University of Antwerp

Artificial Neural Networks

[2500WETANN]

José Oramas

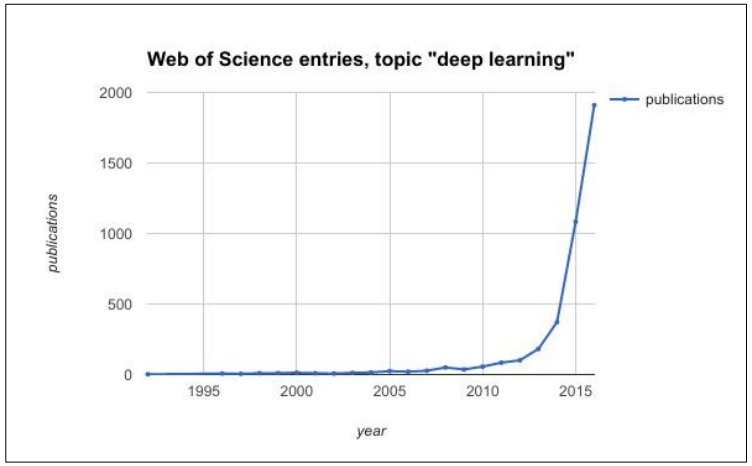
Model Interpretation and Explanation for Deep Neural Networks

José Oramas

University of Antwerp

Background

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

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

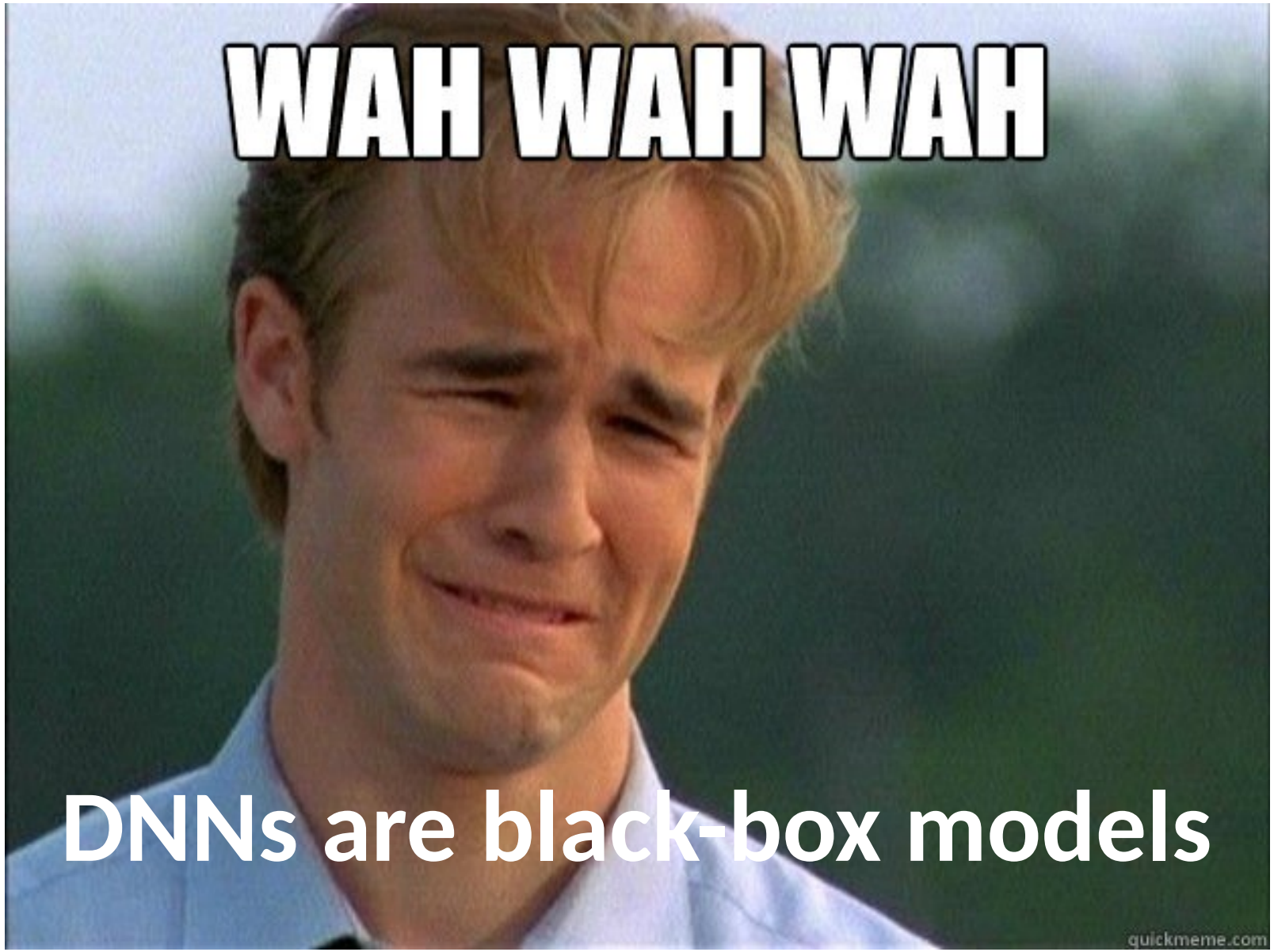
Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride & Weiva Sieh

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

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

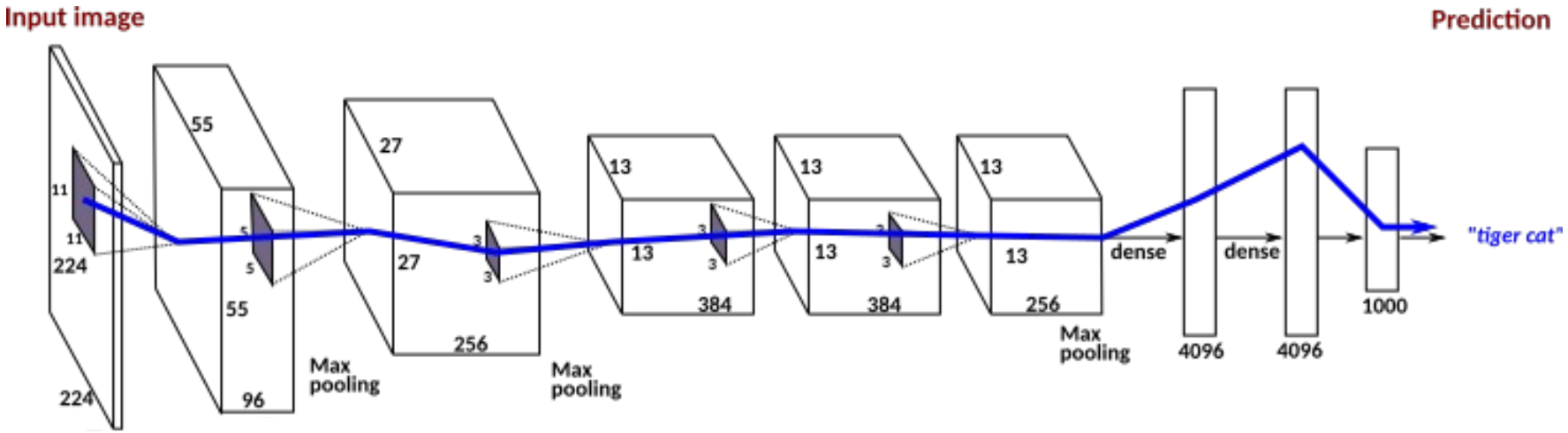
Background

However….



Background

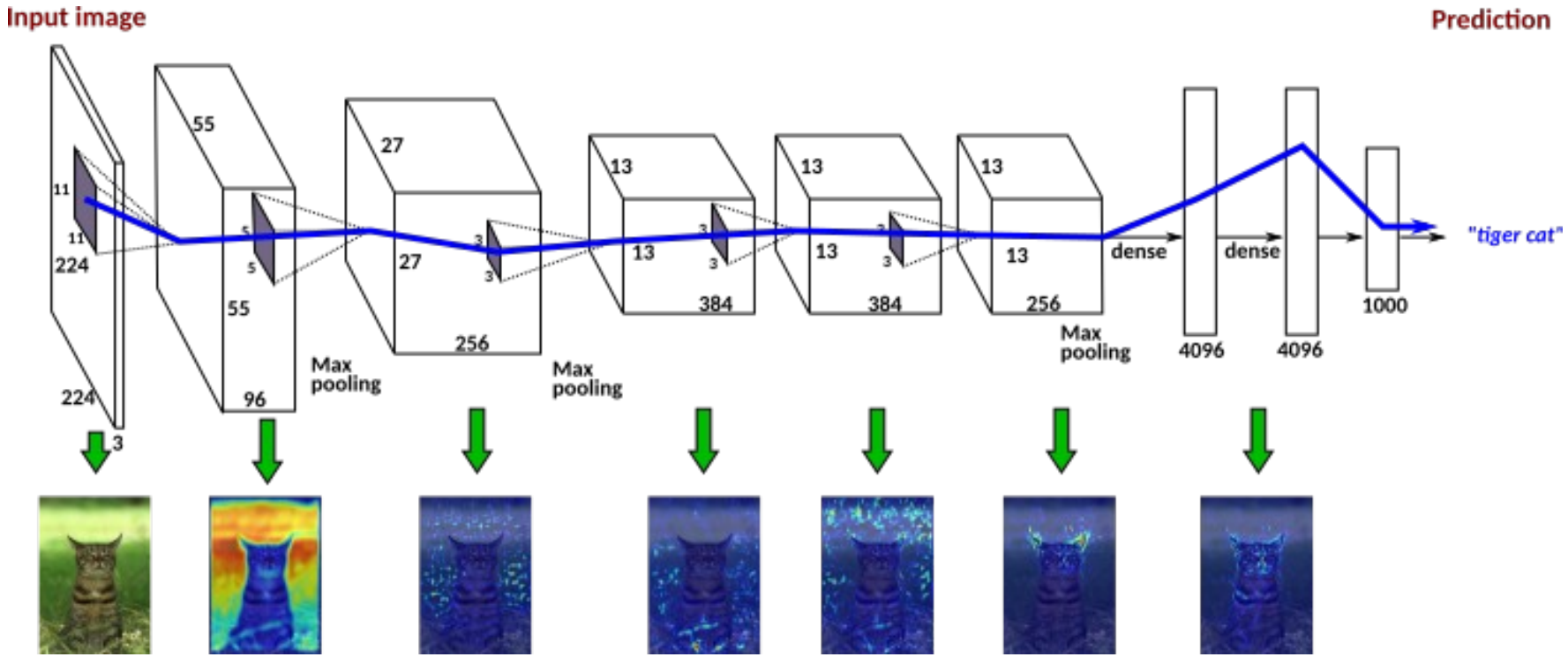
Deep Neural Networks



Background

Research Questions:

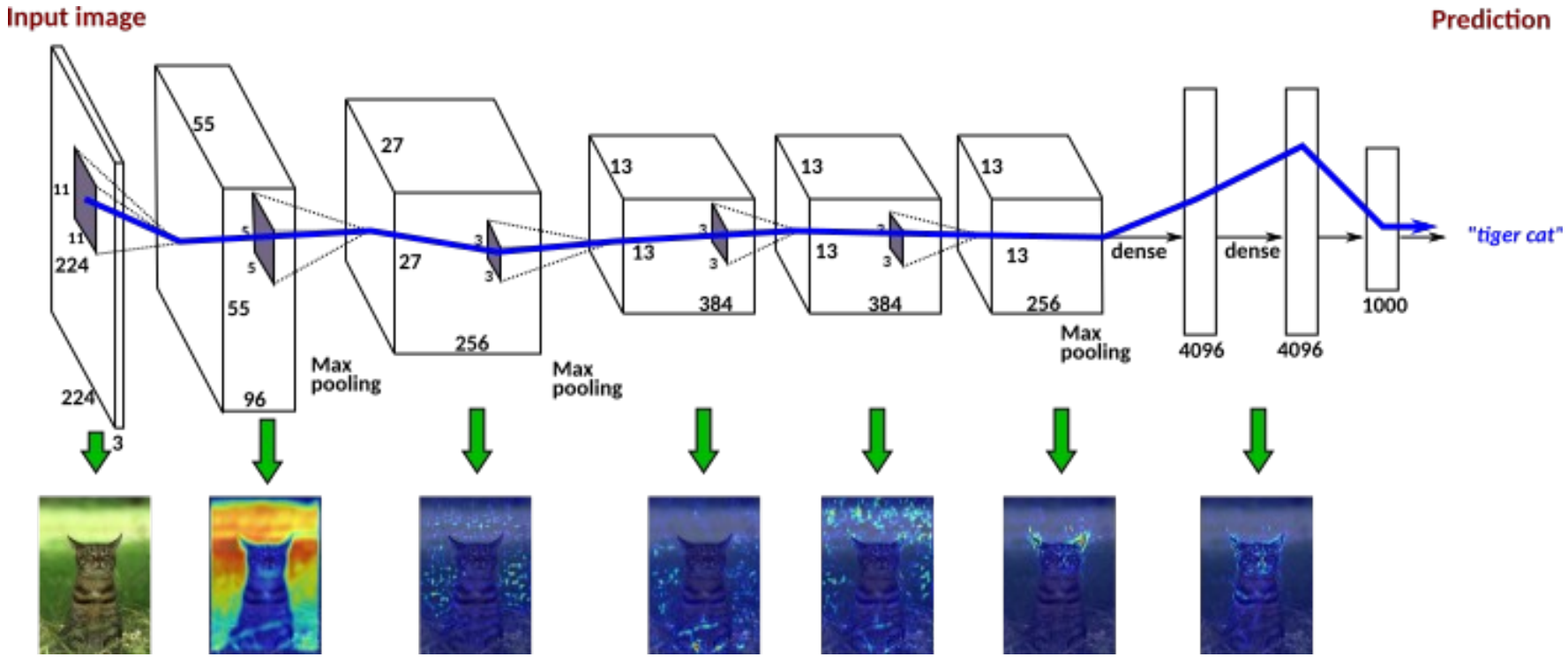
- Q_1: What the model has actually learned?



Background

Research Questions:

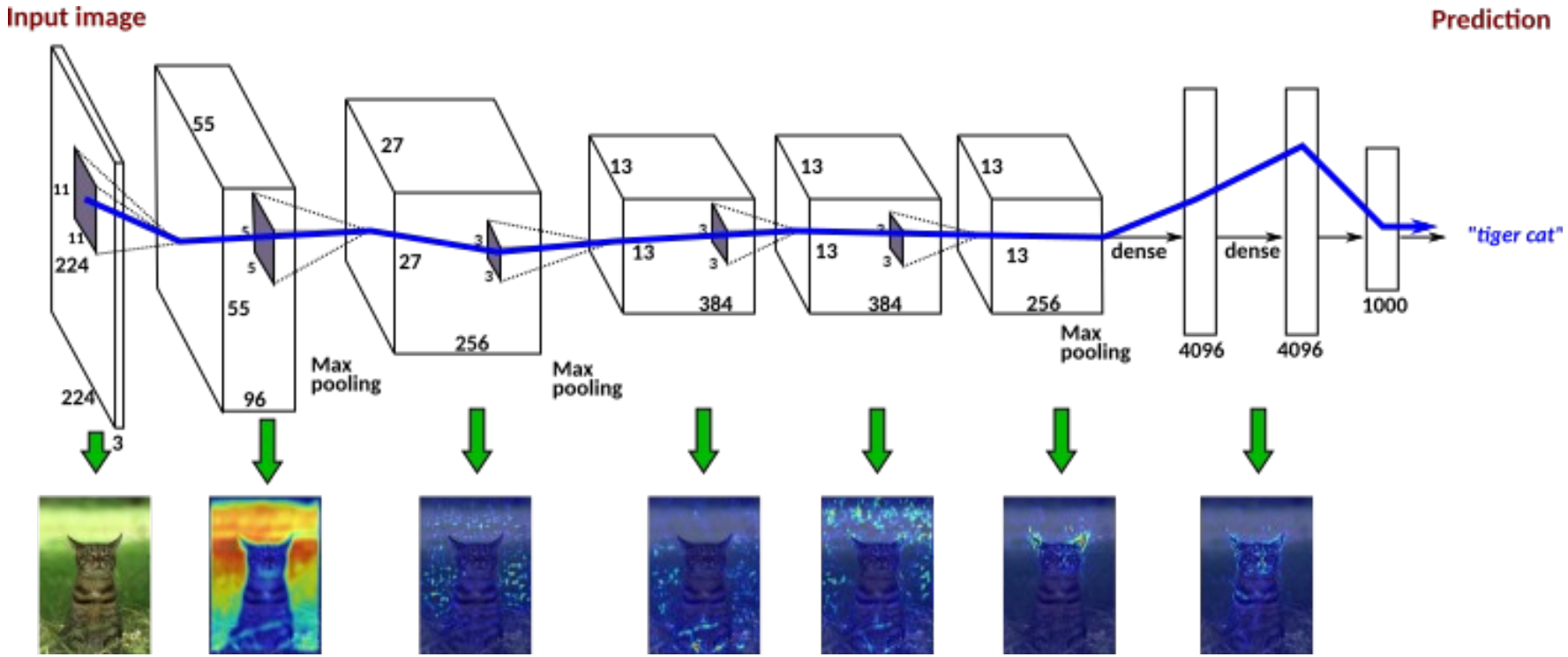
- Q_1: What the model has actually learned? → ***interpretation***



Background

Research Questions:

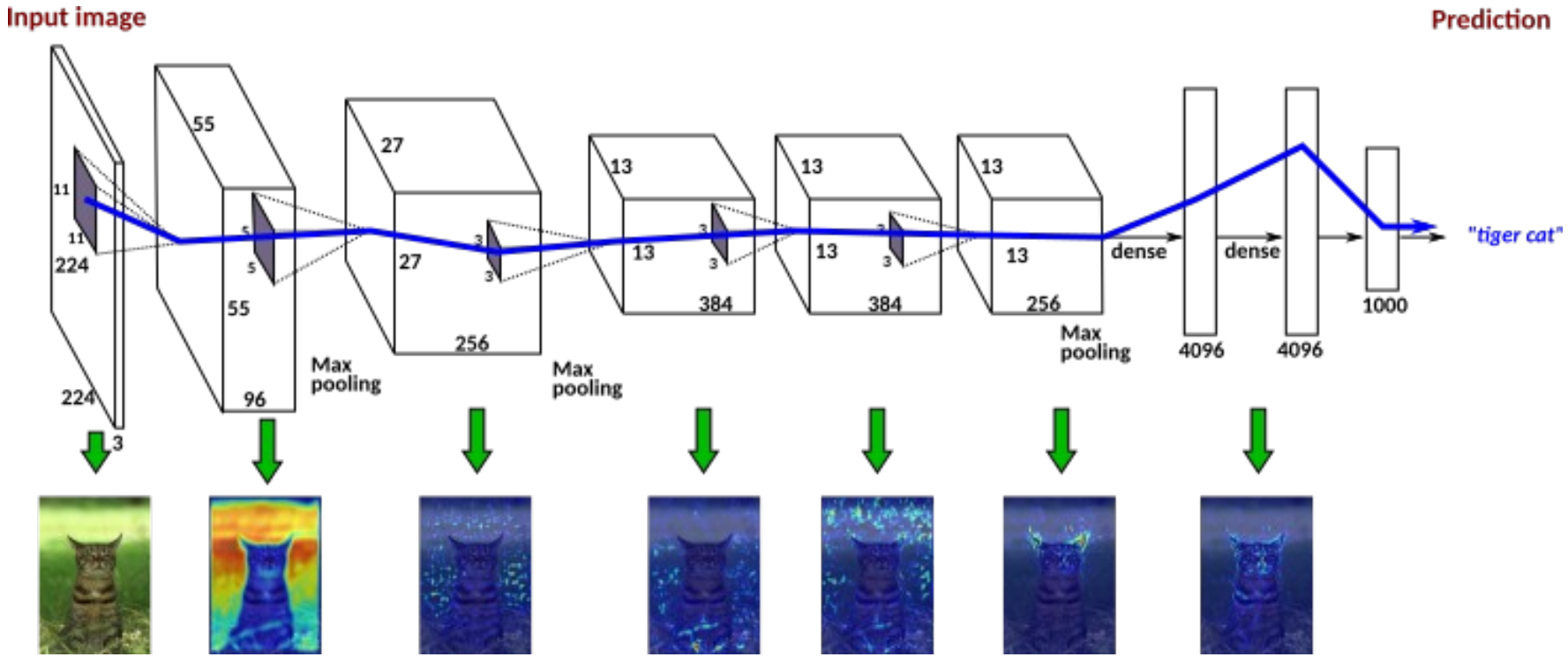
- Q_1: What the model has actually learned? → ***interpretation***
- Q_2: What information from the input is using to make predictions?



Background

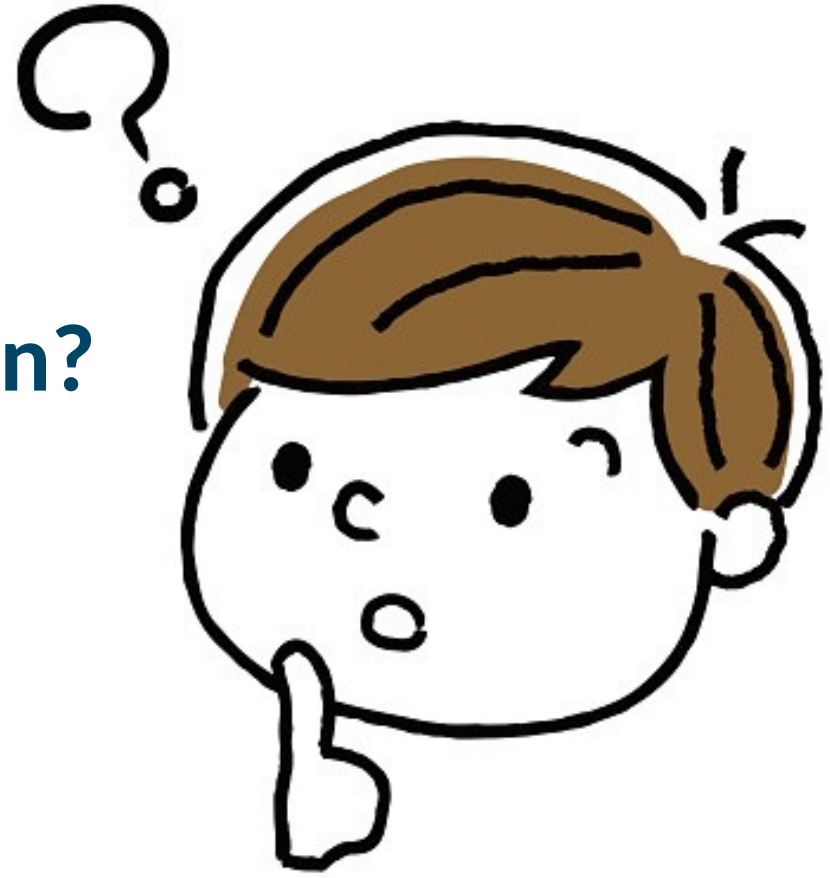
Research Questions:

- Q_1: What the model has actually learned? → ***interpretation***
- Q_2: What information from the input is using to make predictions? → ***explanation***



Nice but…

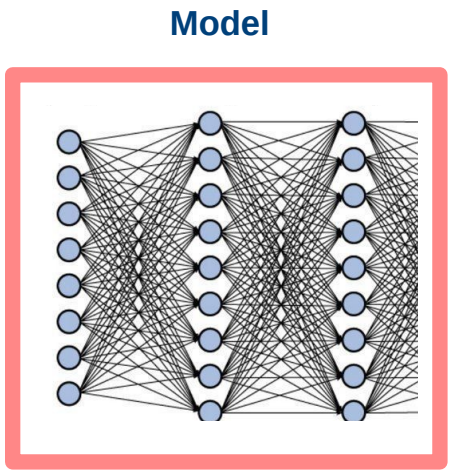
What do you really mean?

