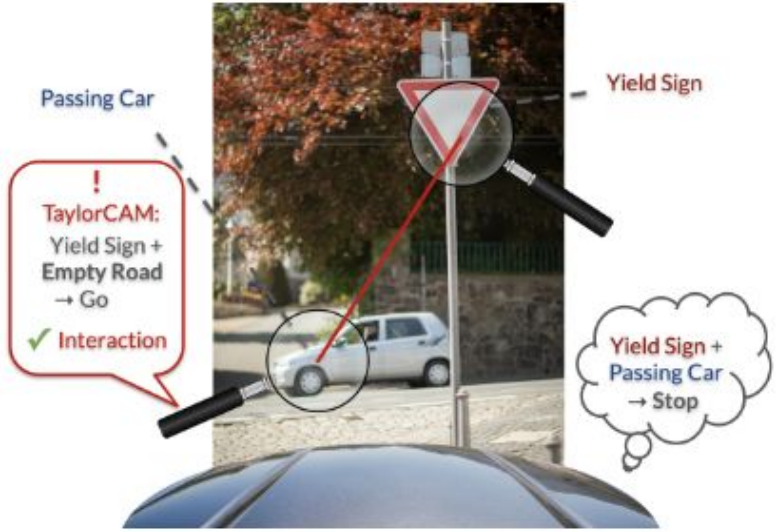
CAM currently: Novel Ideas

LFI-CAM



Lee, Park, Oh, & Kwak 2021

Taylor CAM



Lerman, Xu, Venuto, & Kautz, 2020

Novelty Penalizations.

- Interpretability is a buzz word, alike snake oil.
- What?, How?, Why? How does it compare?
- Current approaches do not present Interpretable metrics!
- Interpretability is often seen as an ablation to a recognition/localization method, not a main point.

What can we do with CAM?

- We have observed a switch of the paradigm with Score-CAM removing the reliance on gradients.
- Conversely, the way in which we compute weights is not finite; thus CAM-methods could have a limit on utilization of network parameters.
- Training a neural network with CAM as attention could yield good results.
- We can a different method to yield good explanations on terms with metrics.

Metrics still present a point of debate.

Transformers are beyond the scope of today’s talk.

Recommended reading

- Ablation CAM.(Desai & Ramaswamy, 2020)
- Integrated Gradients (Sundararajan, Taly, & Yan, 2017)
- Smoothed Score-CAM. (Naidu & Michael, 2020)
- Jacobgil's Pytorch CAM library (Gildenblat & contributors, 2021)

References:

- Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2017). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. CoRR, abs/1710.11063 . Retrieved from http://arxiv.org/abs/1710.11063
- Desai, S., & Ramaswamy, H. G. (2020). Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In 2020 ieee winter conference on applications of computer vision (wacv) (p. 972-980). doi: 10.1109/WACV45572.2020.9093360
- Gildenblat, J., & contributors. (2021). Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam . GitHub.
- Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., & Wei, Y. (2021). LayerCAM: Exploring hierarchical class activation maps for localization. Proc. IEEE Trans. Image. Process.(TIP), 30 , 5875–5888.
- Lee, K. H., Park, C., Oh, J., & Kwak, N. (2021). LFI-CAM: learning feature importance for better visual explanation. CoRR, abs/2105.00937 . Retrieved from https://arxiv.org/abs/2105.00937
- Lerman, S., Xu, C., Venuto, C., & Kautz, H. A. (2020). Explaining local, global, and higher-order interactions in deep learning. CoRR, abs/2006.08601 . Retrieved from https://arxiv.org/abs/2006.08601

References

- Lipton, Z. C. (2018, jun). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16 (3), 31–57. Retrieved from https://doi.org/10.1145/3236386.3241340 doi: 10.1145/3236386.3241340
- Naidu, R., & Michael, J. (2020). SS-CAM: smoothed score-cam for sharper visual feature localization. CoRR, abs/2006.14255 . Retrieved from https://arxiv.org/abs/2006.14255
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323 (6088), 533–536.
- Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-cam: Why did you say that? Visual explanations from deep networks via gradient-based localization. CoRR, abs/1610.02391 . Retrieved from http://arxiv.org/abs/1610.02391
- Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. CoRR, abs/1706.03825 . Retrieved from http://arxiv.org/abs/1706.03825
- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 .
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. CoRR, abs/1703.01365 . Retrieved from http://arxiv.org/abs/1703.01365

References

- Wang, H., Du, M., Yang, F., & Zhang, Z. (2019). Score-cam: Improved visual explanations via score-weighted class activation mapping. CoRR, abs/1910.01279 . Retrieved from http://arxiv.org/abs/1910.01279
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2921–2929).