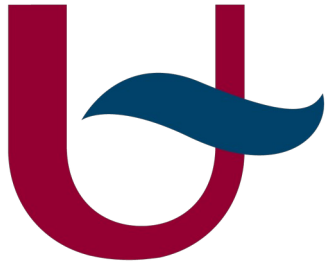


Universiteit
Antwerpen

# MAKING LEARNING-BASED VISUAL REPRESENTATIONS MORE INTELLIGIBLE

JOSÉ ORAMAS MOGROVEJO

Email: Jose.Oramas@uantwerpen.be
Twitter: @jaom7

# OUTLINE
## WHAT I WILL TALK ABOUT…

- Intro + Context
- Model Interpretation & Explanation
- Current Research + Evaluation
- Conclusion

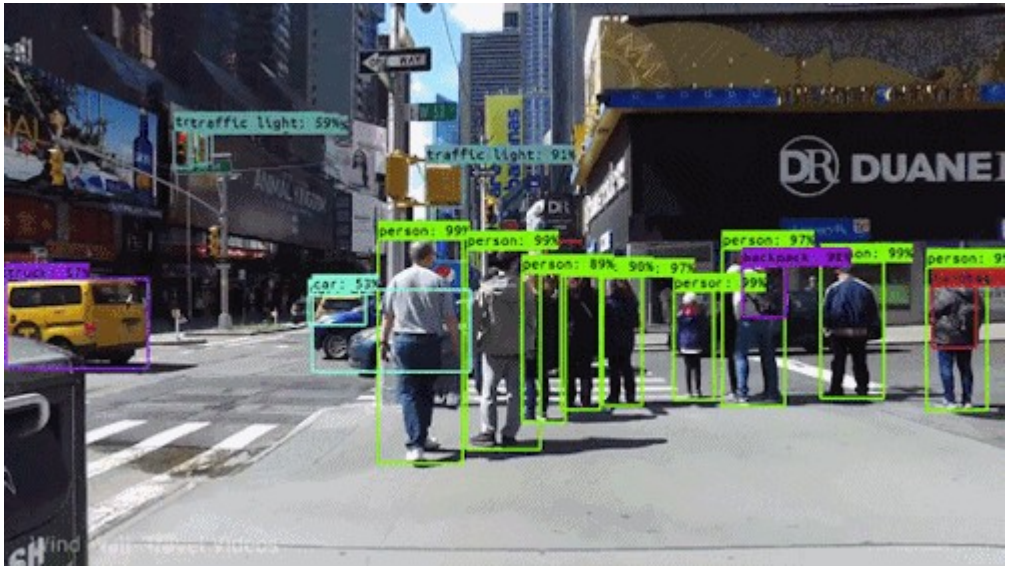
# RESEARCH BACKGROUND

# RESEARCH BACKGROUND
## COMPUTER VISION

- **Goal:**
  Provide Computer Systems/Algorithms with the sense of Sight

# RESEARCH BACKGROUND
## COMPUTER VISION

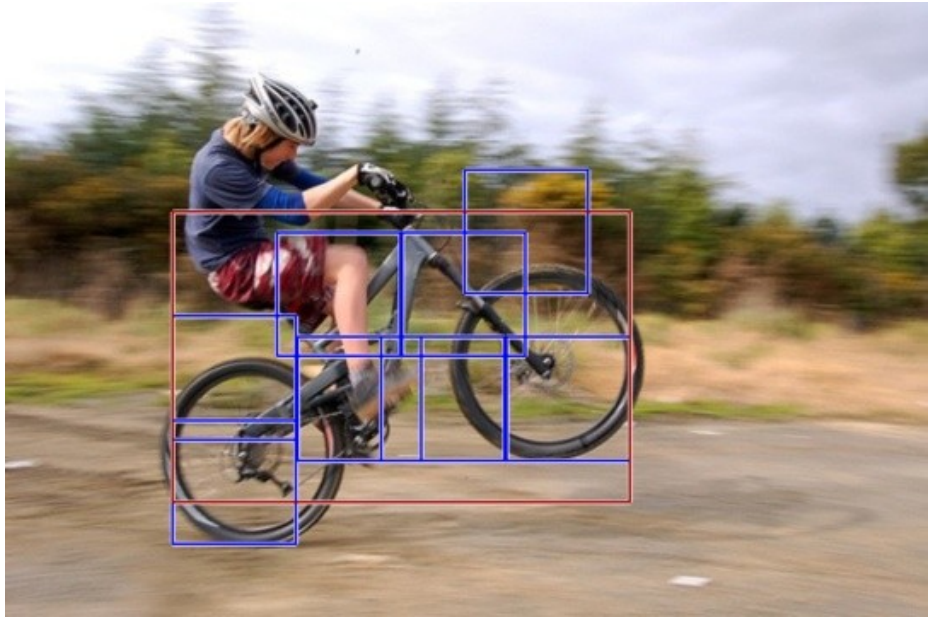
- Recognize and localize objects, actions, etc. on visual data ( images and video )

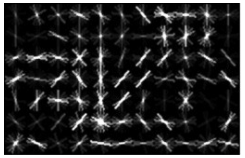


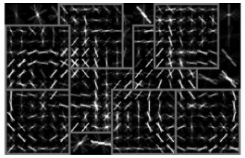
# RESEARCH BACKGROUND
## COMPUTER VISION

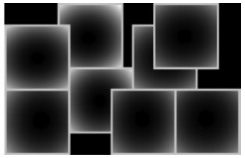
- **How to do that?**

  **Deformable-Parts Model (DPM)**
  [ Felzenszwalb et al. TPAMI'10 ]







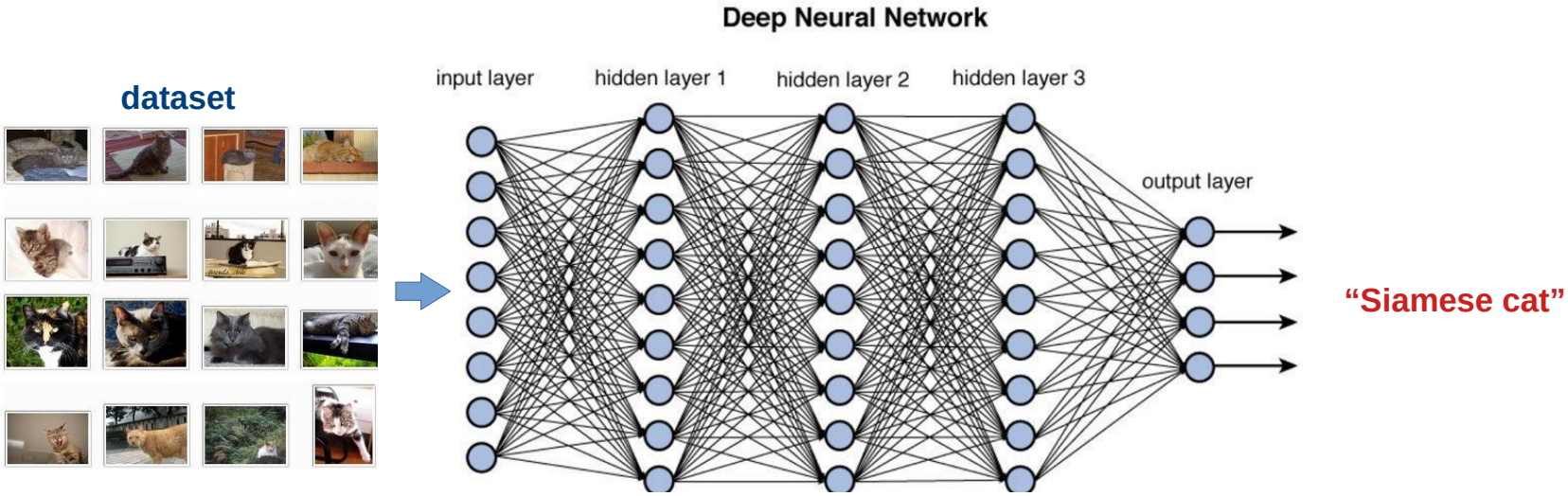


# THE DEEP LEARNING AGE

## ( LEARNING-BASED REPRESENTATIONS )

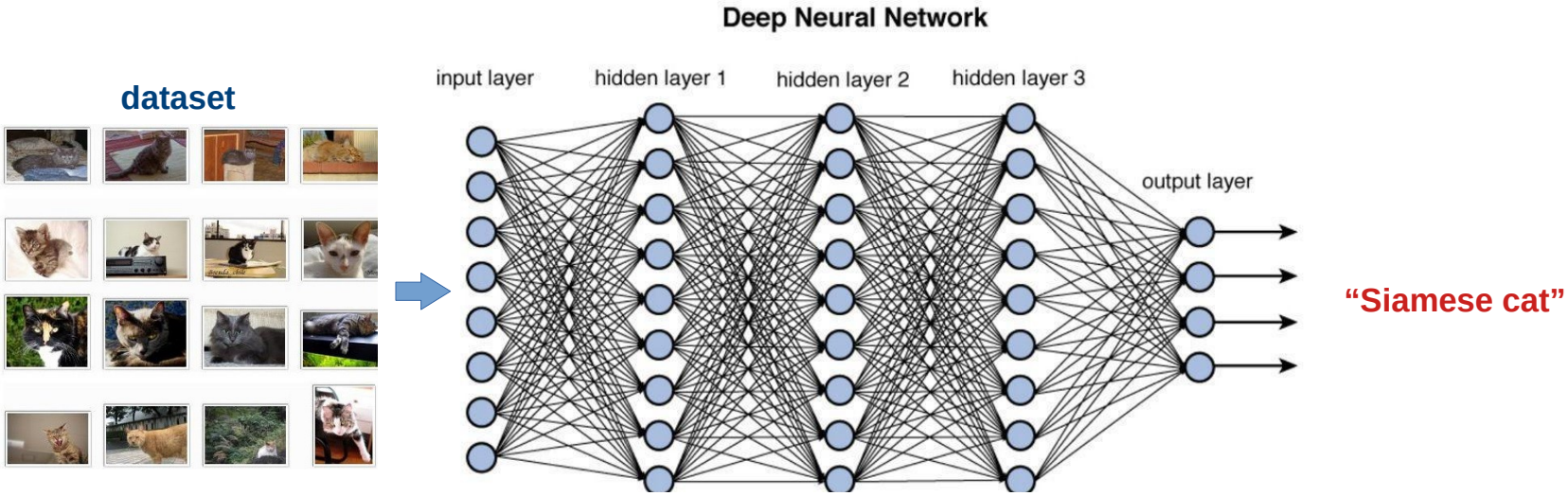
# THE AGE OF DEEP LEARNING

## LEARNING-BASED REPRESENTATIONS



# THE AGE OF DEEP LEARNING

## LEARNING-BASED REPRESENTATIONS



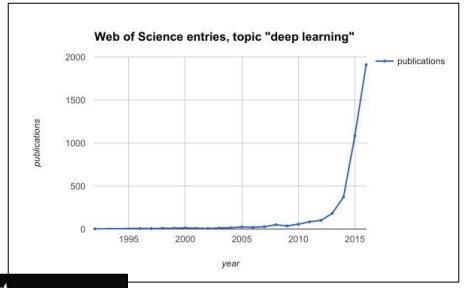

Idea: Let the model figure out what features are important
( i.e. representation learning )

# THE AGE OF DEEP LEARNING
## IN THE NEWS

**Microsoft's speech recognition engine listens as well as a human**

"This is an historic achievement" - Xuedong Huang

Andrew Tarantola, @terrortola
10.18.16 in Personal Computing




The Big Read **Driverless vehicles** + Add to myFT

**Driverless cars inspire a new gold rush in California**

MAY 23, 2017 by: Leslie Hook and Tim Bradshaw

The Washington Post
*Democracy Dies in Darkness*

Innovations

**Google's AlphaGo beats the world's best Go player — again**

By Hamza Shaban May 26

WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

TOM SIMONITE BUSINESS 10.25.2019 03:00 AM

**Google Search Now Reads at a Higher Level**

The company is incorporating new software that better understands subtleties of language, with the biggest changes for queries outside the US.

Intelligent Machines

**Deep-Learning Machine Listens to Bach, Then Writes Its Own Music in the Same Style**

Can you tell the difference between music composed by Bach and by a neural network?

by Emerging Technology from the arXiv December 14, 2016

Article | Open Access | Published: 29 August 2019

**Deep Learning to Improve Breast Cancer Detection on Screening Mammography**

Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride & Weiva Sieh

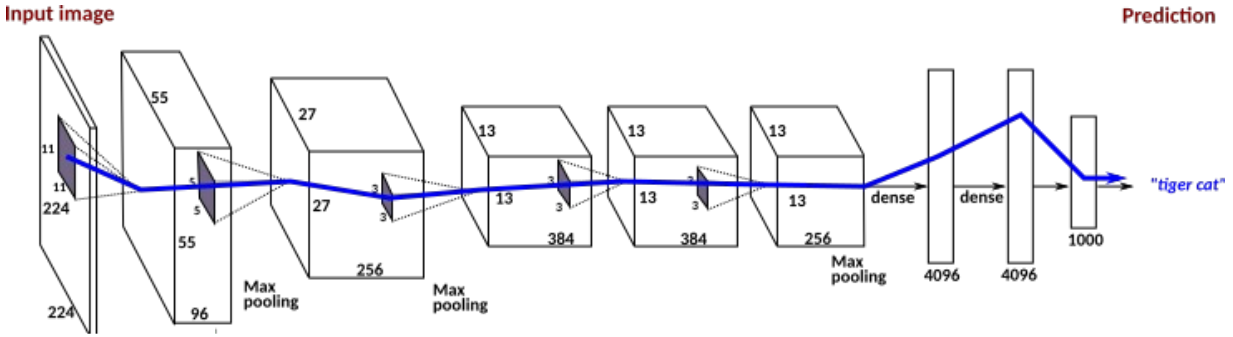
*Scientific Reports* **9**, Article number: 12495 (2019) | Cite this article

**9229** Accesses | **2** Citations | **27** Altmetric | Metrics

SO FAR SO GOOD,
BUT...

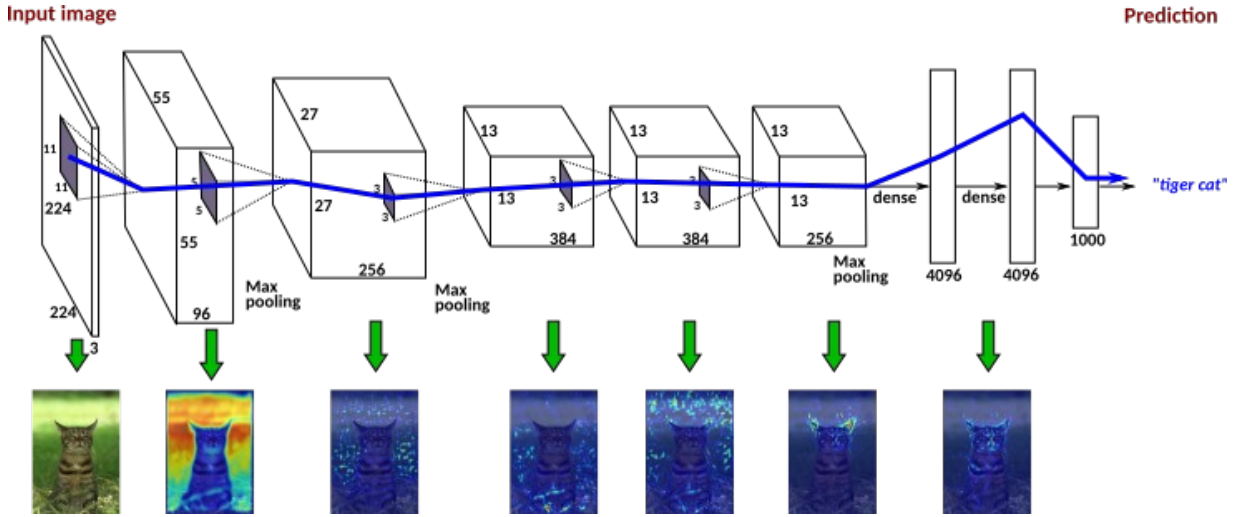
# PROBLEM STATEMENT

## RESEARCH QUESTIONS:



# PROBLEM STATEMENT
## RESEARCH QUESTIONS:

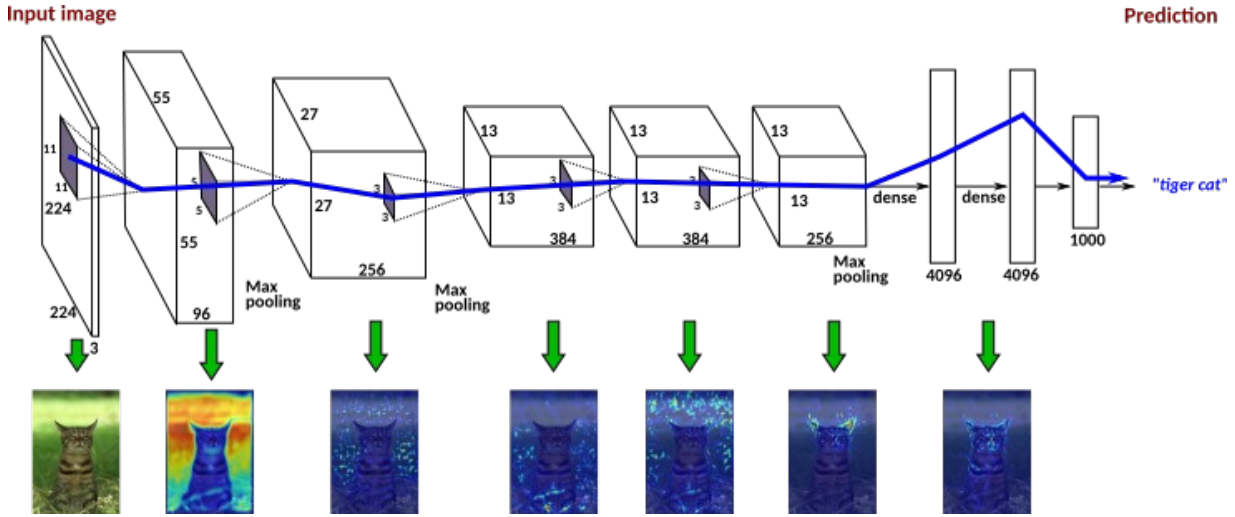
- Q1 : What the model has actually learned? → ***Interpretation***



# PROBLEM STATEMENT

## RESEARCH QUESTIONS:

- Q1 : What the model has actually learned? → ***Interpretation***
- Q2 : What information from the input is using to make predictions? → ***explanation***



BUT, ...
WHY IS THIS DESIRABLE?

