FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Computer
Graphics
Charles
University

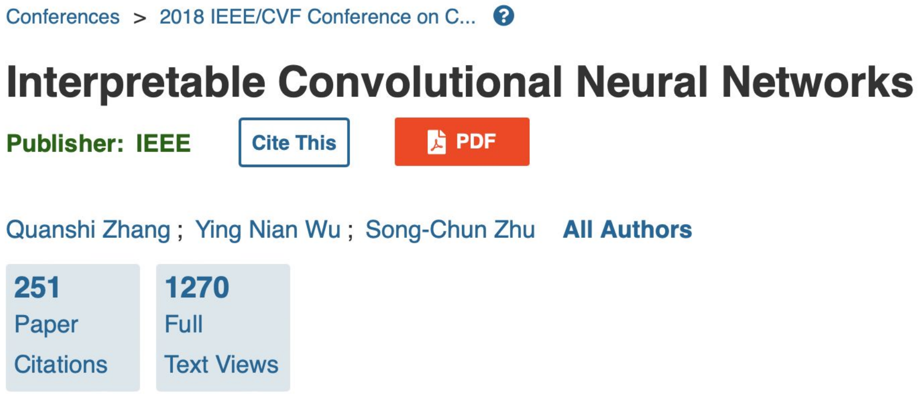
Interpretable Convolutional Neural Networks

2022.12.09

Ph.D. Student : Neşe Güneş, M.Sc.
Thesis Advisor : Doc. RNDr. Elena Šikudová, Ph.D.

About article

IEEE Conference on Computer Vision and Pattern Recognition, CVPR



Conferences > 2018 IEEE/CVF Conference on C...

Interpretable Convolutional Neural Networks

Publisher: IEEE

Quanshi Zhang ; Ying Nian Wu ; Song-Chun Zhu All Authors

251 Paper Citations

1270 Full Text Views

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

Clever Hans

Right for the wrong reasons

Clever Hans performing in 1904

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

Why interpretability?

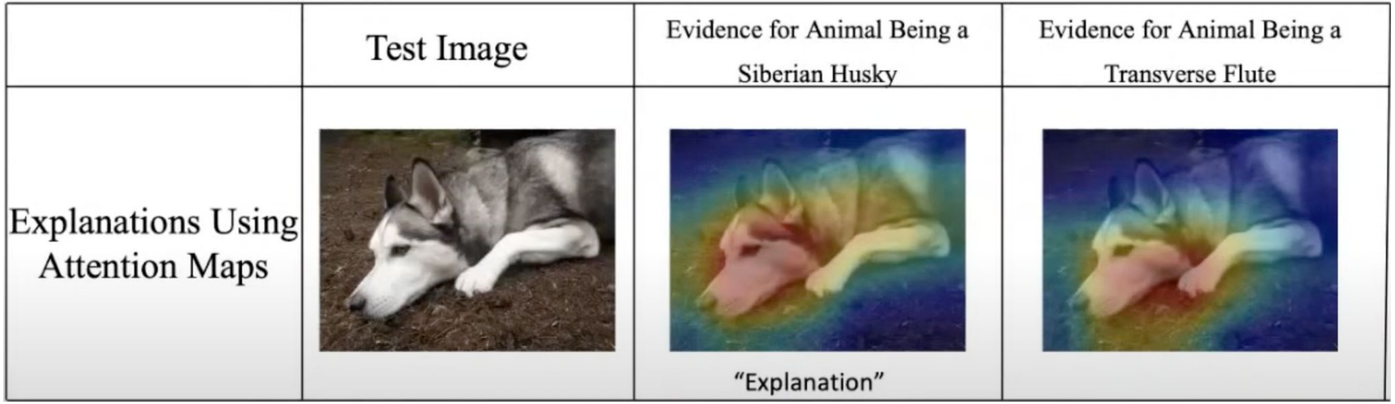
- Confounding: Right for the wrong reasons – Clever Hans
- High-stakes decisions: Should patient get a biopsy?
- Responsibility: It’s the doctor’s responsibility to make a good decision
- Black box models turn computer-aided decisions into automated decisions
  - Doctors won’t have the classification results
  - But explanations

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

ABSTRACT

What about explanations?