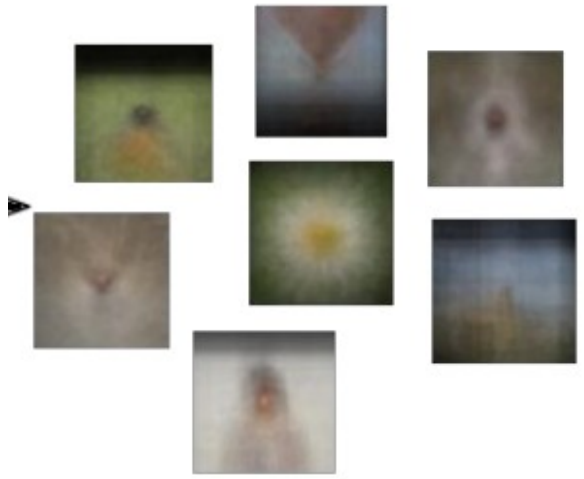


Model Explanation and Interpretation

Model Interpretation

Discovery

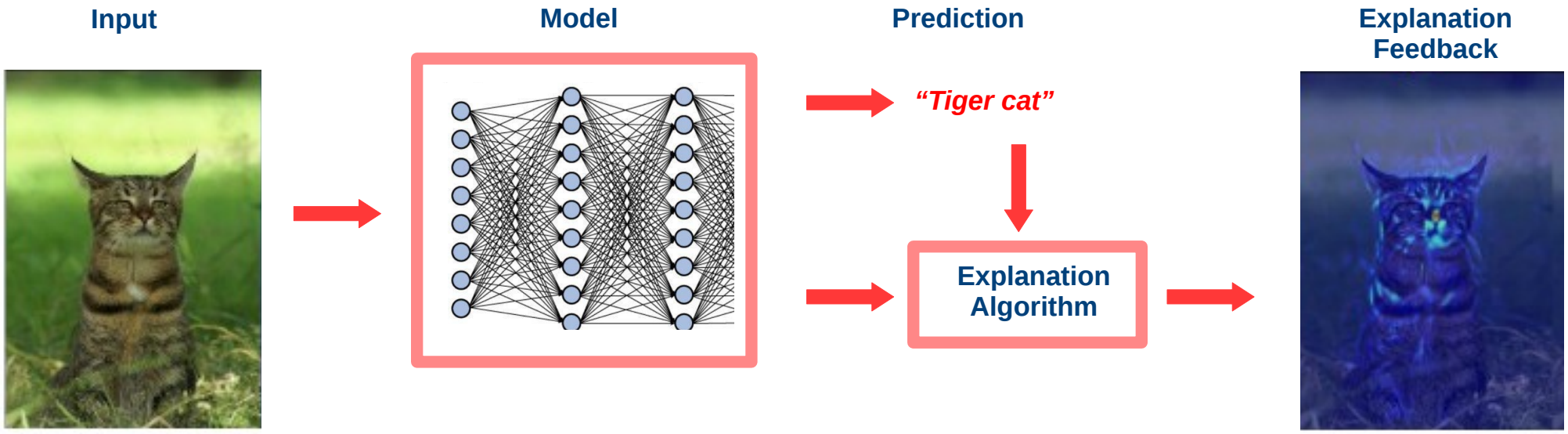




Interpretation Feedback

Model Explanation and Interpretation

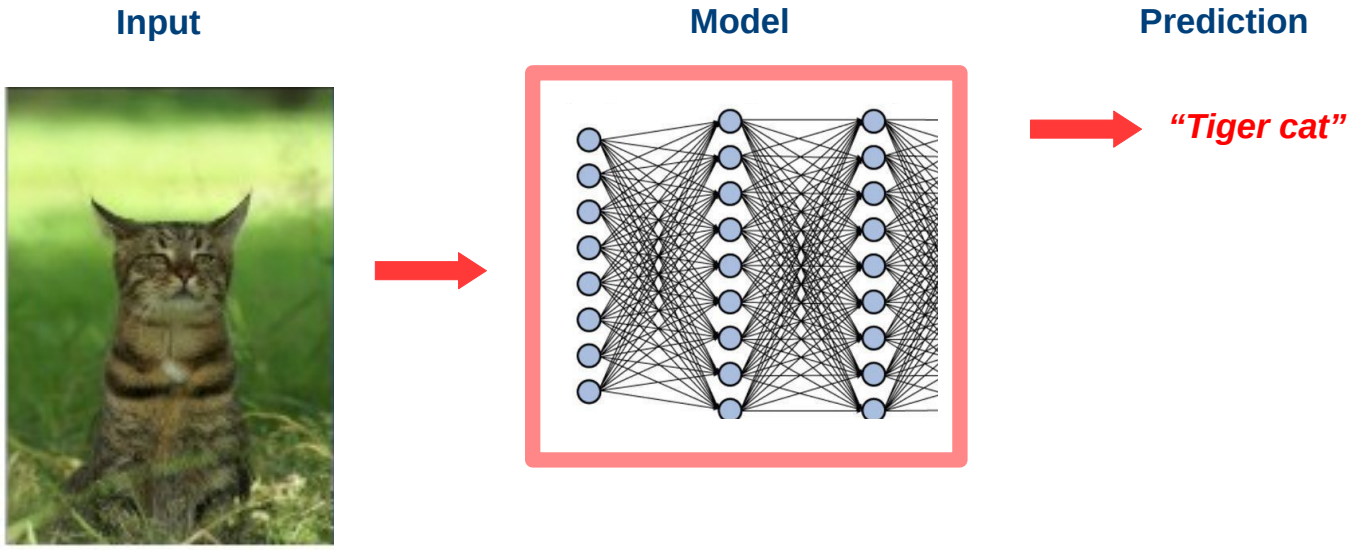
Model Explanation



Justification

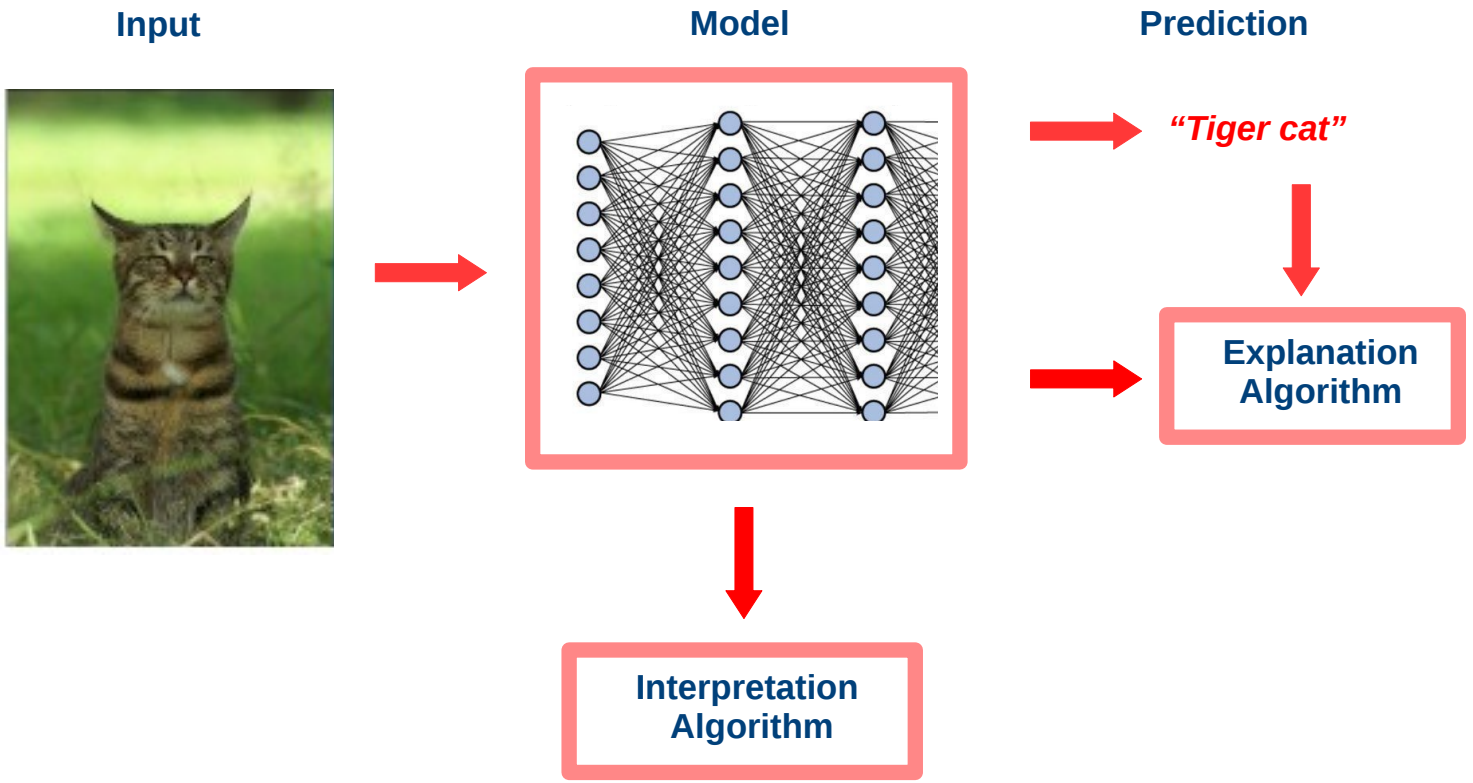
Model Interpretation and Explanation

It’s all about Transparency



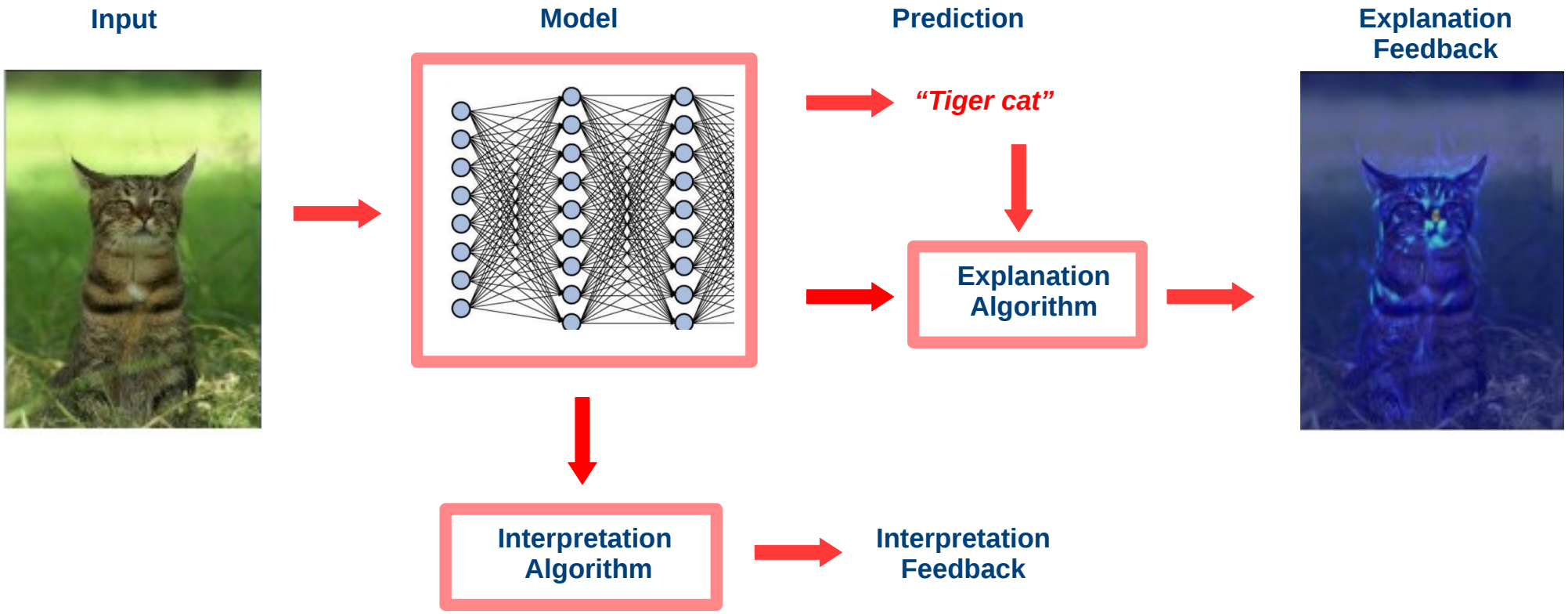
Model Interpretation and Explanation

It's all about Transparency



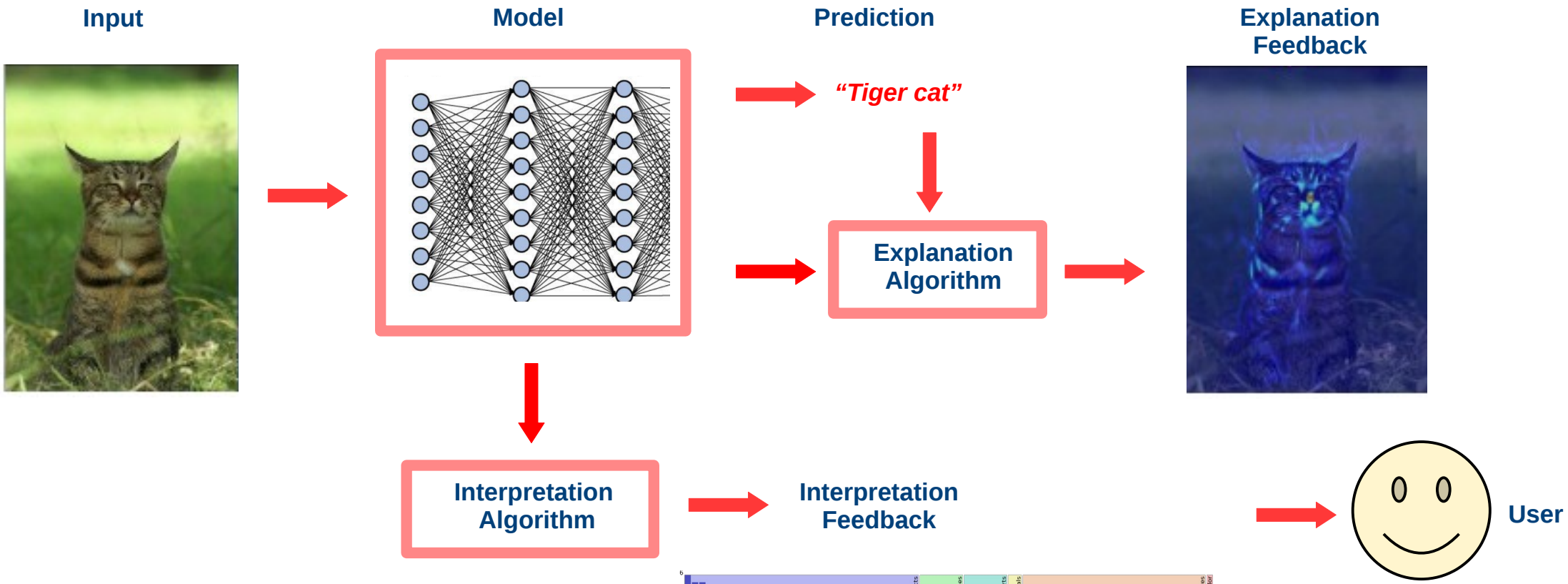
Model Interpretation and Explanation

It’s all about Transparency



Model Interpretation and Explanation

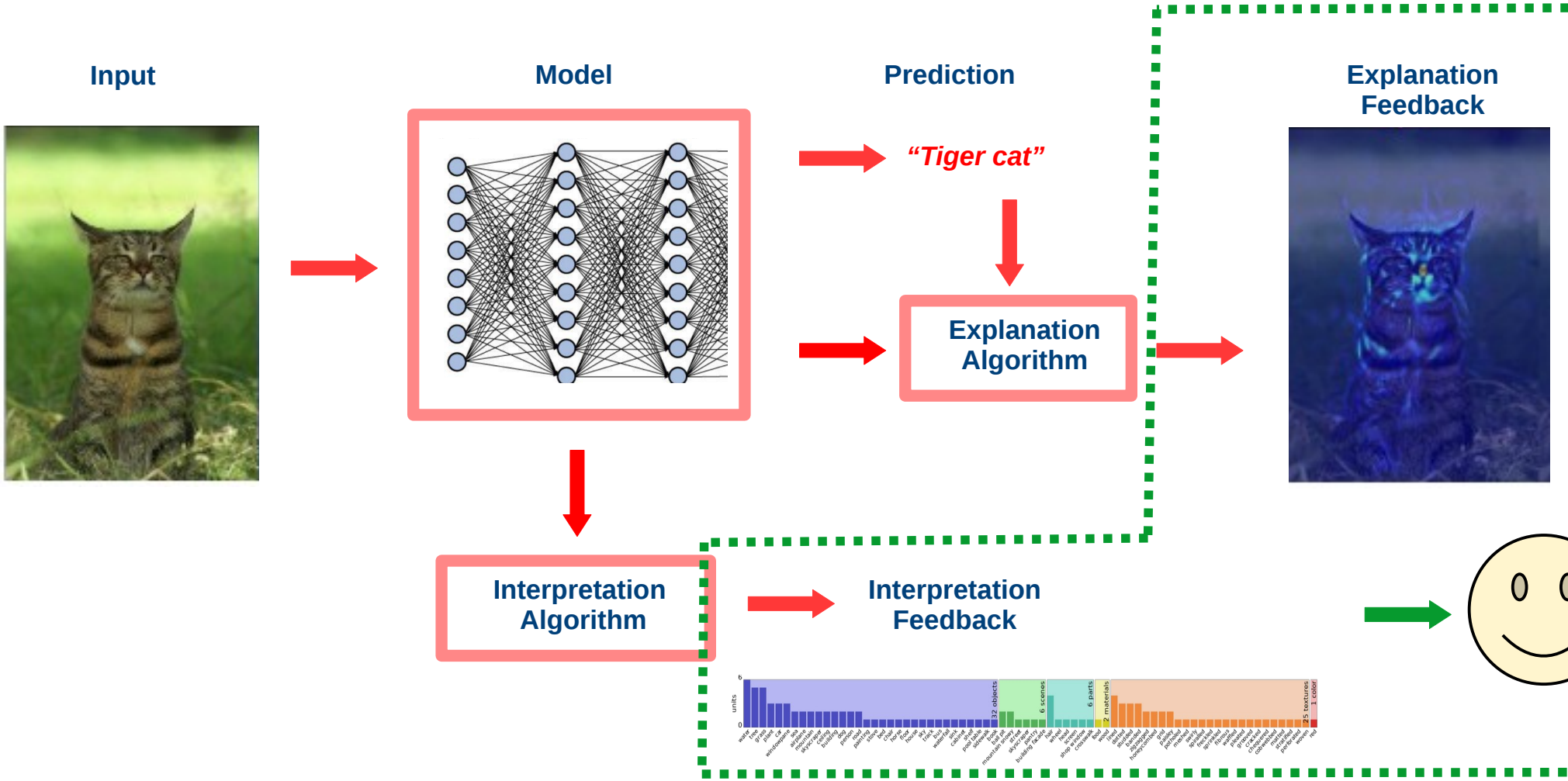
It's all about Transparency



Model Interpretation and Explanation

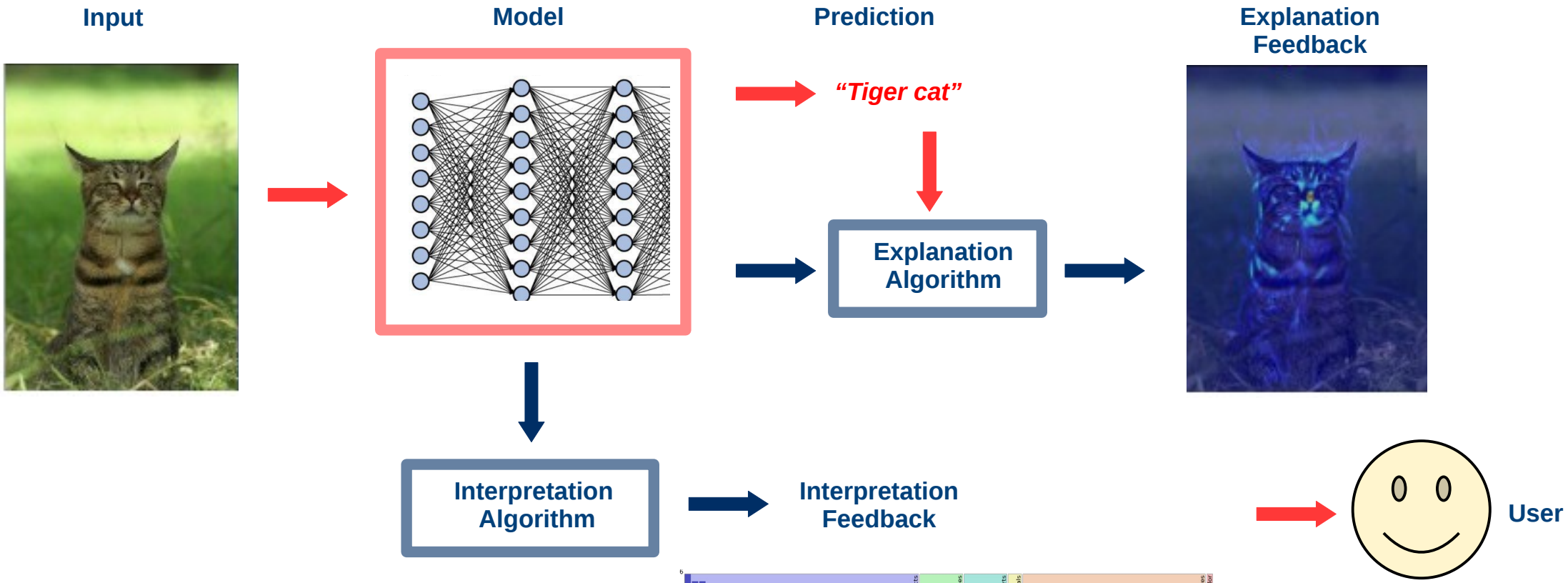
It's all about Transparency

HCI



Model Interpretation and Explanation

It's all about Transparency



But if DNNs have high performance…

Why is this desirable?

Interpretation & Explanation of DNNs

Motivation

- Detection of undesirable properties in the model (debugging)

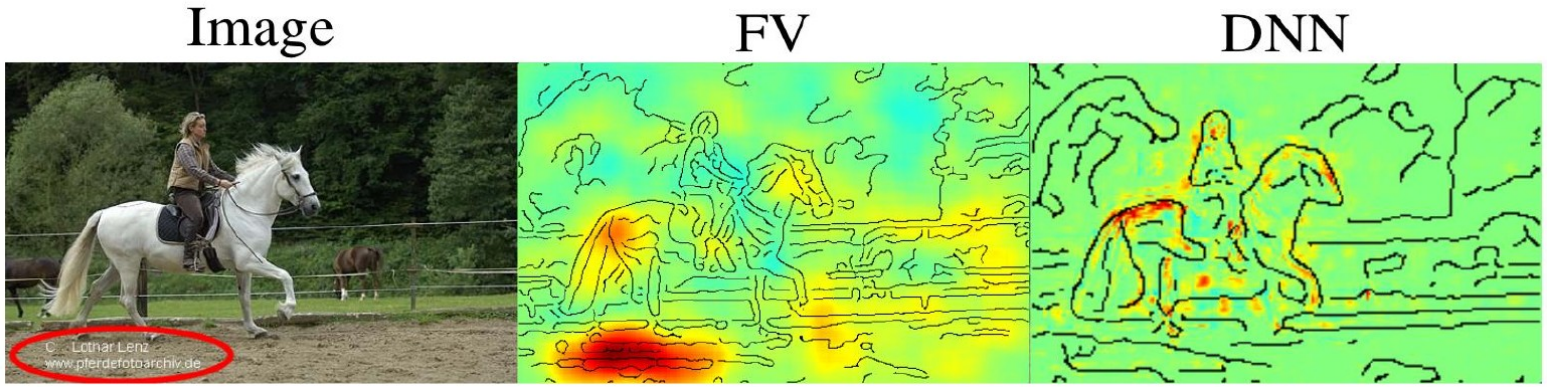
	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

Interpretation & Explanation of DNNs

Motivation

- Detection of undesirable properties in the model (debugging)

	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%



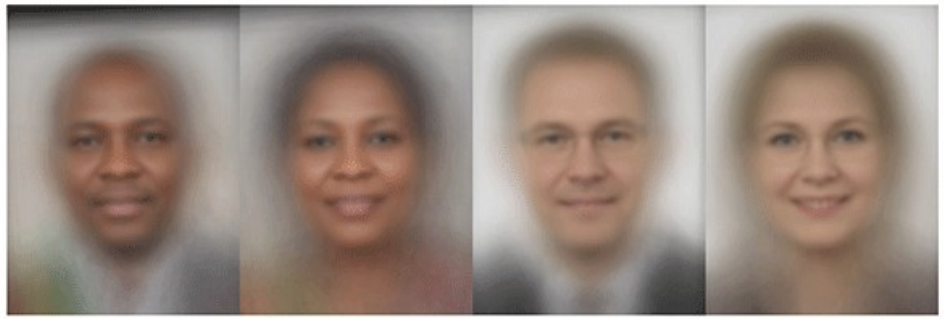
Interpretation & Explanation of DNNs

Motivation

- Enable the detection of possible biases

Face Detection

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

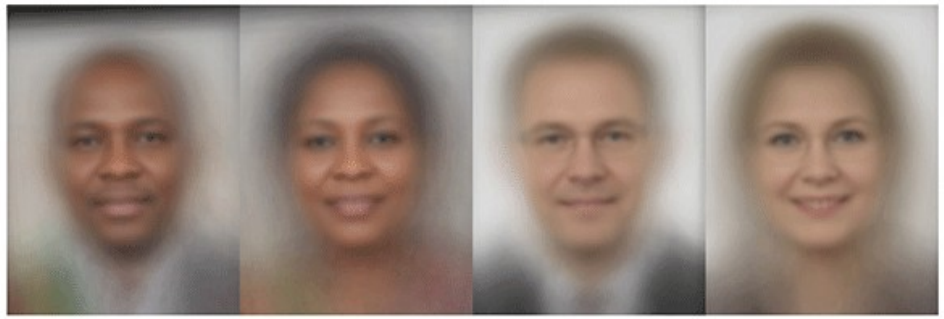
Interpretation & Explanation of DNNs

Motivation

- Enable the detection of possible biases

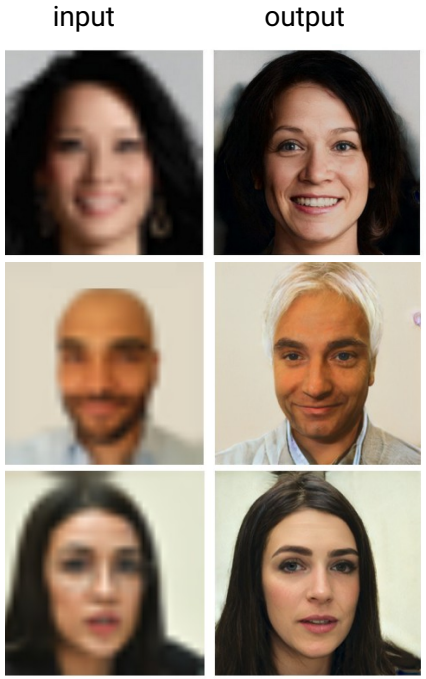
Face Detection

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

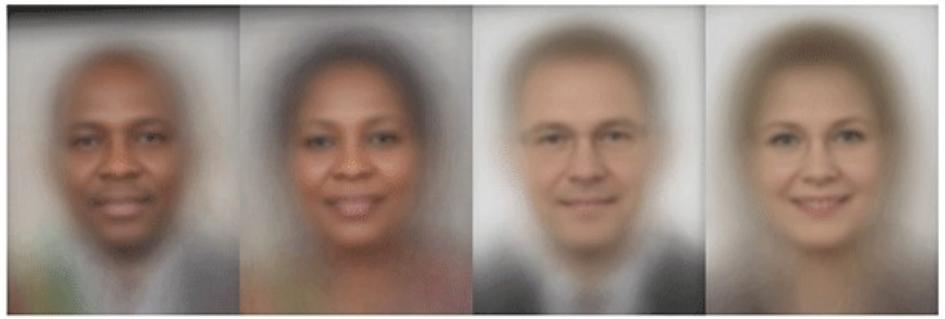
Interpretation & Explanation of DNNs

Motivation

- Enable the detection of possible biases

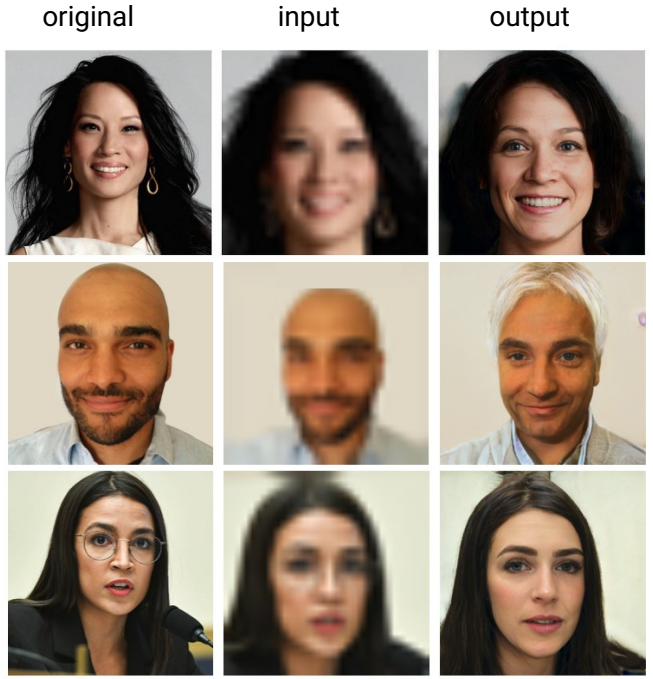
Face Detection

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

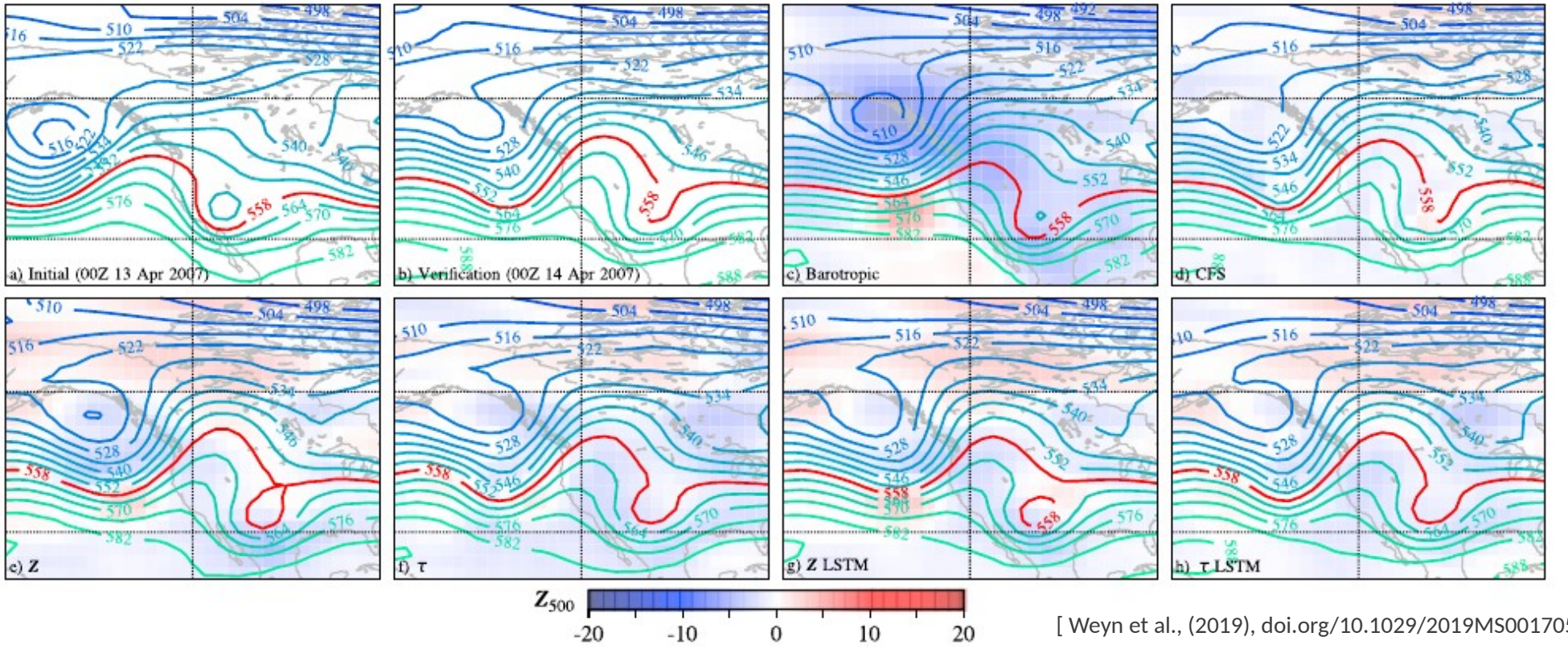


Twitter: @osazuwa

Interpretation & Explanation of DNNs

Motivation

- Obtain Additional Insights on Existing Problems (e.g. weather forecasting)



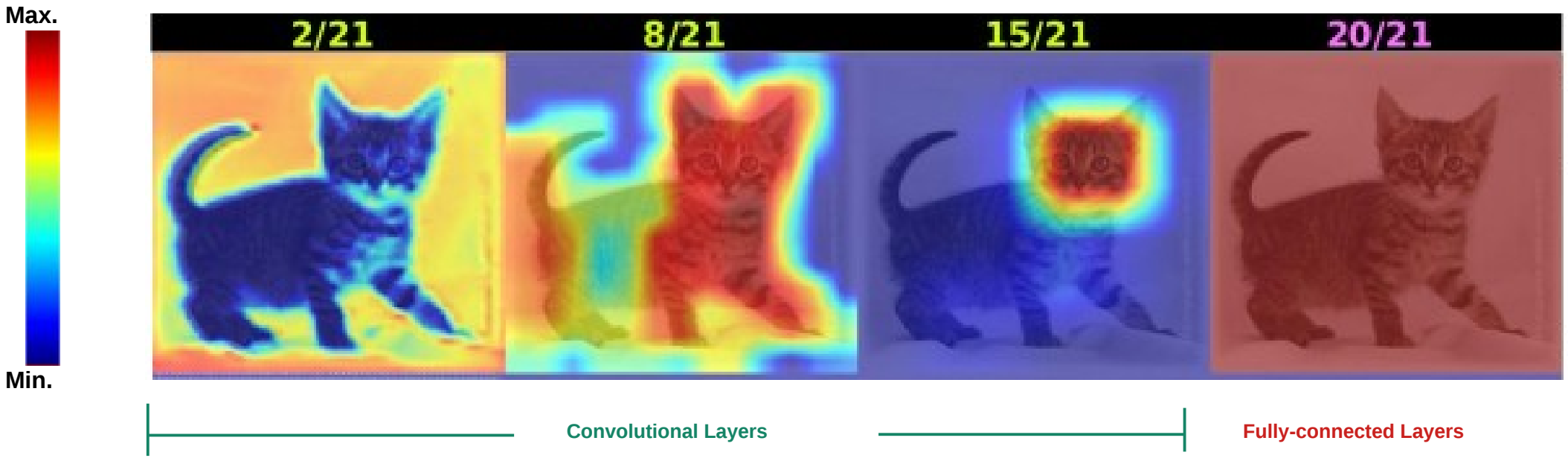
[Weyn et al., (2019), doi.org/10.1029/2019MS001705]

Model Interpretation

[Figuring out what a model has learned]