# PROBLEM STATEMENT
## MOTIVATION

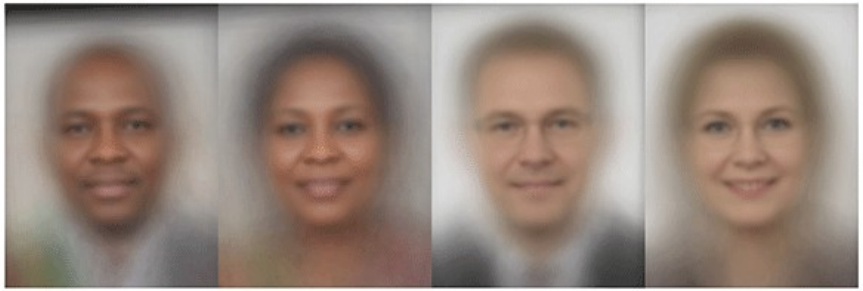
- Enable the detection of undesirable behaviors (e.g. biases, bugs, etc.)

# PROBLEM STATEMENT
## MOTIVATION

- Enable the detection of undesirable behaviors (e.g. biases, bugs, etc.)

### Gender Classification

| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE++ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |



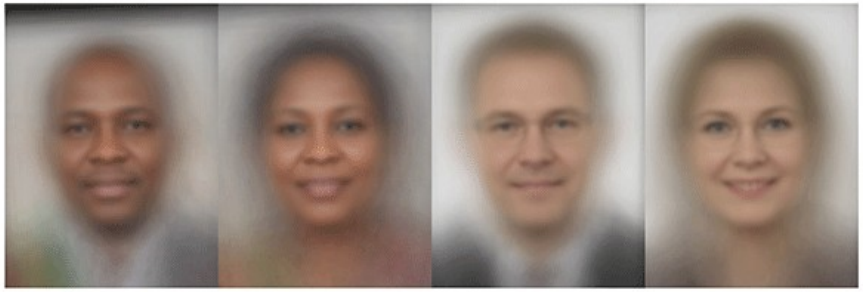
http://gendershades.org

# PROBLEM STATEMENT
## MOTIVATION

- Enable the detection of undesirable behaviors (e.g. biases, bugs, etc.)

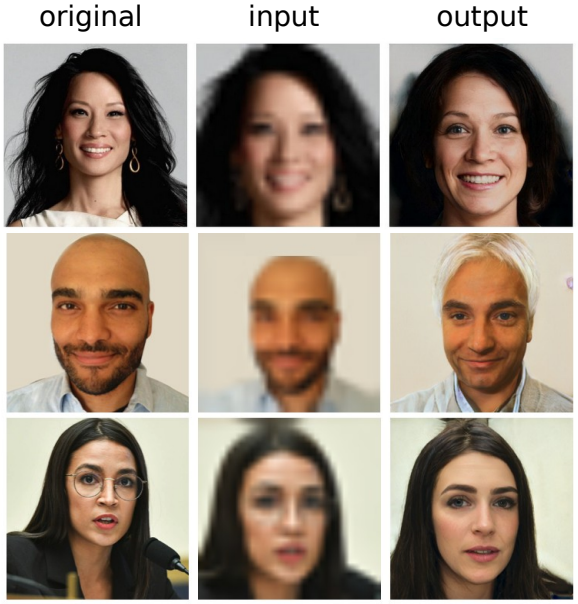
### Gender Classification

| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE++ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |



http://gendershades.org

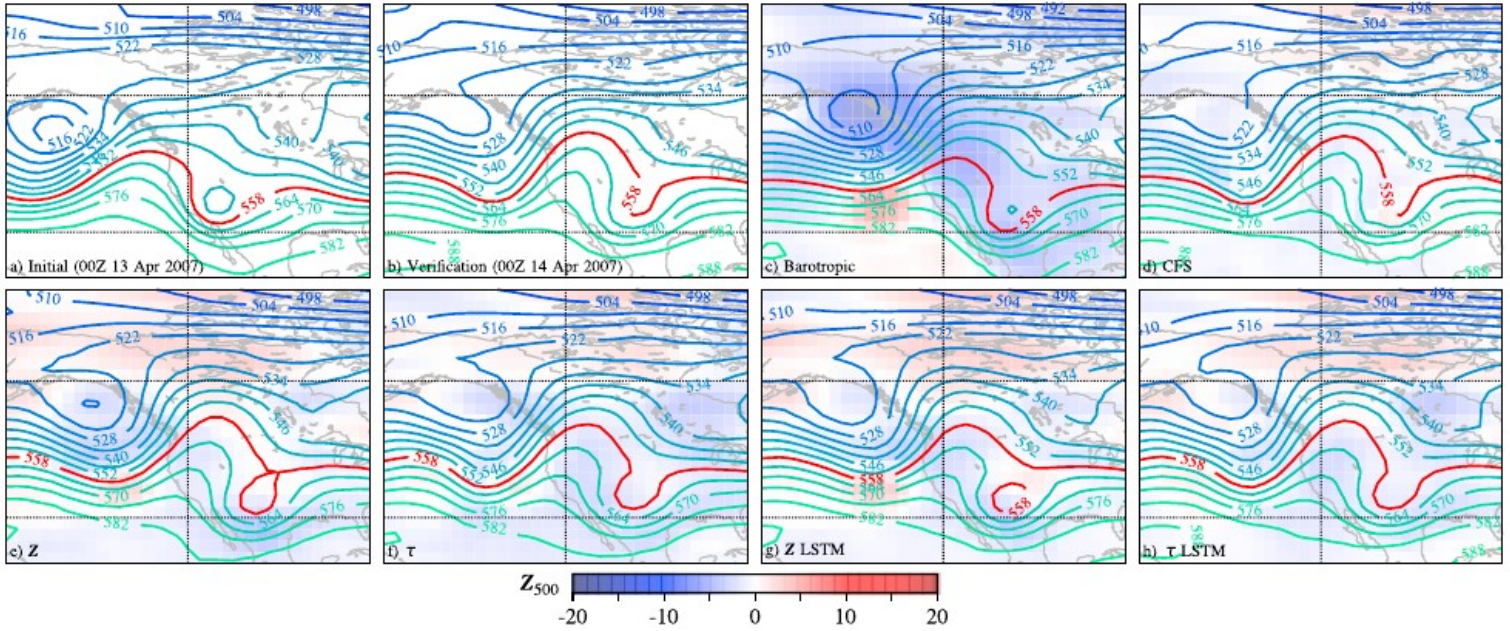
### Super resolution




Twitter: @osazuwa

# PROBLEM STATEMENT
## MOTIVATION

- Obtain additional insights on existing problems (e.g. weather forecasting)

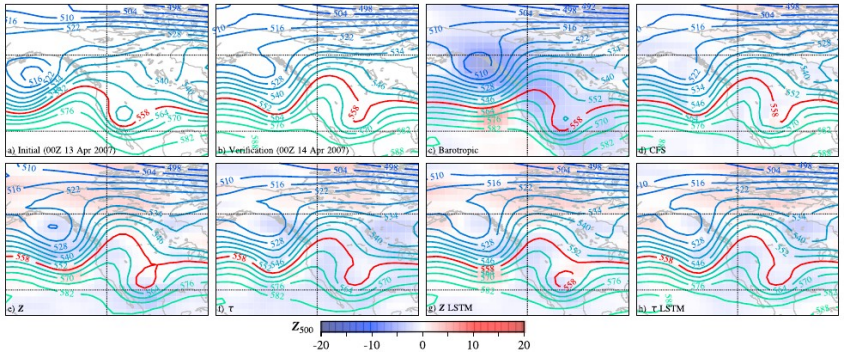





[ Weyn et al., (2019), doi.org/10.1029/2019MS001705 ]

# PROBLEM STATEMENT
## APPLICATION

- Having these additional insights may enable:




[ Weyn et al., (2019), doi.org/10.1029/2019MS001705 ]


Twitter: @osazuwa

**Discovery of novel factors or relationships between factors that can improve performance**

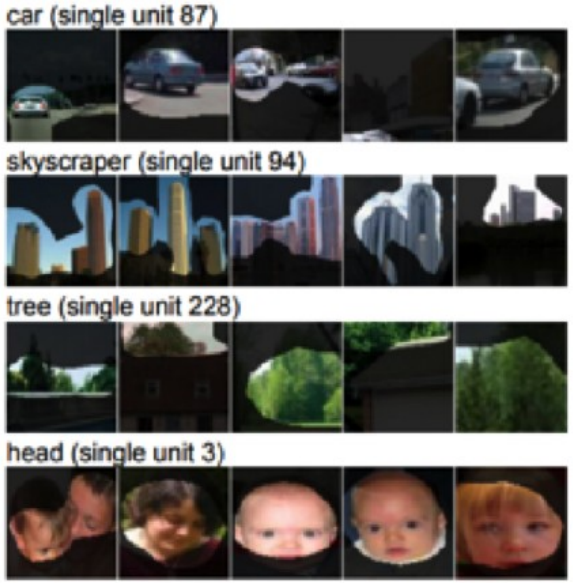
**Identification potential biases before deploying a model**

# WHAT HAS BEEN DONE SO FAR?

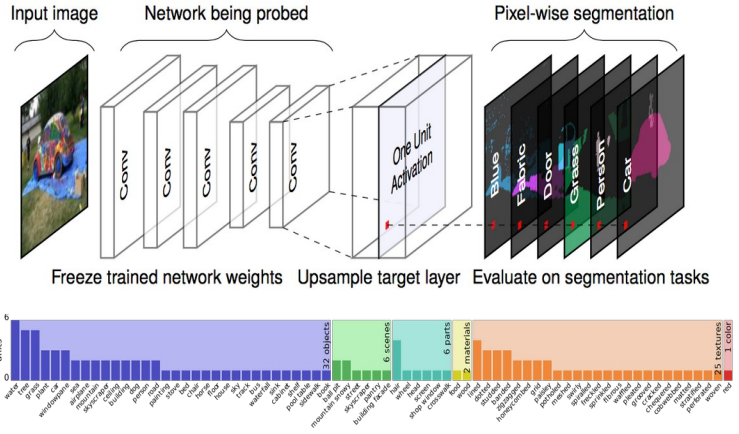
## [ RELATED WORK ]

# RELATED WORK
## VISUAL INTERPRETATION

- Require inspection of a large amount of visualizations
- Require expensive annotations



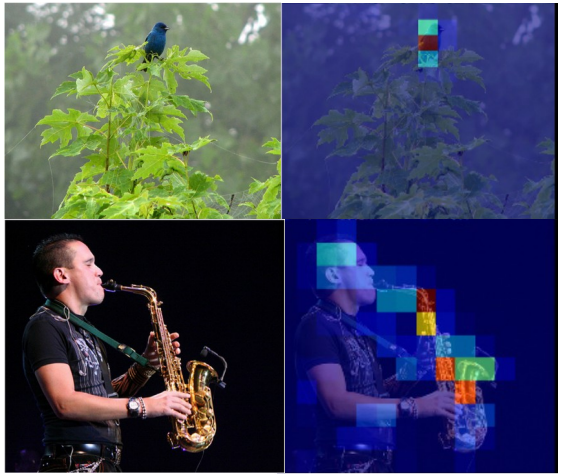

- Zhang et al., CVPR'18
- Zhou et al, CVPR'16



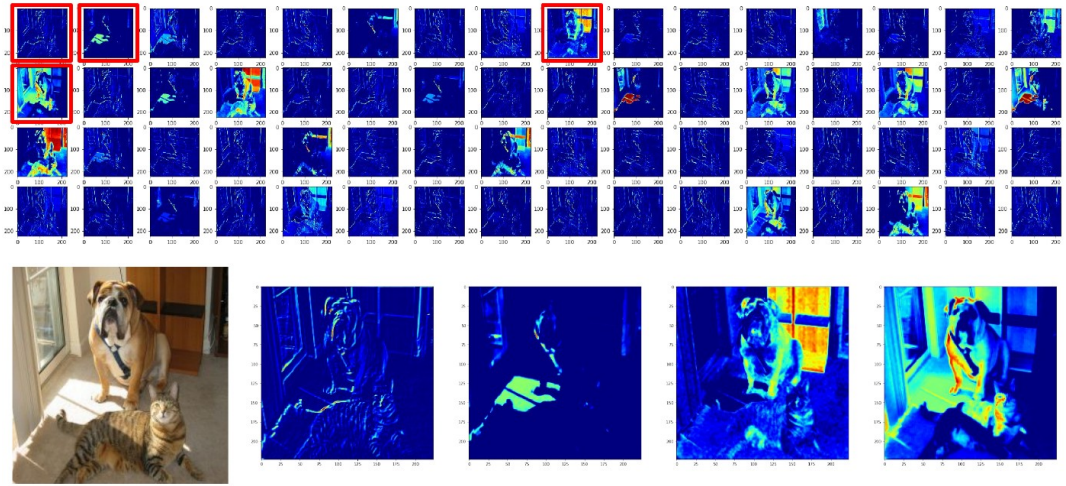

- Bau et al, CVPR'17

# RELATED WORK
## VISUAL EXPLANATION

- Coarse and computational expensive
- [Some] Not faithful to the model prediction



- Zeiler et al. ICCV'11
- Zeiler et al. ECCV'14
- Zhou et al., ICLR'15

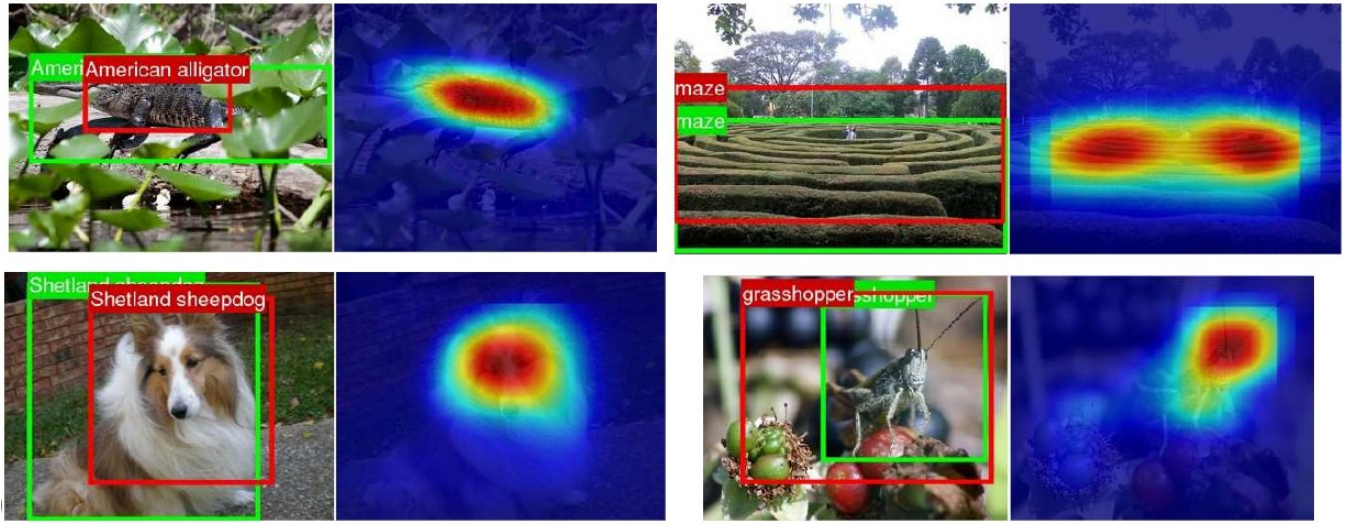


- Zeiler et al. ECCV'14
- Springenberg et al., ICLR'15.
- Grun et al., ICML'16.