- Explaining deep NNs with saliency maps does not work



	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps		“Explanation”	

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

ABSTRACT

Problem scope

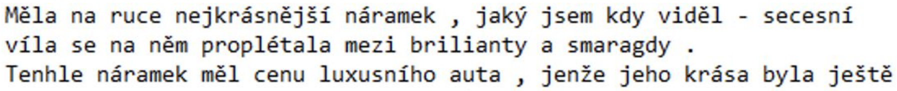
Tabular Data

- All features are interpretable
- Features include numerical and categorical data

Age	36
Gender	F
Exercise?	Yes
Smoking?	No
Diabetes?	No

Raw data

- Features are individually uninterpretable
- Pixels, voxels, words, a bit of sound wave

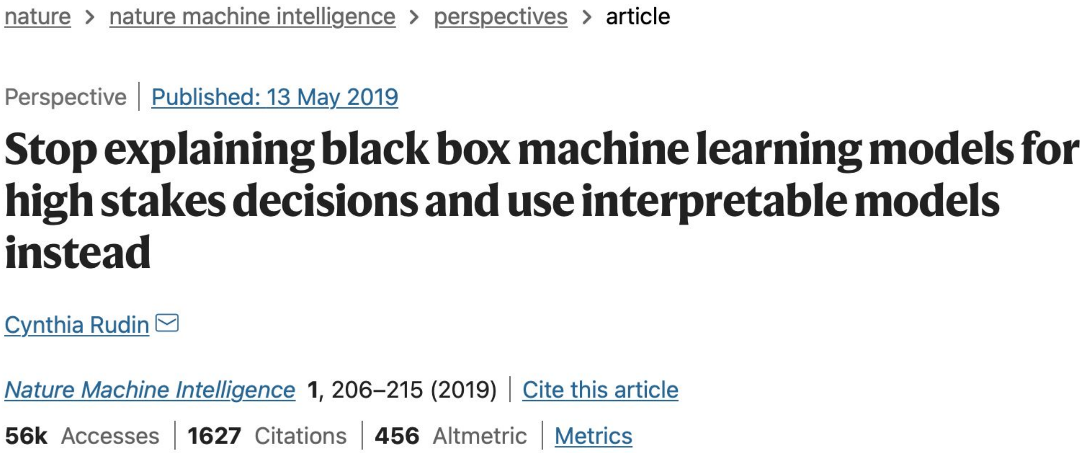


```
Měla na ruce nejkrásnější náramek , jaký jsem kdy viděl - secesní
víla se na něm proplétala mezi brilianty a smaragdy .
Tenhle náramek měl cenu luxusního auta , jenže jeho krása byla ještě
```

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

ABSTRACT

Explainable or Interpretable?