Model Interpretation

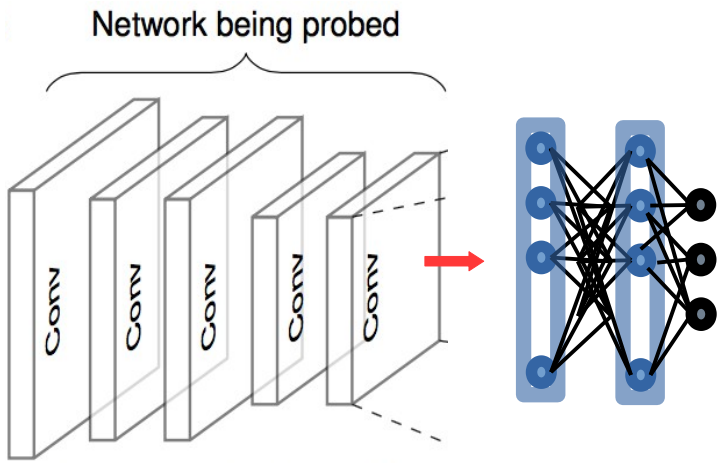
Internal Network Activations



- Response of the network at a given layer-filter location

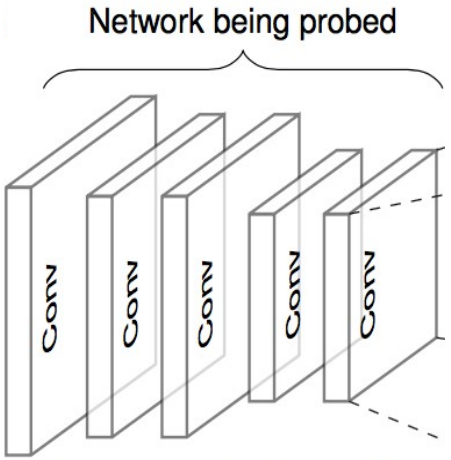
Model Interpretation

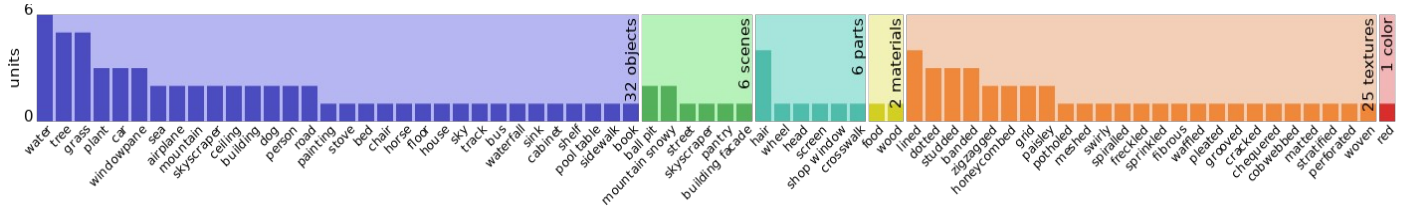
Probing Internal Network Activations [Bau et al., 2017]



Model Interpretation

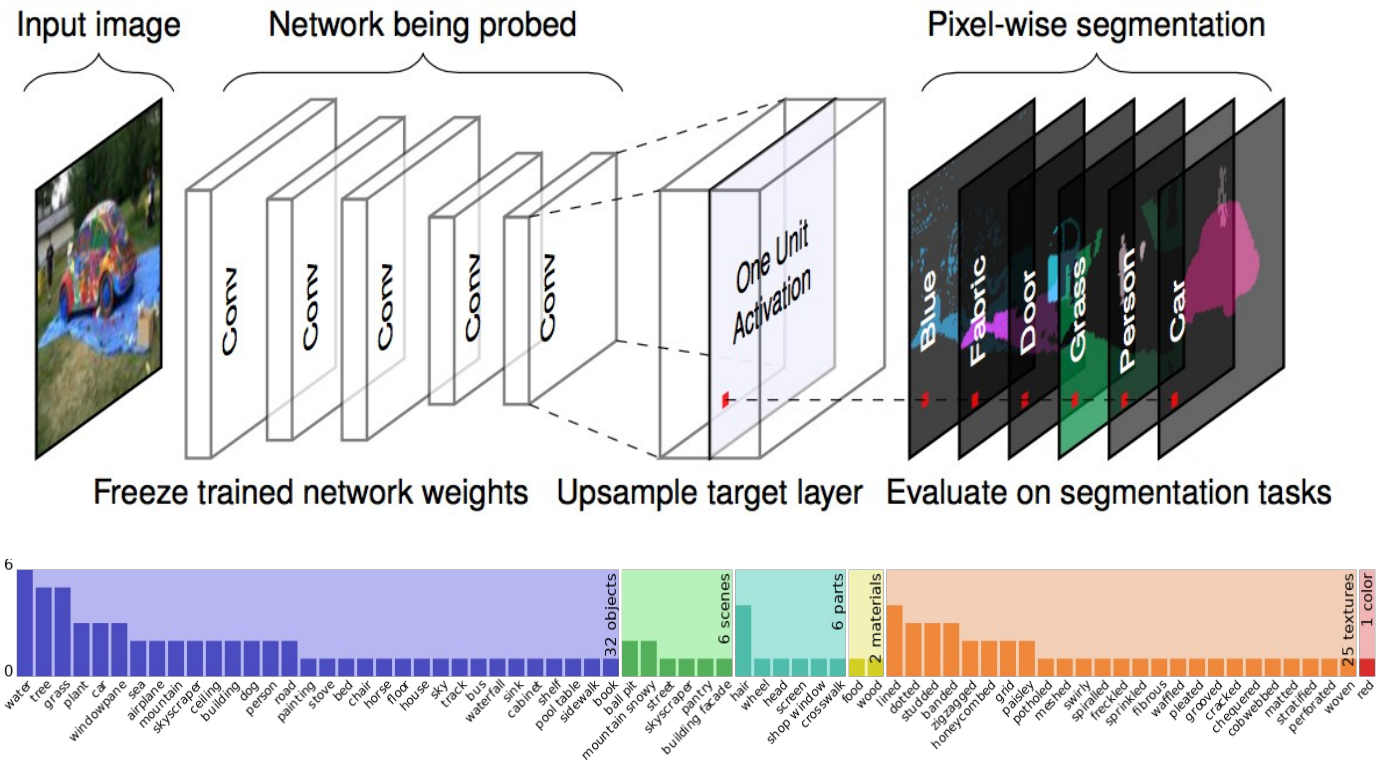
Probing Internal Network Activations [Bau et al., 2017]





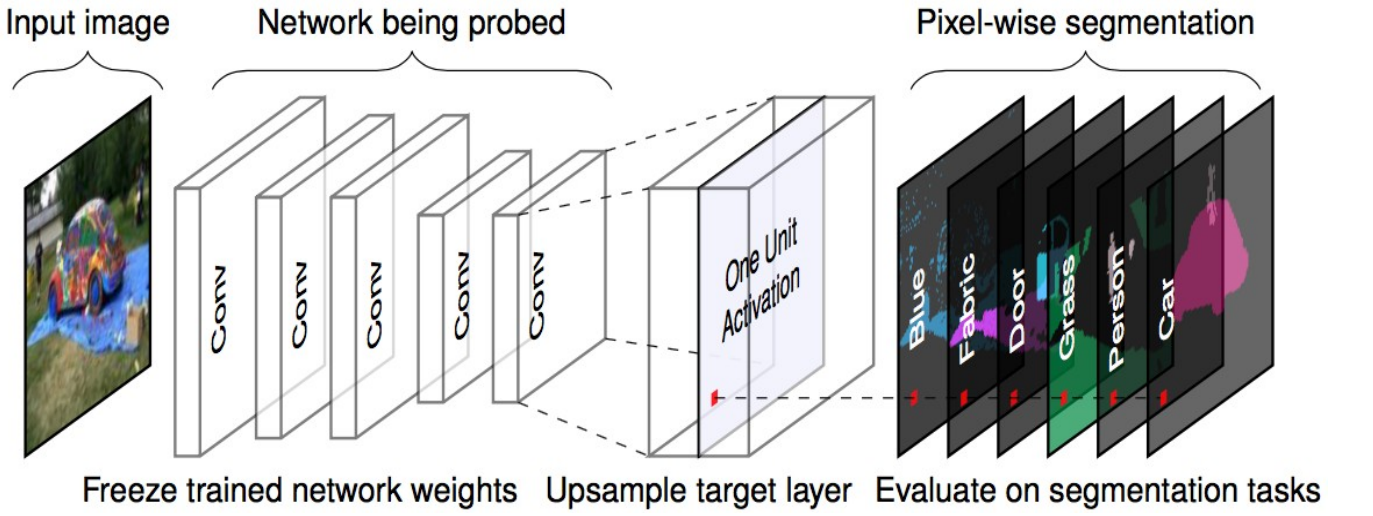
Model Interpretation

Probing Internal Network Activations [Bau et al., 2017]

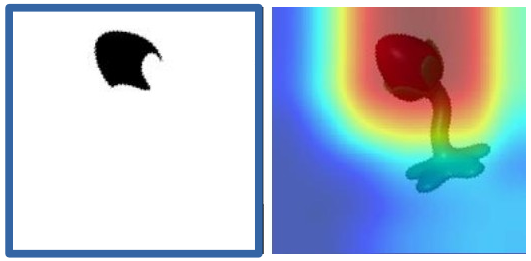


Model Interpretation

Probing Internal Network Activations [Bau et al., 2017]



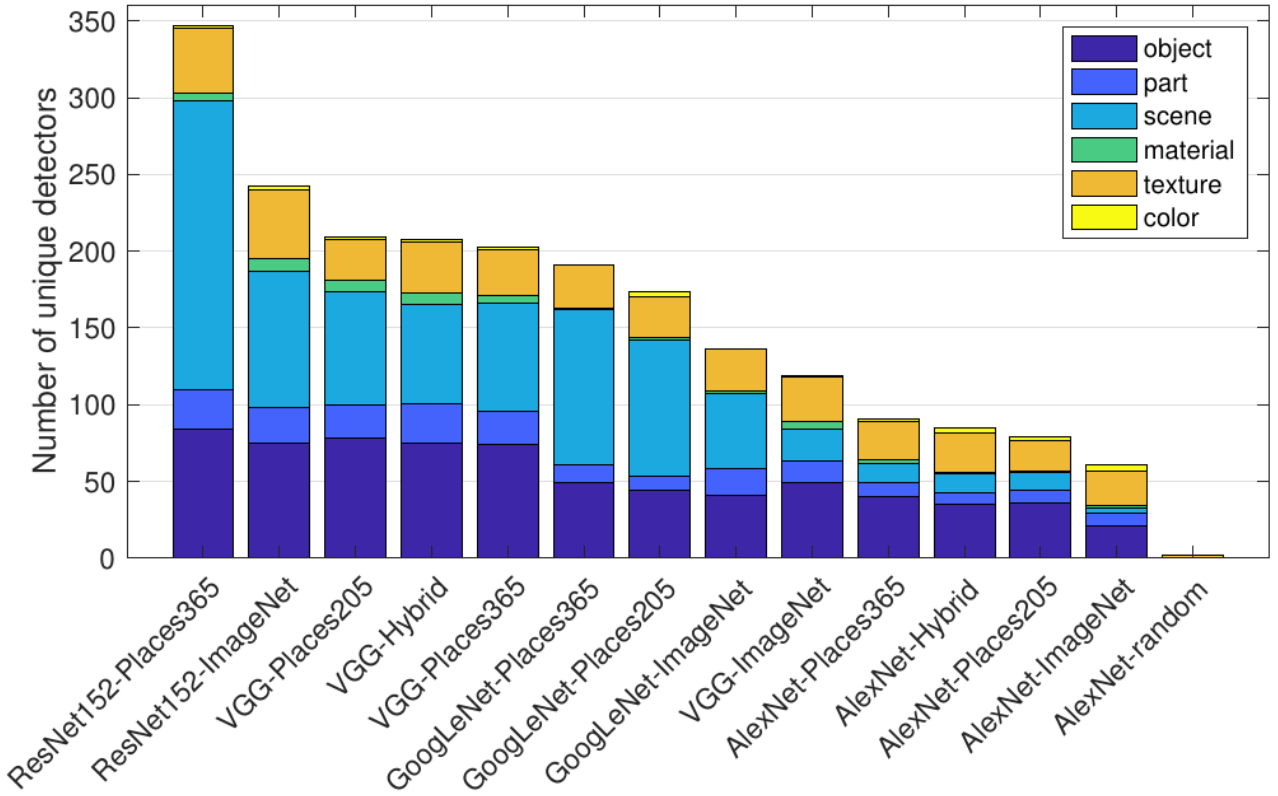
Measuring Overlap



$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|}$$

Model Interpretation

Probing Internal Network Activations [Bau et al., 2017]

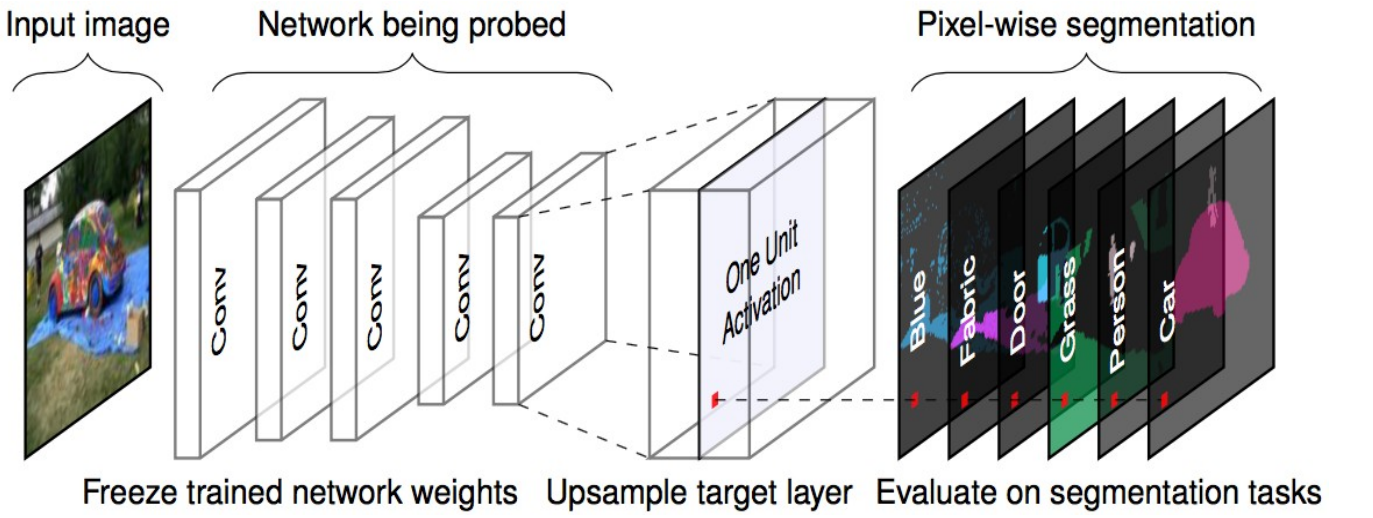


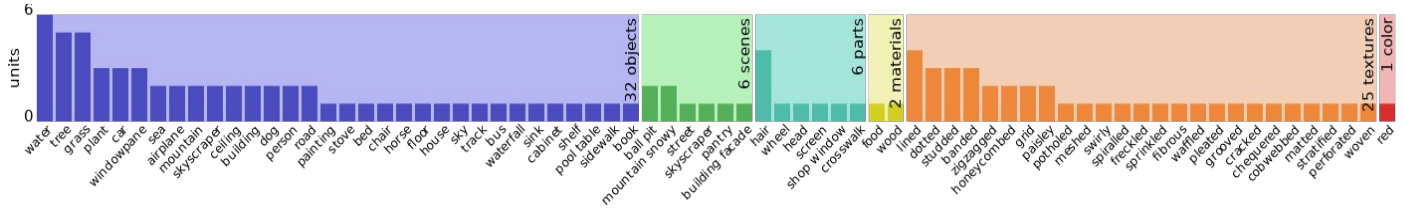
Output

- Histograms of semantic concepts covered by the network
- Quantified interpretation

Model Interpretation

Probing Internal Network Activations [Bau et al., 2017]

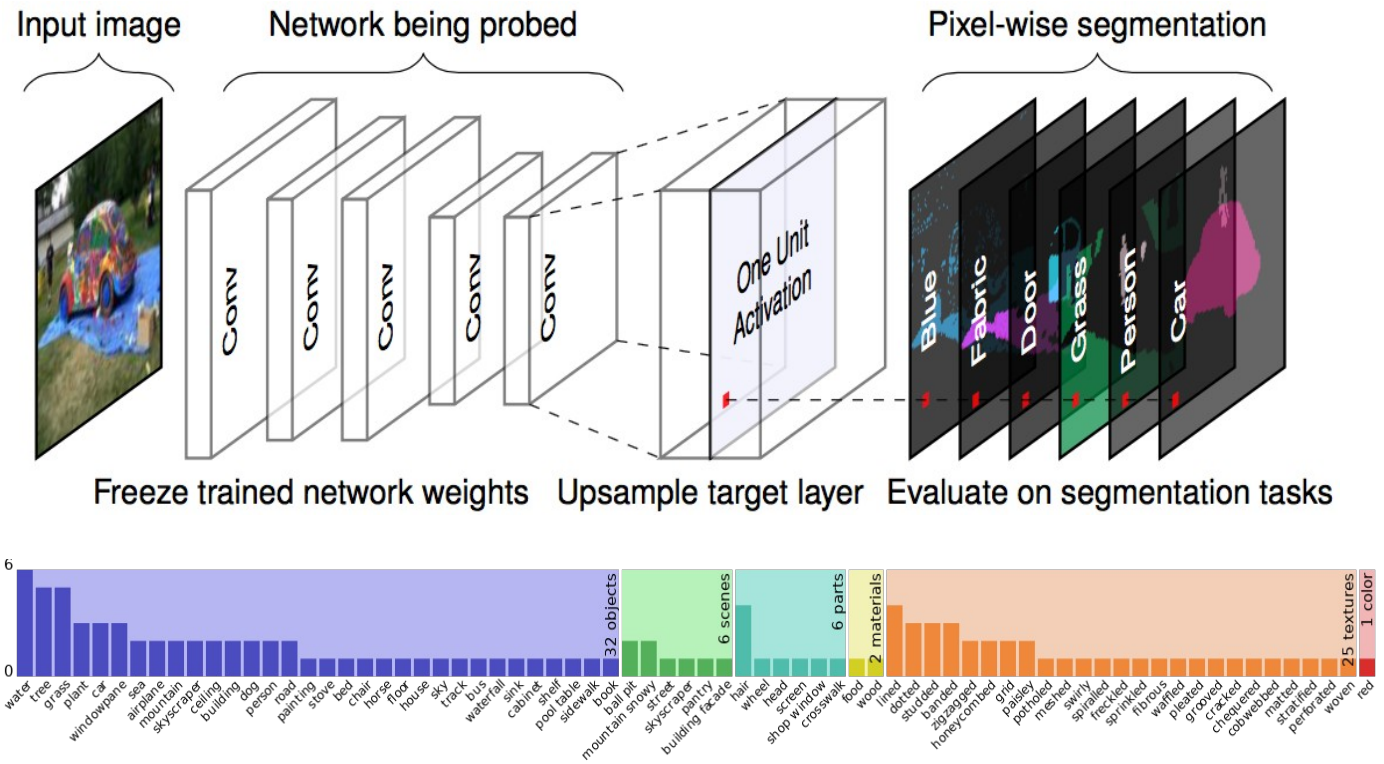




Q: Any potential problem or limitation?

Model Interpretation

Probing Internal Network Activations [Bau et al., 2017]



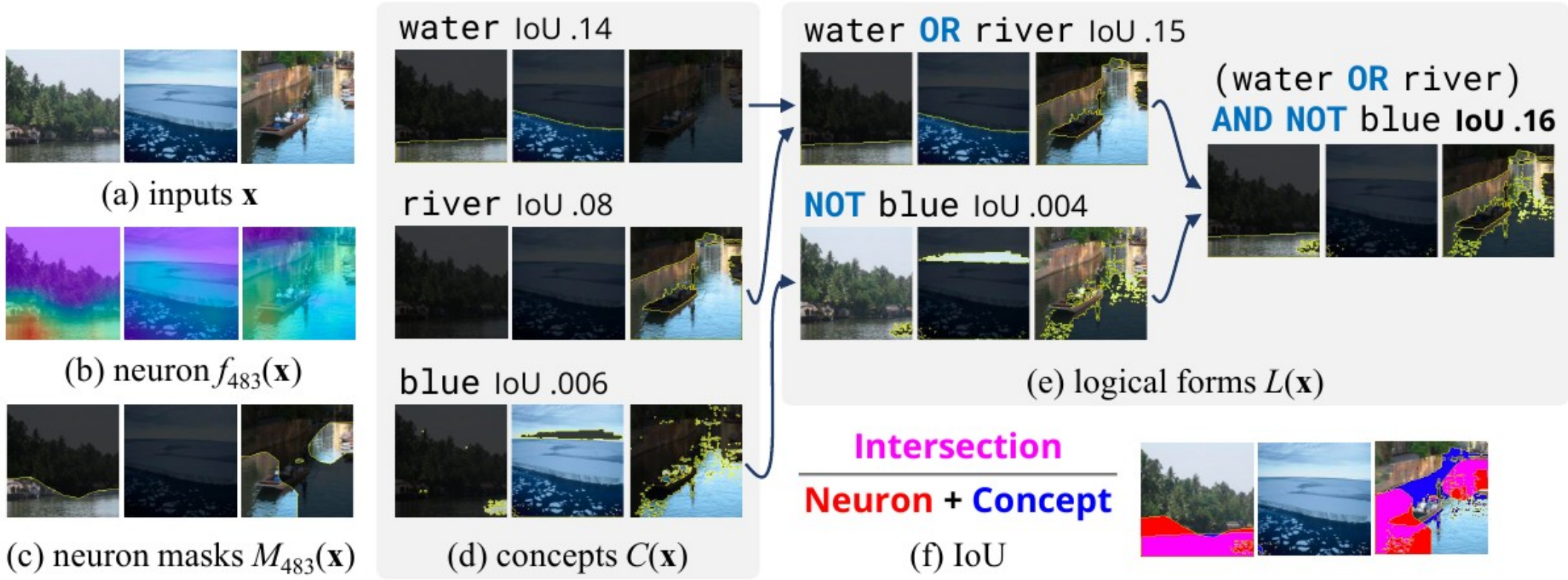
Weakness

- High computational costs
- Interpretation limited to annotated concepts.

Q: Any potential problem or limitation?

Model Interpretation

Compositional Neuron Analysis [Mu & Andreas, 2020]



$$\delta(n, C) \triangleq \text{IoU}(n, C) = \left[\sum_{\mathbf{x}} \mathbb{1}(M_n(\mathbf{x}) \wedge C(\mathbf{x}))\right] / \left[\sum_{\mathbf{x}} \mathbb{1}(M_n(\mathbf{x}) \vee C(\mathbf{x}))\right]$$

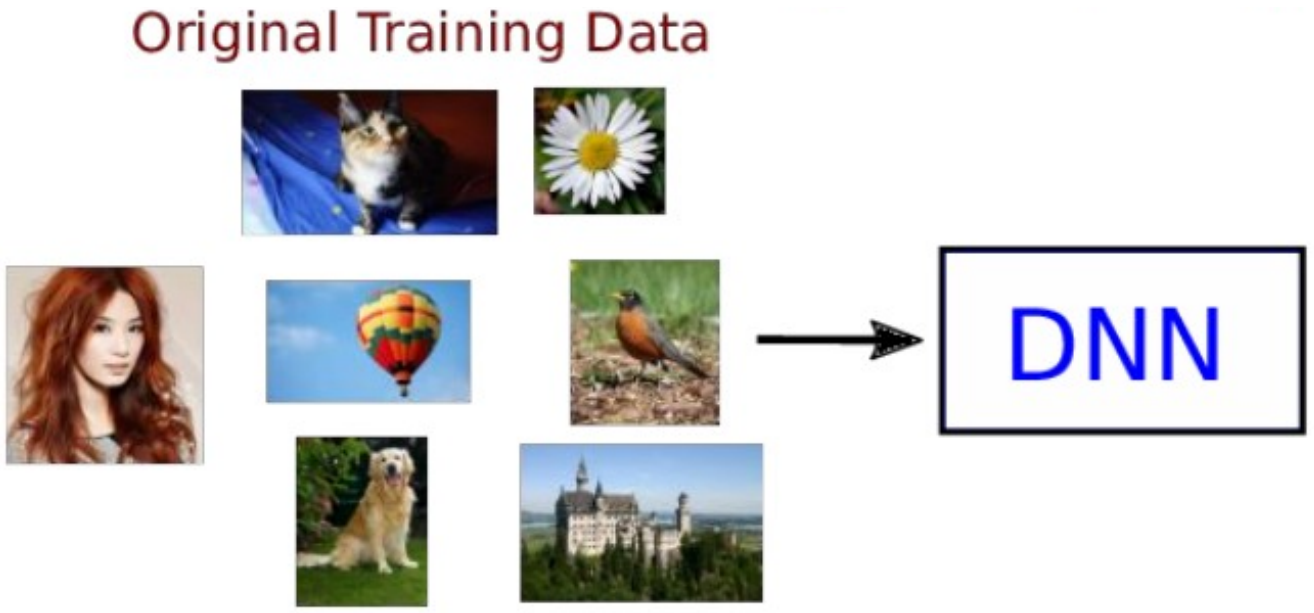
Model Interpretation

Characterizing Internal Network Activations [Oramas et al., 2019]

- **Assumption:** there exist a subset of features that can serve as indicators of the representation encoded by the model.

Model Interpretation

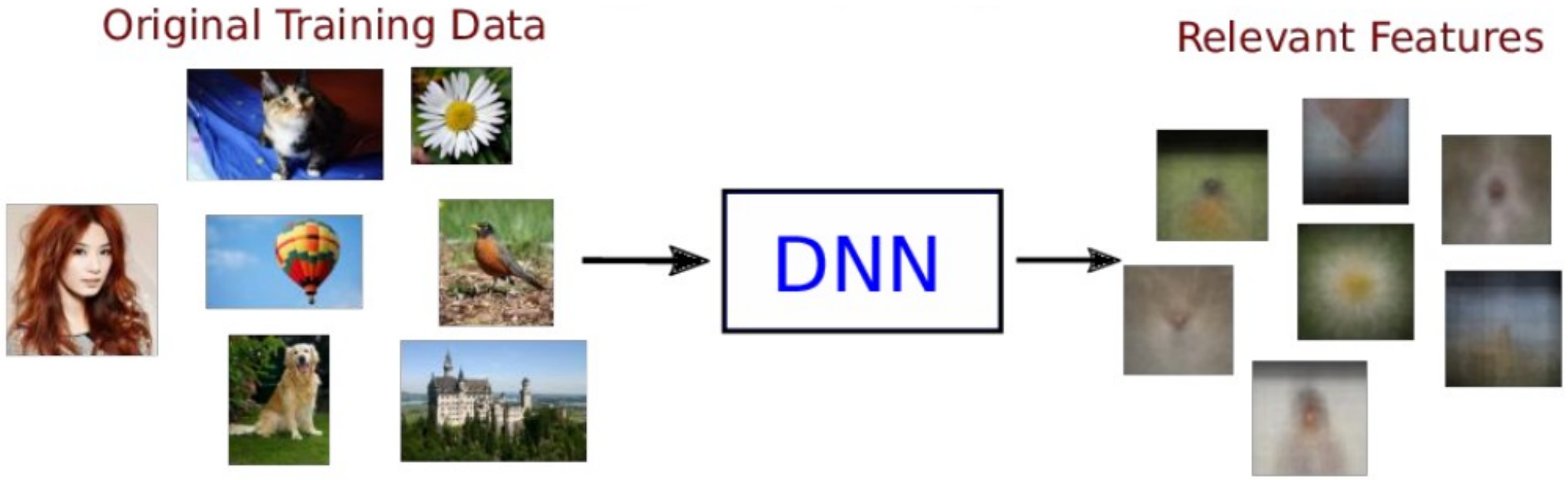
Characterizing Internal Network Activations [Oramas et al., 2019]



- **Assumption:** there exist a subset of features that can serve as indicators of the representation encoded by the model.

Model Interpretation

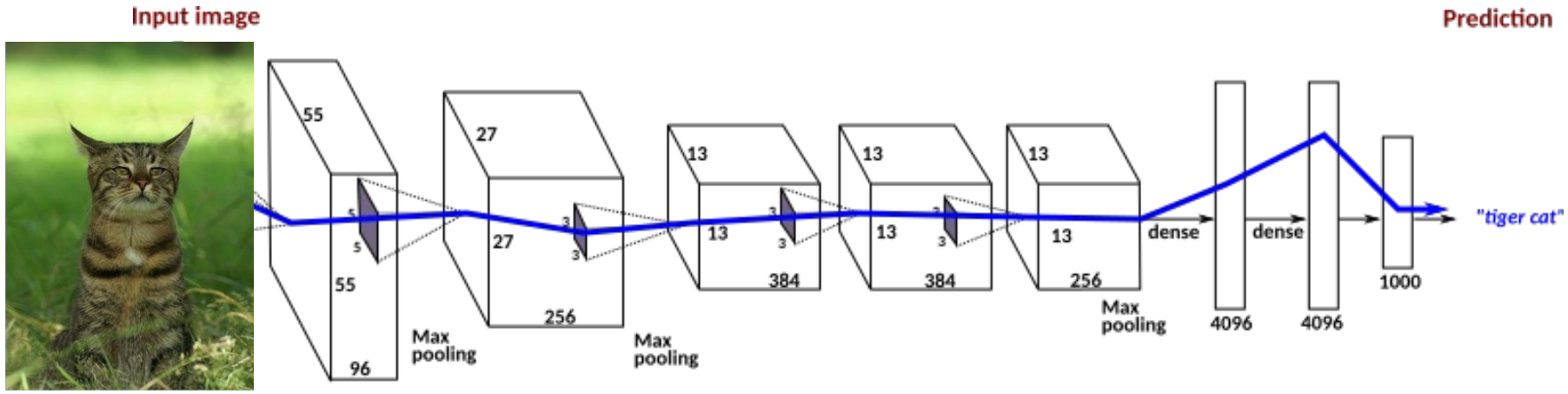
Characterizing Internal Network Activations [Oramas et al., 2019]



- **Assumption:** there exist a subset of features that can serve as indicators of the representation encoded by the model.
- Provide visualizations of these features as means of model interpretation.