- Zhou et al., CVPR'16.
- Zhang et al., ECCV'16
- Selvaraju et al., ICCV'17.
- Chattopadhyay et al., WACV'18.
- Zhang. et al., CVPR'18.

# RELATED WORK
## EVALUATION OF VISUAL EXPLANATIONS

- Proxy tasks → based on assumptions
- User studies → subjective.




- Zhou et al., CVPR'16
- Zhang et al., ECCV'16.

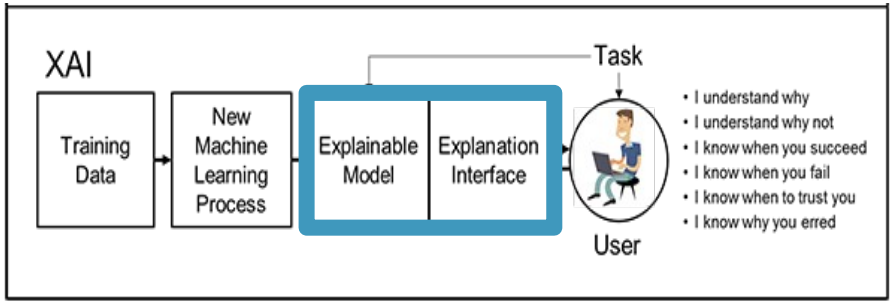
- Zeiler & Fergus, ECCV'14
- Selvaraju et al., ICCV'17.

[ ON-GOING RESEARCH ]

# POST-HOC EXPLANATION & INTERPRETATION

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## PRE-CONDITIONS (I)

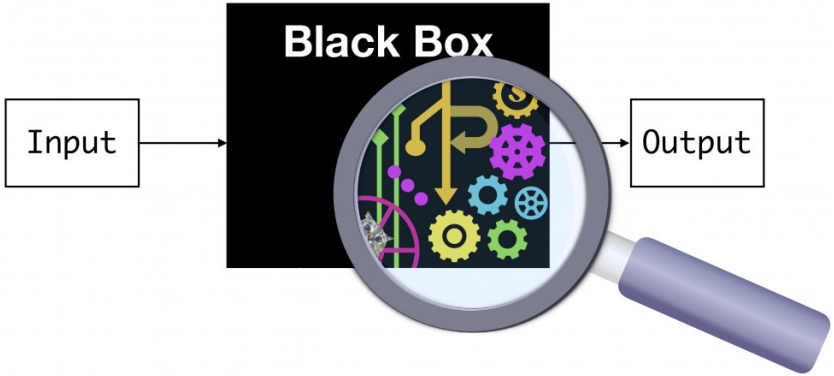
- Complement a pre-trained model




src: darpa.mil

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## PRE-CONDITIONS (II)

- Accessible representation




src: CMU ML Blog

# POST-HOC MODEL INTERPRETATION/EXPLANATION

## INTERPRETATION OF DEEP NEURAL NETWORKS

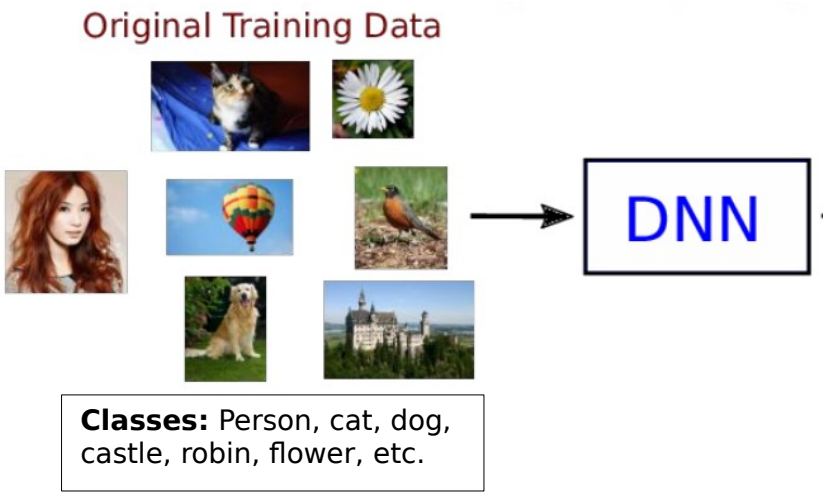
- Understanding what a deep model actually learned




[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## INTERPRETATION OF DEEP NEURAL NETWORKS

- Understanding what a deep model actually learned

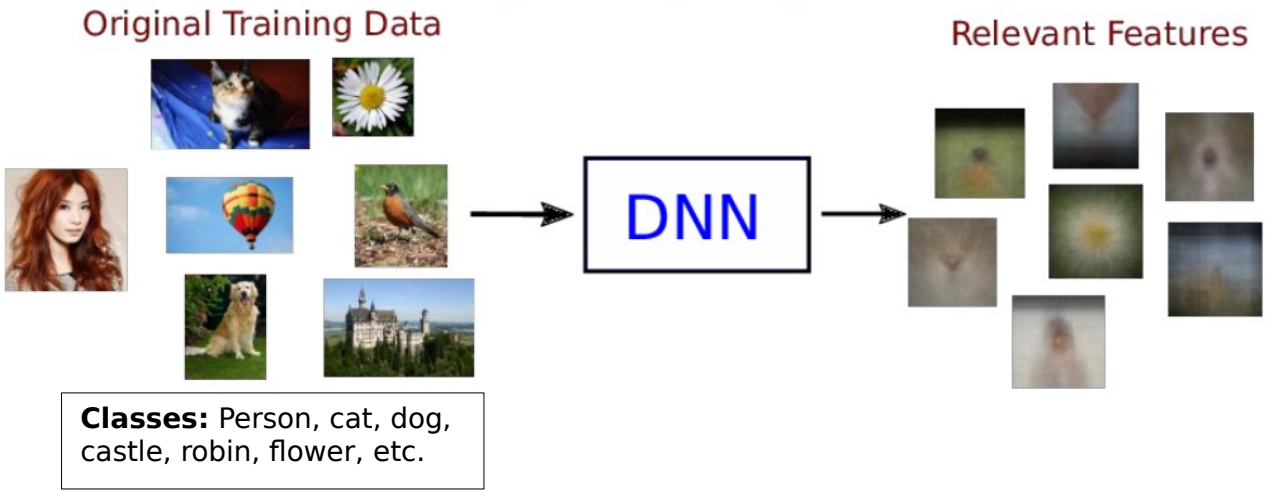





[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## INTERPRETATION OF DEEP NEURAL NETWORKS

- Understanding what a deep model actually learned




[ Oramas et al., ICLR'19 ]

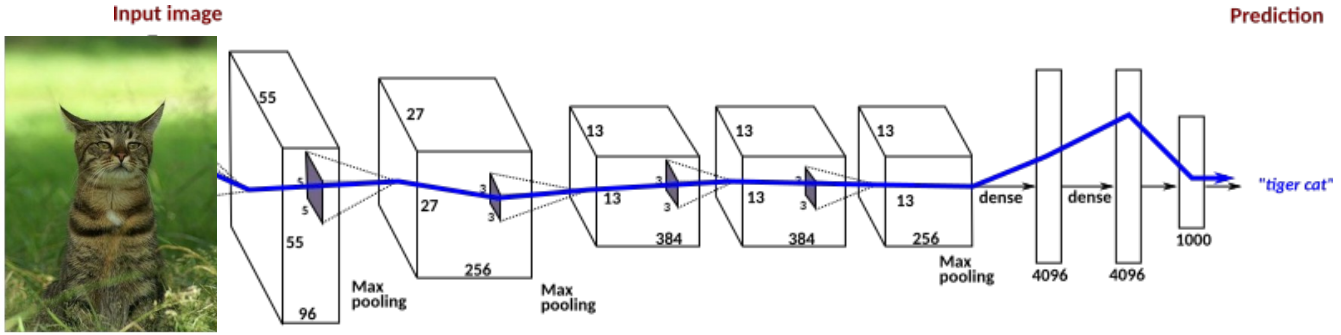
# POST-HOC MODEL INTERPRETATION/EXPLANATION

## INTERPRETATION OF DEEP NEURAL NETWORKS

**Identifying Relevant Features**

Given a pre-trained model *F* for *C* classes of interest

- Identify the subset of relevant features $W^*$ for class *j*

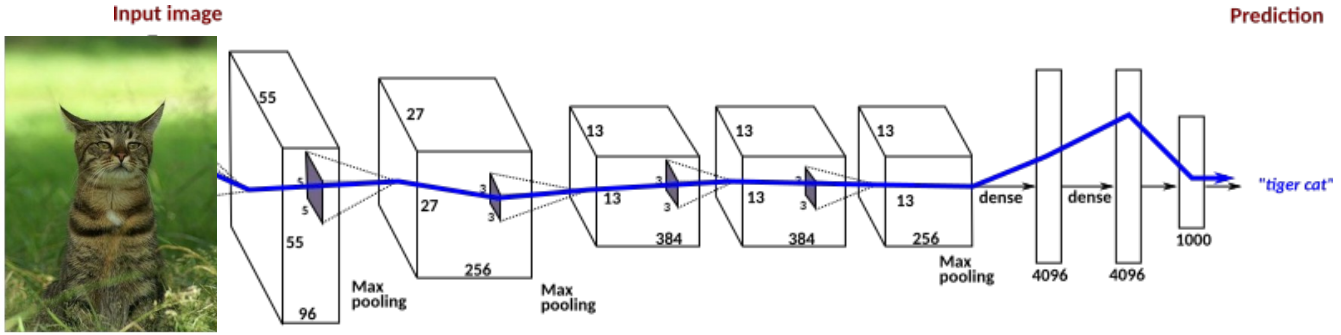



imec IDLab INTERNET & DATA LAB Universiteit Antwerpen

[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## INTERPRETATION OF DEEP NEURAL NETWORKS

**Identifying Relevant Features**
Given a pre-trained model *F* for *C* classes of interest
- Identify the subset of relevant features $W^*$ for class *j*



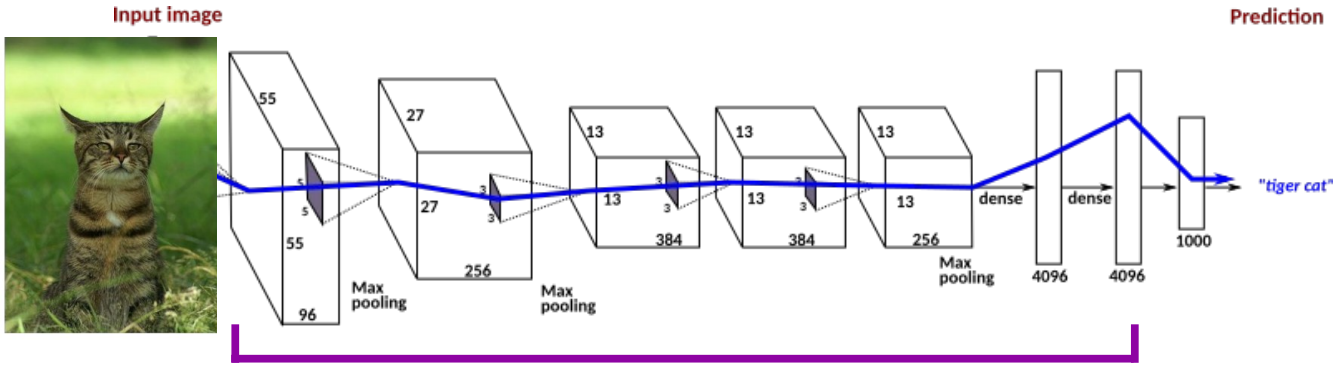

$$W^* = argmin_W \, ||X^T W - L^T||_F^2$$



[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## INTERPRETATION OF DEEP NEURAL NETWORKS

**Identifying Relevant Features**
Given a pre-trained model *F* for *C* classes of interest
- Identify the subset of relevant features $W^*$ for class *j*



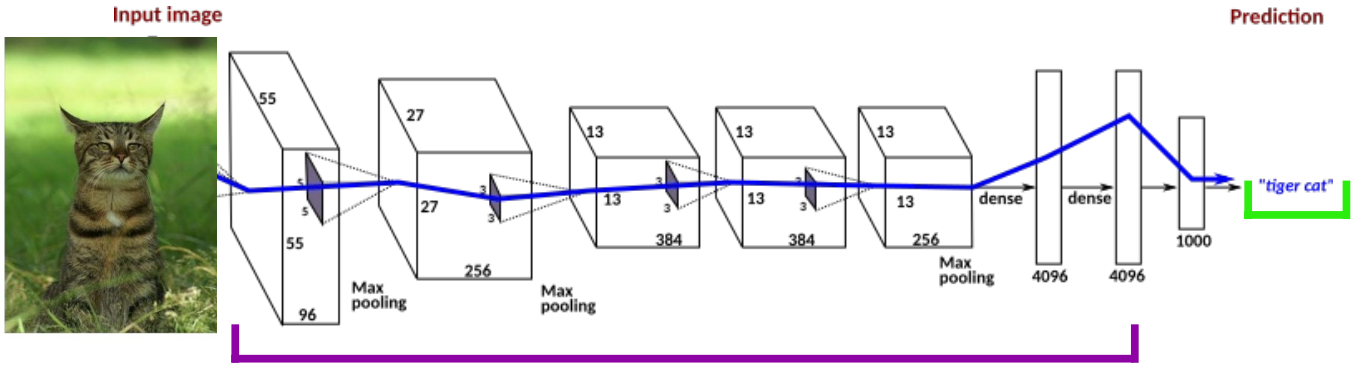

$$W^* = argmin_W ||X^T W - L^T||_F^2$$



[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## INTERPRETATION OF DEEP NEURAL NETWORKS

**Identifying Relevant Features**
Given a pre-trained model *F* for *C* classes of interest
- Identify the subset of relevant features $W^*$ for class *j*



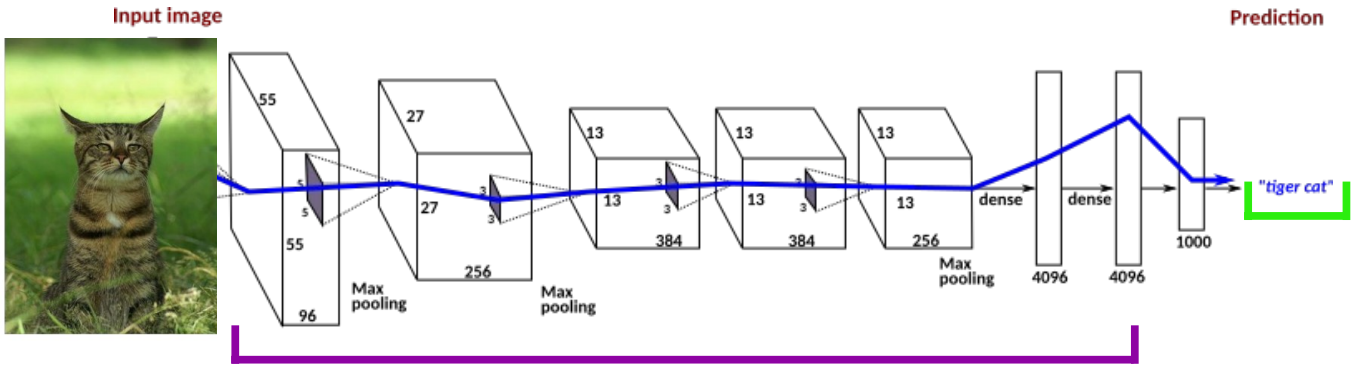

$$W^* = argmin_W \, ||X^T W - L^T||_F^2$$



[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## INTERPRETATION OF DEEP NEURAL NETWORKS

**Identifying Relevant Features**
Given a pre-trained model *F* for *C* classes of interest
- Identify the subset of relevant features $W^*$ for class *j*