nature › nature machine intelligence › perspectives › article

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article

56k Accesses | **1627** Citations | **456** Altmetric | Metrics

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

From feature maps to image regions

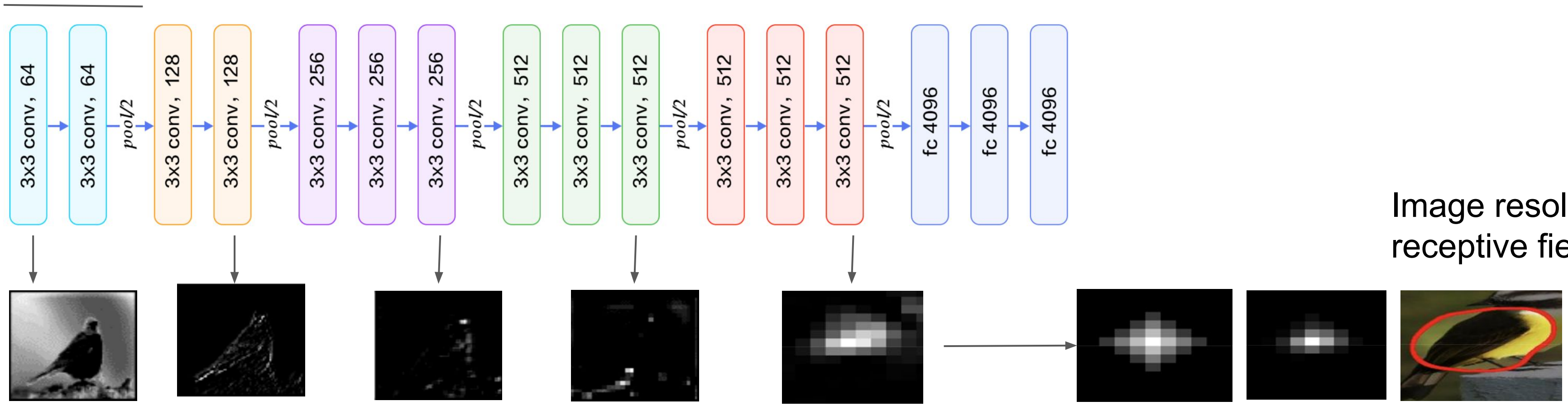


Image resolution receptive field

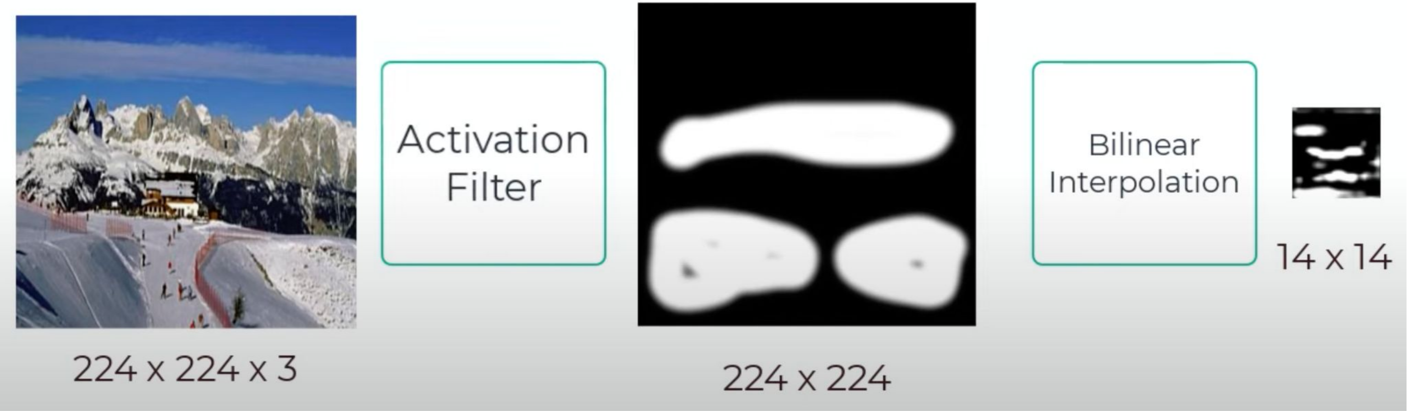
Raw map

Mask

Masked map

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

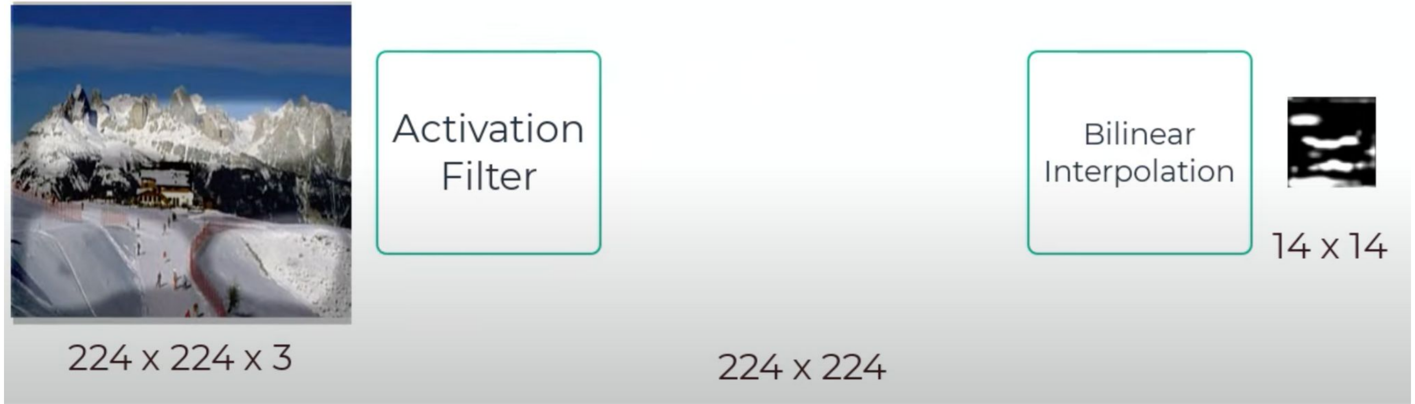
From feature maps to “segmentation” based on filters