Model Interpretation

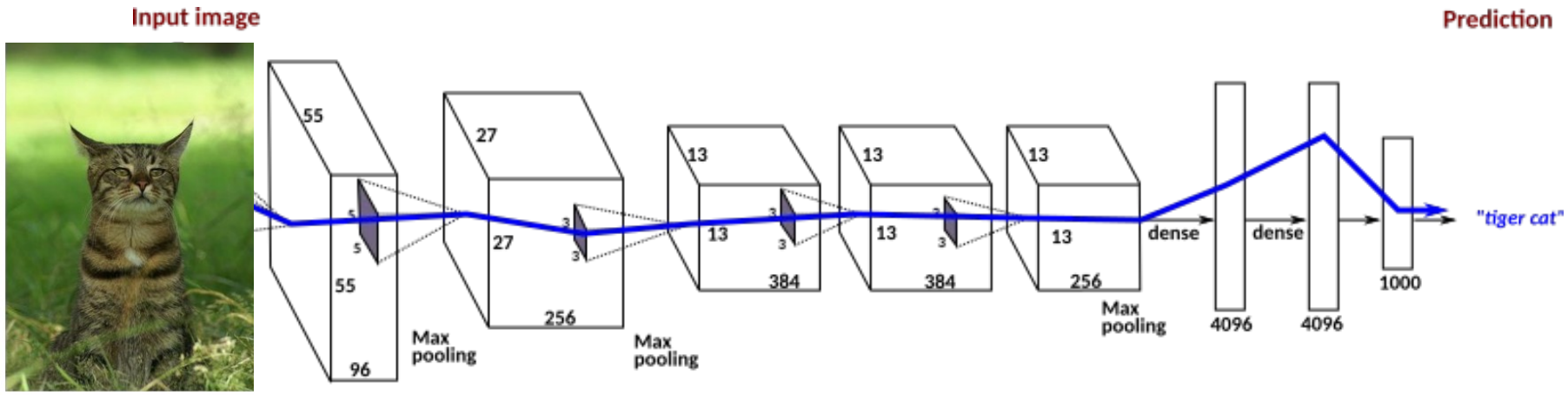
Characterizing Internal Network Activations [Oramas et al., 2019]



- Identifying relevant features (inspired by Escorcia et al., 2015)

Model Interpretation

Characterizing Internal Network Activations [Oramas et al., 2019]

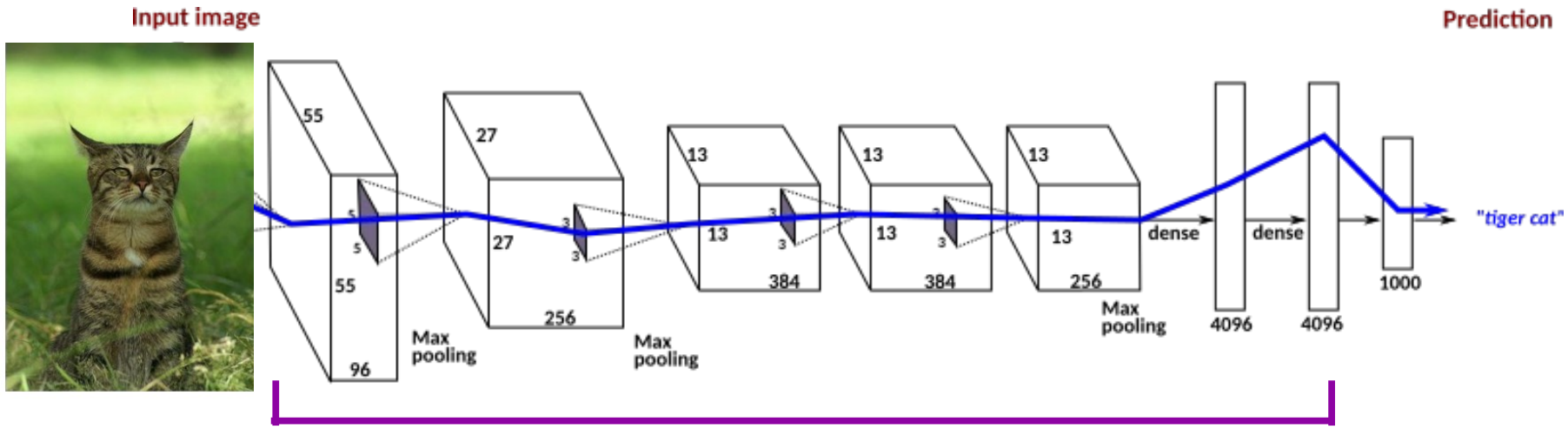


$$W^* = argmin_W \, ||X^T W - L^T||_F^2$$

- Identifying relevant features (inspired by Escorcia et al., 2015)

Model Interpretation

Characterizing Internal Network Activations [Oramas et al., 2019]

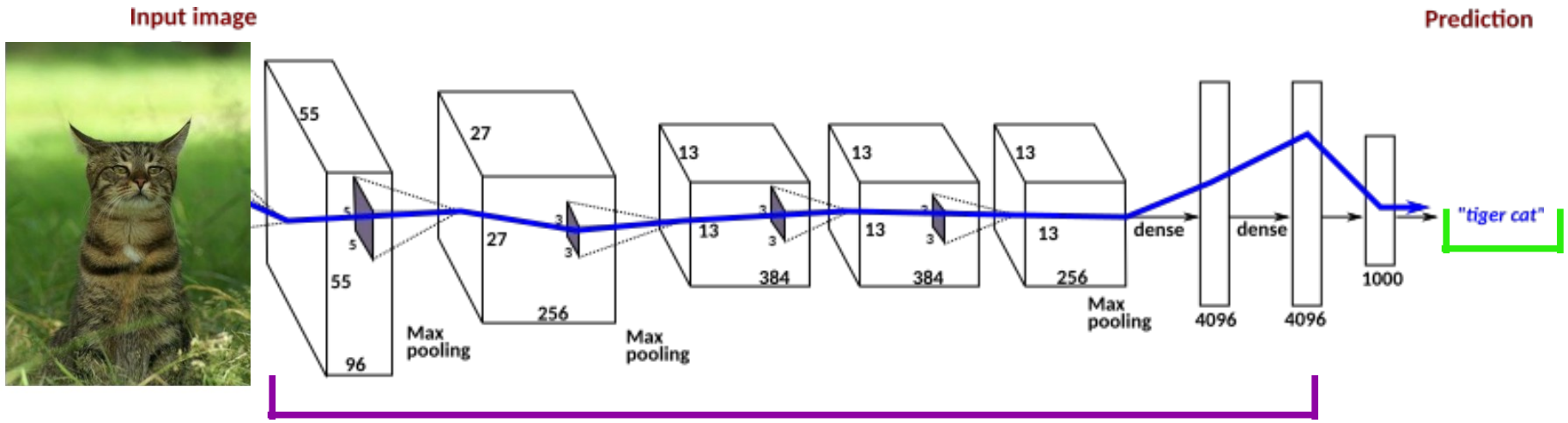


$$W^* = argmin_W \, ||X^T W - L^T||_F^2$$

- Identifying relevant features (inspired by Escorcia et al., 2015)

Model Interpretation

Characterizing Internal Network Activations [Oramas et al., 2019]

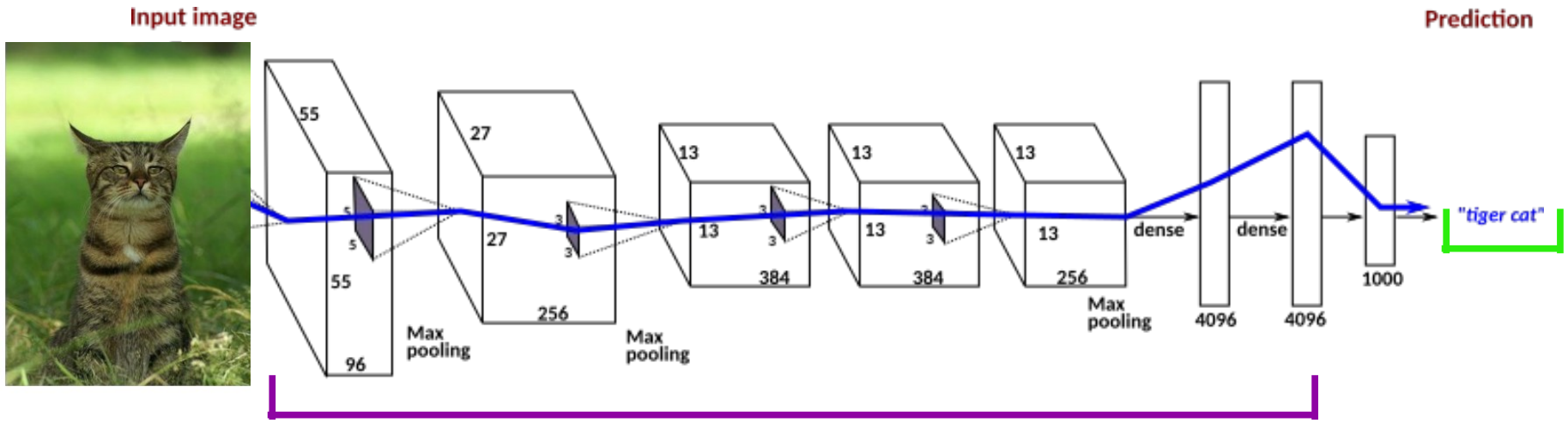


$$W^* = argmin_W \, ||X^T W - L^T||_F^2$$

- Identifying relevant features (inspired by Escorcia et al., 2015)

Model Interpretation

Characterizing Internal Network Activations [Oramas et al., 2019]

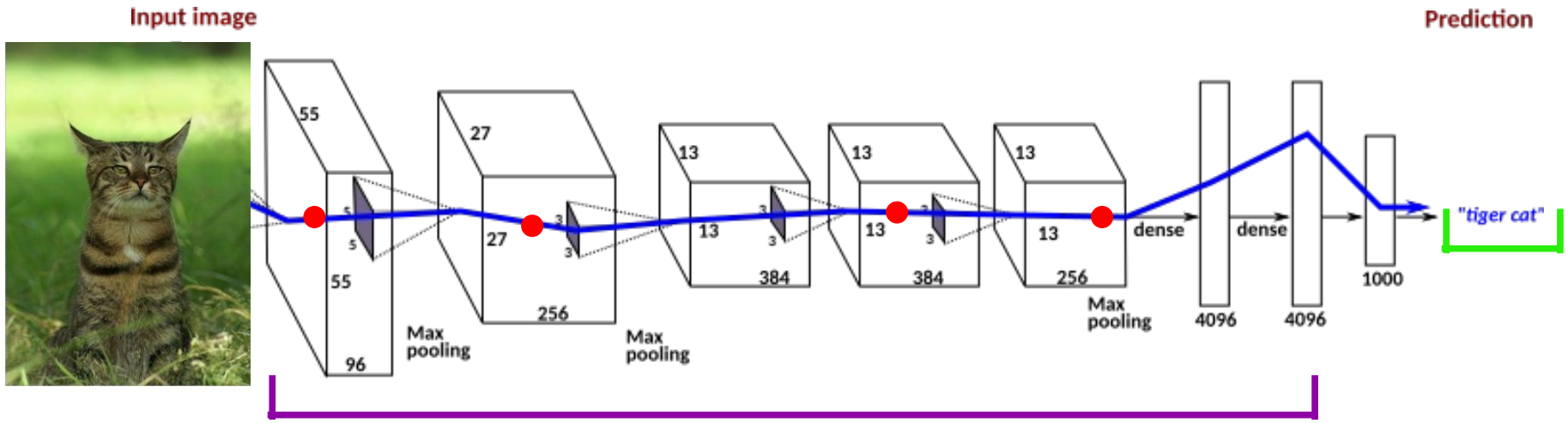


$$W^* = argmin_W \; ||X^T W - L^T||_F^2$$

- Identifying relevant features (inspired by Escorcia et al., 2015)

Model Interpretation

Characterizing Internal Network Activations [Oramas et al., 2019]



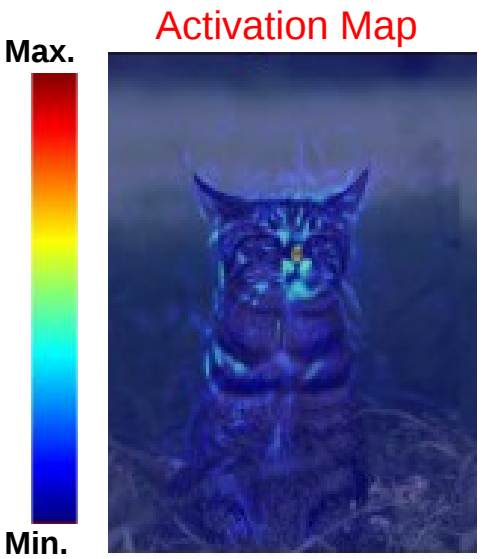
$$W^* = argmin_W \; ||X^T W - L^T||_F^2$$

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

- Identifying relevant features (inspired by Escorcia et al., 2015)

Model Interpretation

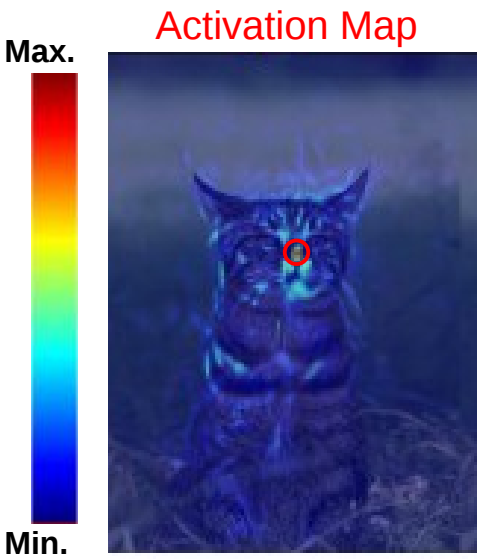
Characterizing Internal Network Activations [Oramas et al., 2019]



- Generating Intepretation Visualizations (inspired by Rematas et al., 2015)

Model Interpretation

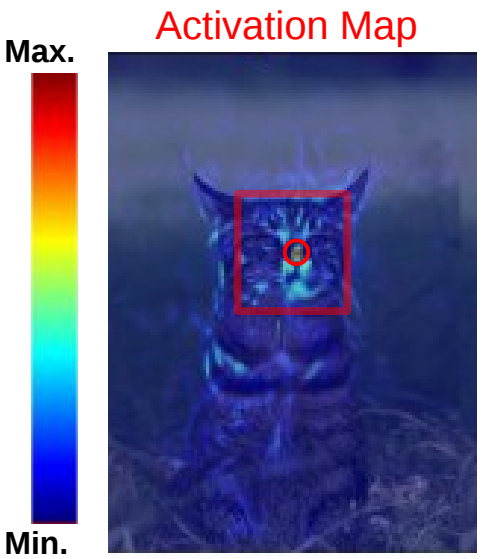
Characterizing Internal Network Activations [Oramas et al., 2019]



- Generating Intepretation Visualizations (inspired by Rematas et al., 2015)

Model Interpretation

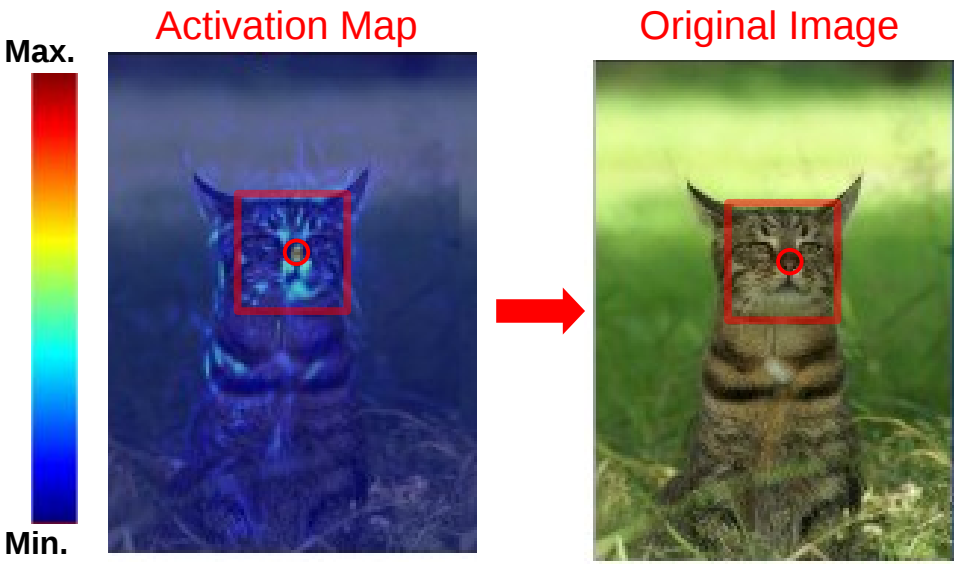
Characterizing Internal Network Activations [Oramas et al., 2019]



- Generating Intepretation Visualizations (inspired by Rematas et al., 2015)

Model Interpretation

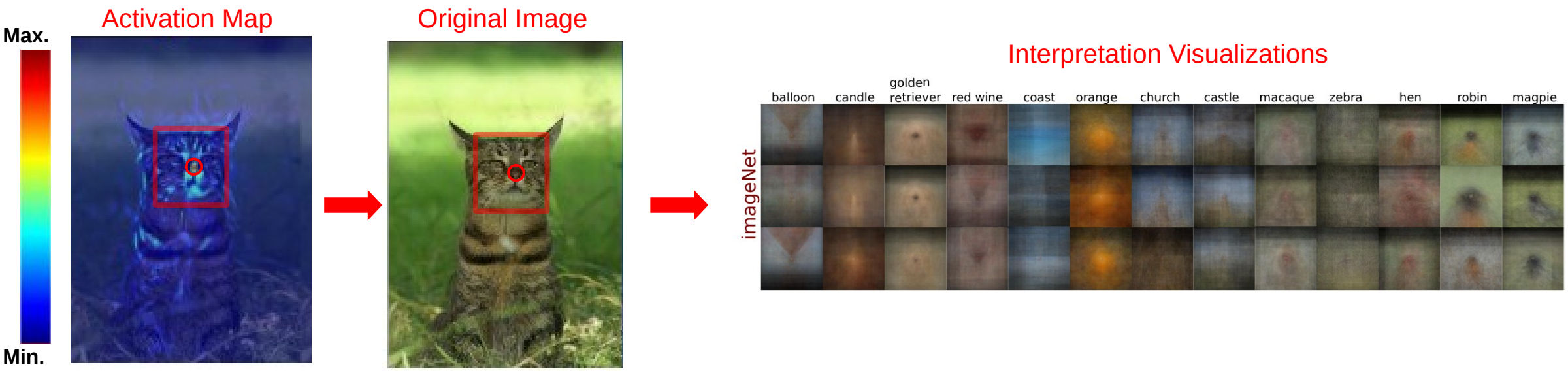
Characterizing Internal Network Activations [Oramas et al., 2019]



- Generating Intepretation Visualizations (inspired by Rematas et al., 2015)

Model Interpretation

Characterizing Internal Network Activations [Oramas et al., 2019]

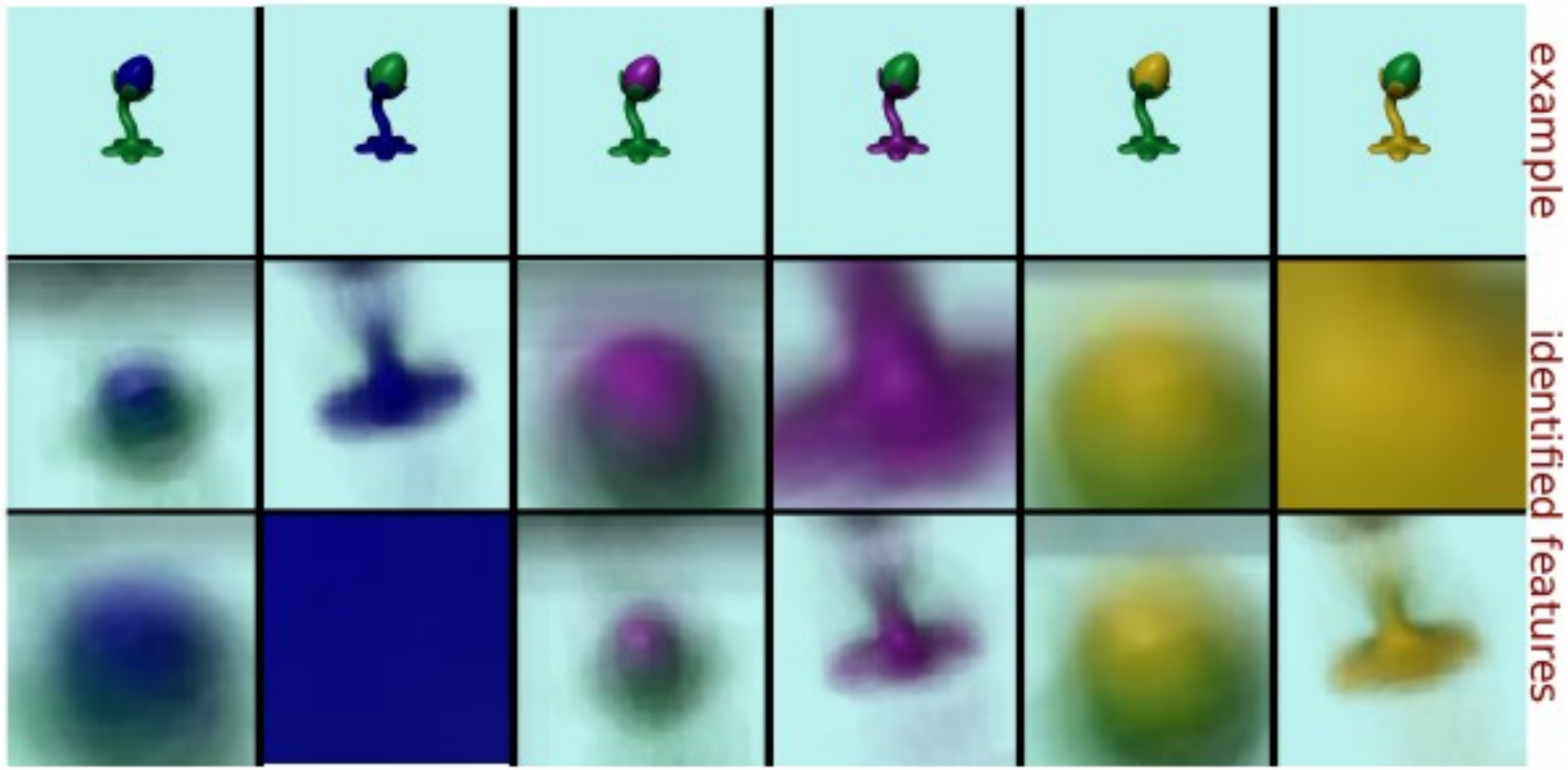


- Generating Intepretation Visualizations (inspired by Rematas et al., 2015)

Model Interpretation

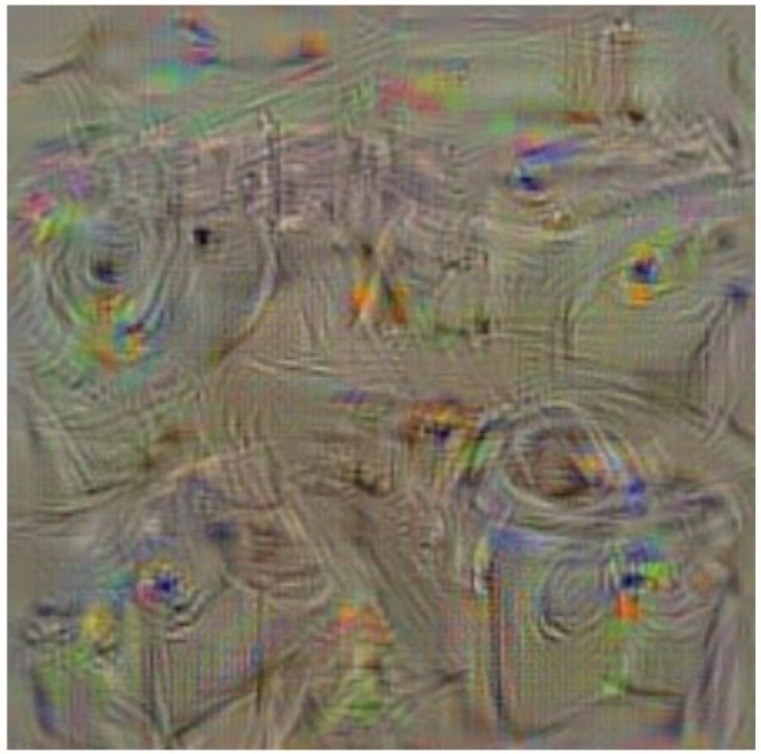
Characterizing Internal Network Activations [Oramas et al., 2019]

Interpretation Feedback

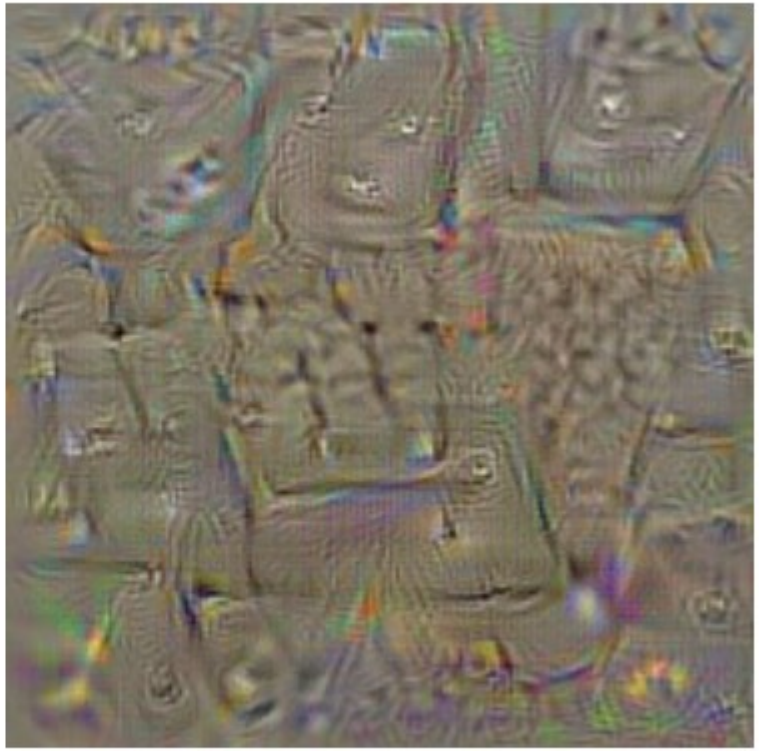


Model Interpretation

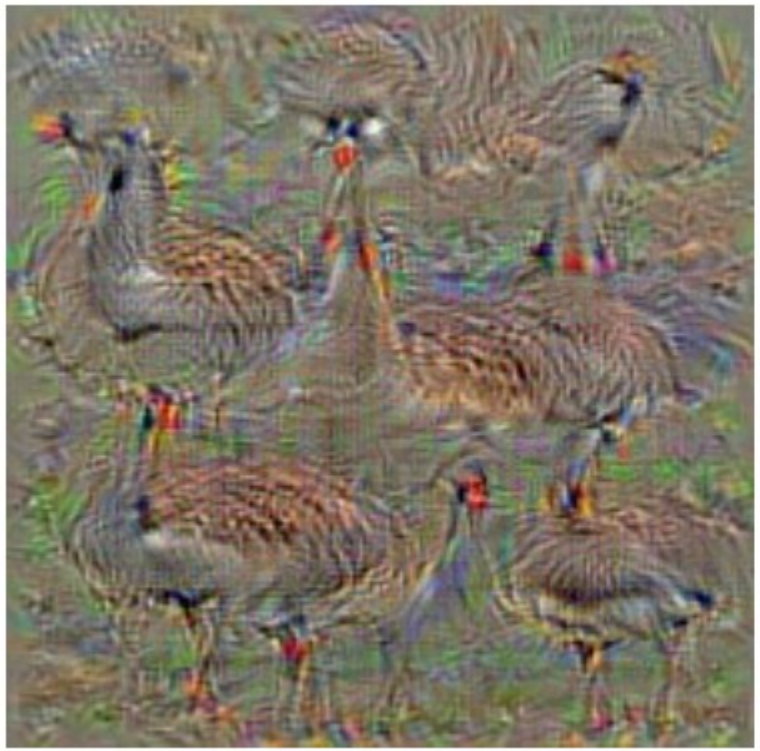
Input Reconstruction [Simonyan et al, 2013]



washing machine



computer keyboard

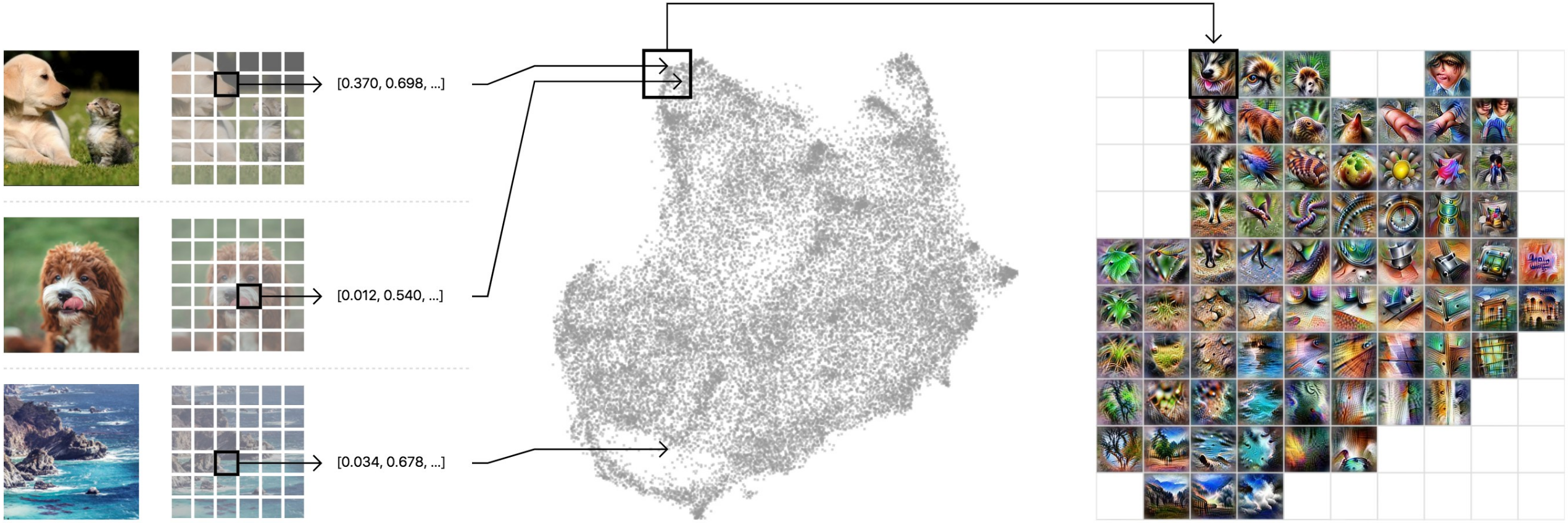


goose

- Maximizing the activations of **the output unit** by changing the input

Model Interpretation

Model Inversion – Activation Atlas [Olah & Schubert, 2019]