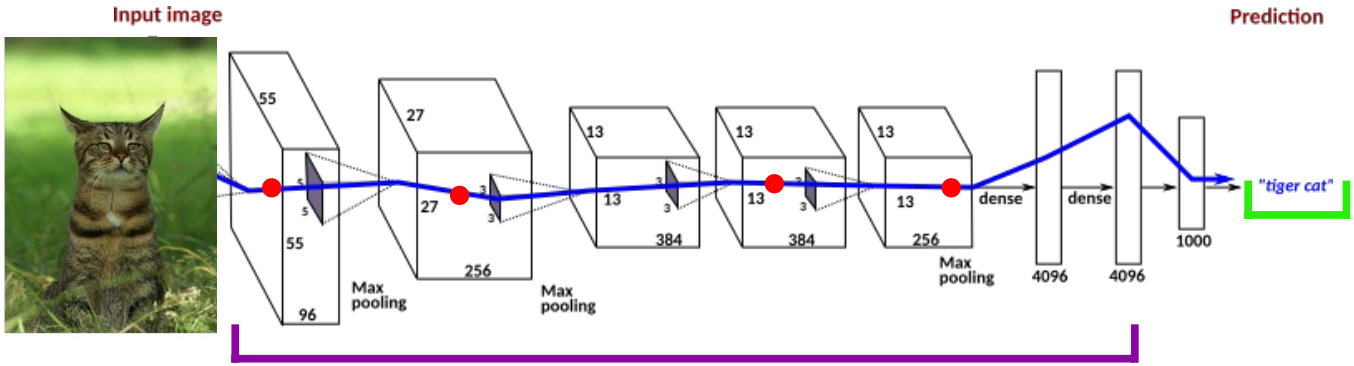

$$W^* = argmin_W \; ||X^T W - L^T||_F^2$$



[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## INTERPRETATION OF DEEP NEURAL NETWORKS

**Identifying Relevant Features**
Given a pre-trained model *F* for *C* classes of interest
- Identify the subset of relevant features $W^*$ for class *j*



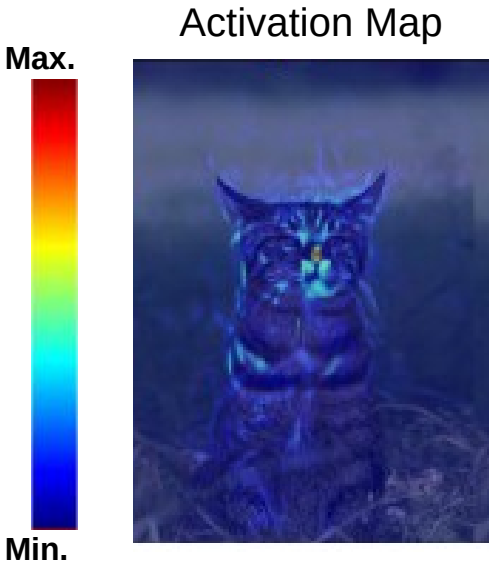

$$W^* = argmin_W \; ||X^T W - L^T||_F^2$$

$$subject\ to: \; ||w_j||_1 \leq \mu, \; \forall_j = 1, ..., C$$

[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION

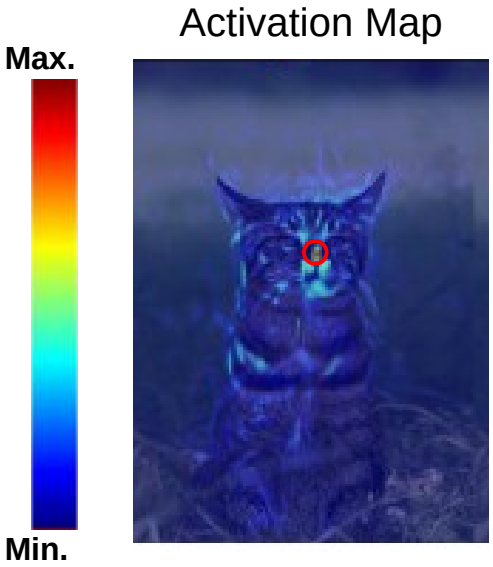
## INTERPRETATION OF DEEP NEURAL NETWORKS




[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION

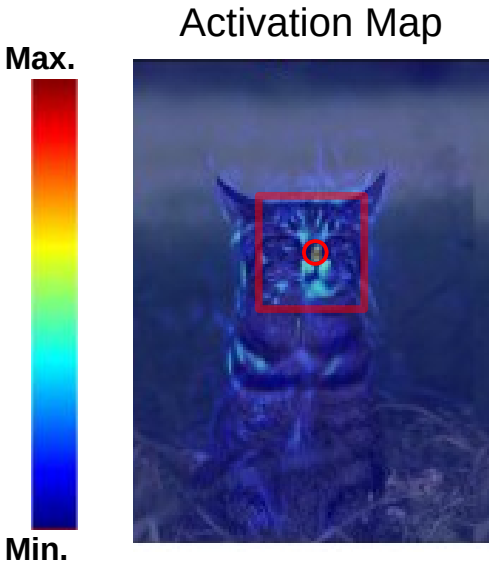
## INTERPRETATION OF DEEP NEURAL NETWORKS




[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION

## INTERPRETATION OF DEEP NEURAL NETWORKS

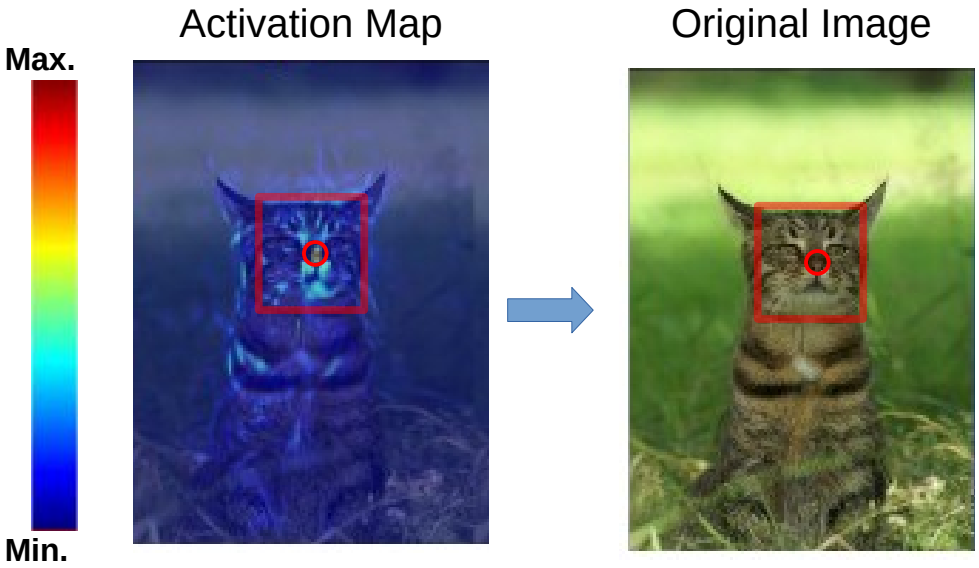





[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION

## INTERPRETATION OF DEEP NEURAL NETWORKS

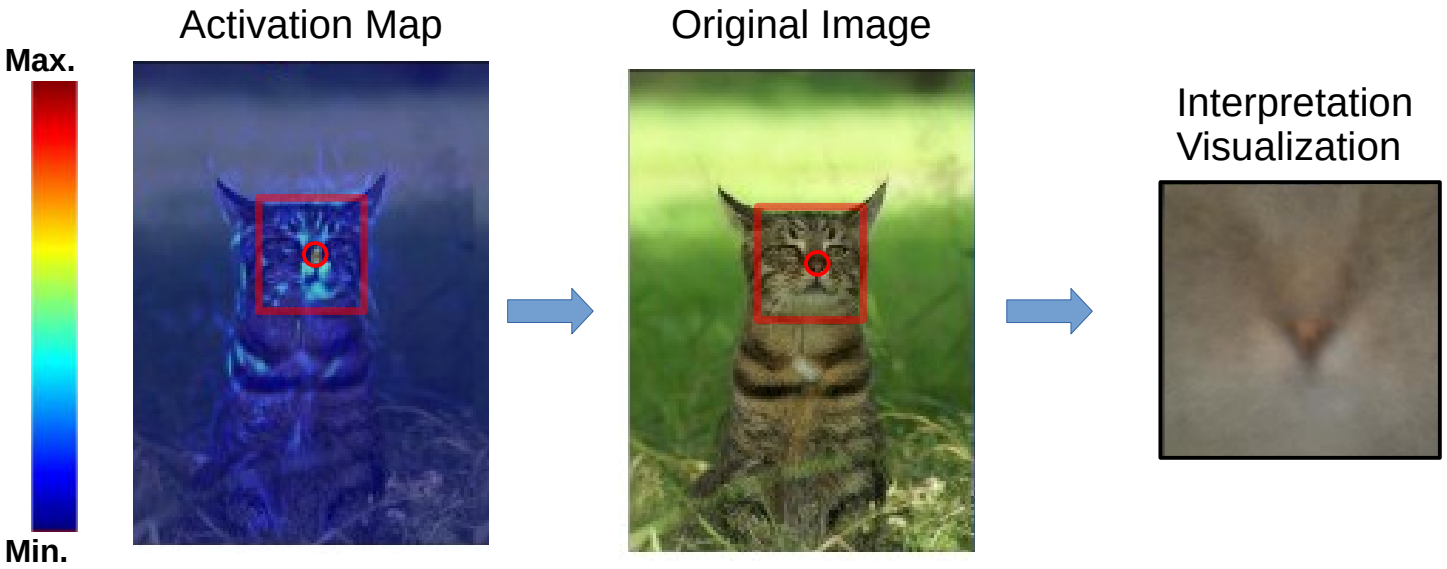





[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION

## INTERPRETATION OF DEEP NEURAL NETWORKS




For every identified relevant feature of each class:

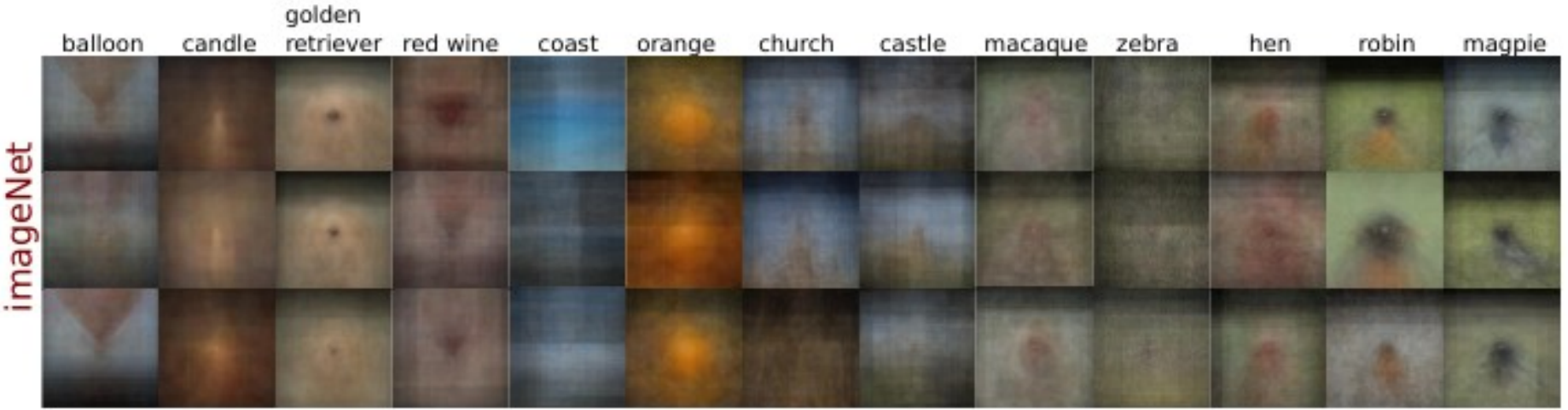
- Select top-*k* images with highest response
- Crop every image based on the receptive field
- Compute average image

[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION

## INTERPRETATION OF DEEP NEURAL NETWORKS

- Understanding what a deep model actually learned



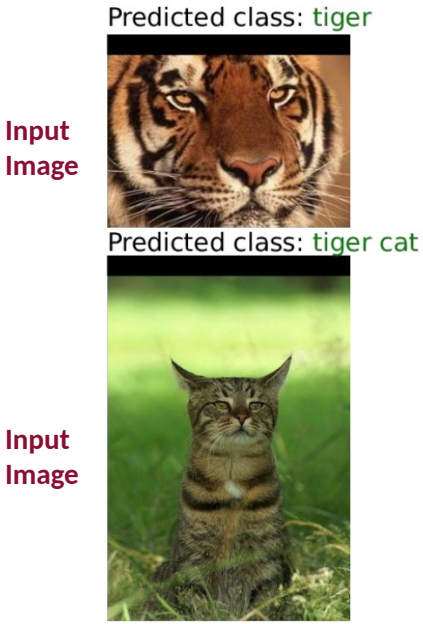

Random sample from 1000 general categories



[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## MODEL EXPLANATION

- Identifying features that are used by deep models when making predictions

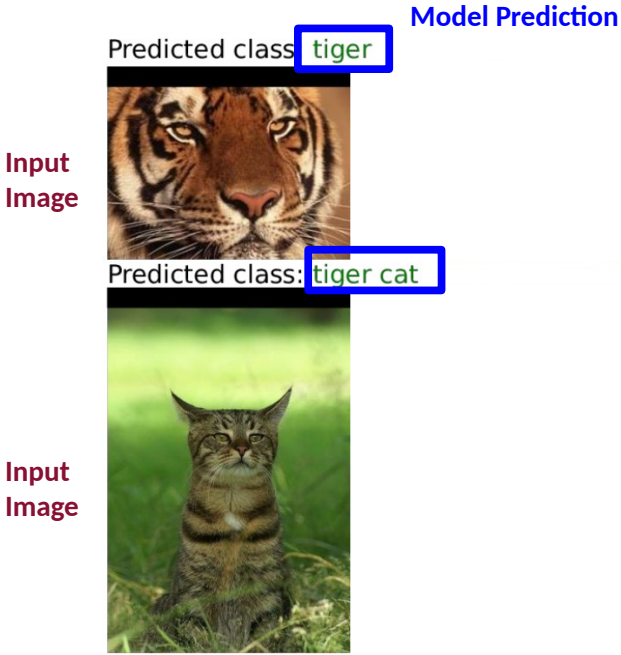





[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## MODEL EXPLANATION

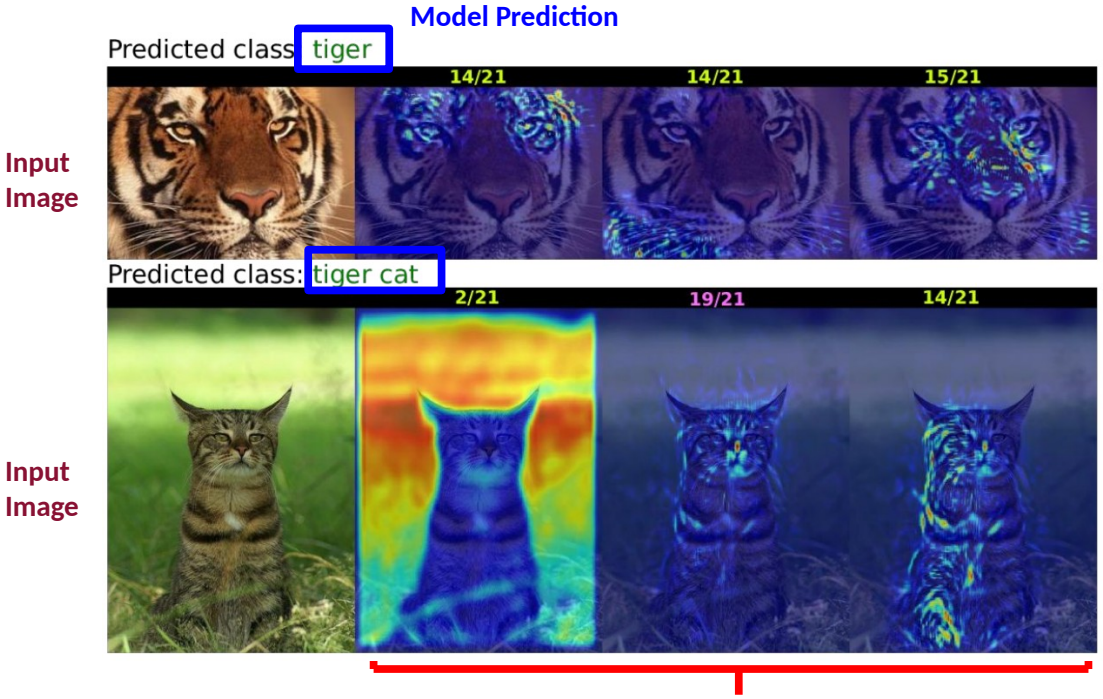
- Identifying features that are used by deep models when making predictions




[ Oramas et al., ICLR'19 ]

# POST-HOC MODEL INTERPRETATION/EXPLANATION
## MODEL EXPLANATION

- Identifying features that are used by deep models when making predictions

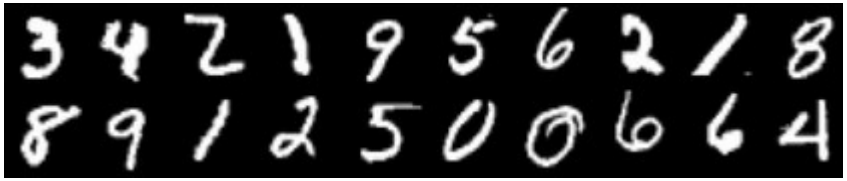



[ Oramas et al., ICLR'19 ]