[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

INTRODUCTION

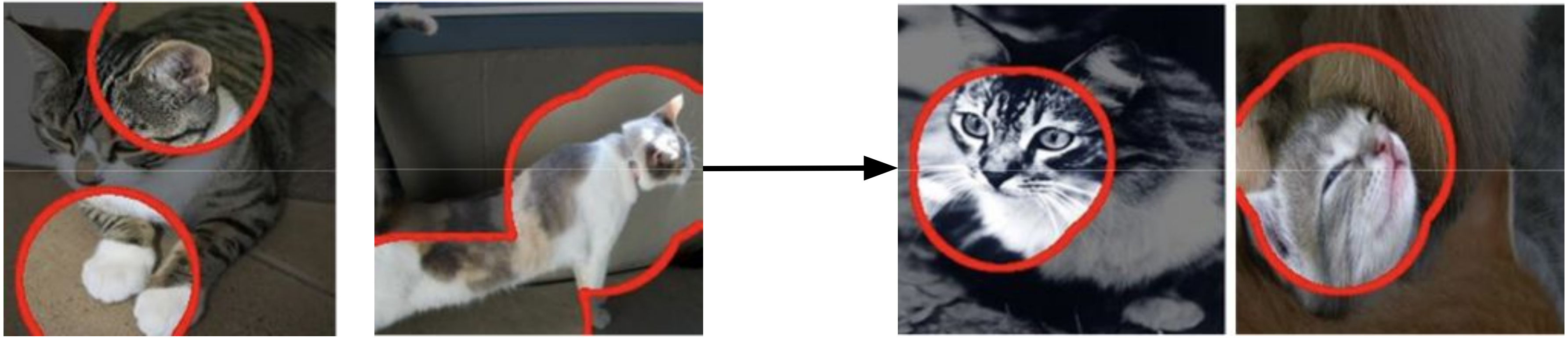
From feature maps to “segmentation” based on filters



[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

In a nutshell

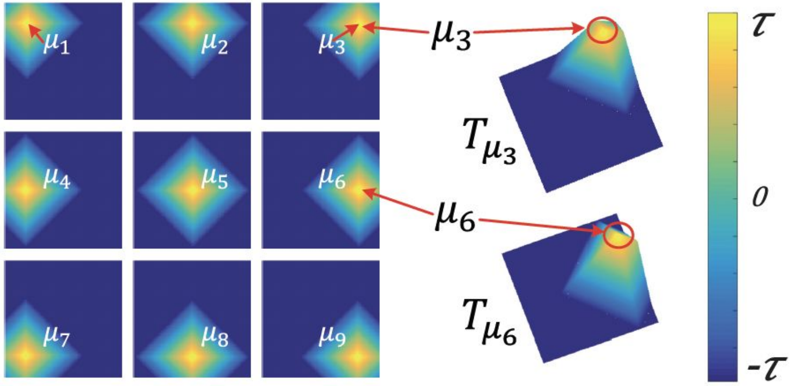
- From mixture of patterns to object-parts



[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

How to regularize filters?

- ❑ Understanding the local filter loss
  - ❑ Forgetting irrelevant information
  - ❑ Filter out noisy activations
- ❑ Forward propagation
  - ❑ Part template selection
- ❑ Backpropagation
  - ❑ Determining a target category for each filter



[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

EXPERIMENTS

Time to experiment

- ❏ Single-category, multi-category classification
  - ❏ Back to binary classification!
- ❏ Metrics: object-part interpretability, location stability: Avoid Picasso-filters
- ❏ Ground truth annotations for evaluation: GTs are still necessary
- ❏ Four types of CNNs
  - ❏ No ResNet: Who explains the ResNet wins the game

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

Evaluation metrics

- ❑ Object-part interpretability
- ❑ Stability of object-part locations
  - ❑ Categories to filters
  - ❑ Location deviation of object-parts

Object-parts of a cat
Location stability

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

CONCLUSION

Discussion

- Designing filters only for top conv layer
  - Future work: designing filters for low conv layers
- Shared object-parts by different categories
- Segmentation datasets and annotations

[1] Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

FACULTY OF MATHEMATICS AND PHYSICS
Charles University

Computer Graphics
Charles University

Neşe Güneş, Ph.D. Student in Ophthalmic Medical Imaging
Charles University, Prague, Czech Republic
E-mail: gunes.nese@matfyz.cuni.cz
Phone: +420 606 318 809

Na zdraví!