Info + demo: https://distill.pub/2019/activation-atlas/

Model Interpretation

Sparse Dictionary Learning via Autoencoders [Cunninghan et al., 2024]

- *Superposition* problem → more features than neurons

  → Model learned an overcomplete set of non-orthogonal features which activate sparsely

SAE:

$$\begin{aligned} \mathbf{c} &= \mathrm{ReLU}(M\mathbf{x}+\mathbf{b}) \\ \hat{\mathbf{x}} &= M^T\mathbf{c} \\ &= \sum_{i=0}^{d_{\text{hid}}-1} c_i \mathbf{f}_i \end{aligned}$$

Loss Function:

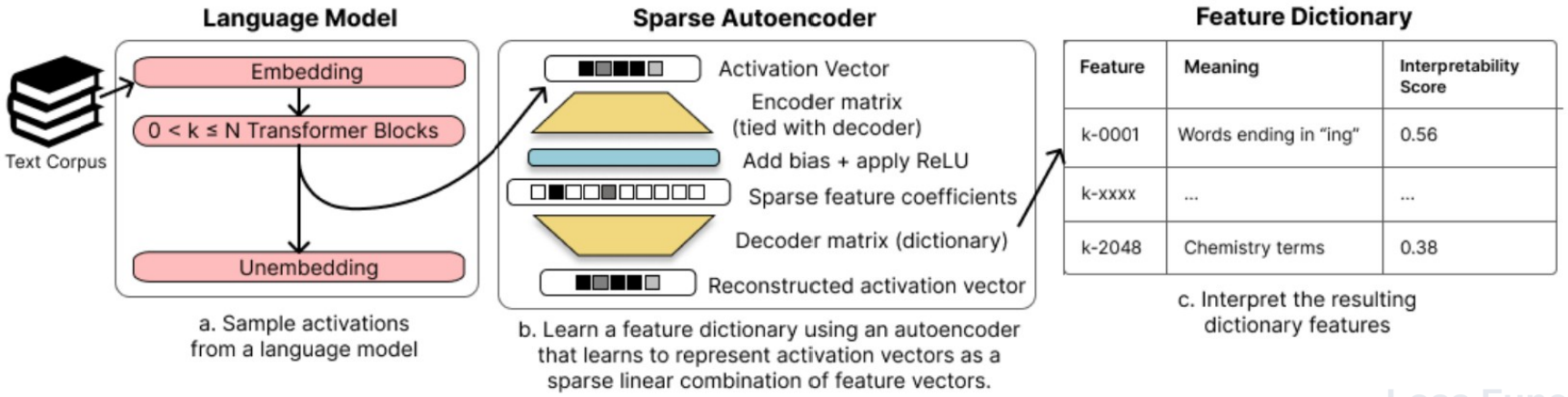
$$\mathcal{L}(\mathbf{x}) = \underbrace{\frac{||\mathbf{x}-\hat{\mathbf{x}}||_2^2}{\dim(\mathbf{x})}}_{\text{Reconstruction loss}} + \underbrace{\alpha||\mathbf{c}||_1}_{\text{Sparsity loss}}$$

Model Interpretation

Sparse Dictionary Learning via Autoencoders [Cunninghan et al., 2024]

- *Superposition* problem → more features than neurons

  → Model learned an overcomplete set of non-orthogonal features which activate sparsely



SAE:

$$\begin{aligned} \mathbf{c} &= \mathrm{ReLU}(M\mathbf{x}+\mathbf{b}) \\ \hat{\mathbf{x}} &= M^T\mathbf{c} \\ &= \sum_{i=0}^{d_{\text{hid}}-1} c_i \mathbf{f}_i \end{aligned}$$

Loss Function:

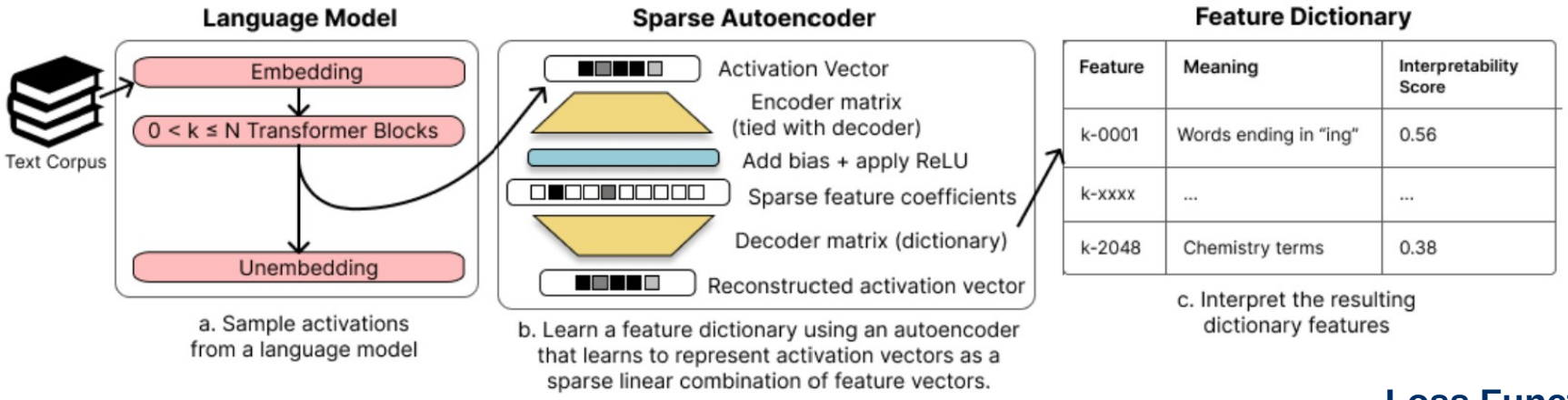
$$\mathcal{L}(\mathbf{x}) = \underbrace{\frac{||\mathbf{x}-\hat{\mathbf{x}}||_2^2}{\dim(\mathbf{x})}}_{\text{Reconstruction loss}} + \underbrace{\alpha||\mathbf{c}||_1}_{\text{Sparsity loss}}$$

Model Interpretation

Sparse Dictionary Learning via Autoencoders [Cunninghan et al., 2024]

- *Superposition* problem → more features than neurons

  → Model learned an overcomplete set of non-orthogonal features which activate sparsely



SAE:

$$\begin{aligned} \mathbf{c} &= \mathrm{ReLU}(M\mathbf{x}+\mathbf{b}) \\ \hat{\mathbf{x}} &= M^T\mathbf{c} \\ &= \sum_{i=0}^{d_{\text{hid}}-1} c_i \mathbf{f}_i \end{aligned}$$

Loss Function:

$$\mathcal{L}(\mathbf{x}) = \underbrace{\frac{||\mathbf{x}-\hat{\mathbf{x}}||_2^2}{\dim(\mathbf{x})}}_{\text{Reconstruction loss}} + \underbrace{\alpha||\mathbf{c}||_1}_{\text{Sparsity loss}}$$

Questions?

Model Explanation

[Justifying the predictions made by the model]

Model Explanation

Visual Explanations

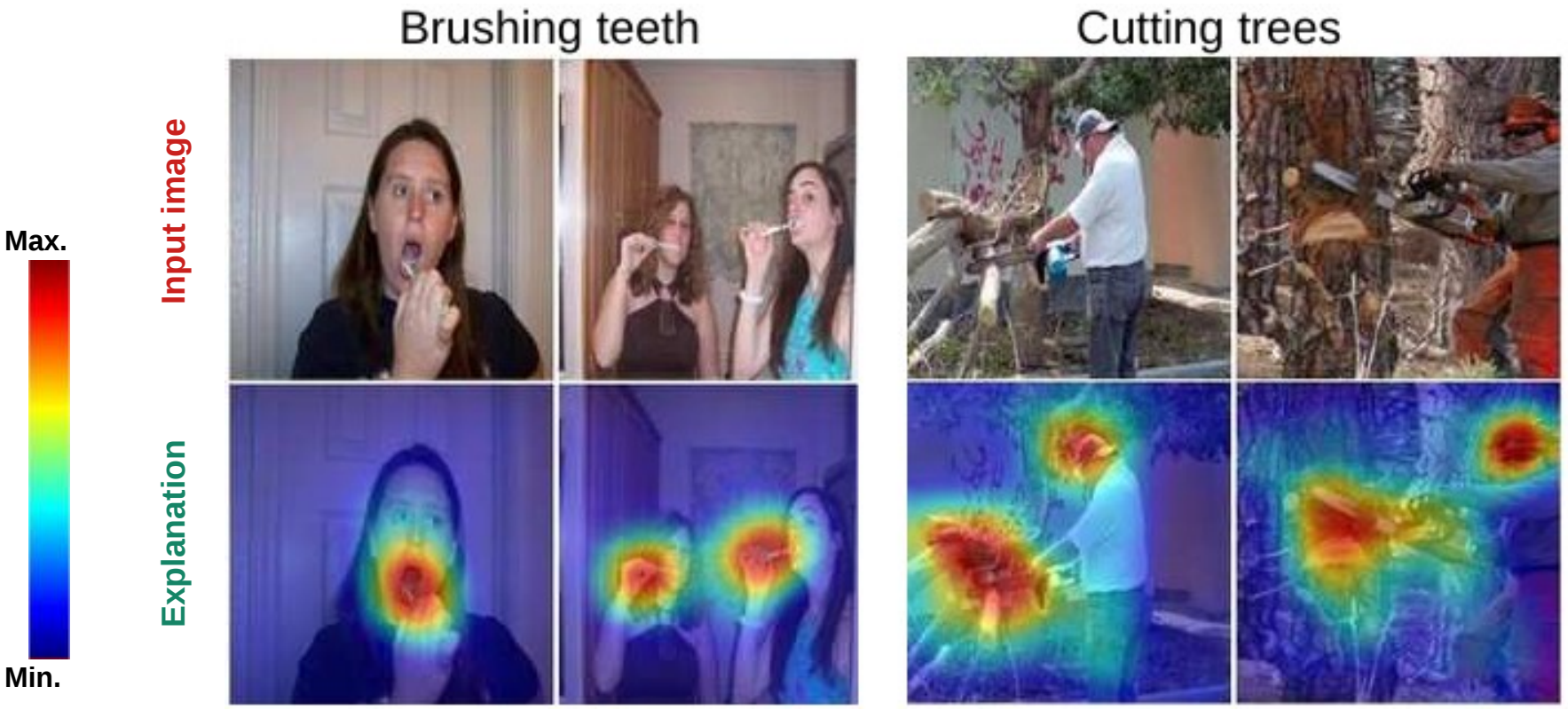
Input image

Brushing teeth

Cutting trees

Model Explanation

Visual Explanations

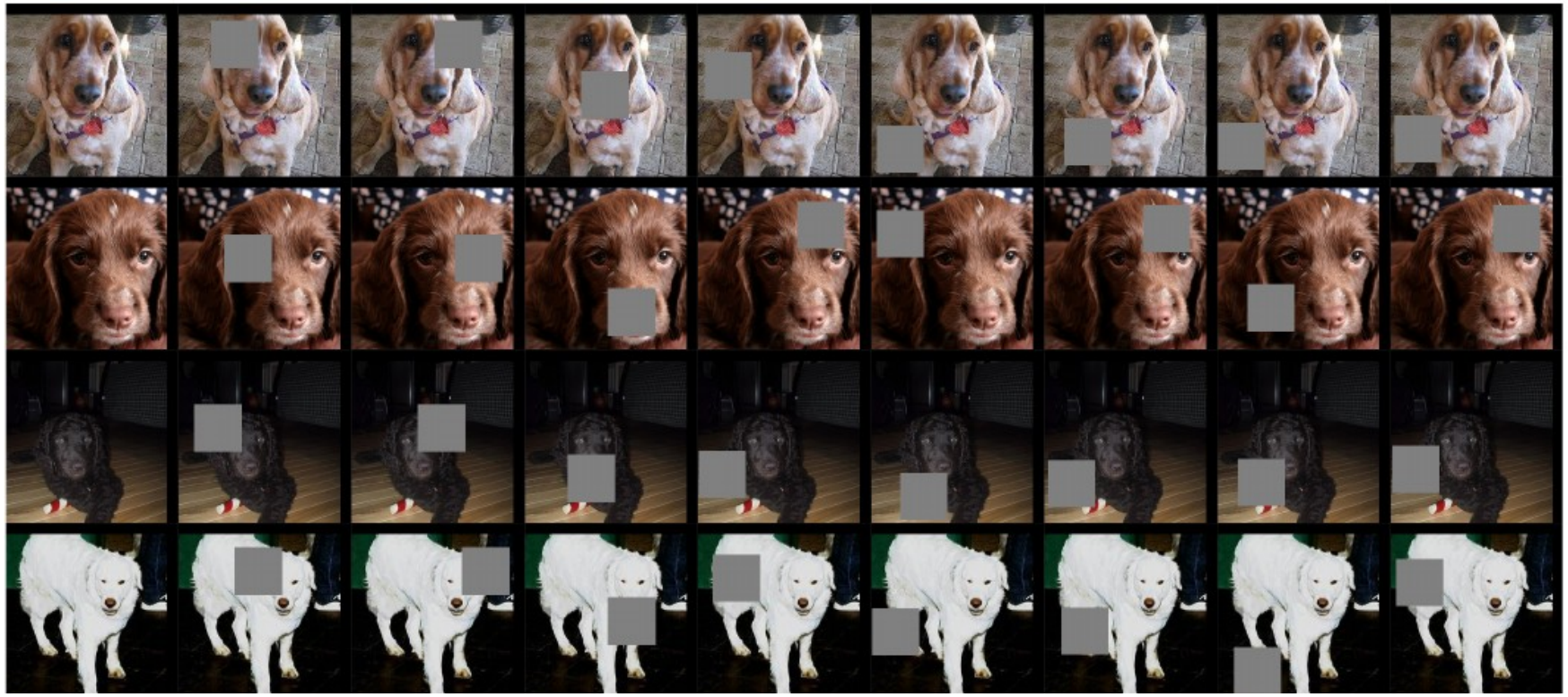


[Zhou et al.,2016]

- Heatmap visualization highlighting parts of the input supporting the prediction

Model Explanation

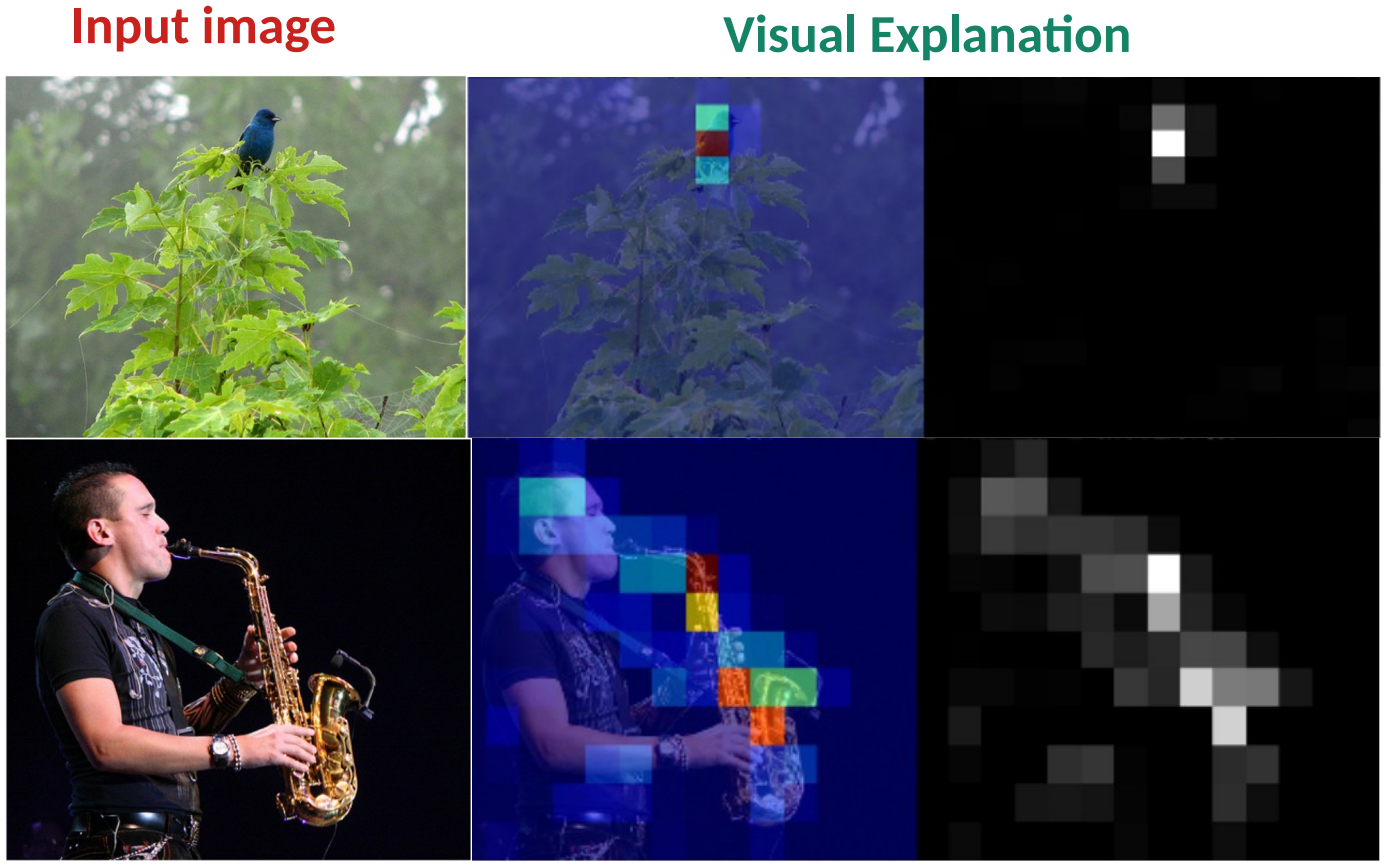
Input Modification Methods



- **Goal:** Identify relevant regions from the input.
- **Idea:** Modify the input, then measure changes in the output.

Model Explanation

Input Modification Methods



[Grun et al.,2016]

- Color-codes in the visual explanation indicate the effect regions has on the prediction (when modified)

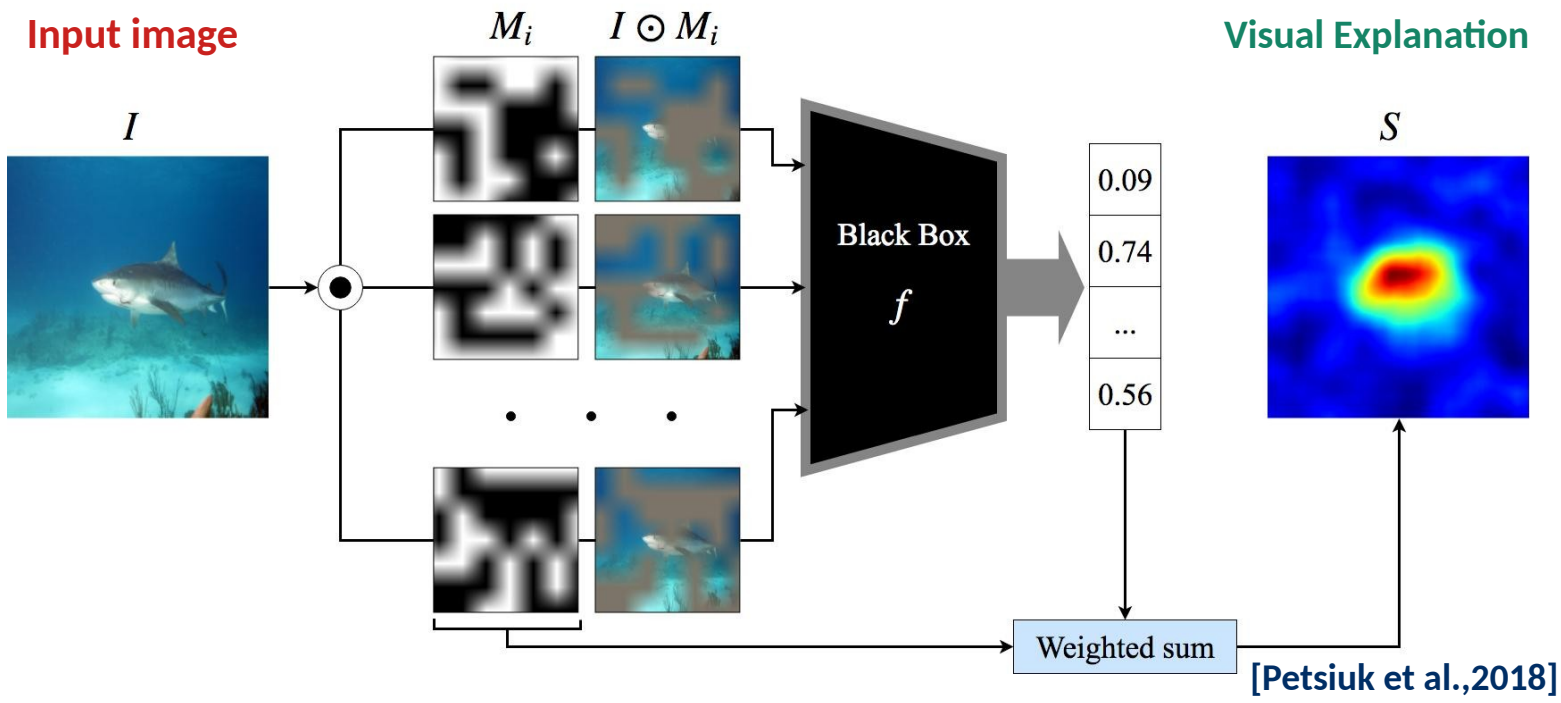
Related Work

- Zeiler et al., 2011.
- Zhou et al., 2015.
- Fong & Vedaldi, 2017.

Model Explanation

Input Modification Methods

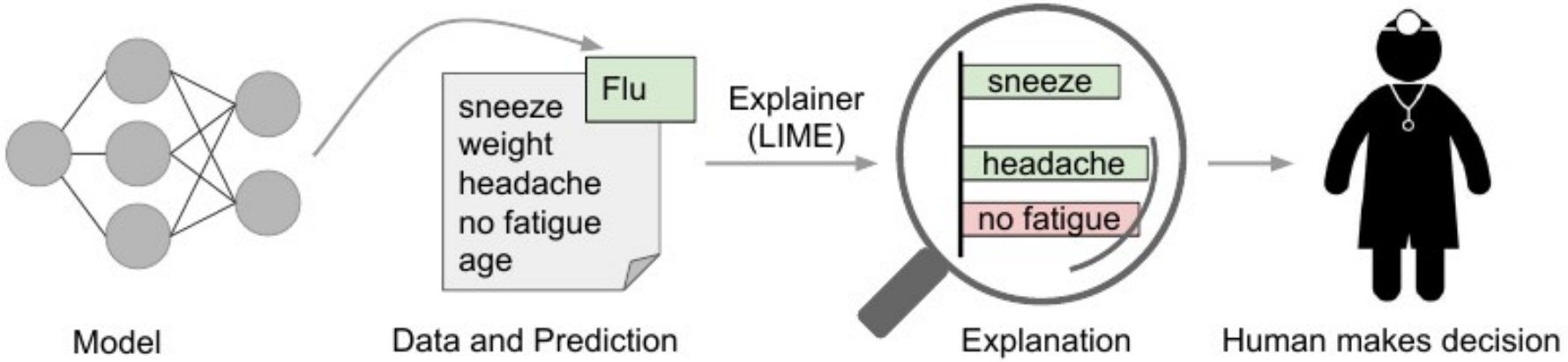
– RISE: Randomized Input Sampling for Explanation [Pesiuk et al., 2018]



- Generate random masks M_i
- Mask the input $I \rightarrow (I \odot M_i)$
- Evaluate the masked input
- Aggregate the masks

Model Explanation

Surrogate Methods - LIME [Ribeiro et al., 2016]

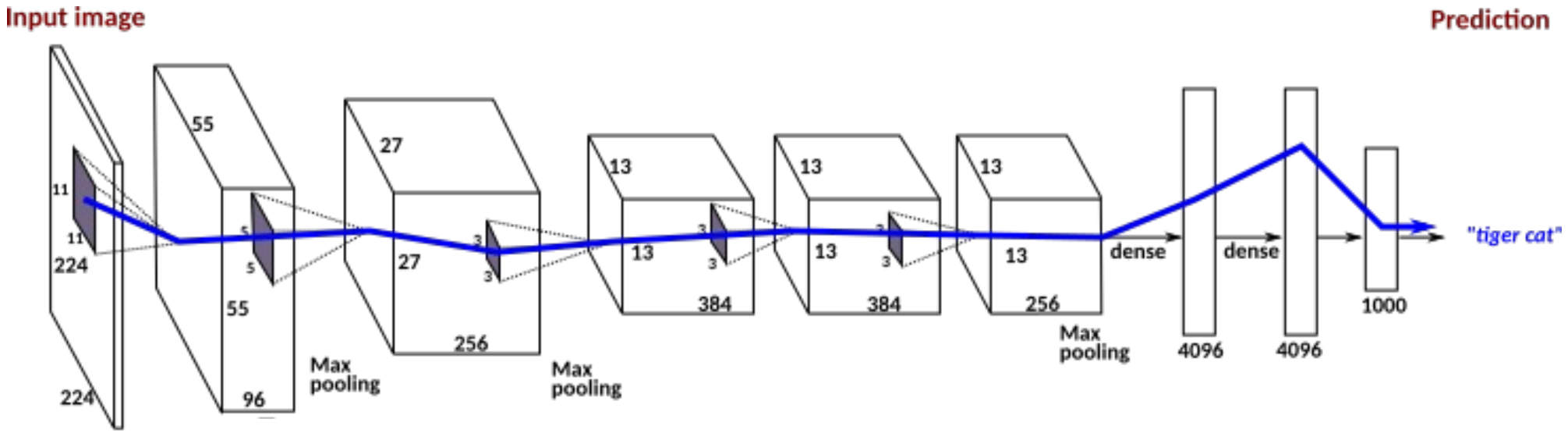


[Ribeiro et al.,2016]

- Train a simpler interpretable model
- Collect [perturbation, output] pairs
- Train a linea classifier or model of interest.