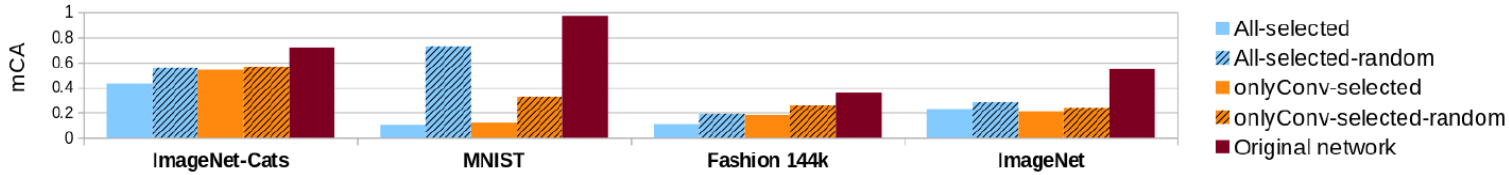
# EVALUATION

# EVALUATION
## DATASETS

- **MNIST**
  - 10 classes
  - 70k images
- **ImageNet (ILSVRC'12)**
  - 1000 classes
  - 1M images
- **ImageNet-Cats**
  - 13 classes
  - 18k images
- **Fashion-144k**
  - 12 classes
  - 12k images



# EVALUATION
## MEASURING IMPORTANCE OF THE SELECTED FEATURES

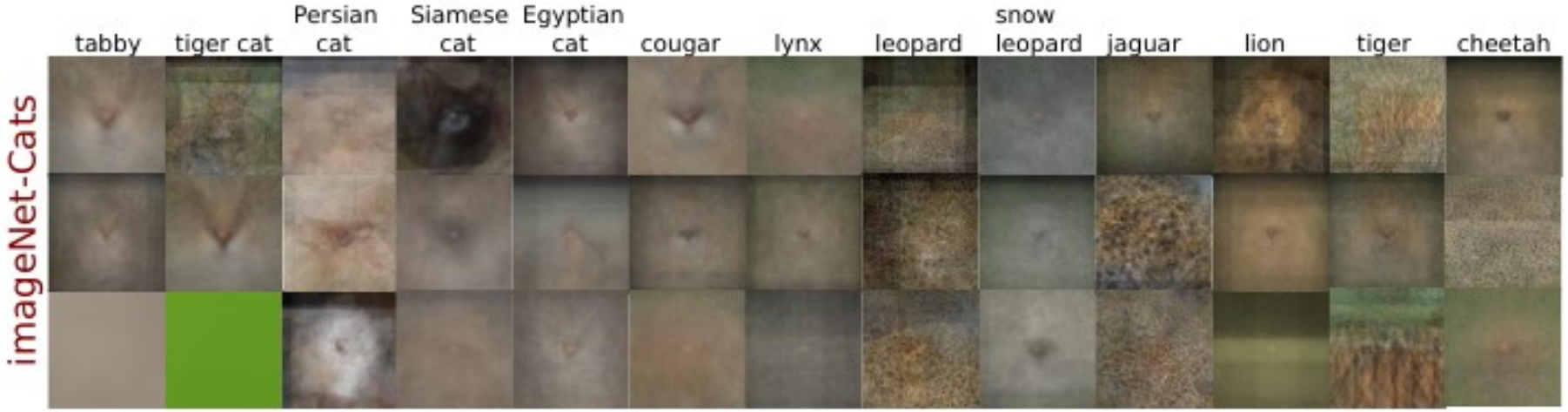



Changes in mean classification accuracy (mCA) as the identified relevant features/filters are ablated.



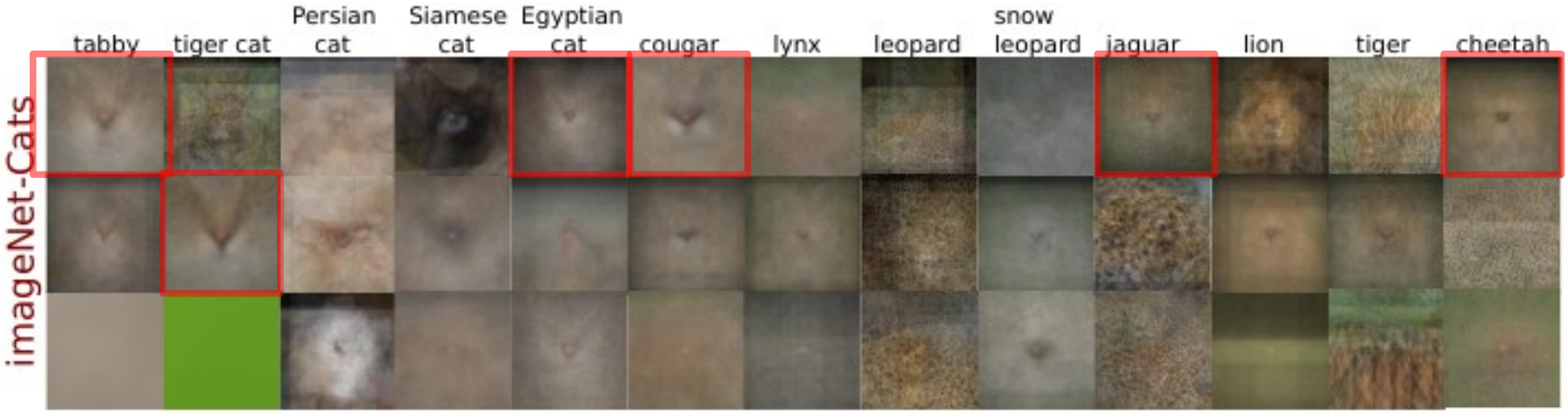
[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL INTERPRETATION



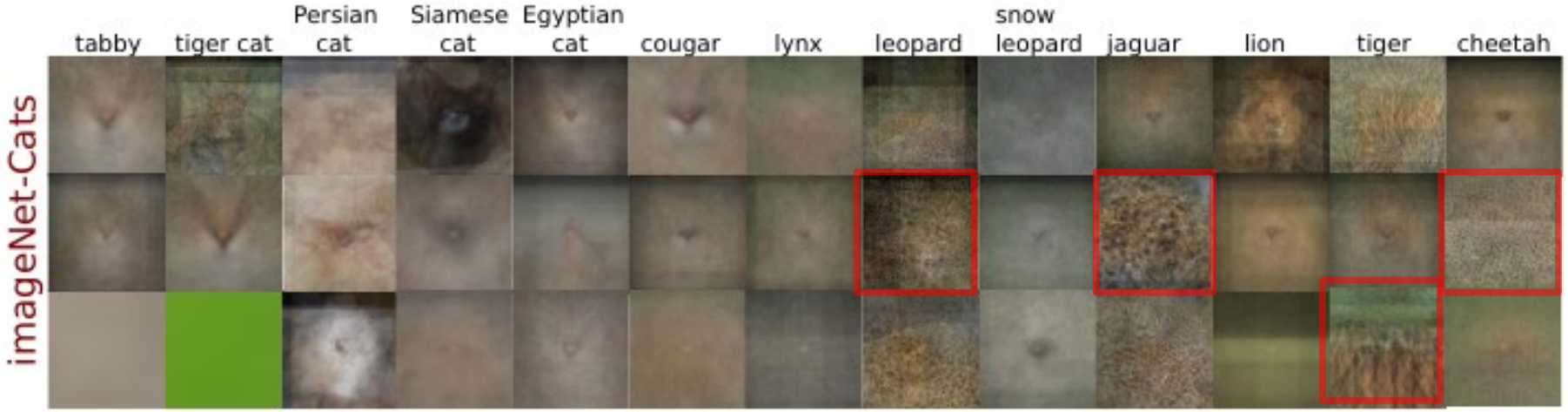

[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL INTERPRETATION




[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL INTERPRETATION






[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL INTERPRETATION

- Human-centered Geolocation

  → **Predict the location** where a photo was taken (classification)

Fashion-oriented data

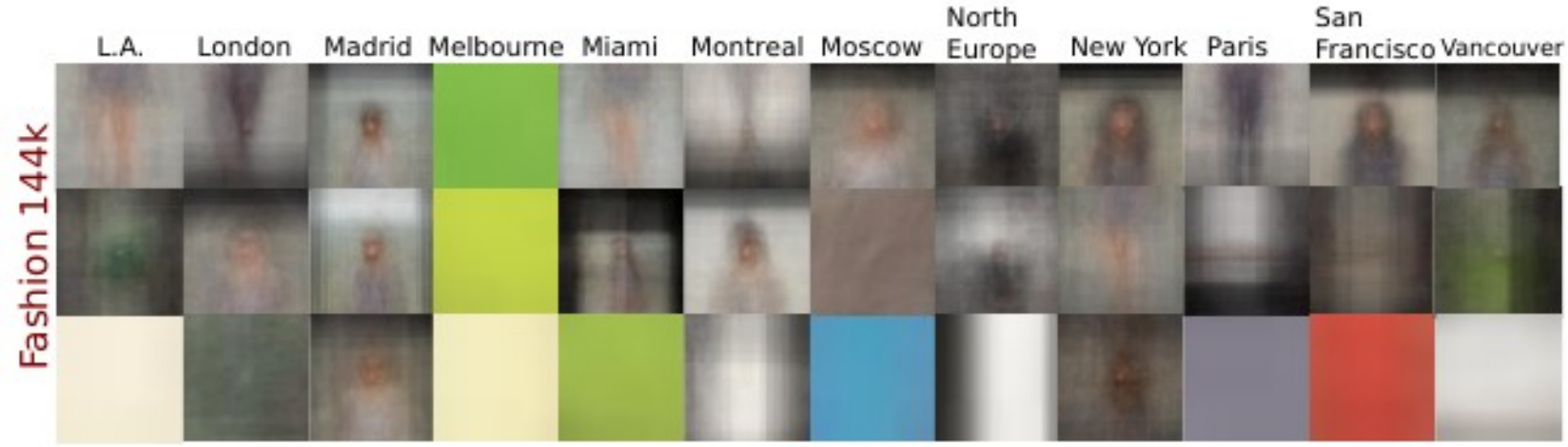
**Fashion144k dataset**

[ Simo-Serra et al., CVPR'15 ]

[ Wang et al., WACV'18 ]

# EVALUATION
## VISUAL INTERPRETATION

- Understanding what a deep model actually learned

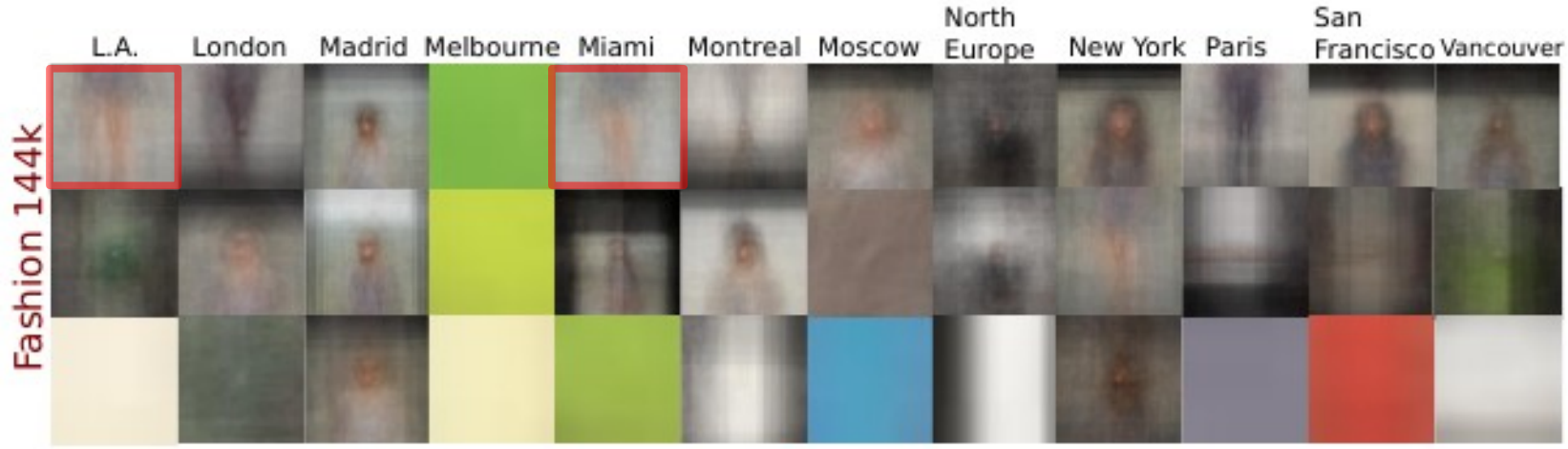



Fashion-oriented data →

[ Wang et al., WACV'18 ]
[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL INTERPRETATION

- Understanding what a deep model actually learned




Fashion-oriented data →

[ Wang et al., WACV'18 ]
[ Oramas et al., ICLR'19 ]

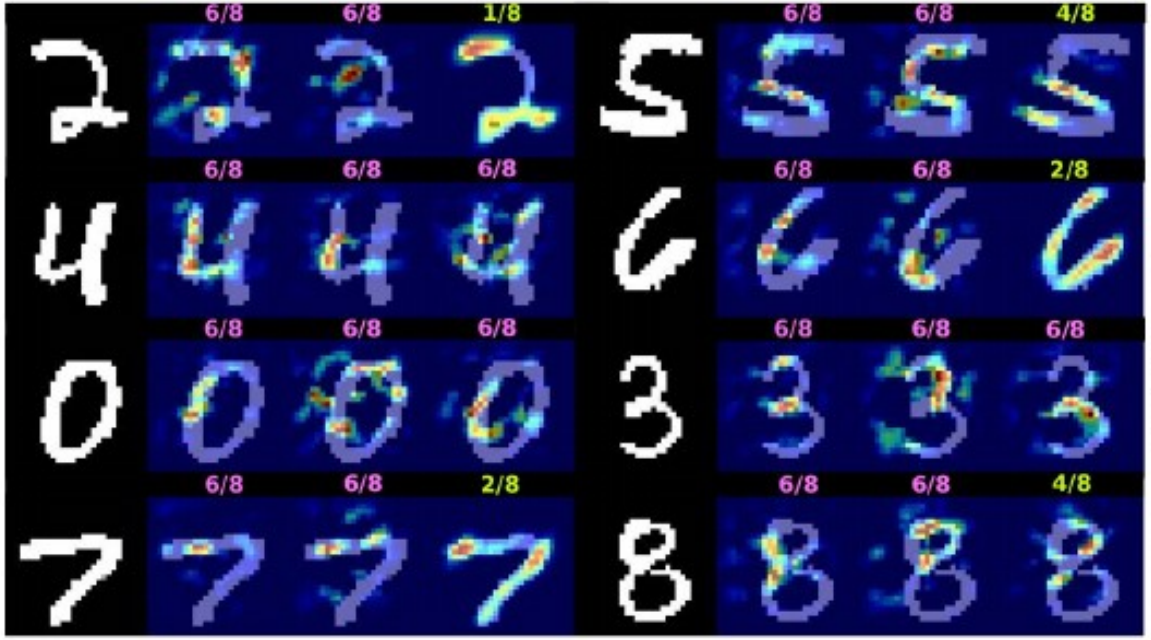
# EVALUATION
## VISUAL INTERPRETATION

- Understanding what a deep model actually learned


Fashion-oriented data →

[ Wang et al., WACV'18 ]
[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL EXPLANATION