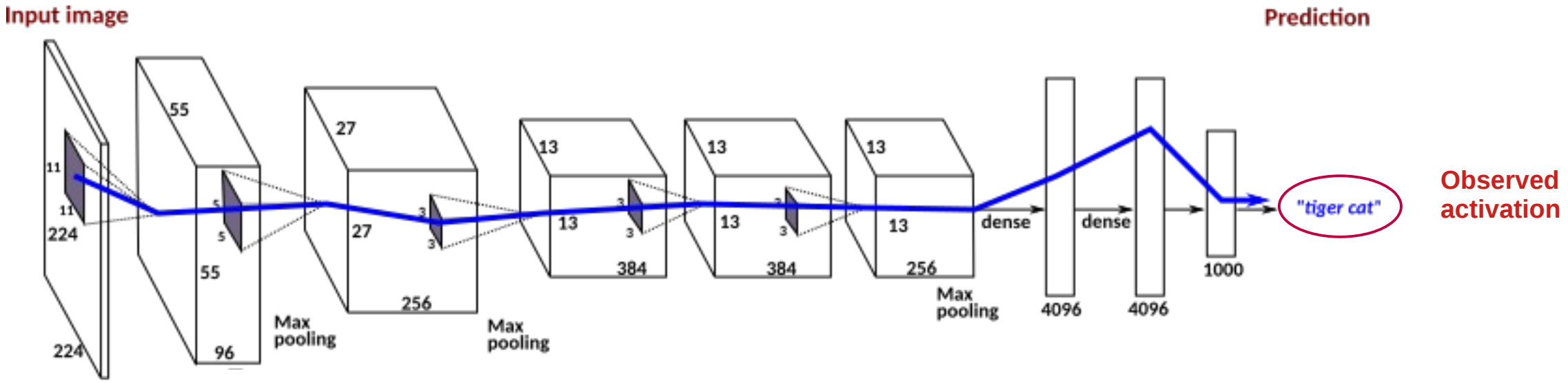
Model Explanation

Deconvolution-based Methods



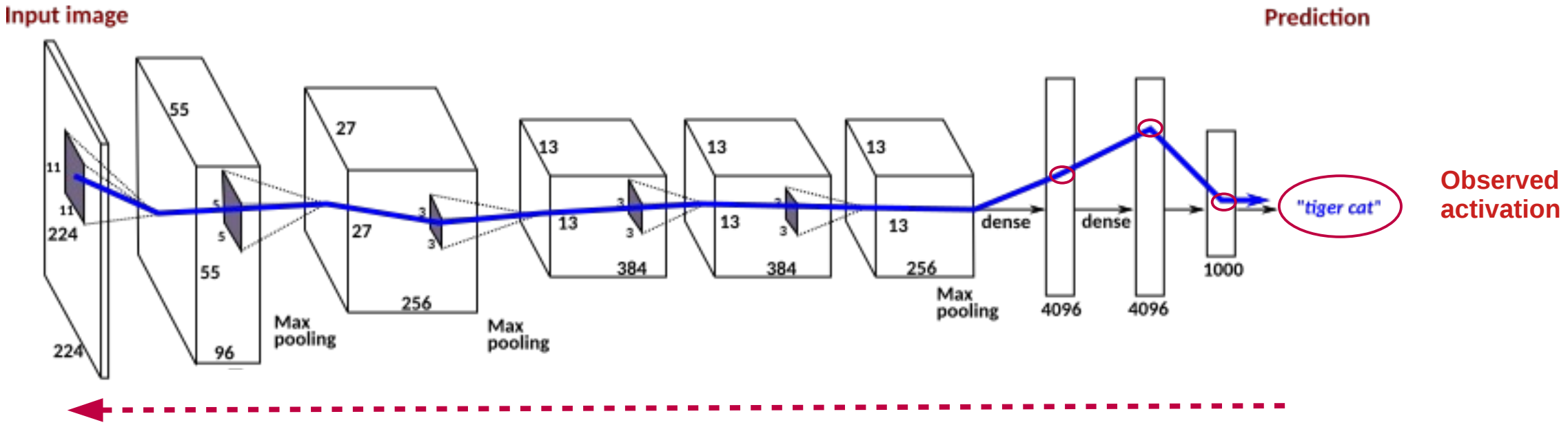
Model Explanation

Deconvolution-based Methods



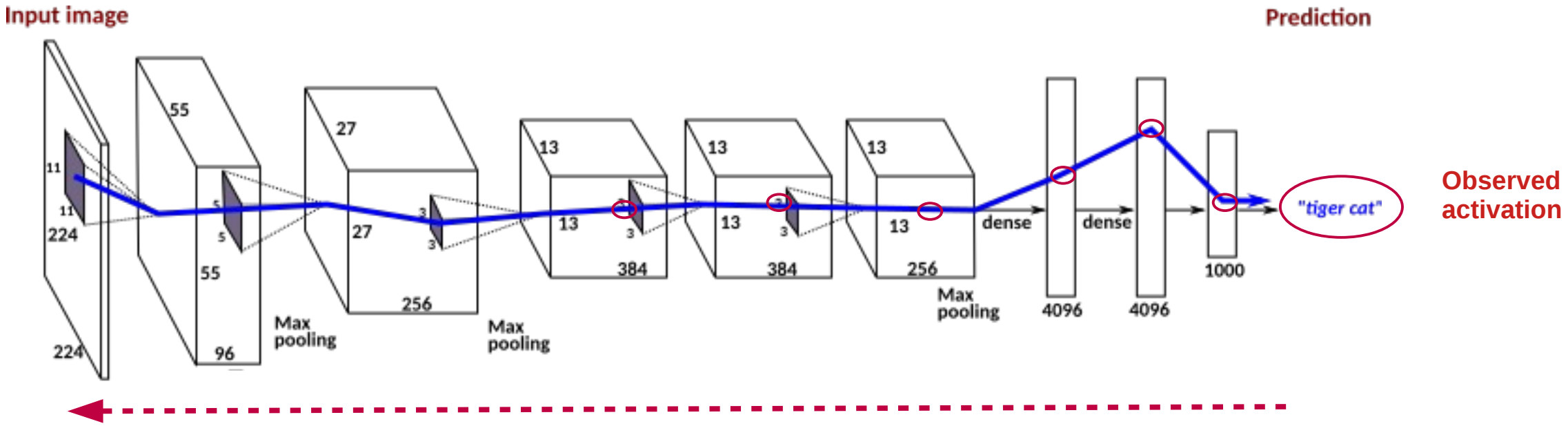
Model Explanation

Deconvolution-based Methods



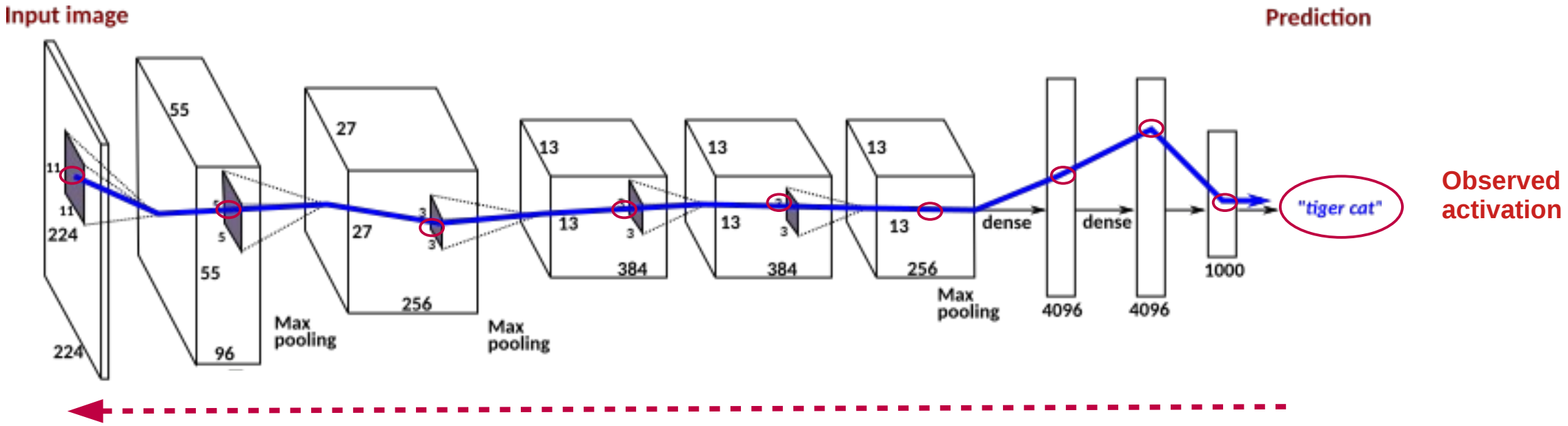
Model Explanation

Deconvolution-based Methods



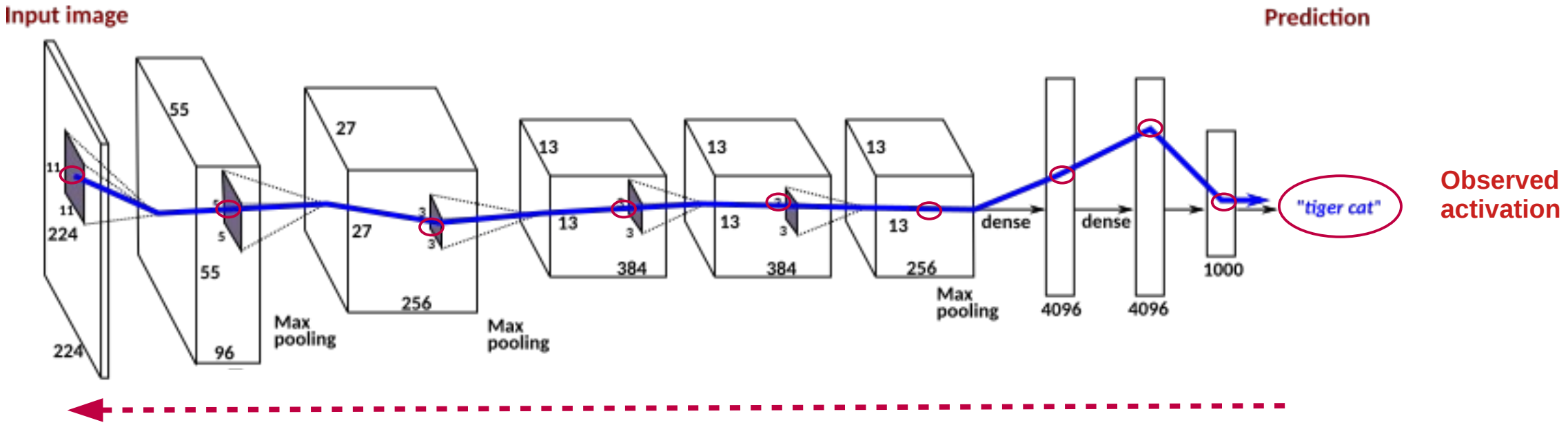
Model Explanation

Deconvolution-based Methods



Model Explanation

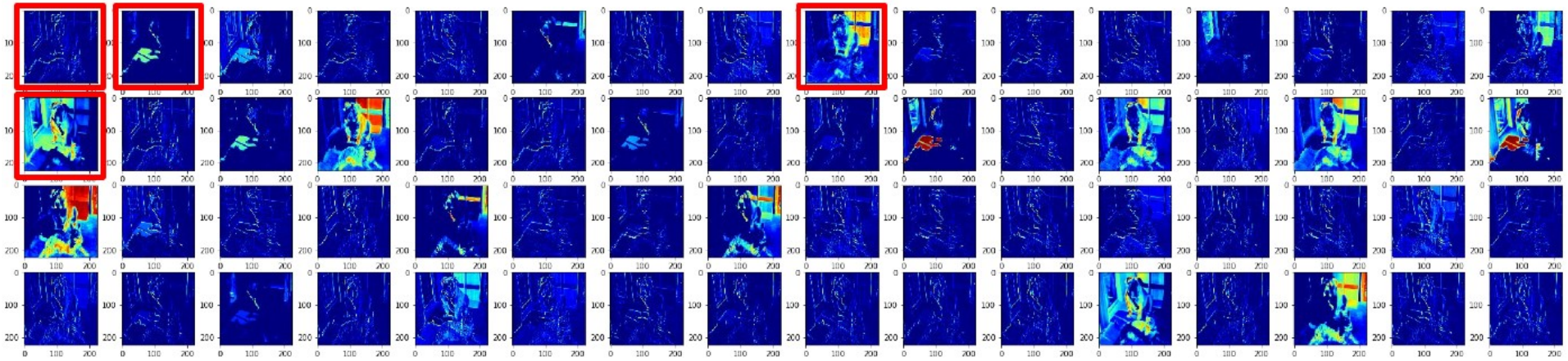
Deconvolution-based Methods

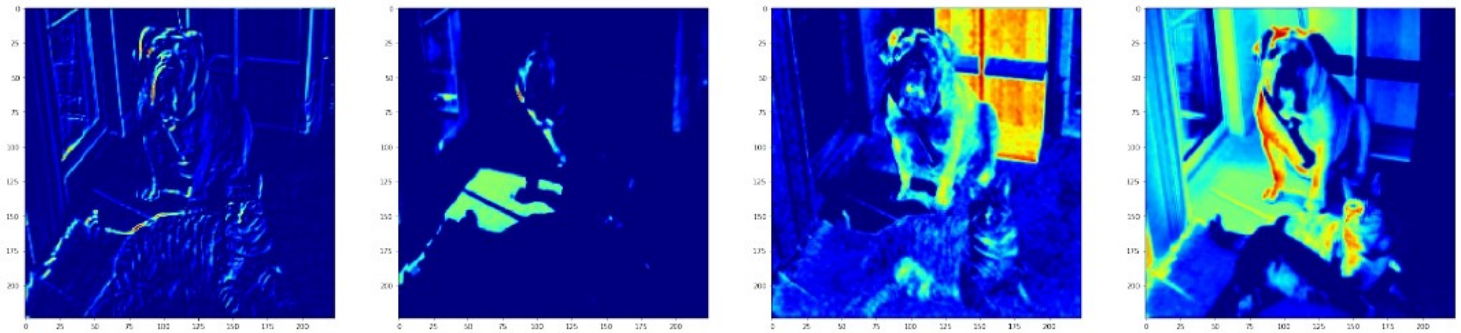


- **Goal:** Identify evidence from the input supporting an observed activation

Model Explanation

Deconvolution-based Methods



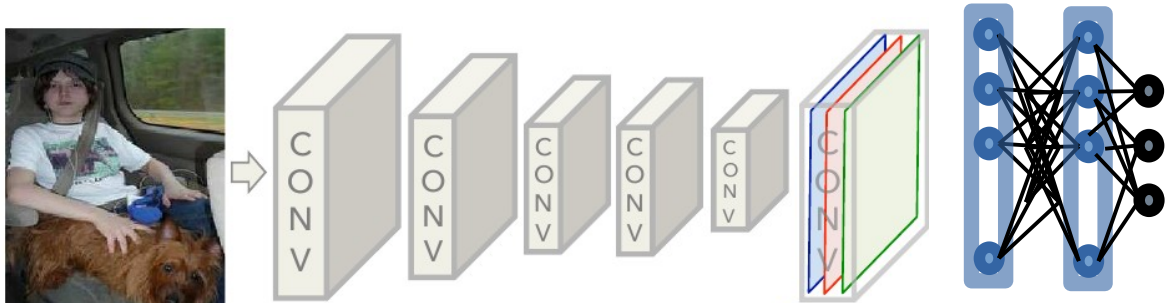


Related Work

- Zeiler et al., 2014.
- Springenberg et al., 2015
- Grun et al., 2016.

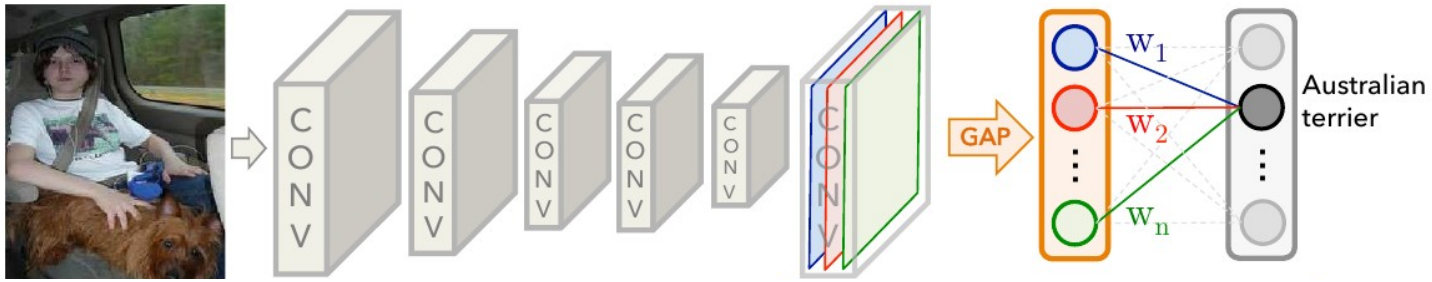
Model Explanation

Class Activation Mapping (CAM) [Zhou et al., 2016]



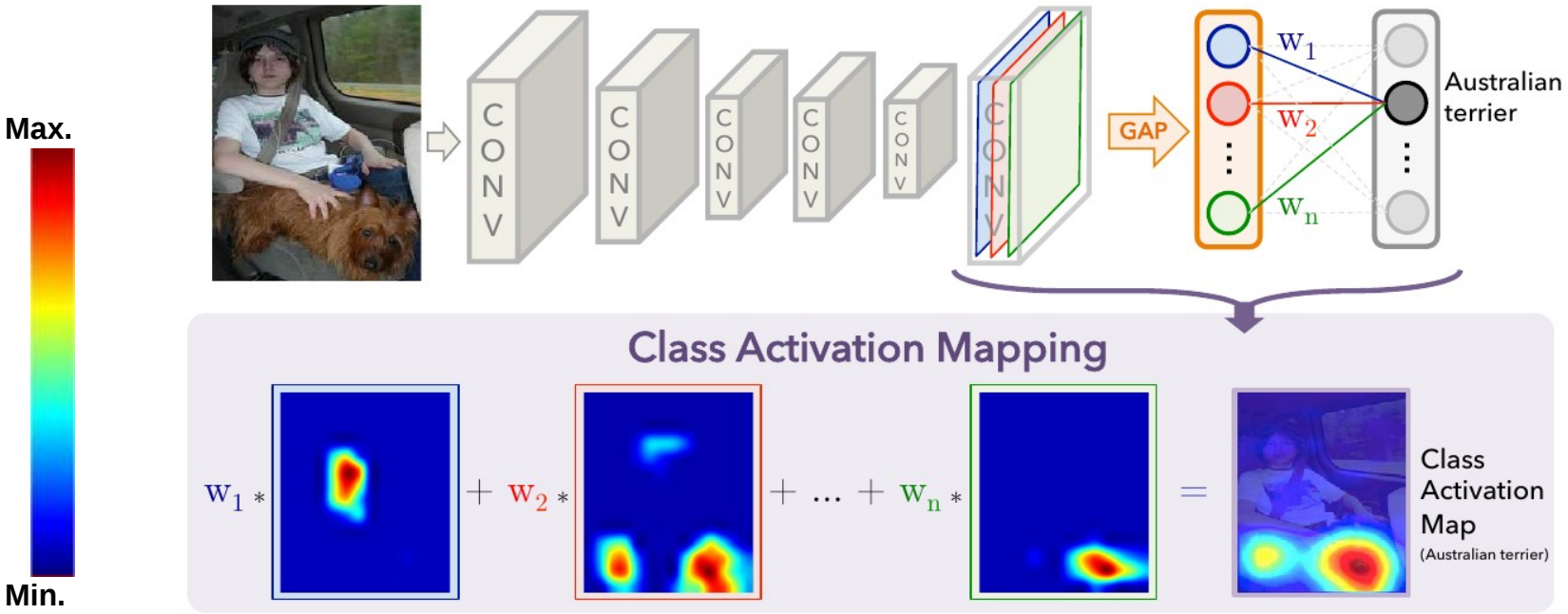
Model Explanation

Class Activation Mapping (CAM) [Zhou et al., 2016]



Model Explanation

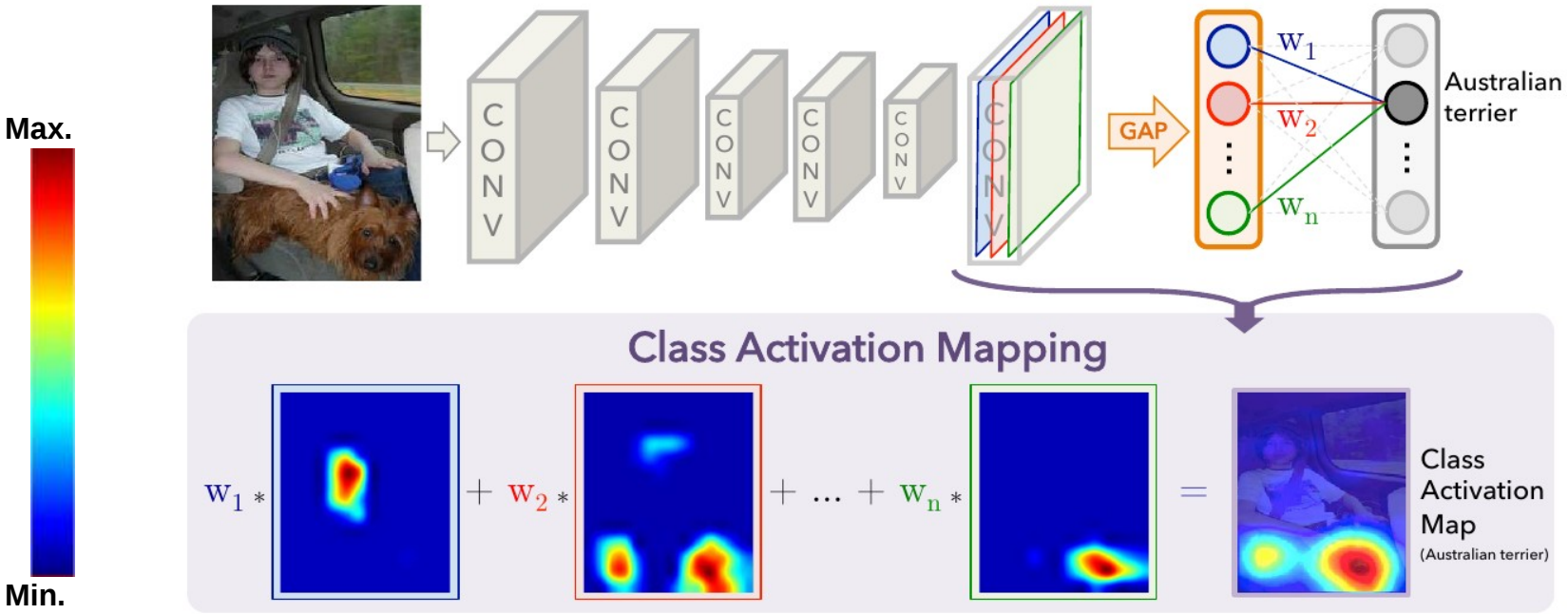
Class Activation Mapping (CAM) [Zhou et al., 2016]



- Linear combination of channel-wise activations of the last convolutional network
- Wi terms are learned by modifying the original architecture

Model Explanation

Class Activation Mapping (CAM) [Zhou et al., 2016]



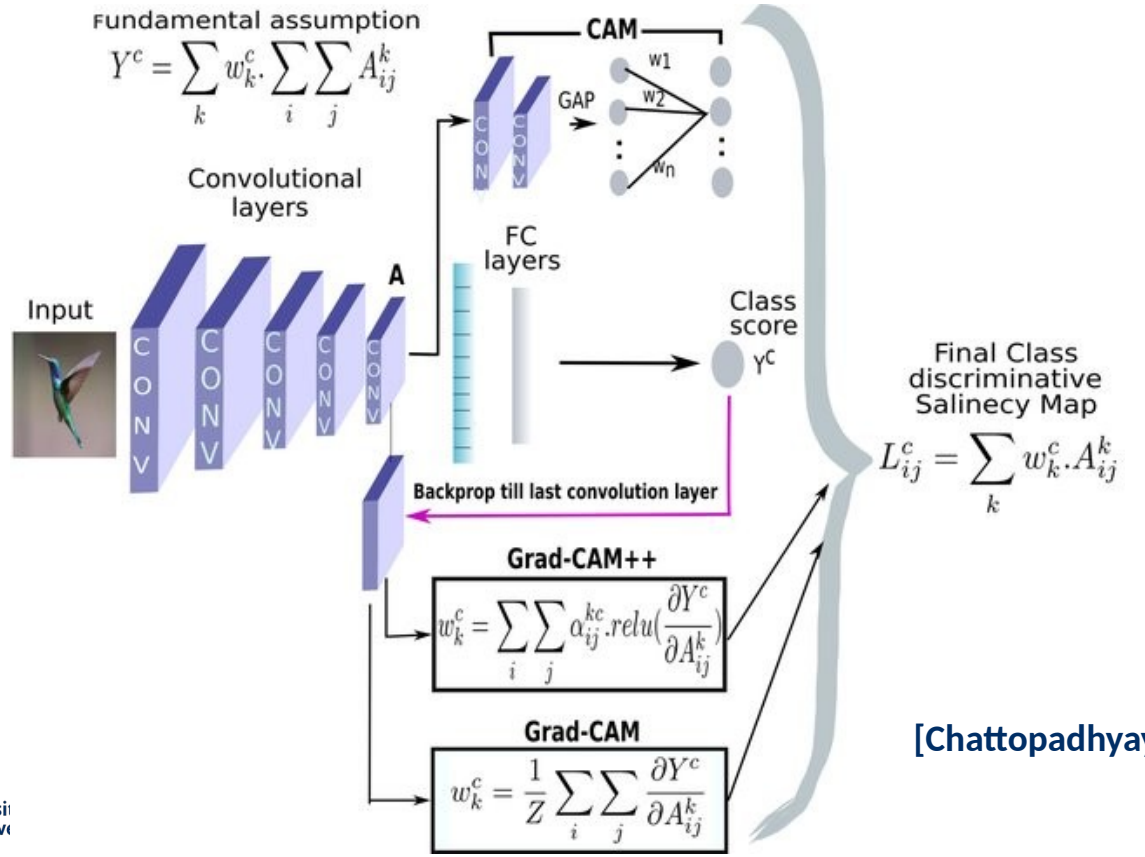
- Linear combination of channel-wise activations of the last convolutional network
- Wi terms are learned by modifying the original architecture

Q: Any potential problem or limitation?

Model Explanation

Class Activation Mapping (CAM) [Zhou et al., 2016]

- Existing alternatives for computing W_i



[Chattopadhyay et al., 2018.]

Related Work

- Zhou et al., 2016.
- Zhang et al., 2016.
- Selvaraju et al., 2017.
- Chattopadhyay et al., 2018.
- Zhang. et al., 2018.

Model Explanation

Maximizing unit activations wrt. to a given input

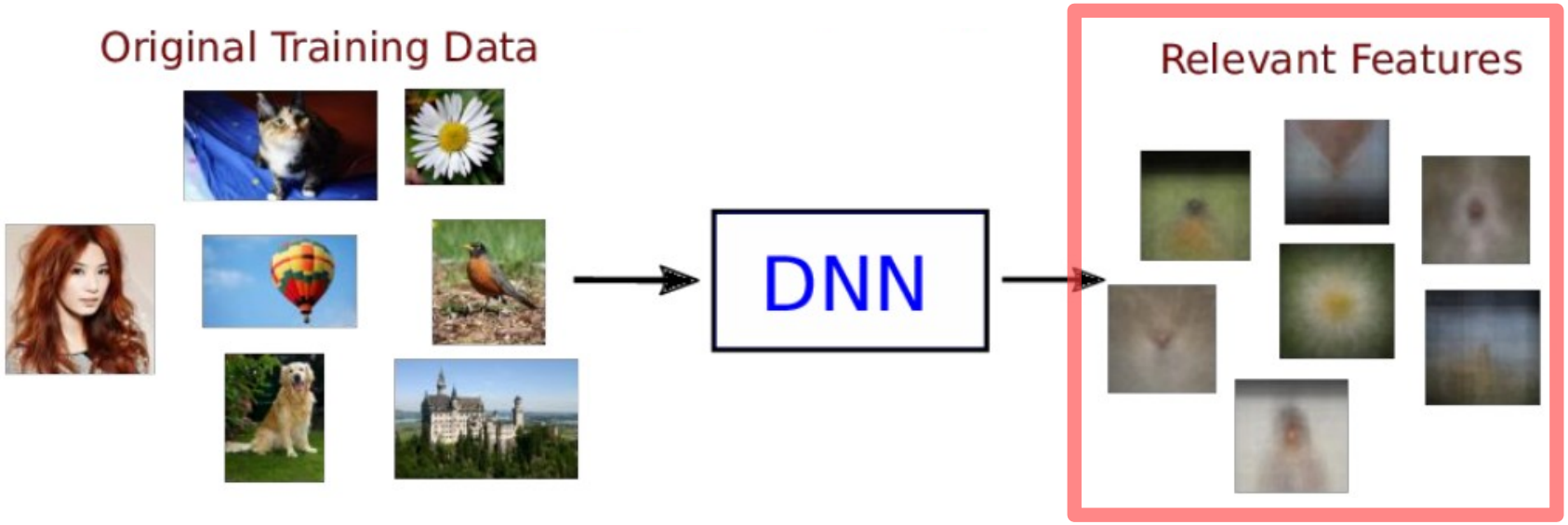
Model Explanation

Maximizing unit activations wrt. to a given input

- Maximizing the activations of **a given unit** by changing the input

Model Explanation

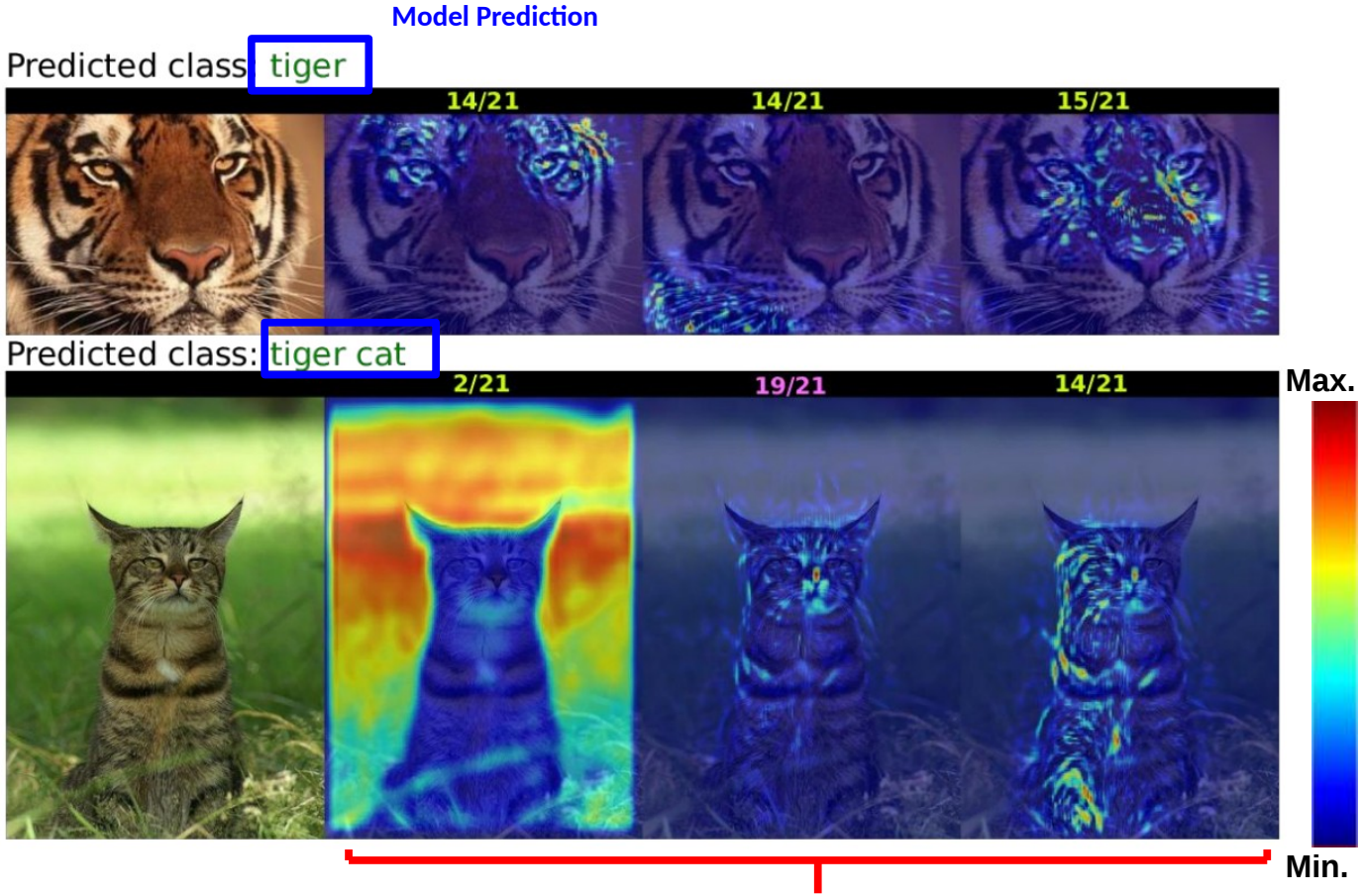
Explanation by Interpretation [Oramas et al., 2019]



- Given identified relevant units for a predicted class
- Explain by considering high-reponse units on a given input

Model Explanation

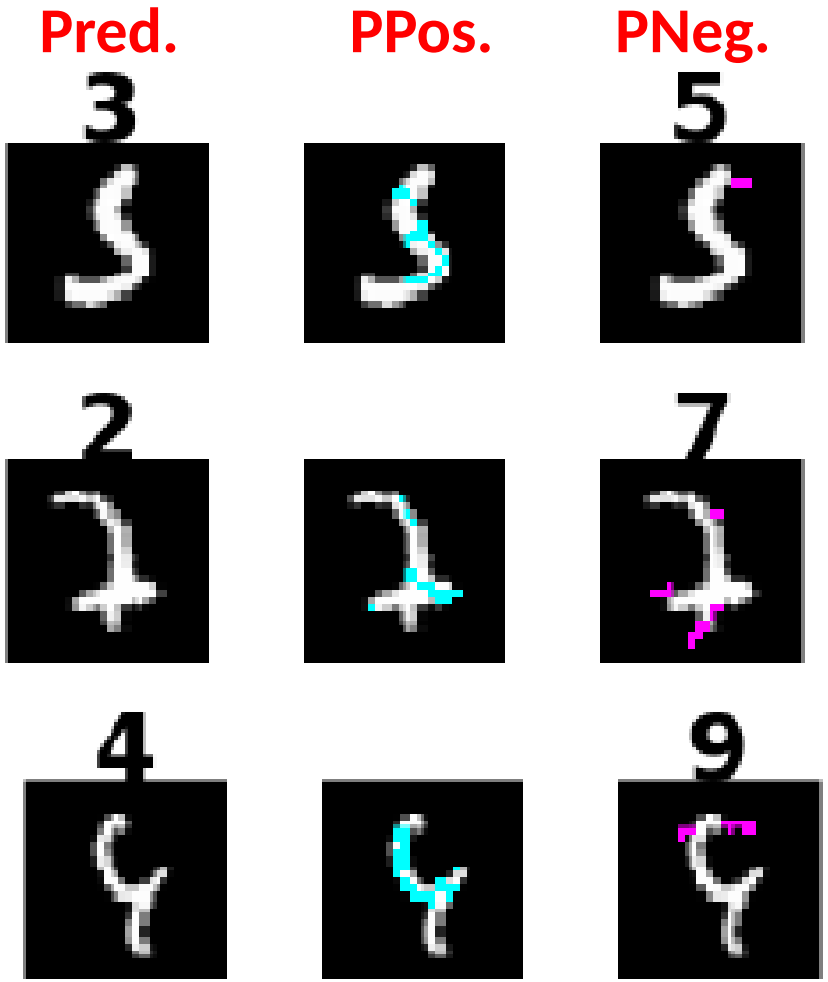
Explanation by Interpretation [Oramas et al., 2019]



- Given identified relevant units for a predicted class
- Explain by considering high-reponse units on a given input

Model Explanation

Contrastive/Counterfactual Explanations [Dhurandhar et al., 2018]



- Why A and not B?
- Highlight aspects that are in favor (class A) or against (class B) for a given input.