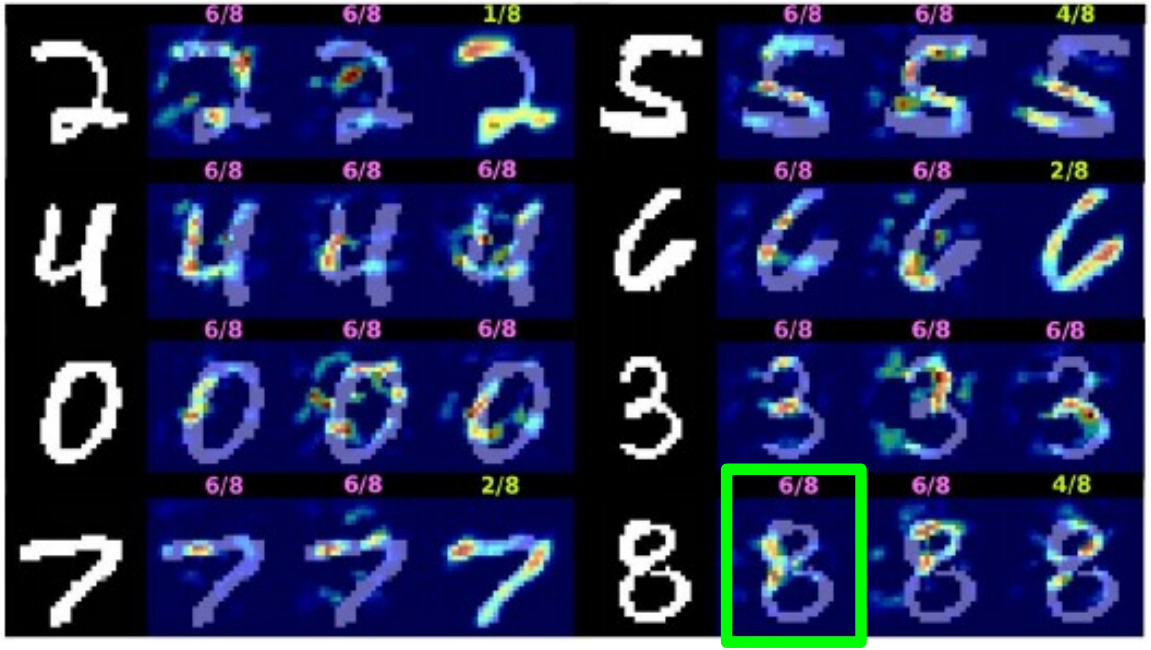


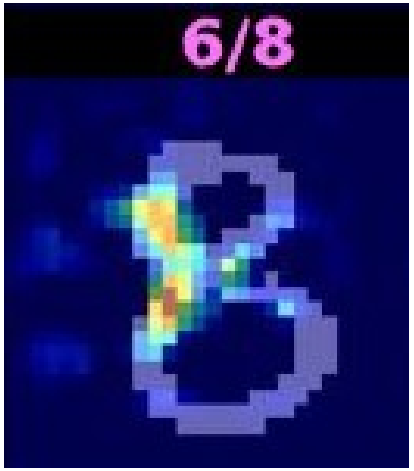


imec IDLab INTERNET & DATA LAB Universiteit Antwerpen

[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL EXPLANATION



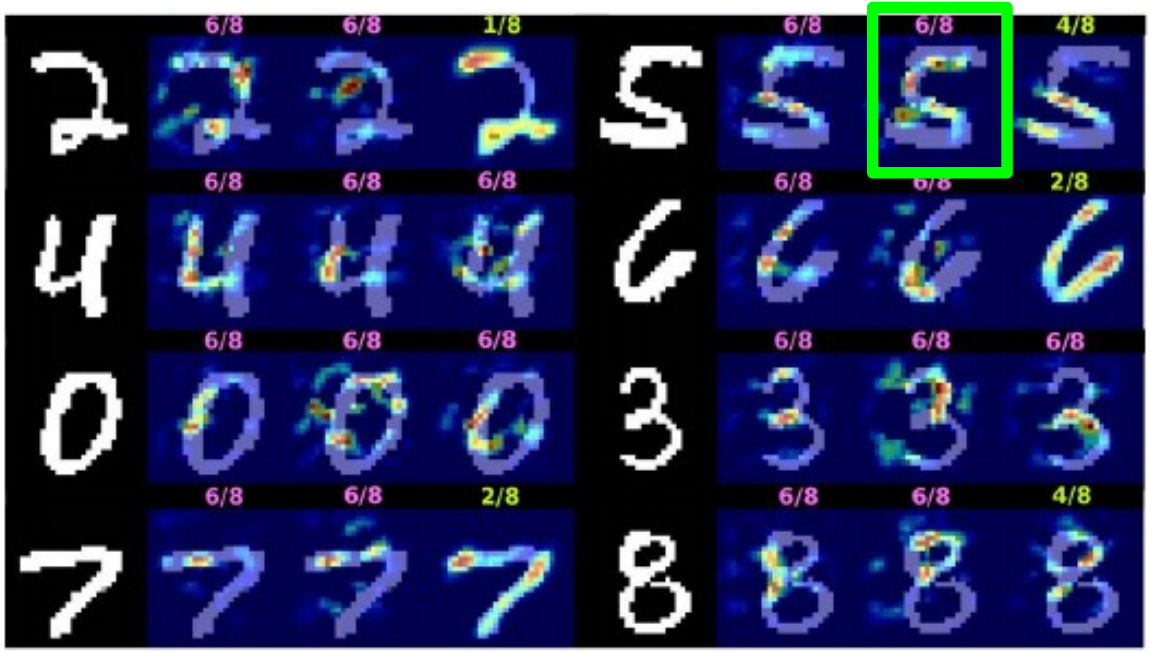




**Highlighting features that must be present**

[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL EXPLANATION

**Highlighting features that must be absent**





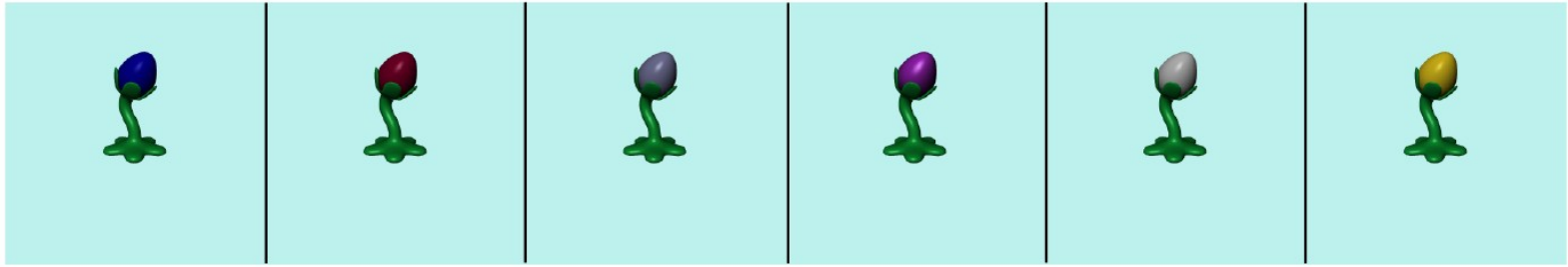


[ Oramas et al., ICLR'19 ]

# EVALUATION
## OBJECTIVE EVALUATION OF VISUAL EXPLANATIONS

- Control the discriminative feature for the classes of interest by design.

an8flower-single-6c



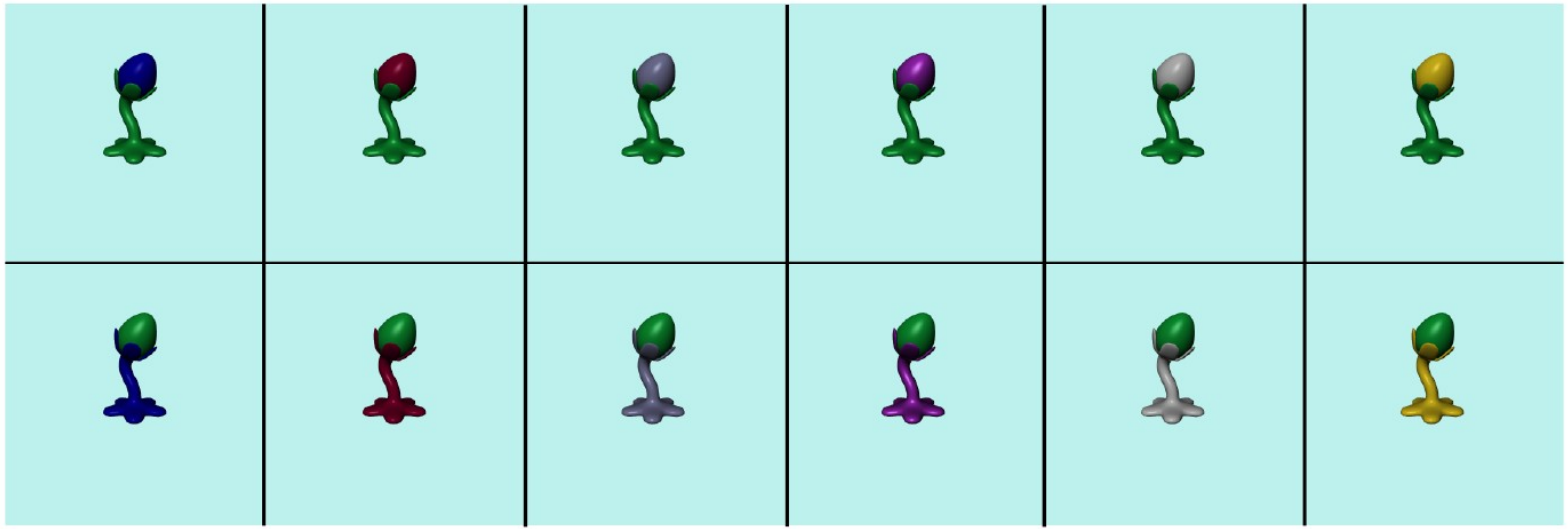


[ Oramas et al., ICLR'19 ]

# EVALUATION
## OBJECTIVE EVALUATION OF VISUAL EXPLANATIONS

- Control the discriminative feature for the classes of interest by design.

an8flower-double-12c

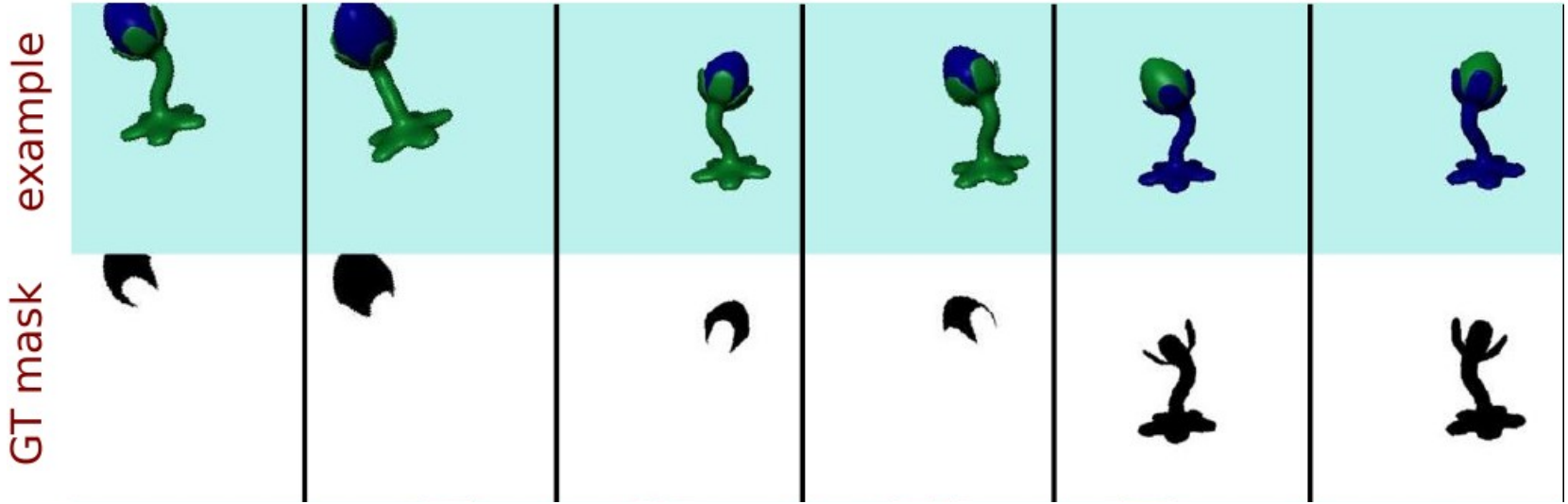




[ Oramas et al., ICLR'19 ]

# EVALUATION
## OBJECTIVE EVALUATION OF VISUAL EXPLANATIONS

- Measure the overlap (IoU) of the visual explanation with the GT-mask

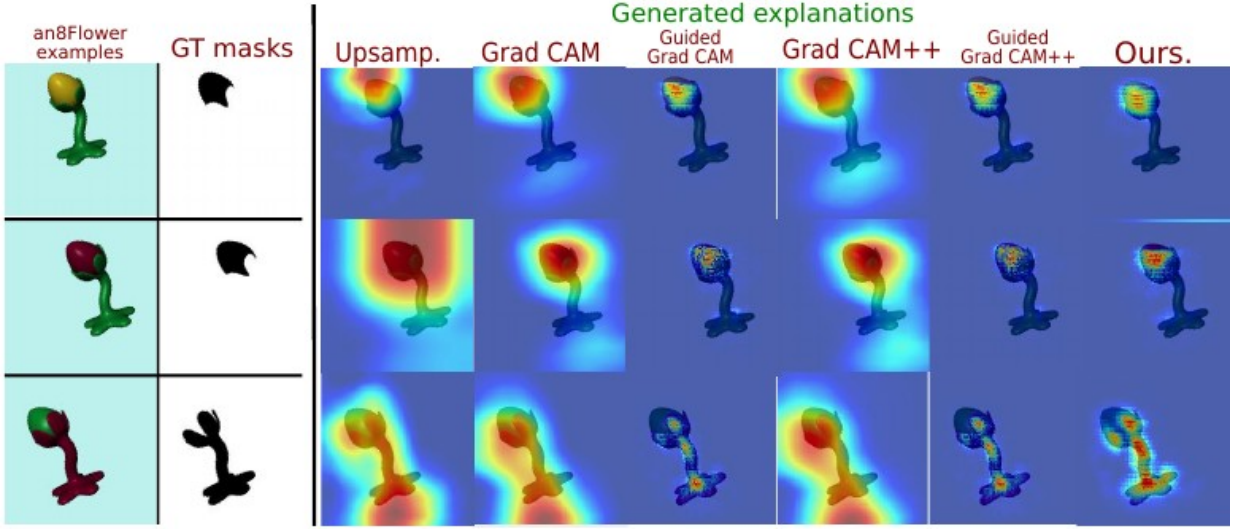





[ Oramas et al., ICLR'19 ]

# EVALUATION
## OBJECTIVE EVALUATION OF VISUAL EXPLANATIONS

- A good balance between coverage of large regions and precision on details.

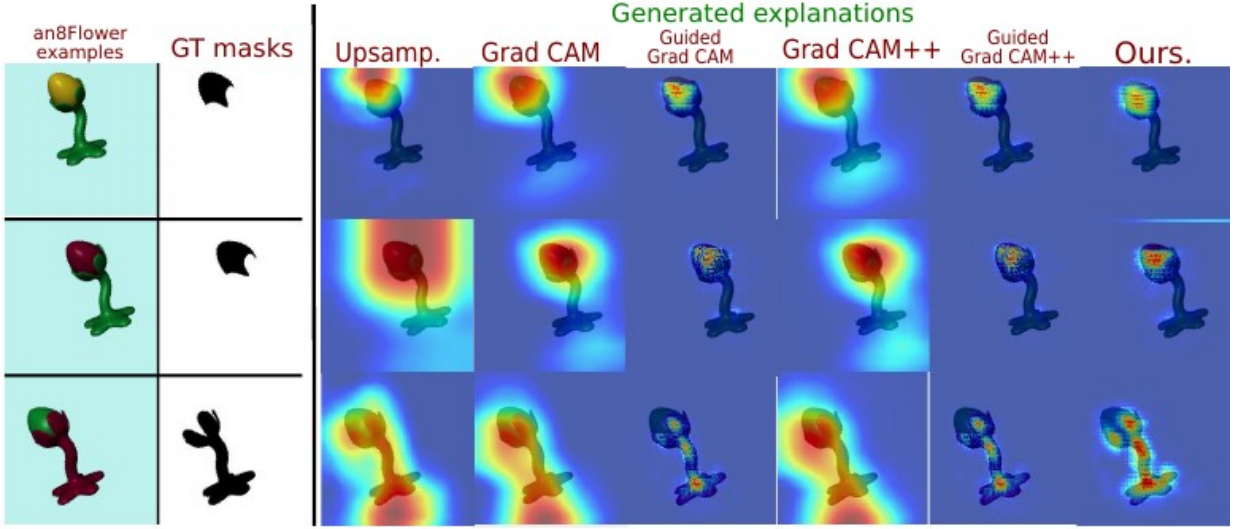





[ Oramas et al., ICLR'19 ]

# EVALUATION
## OBJECTIVE EVALUATION OF VISUAL EXPLANATIONS

- A good balance between coverage of large regions and precision on details.



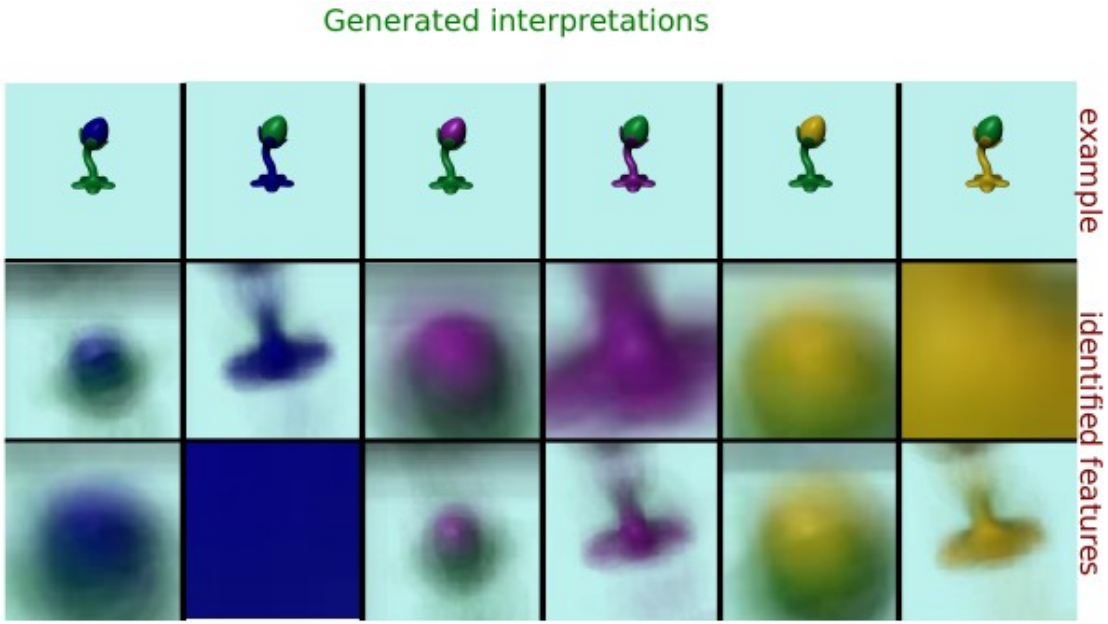

| Method | single-6c |
|---|---|
| Upsam. Act. | 16.8±2.63 |
| Deconv+GB, Springenberg et al. (2015) | 21.3±0.77 |
| Grad-CAM, Das et al. (2016) | 17.5±0.25 |
| Guided Grad-CAM, Das et al. (2016) | 19.9±0.61 |
| Grad-CAM++, Chattopadhyay et al. (2018) | 15.6±0.57 |
| Guided Grad-CAM++, Chattopadhyay et al. (2018) | 19.6±0.65 |
| ***Ours*** | **22.5±0.82** |
| **Method** | **double-12c** |
| Upsam. Act. | 16.1±1.30 |
| Deconv+GB, Springenberg et al. (2015) | 21.9±0.72 |
| Grad-CAM, Das et al. (2016) | 14.8±0.16 |
| Guided Grad-CAM, Das et al. (2016) | 19.4±0.34 |
| Grad-CAM++, Chattopadhyay et al. (2018) | 14.6±0.12 |
| Guided Grad-CAM++, Chattopadhyay et al. (2018) | 19.7±0.27 |
| ***Ours*** | **23.2±0.60** |



[ Oramas et al., ICLR'19 ]

# EVALUATION
## OBJECTIVE EVALUATION OF VISUAL EXPLANATION

- What have the model learned ?

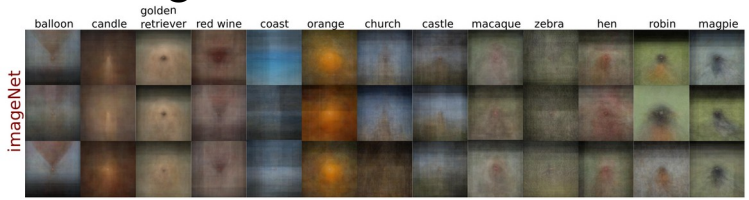



[ Oramas et al., ICLR'19 ]

# TO CONCLUDE

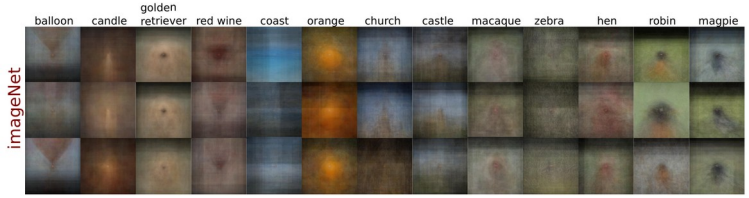
# TO CONCLUDE
## SUMMARIZING…

- Visual Interpretation through visualizations of relevant features

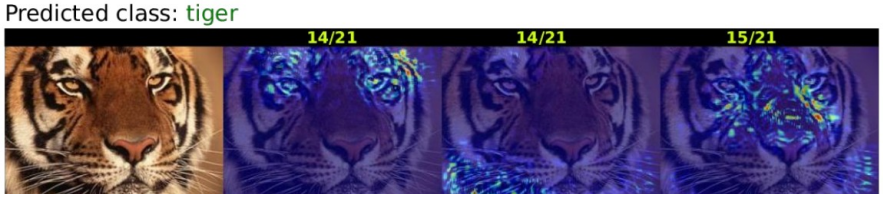


# TO CONCLUDE
## SUMMARIZING…

- Visual Interpretation through visualizations of relevant features

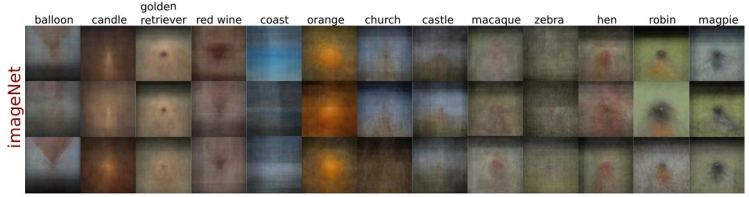



- Visual Explanation verifying the response of relevant features on a given input.

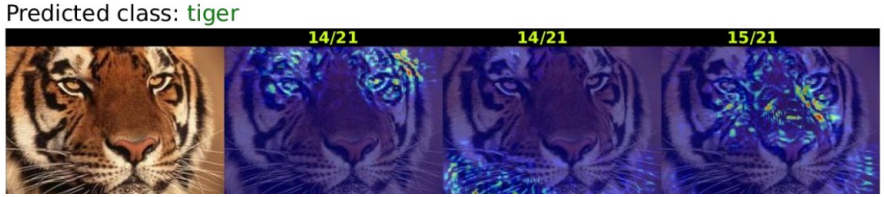


# TO CONCLUDE
## SUMMARIZING…

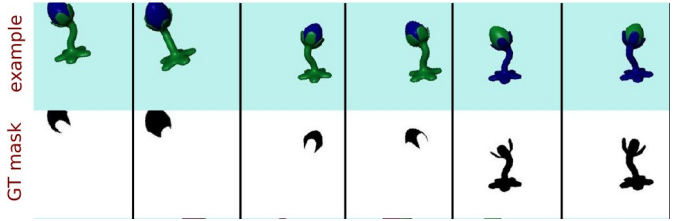
- Visual Interpretation through visualizations of relevant features




- Visual Explanation verifying the response of relevant features on a given input.




- Objective evaluation protocol for visual explanations.



# TAKE-HOME MESSAGE

# TAKE HOME MESSAGE

**A.I. -related technologies have improved significantly…**

**… yet, it is far from being perfect.**

Thus the need for additional steps
( model interpretation/explanation, sanity checks, etc. )

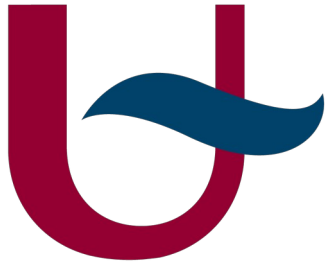
# ACKNOWLEDGMENTS

# PEOPLE INVOLVED

## RESEARCH IS TEAM SPORT

Kaili Wang

Tinne Tuytelaars



Universiteit
Antwerpen

# MAKING LEARNING-BASED VISUAL REPRESENTATIONS MORE INTELLIGIBLE

JOSÉ ORAMAS MOGROVEJO

Email: Jose.Oramas@uantwerpen.be
Twitter: @jaom7

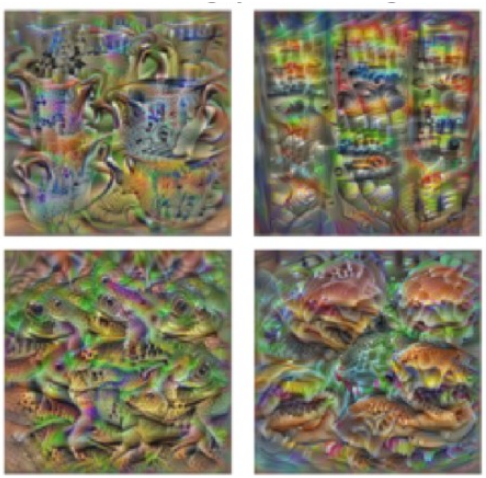
# SUPPLEMENTARY SLIDES

# WHAT HAS BEEN DONE SO FAR?

## [ RELATED WORK ]

# RELATED WORK
## VISUAL INTERPRETATION

### Visualizing pre-images

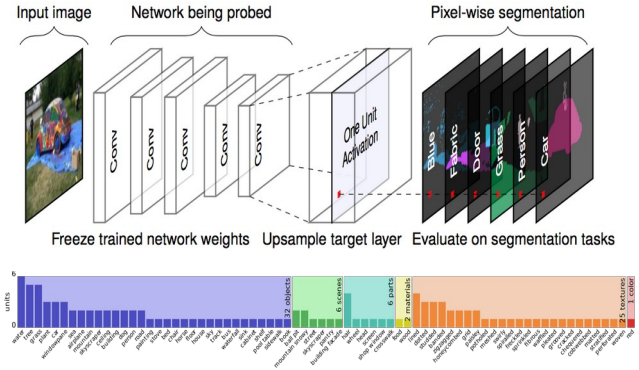


- Mahendran & Vedaldi, CVPR15
- Carter et al., Distill'19.

### Visualizing node activations


- Zhang et al., CVPR'18
- Zhou et al, CVPR'16

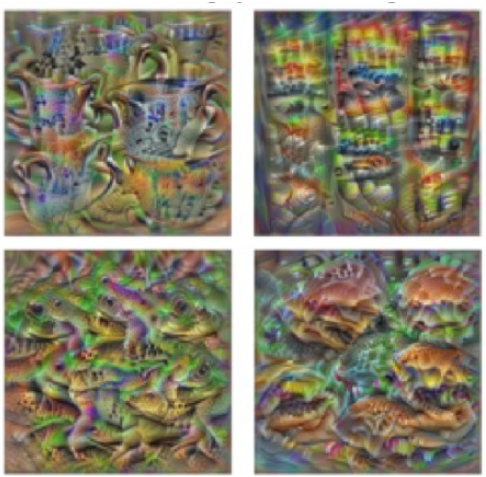
### Link to proxy tasks




- Bau et al, CVPR'17

# RELATED WORK
## VISUAL INTERPRETATION

### Visualizing pre-images

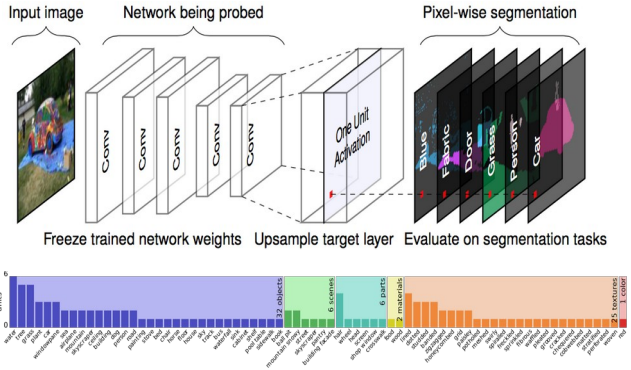


- Mahendran & Vedaldi, CVPR15
- Carter et al., Distill'19.

### Visualizing node activations


- Zhang et al., CVPR'18
- Zhou et al, CVPR'16

### Link to proxy tasks

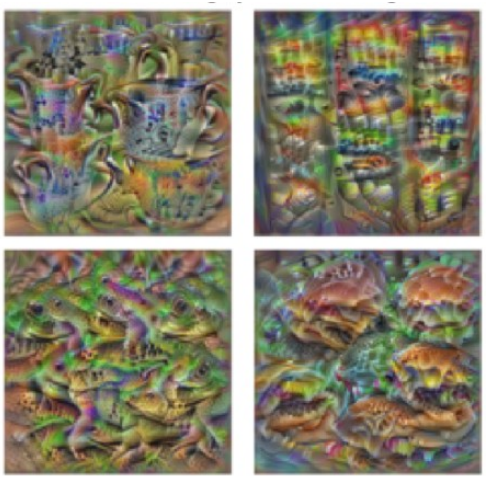



- Bau et al, CVPR'17

**But: subjective**

# RELATED WORK
## VISUAL INTERPRETATION

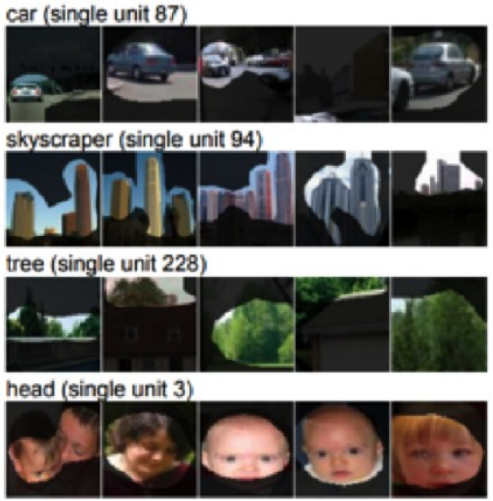
### Visualizing pre-images



- Mahendran & Vedaldi, CVPR15
- Carter et al., Distill'19.

**But: subjective**

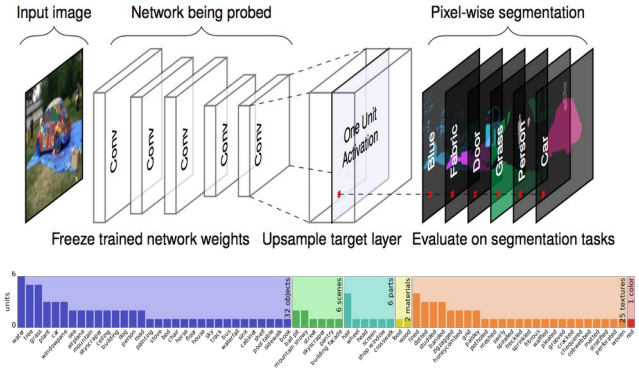
### Visualizing node activations




- Zhang et al., CVPR'18
- Zhou et al, CVPR'16

**But: too many nodes**

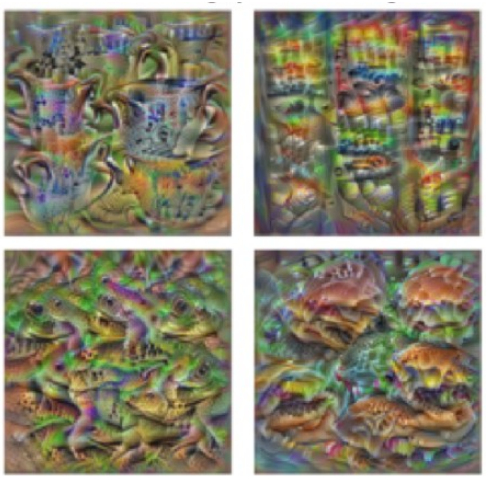
### Link to proxy tasks




- Bau et al, CVPR'17

# RELATED WORK
## VISUAL INTERPRETATION

### Visualizing pre-images



- Mahendran & Vedaldi, CVPR15
- Carter et al., Distill'19.

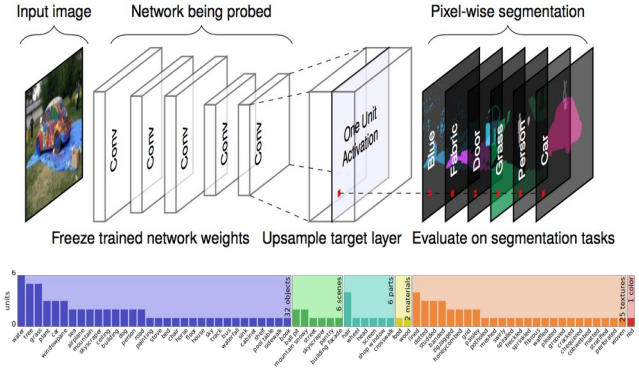
**But: subjective**

### Visualizing node activations


- Zhang et al., CVPR'18
- Zhou et al, CVPR'16

**But: too many nodes**

### Link to proxy tasks

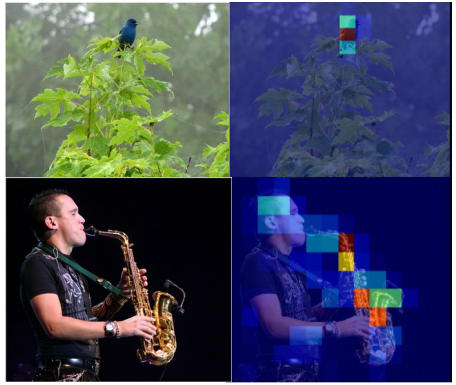



- Bau et al, CVPR'17

**But: limited**

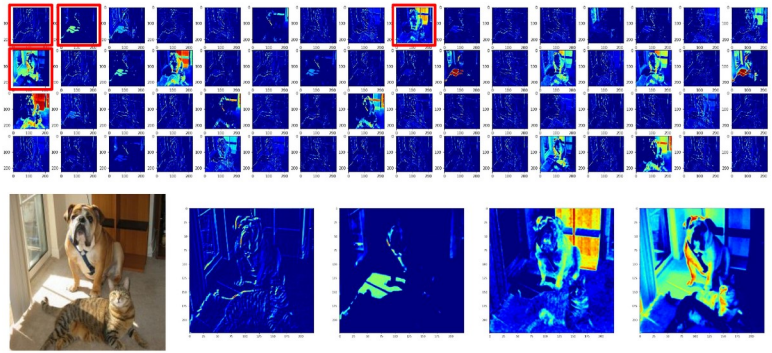
# RELATED WORK
## VISUAL EXPLANATION

### Input-modification



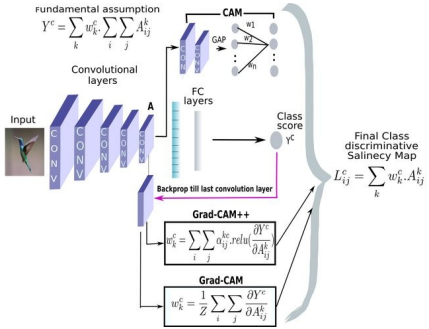
- Zeiler et al. ICCV'11
- Zeiler et al. ECCV'14
- Zhou et al., ICLR'15

### Deconvolution-based



- Zeiler et al. ECCV'14
- Springenberg et al., ICLR'15.
- Grun et al., ICML'16.