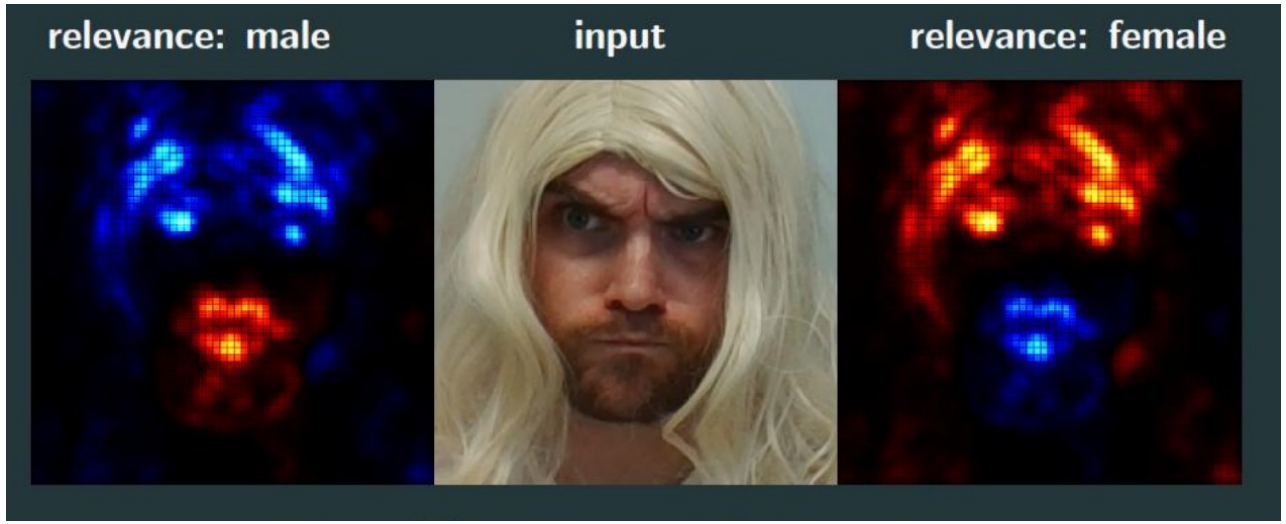
Related Work

- Lipton, 1990.
- Miller et al., 2018.
- Dhurandhar et al., 2018.
- Feghahati et al., 2020.

Model Explanation

Contrastive/Counterfactual Explanations

LRP



Grad-CAM

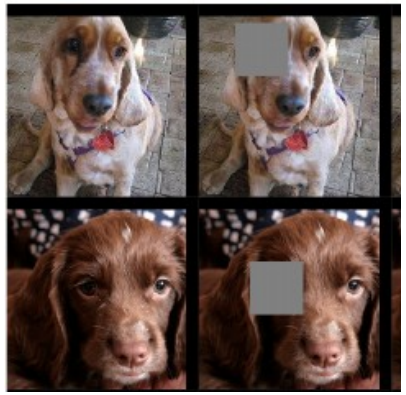
- Via standard explanation methods
- Explain target classes

Model Explanation

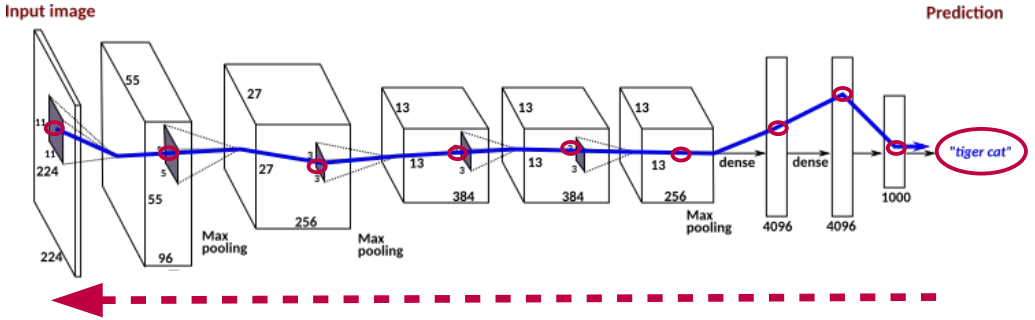
*See [Grun et al., 2016] for more details

So far...

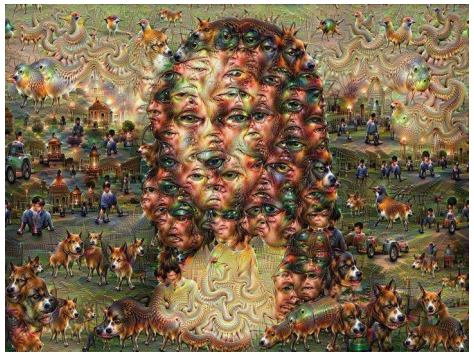
Input Modification



Deconvolution-based



Input Reconstruction

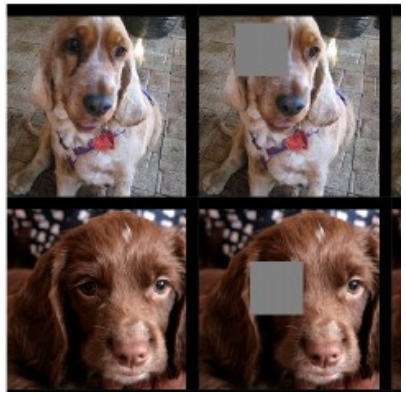


Model Explanation

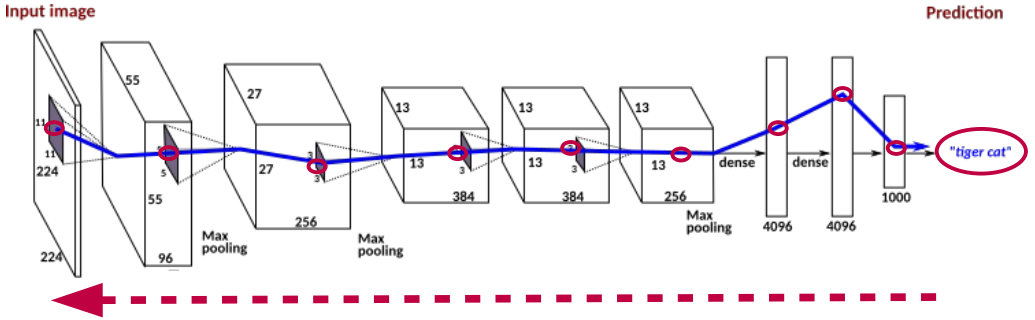
*See [Grun et al., 2016] for more details

Comparison

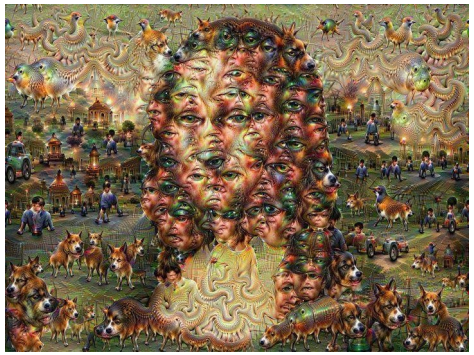
Input Modification



Deconvolution-based



Input Reconstruction



Q1: Which is more computationally expesive?

Q2: Which produces more detailed visualizations?

Q3: Which needs less knowledge about the model being processed?

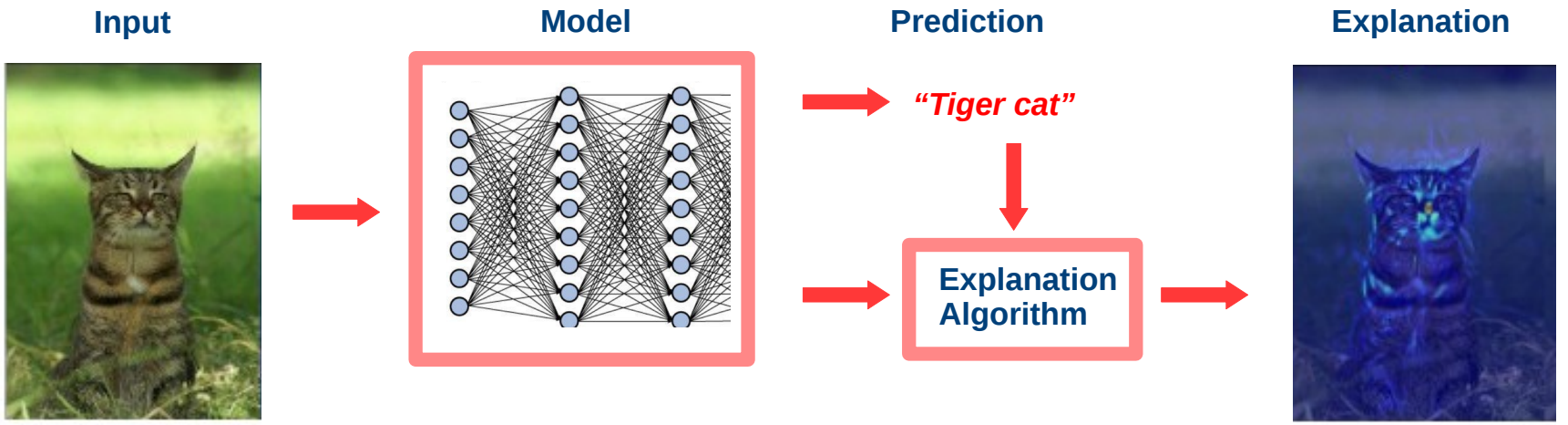
Questions?

Inherently Interpretable Models

[bridging model explanation and interpretation]

Interpretable-by-design Models

What we have assumed so far…

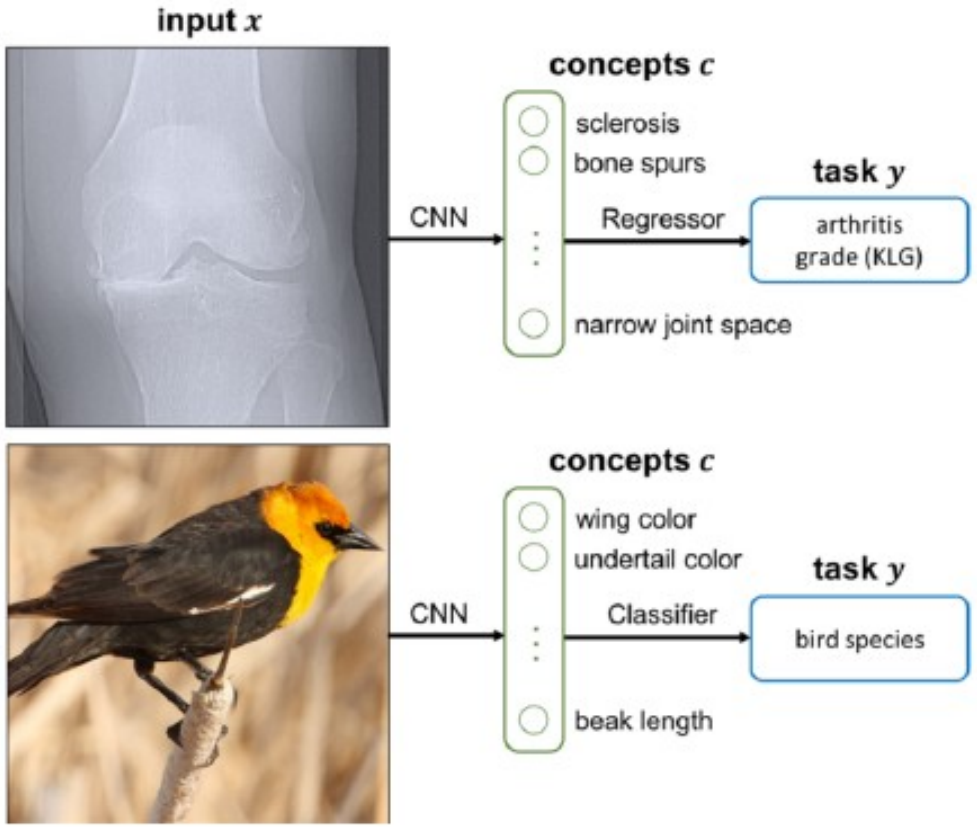


- Assumption: **the model being analyzed was already trained**.

  → **Post-hoc** model interpretation/explanation.

Interpretable-by-design Models

Concept Bottleneck Models [Koh et al., 2020]

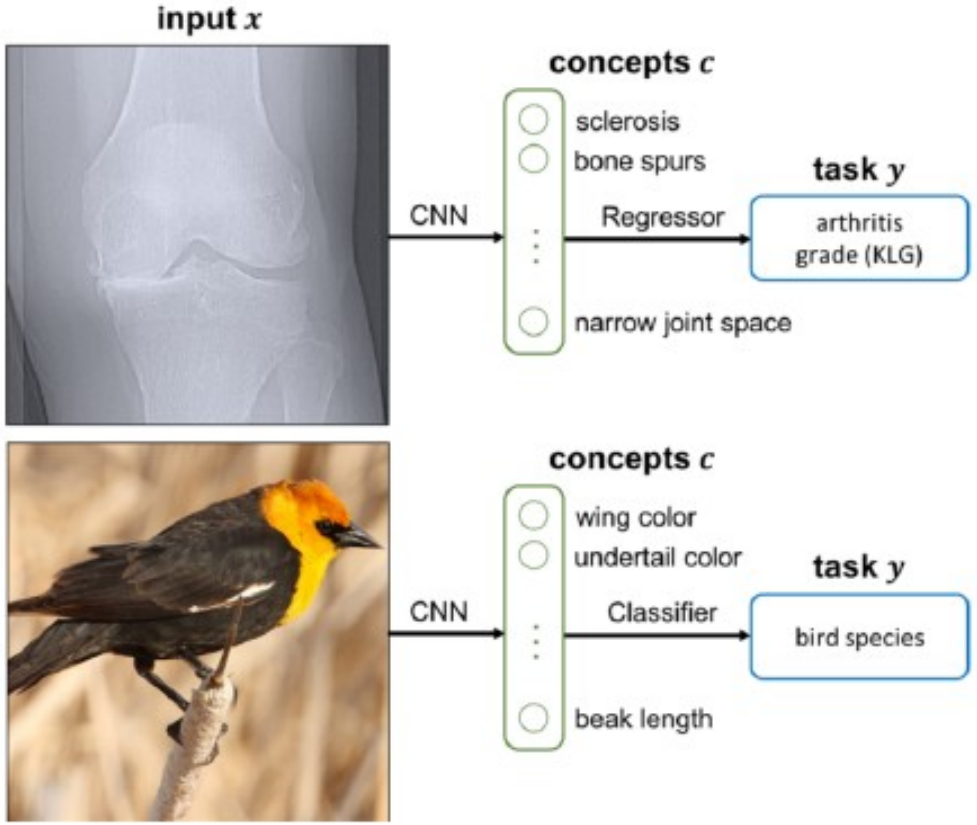


- Enforce the learning of intermediate semantic concepts
- Enable manual intervention

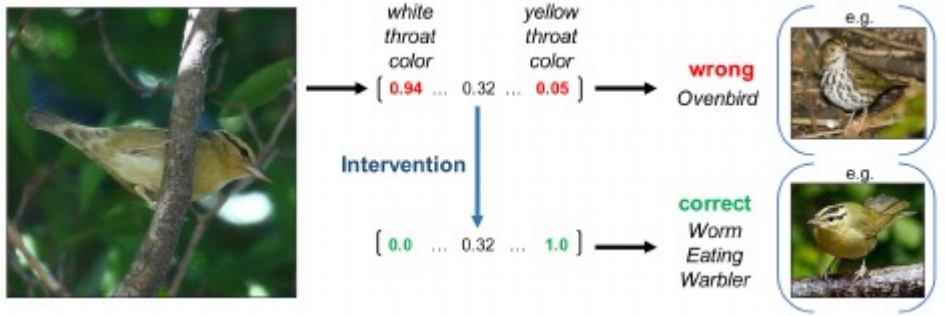
[Koh et al.,2020]

Interpretable-by-design Models

Concept Bottleneck Models [Koh et al., 2020]



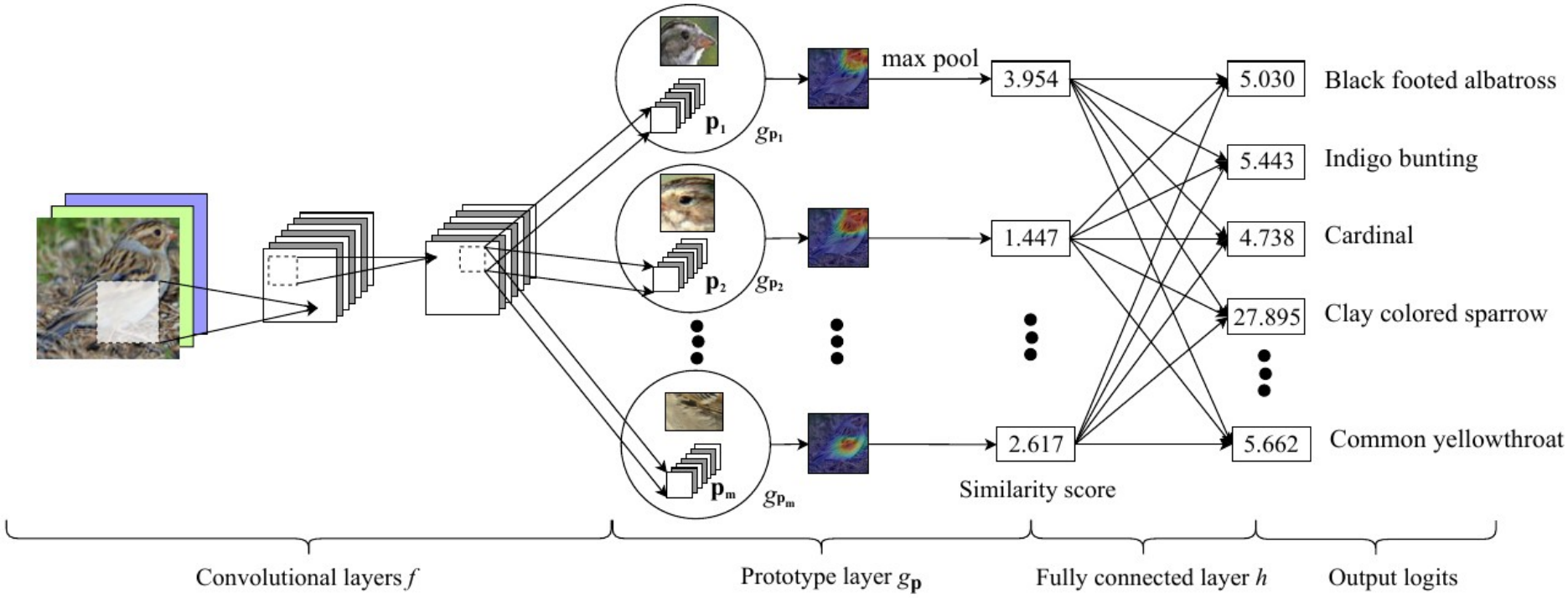
- Enforce the learning of intermediate semantic concepts
- Enable manual intervention



[Koh et al.,2020]

Interpretable-by-design Models

Prototype/Concept -based Learning [Chen et al., 2019]



Related Work

- Chen et al., 2019.
- Wei et al., 2020

- Learn an embedding that links the learned representation with the task of interest (i.e. the predictions).
- Associate the embedding with intermediate concepts of interest (e.g. object parts).

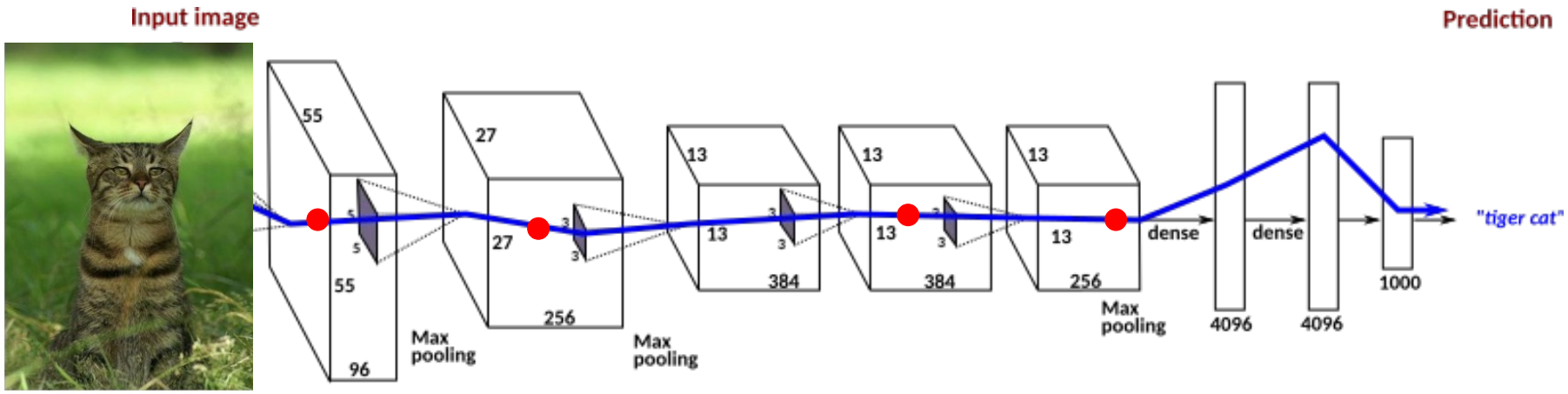
Evaluation

[Assessing how good the provided feedback really is]

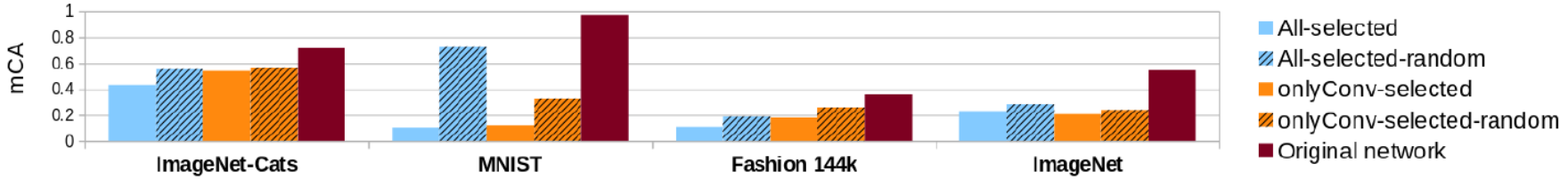
Evaluation – Interpretation

Ablating the Effect of Relevant Units/Neurons in the Model

- Removing relevant neurons from the network itself

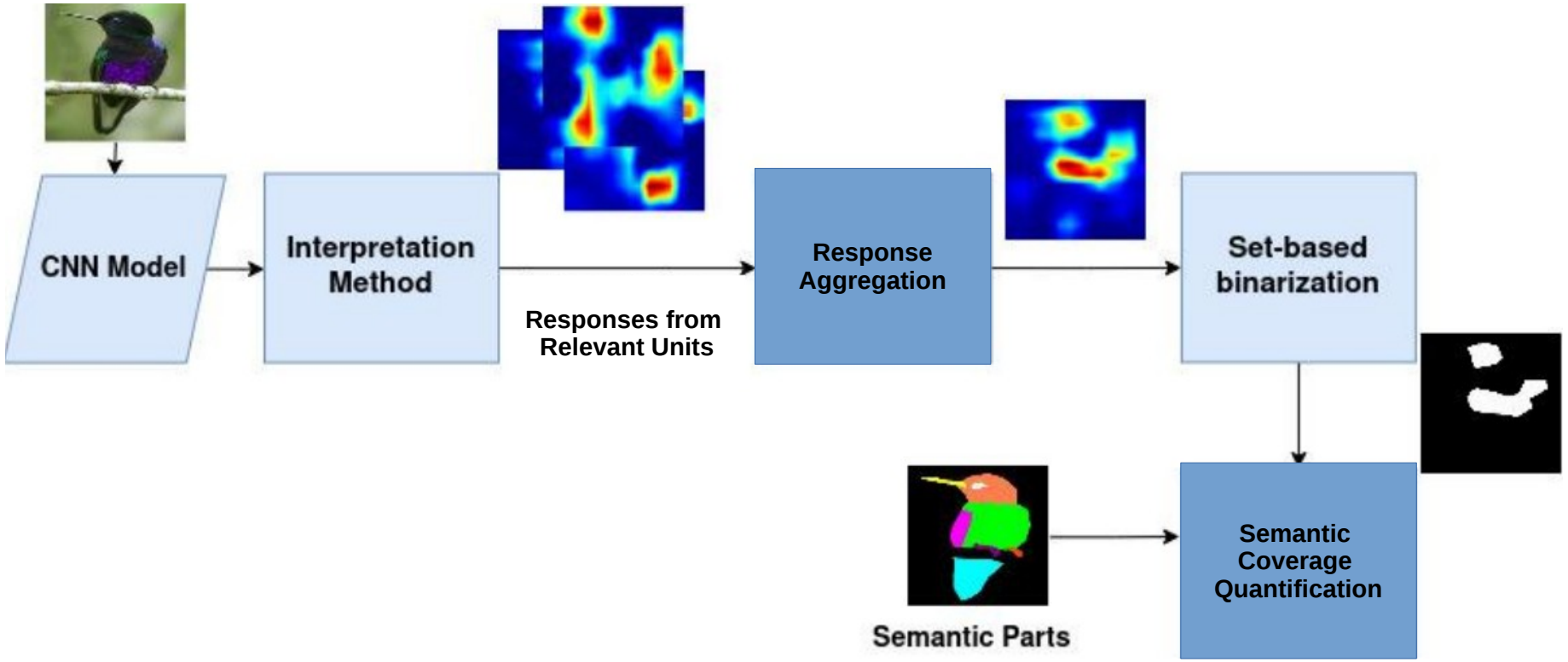


- Iteratively Compute Performance



Evaluation – Interpretation

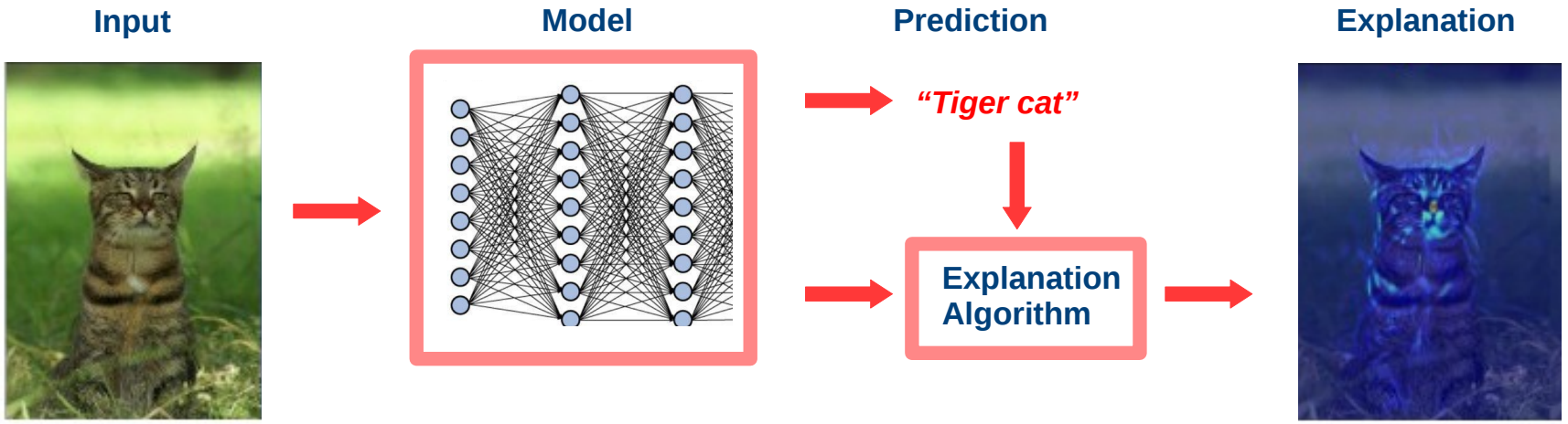
Measuring Semantic Coverage of Relevant Neurons/Units [Behzadi & Oramas, 2023]