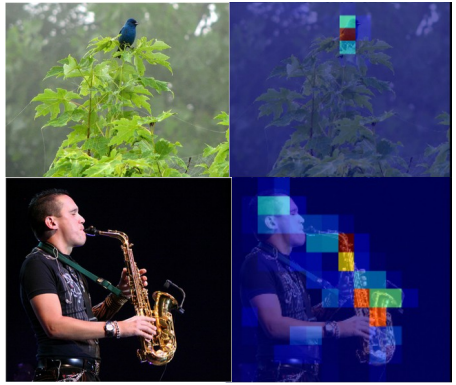
### Class-activation mapping




- Zhou et al., CVPR'16.
- Zhang et al., ECCV'16
- Selvaraju et al., ICCV'17.
- Chattopadhyay et al., WACV'18.
- Zhang. et al., CVPR'18.

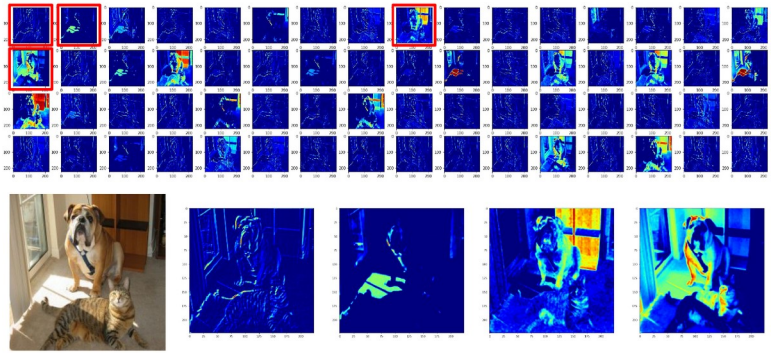
# RELATED WORK
## VISUAL EXPLANATION

### Input-modification



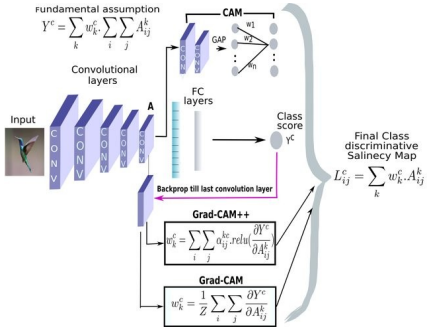
- Zeiler et al. ICCV'11
- Zeiler et al. ECCV'14
- Zhou et al., ICLR'15

### Deconvolution-based



- Zeiler et al. ECCV'14
- Springenberg et al., ICLR'15.
- Grun et al., ICML'16.

### Class-activation mapping

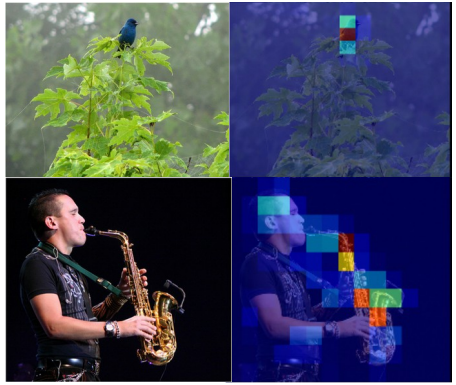



- Zhou et al., CVPR'16.
- Zhang et al., ECCV'16
- Selvaraju et al., ICCV'17.
- Chattopadhyay et al., WACV'18.
- Zhang. et al., CVPR'18.

**But: coarse**
**expensive**

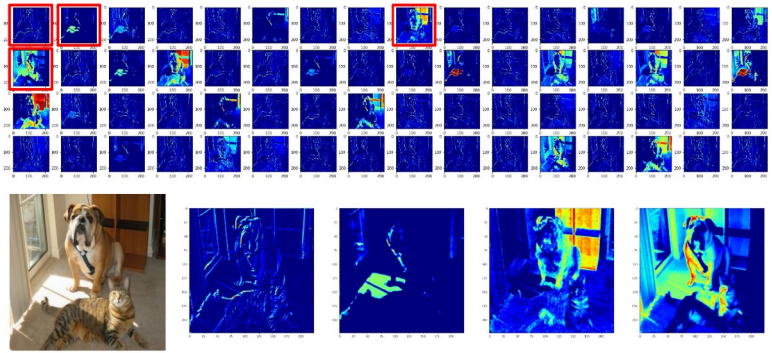
# RELATED WORK
## VISUAL EXPLANATION

### Input-modification



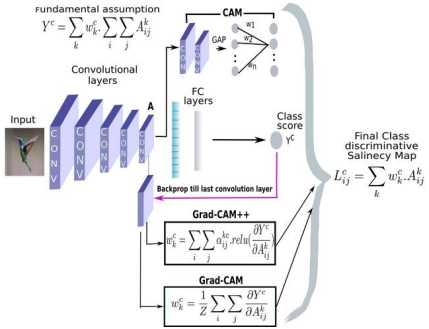
- Zeiler et al. ICCV'11
- Zeiler et al. ECCV'14
- Zhou et al., ICLR'15

### Deconvolution-based



- Zeiler et al. ECCV'14
- Springenberg et al., ICLR'15.
- Grun et al., ICML'16.

### Class-activation mapping



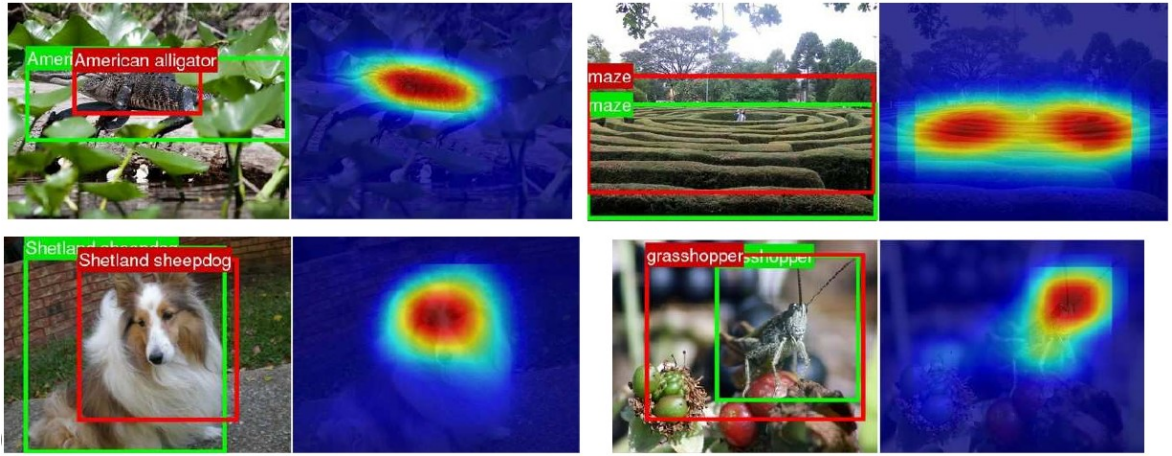

- Zhou et al., CVPR'16.
- Zhang et al., ECCV'16
- Selvaraju et al., ICCV'17.
- Chattopadhyay et al., WACV'18.
- Zhang. et al., CVPR'18.

**But: coarse**
**expensive**

**But: sensible to image edges [some]**

# RELATED WORK

## EVALUATION OF VISUAL EXPLANATIONS

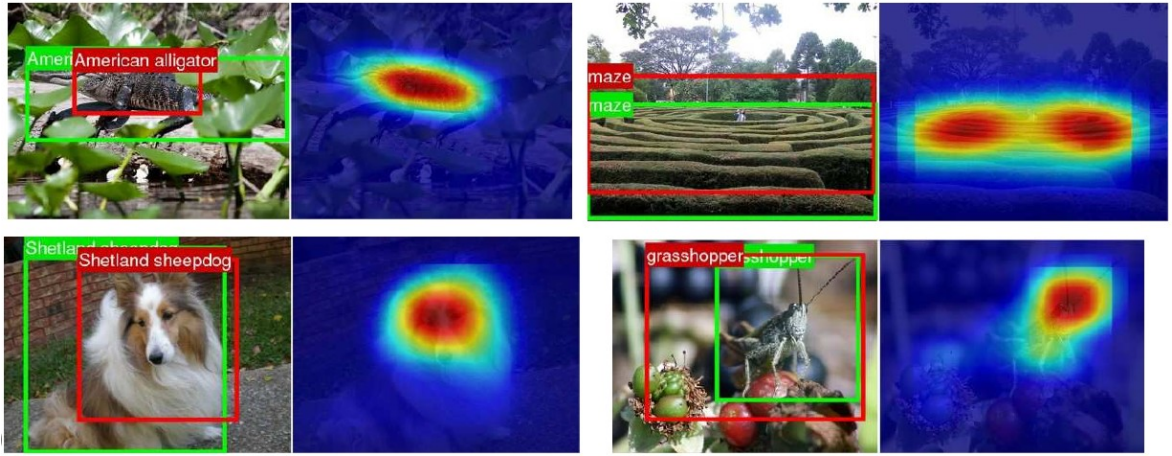



**Via a proxy task**
- Zhou et al., CVPR'16
- Zhang et al., ECCV'16.

**User studies**
- Zeiler & Fergus, ECCV'14
- Selvaraju et al., ICCV'17.

# RELATED WORK

## EVALUATION OF VISUAL EXPLANATIONS



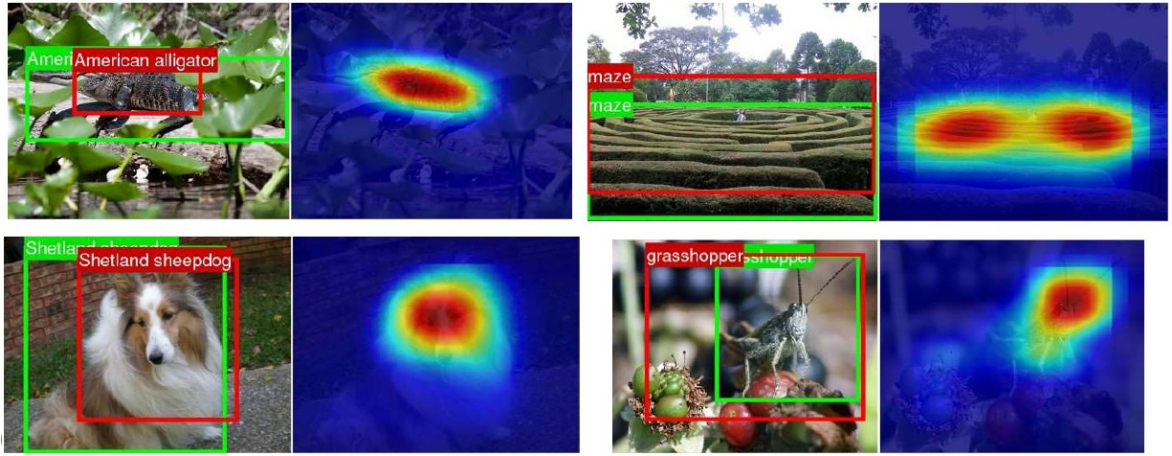

**Via a proxy task**
- Zhou et al., CVPR'16
- Zhang et al., ECCV'16.

**User studies**
- Zeiler & Fergus, ECCV'14
- Selvaraju et al., ICCV'17.

**But: bias towards human interpretation**

# RELATED WORK

## EVALUATION OF VISUAL EXPLANATIONS




**Via a proxy task**
- Zhou et al., CVPR'16
- Zhang et al., ECCV'16.

But: bias towards human interpretation

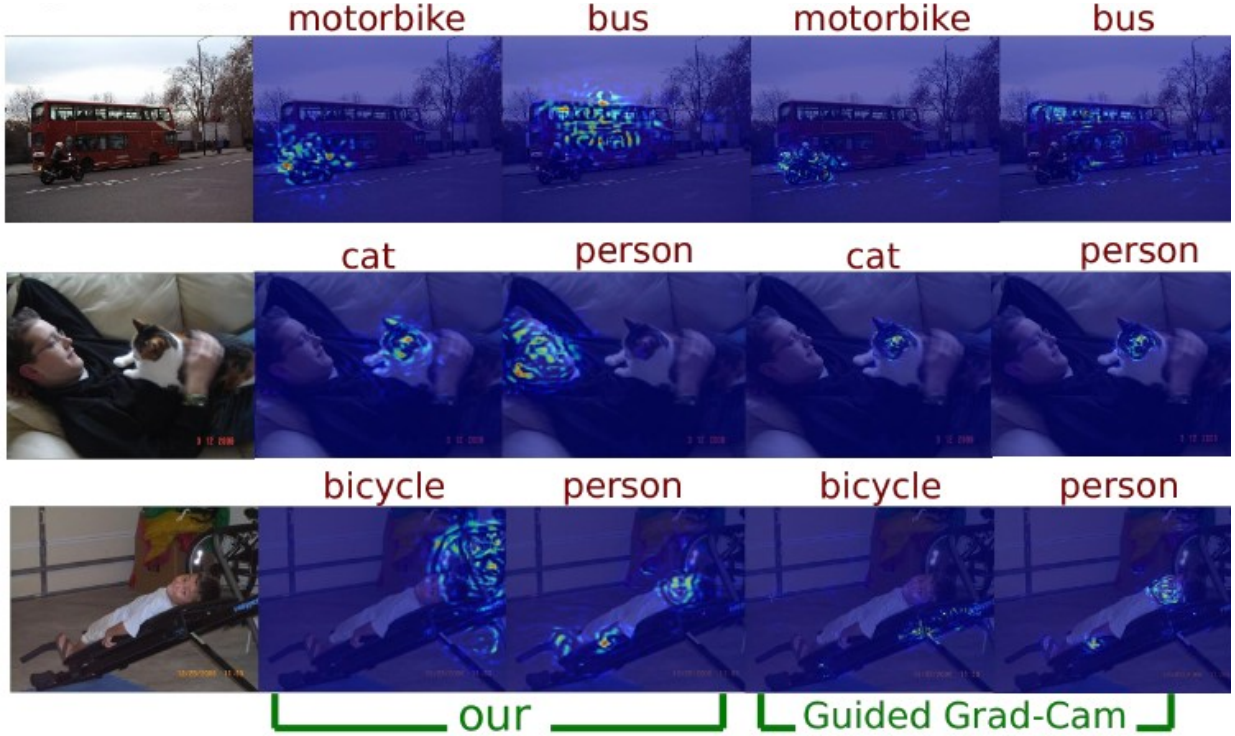
**User studies**
- Zeiler & Fergus, ECCV'14
- Selvaraju et al., ICCV'17.

But: subjective
bias towards human interpretation

# EVALUATION

# EVALUATION
## VISUAL EXPLANATION

**Presence of distracting instances**




[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL EXPLANATION

- Assessing the Sanity of the Generated Explanations


"Generated explanations should be dependent on the model and the classes of interest"

- Kindermans et al., NeurIPS'17.
- Nie et al., ICML'18.
- Adebayo et al., NeurIPS'18.

[ Oramas et al., ICLR'19 ]

# EVALUATION
## VISUAL EXPLANATION

- Assessing the Sanity of the Generated Explanations


“Generated explanations should be dependent on the model and the classes of interest”

- Kindermans et al., NeurIPS’17.
- Nie et al., ICML’18.
- Adebayo et al., NeurIPS’18.



[ Oramas et al., ICLR’19 ]

# EVALUATION
## VISUAL EXPLANATION

- Assessing the Sanity of the Generated Explanations


“Generated explanations should be dependent on the model and the classes of interest”

- Kindermans et al., NeurIPS’17.
- Nie et al., ICML’18.
- Adebayo et al., NeurIPS’18.



[ Oramas et al., ICLR’19 ]