- Aggregate reponses from relevant units
- Measure overlap w.r.t. annotated semantic concepts

Evaluation

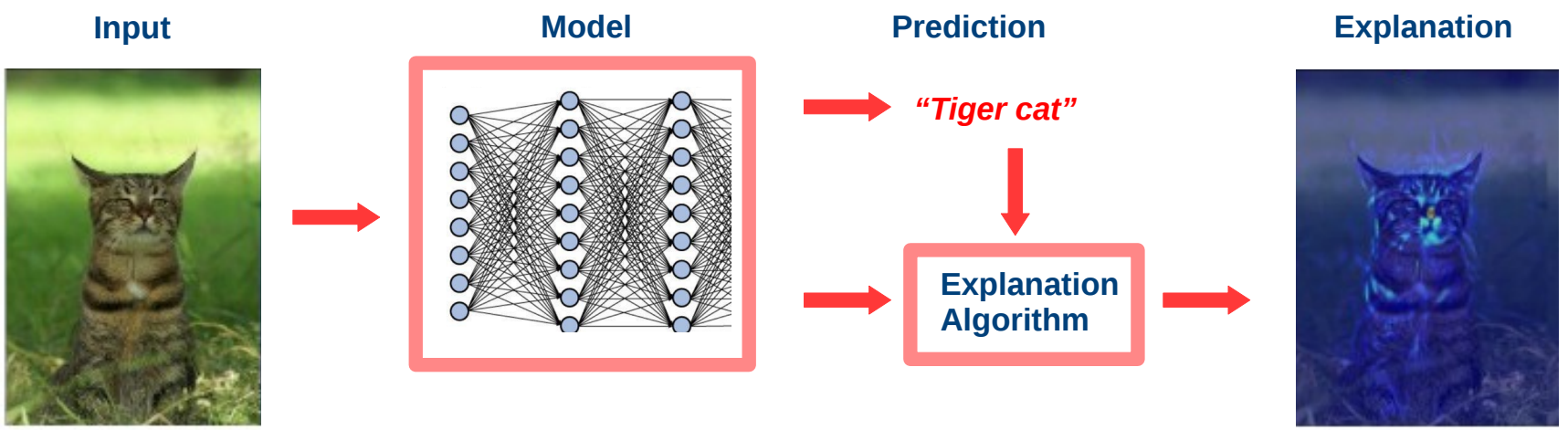
Assessing the Sanity of the Generated Explanations



Q: what factors should be verified?

Evaluation

Assessing the Sanity of the Generated Explanations

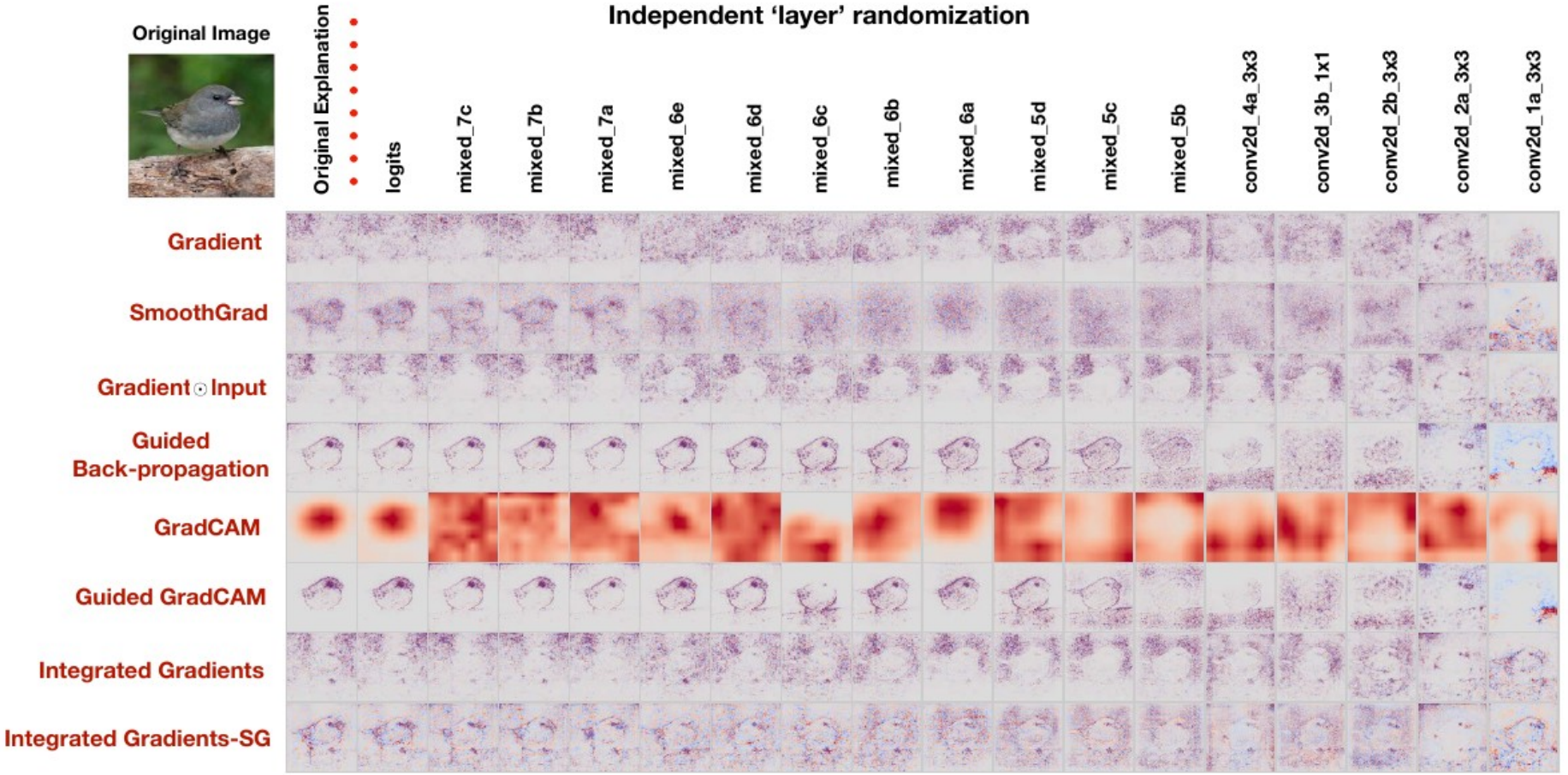


- Ensure the produced explanation is related to the **input**, the **model** and the **prediction**.

Q: what factors should be verified?

Evaluation

Assessing the Sanity of the Generated Explanations



- Ensure the produced explanation is related to the **input**, the **model** and the **prediction**.

Related Work

- Kindermans et al., 2017.
- Nie et al., 2018.
- Adebayo et al., 2018.

Evaluation – Explanation

Ablating the Effect of Relevant Parts of the Input

- Removing relevant parts of the input

Related Work

- Samek et al., 2017.
- Petsiuk et al., 2018.

Evaluation – Explanation

Ablating the Effect of Relevant Parts of the Input

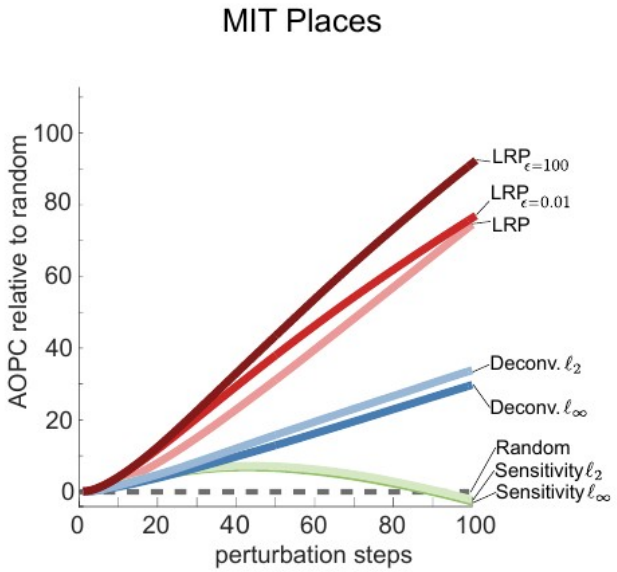
- Removing relevant parts of the input

- Quantify the effect of the perturbation

$$\mathrm{AOPC} = \frac{1}{L+1}\left\langle \sum_{k=0}^{L} f(\boldsymbol{x}^{(0)}_{\mathrm{MoRF}}) - f(\boldsymbol{x}^{(k)}_{\mathrm{MoRF}}) \right\rangle_{p(\boldsymbol{x})}$$

Related Work

- Samek et al., 2017.
- Petsiuk et al., 2018.

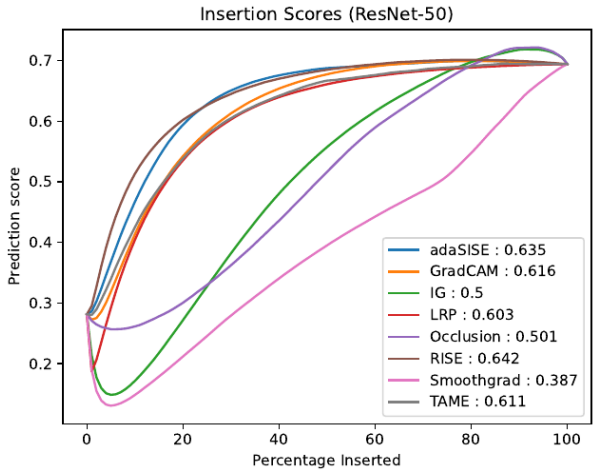


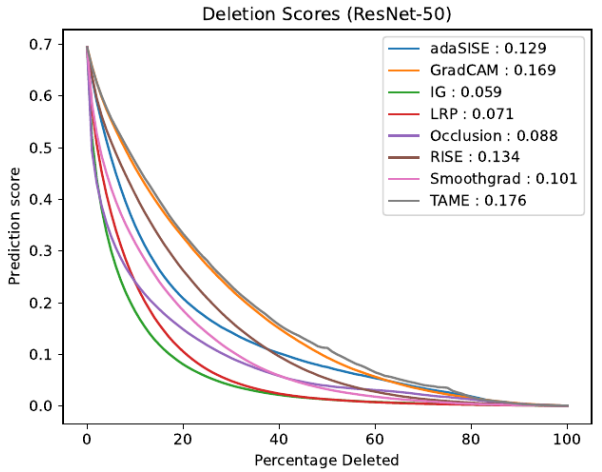
Evaluation – Explanation

Insertion-Deletion Metric [Petsiuk et al., 2018]

- Adding/Removing relevant parts of the input

- Quantify the effect of the perturbation



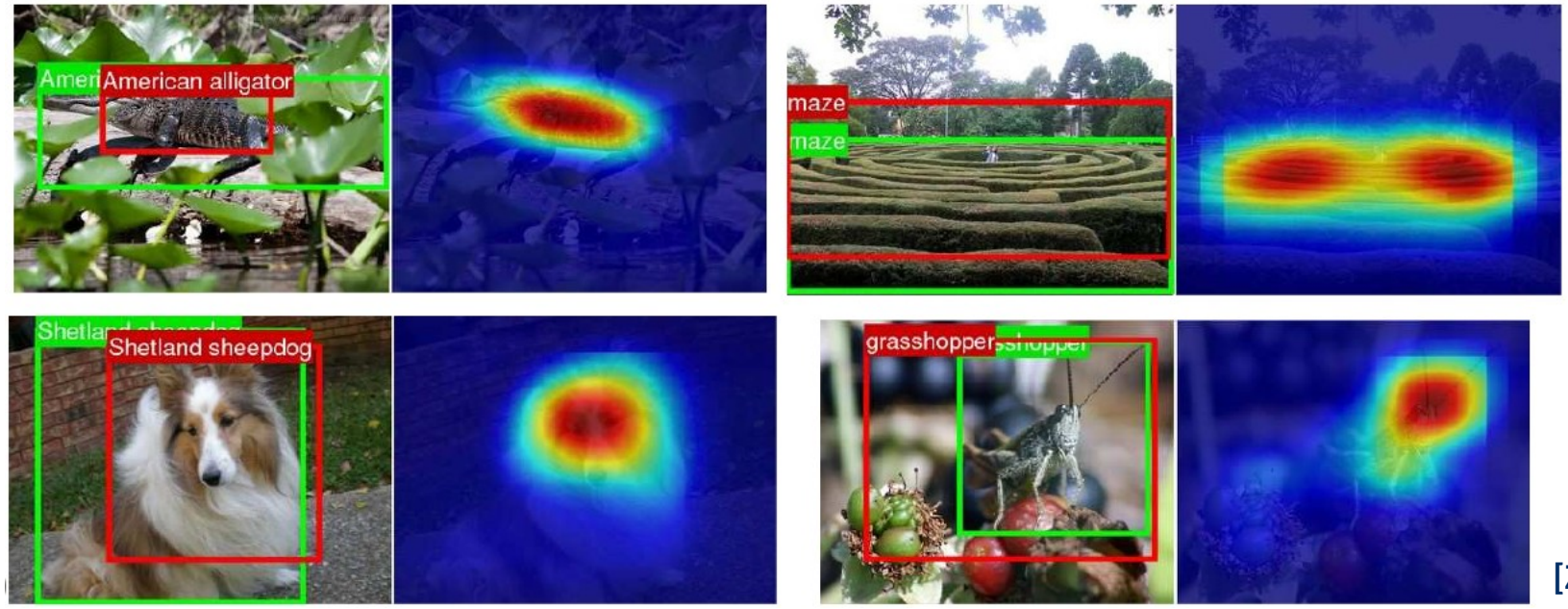


Related Work

- Samek et al., 2017.
- Petsiuk et al., 2018.

Evaluation

Measuring Performance through a Proxy Task



[Zhou et al.,2016]

Related Work

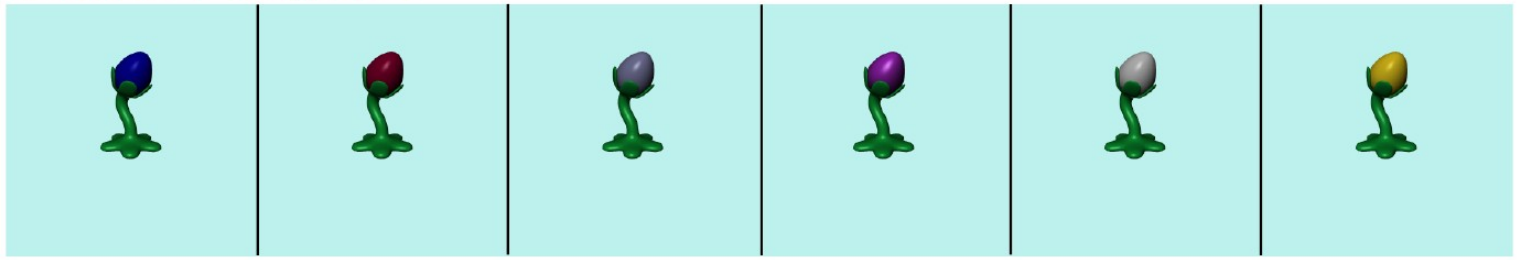
- Zhou et al., 2016.
- Zhang et al., 2016.

- Make an assumption (e.g. object should be localized) that explanation heatmap should follow together with an existing task (e.g. object detection).
- Measure performance using the protocol of the existing task

Evaluation – Explanation

Discriminative Feature Coverage

an8flower-single-6c



- Generate classes with controlled discriminative features

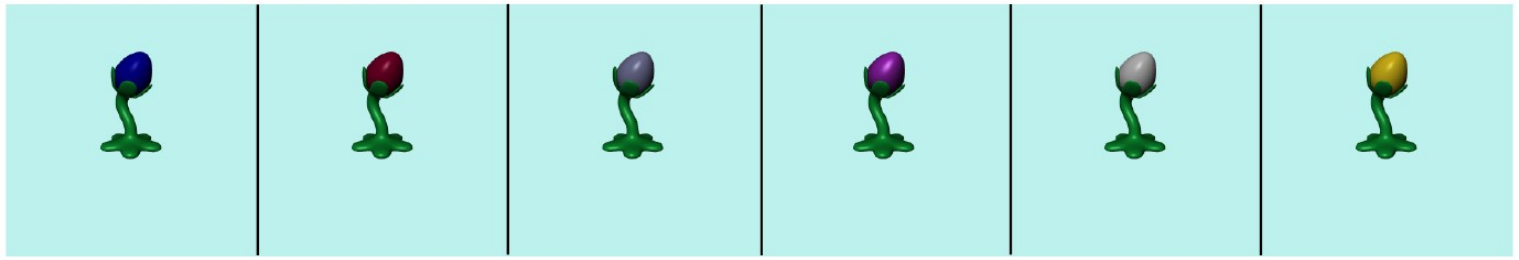
Related Work

- Oramas et al., 2019.
- Kim et al., 2018.
- Yang & Kim., 2019.
- Arras et al., 2020.

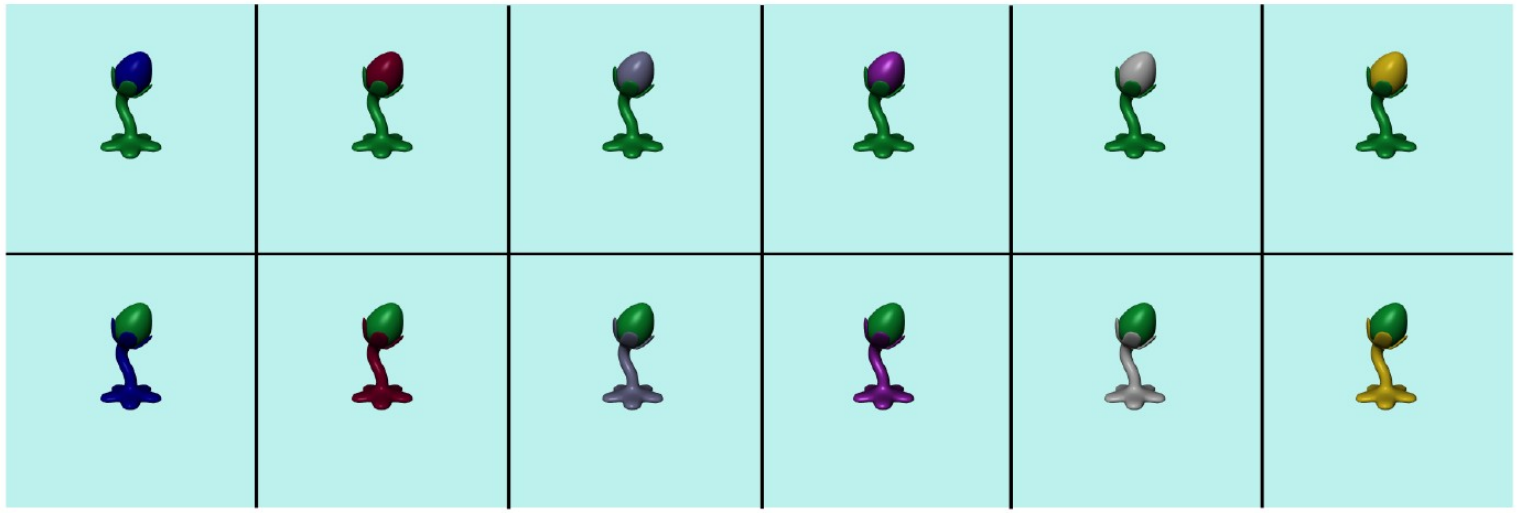
Evaluation – Explanation

Discriminative Feature Coverage

an8flower-single-6c



an8flower-double-12c



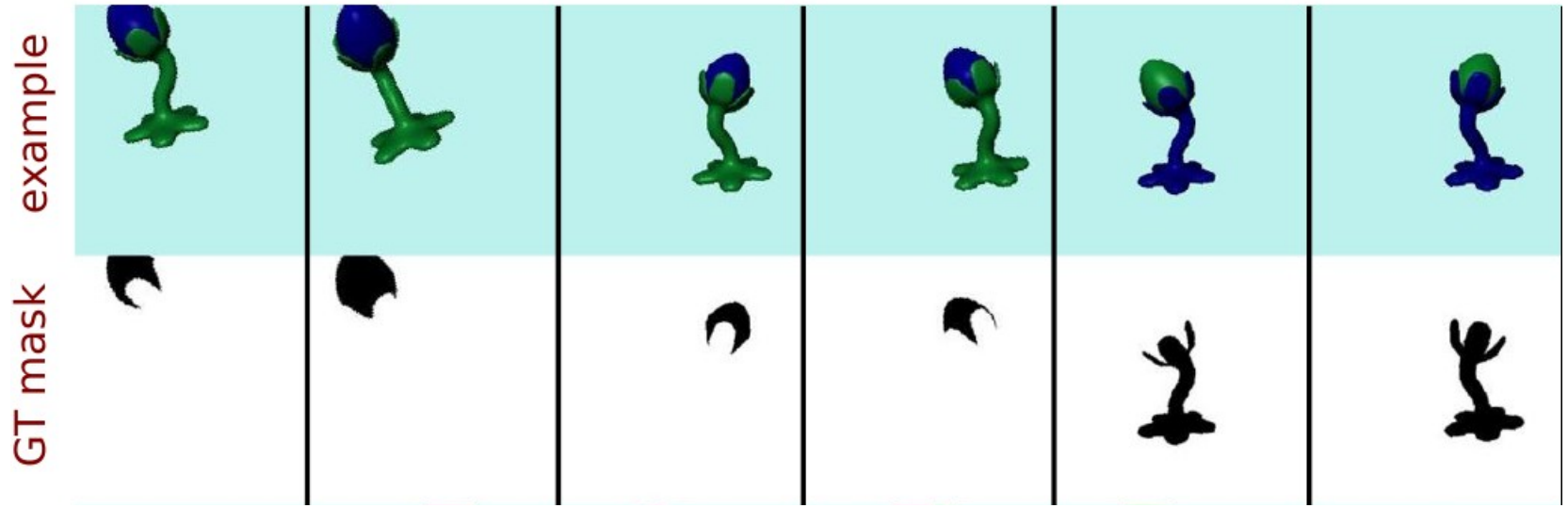
- Generate classes with controlled discriminative features

Related Work

- Oramas et al., 2019.
- Kim et al., 2018.
- Yang & Kim., 2019.
- Arras et al., 2020.

Evaluation – Explanation

Discriminative Feature Coverage



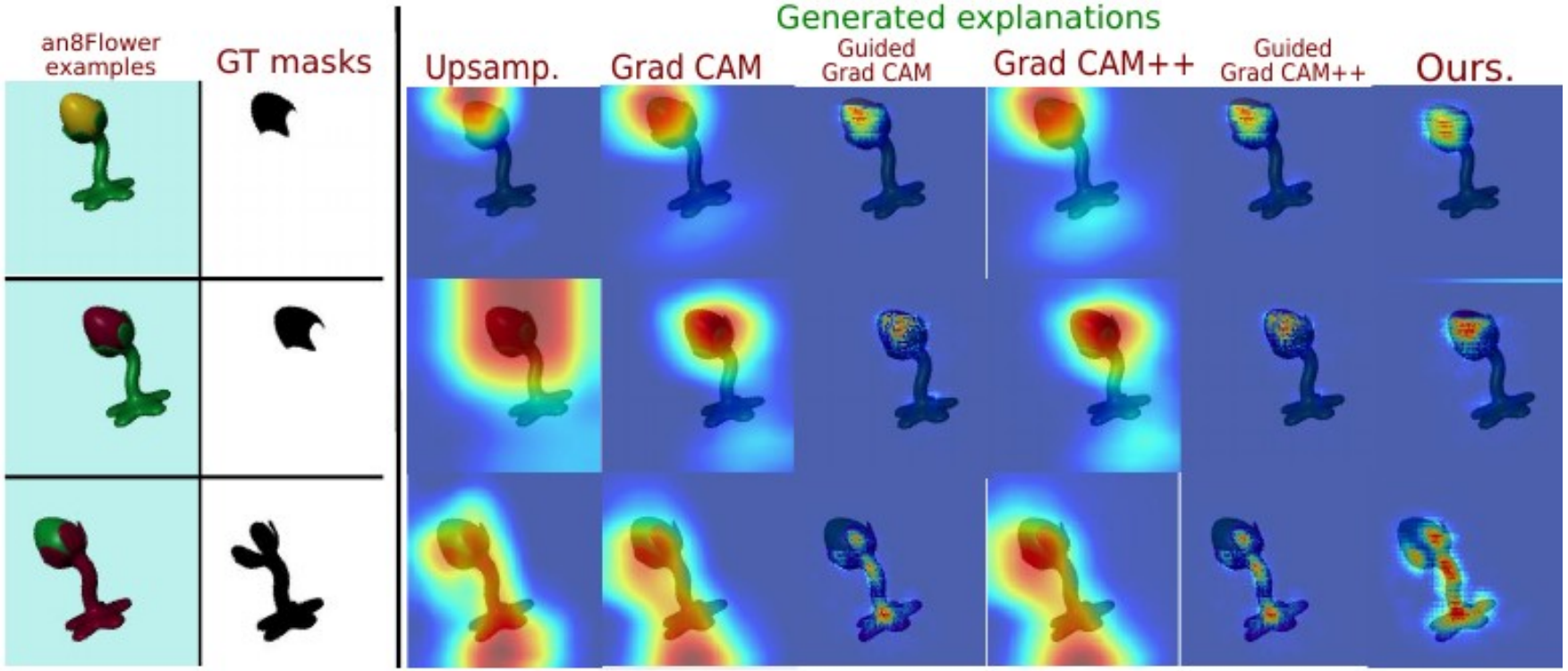
- Generate the ground-truth mask of the controlled features

Related Work

- Oramas et al., 2019.
- Kim et al., 2018.
- Yang & Kim., 2019.
- Arras et al., 2020.

Evaluation – Explanation

Discriminative Feature Coverage



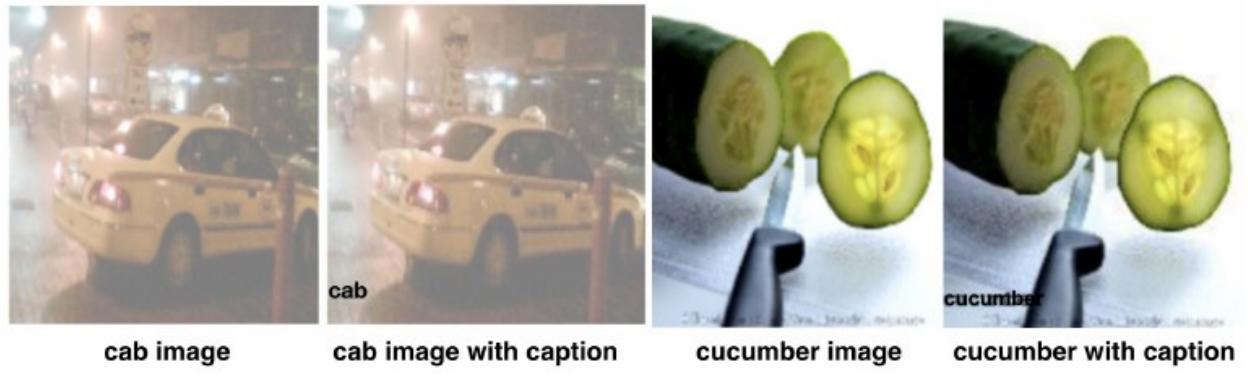
- Measure the overlap (IoU) of the visual explanation with the GT-mask

Related Work

- Oramas et al., 2019.
- Kim et al., 2018.
- Yang & Kim., 2019.
- Arras et al., 2020.

Evaluation – Explanation

Discriminative Feature Coverage



[Kim et al.,2018]

[Yang & Kim, 2019]

Related Work

- Oramas et al., 2019.
- Kim et al., 2018.
- Yang & Kim., 2019.
- Arras et al., 2020.

Summarizing

[Finally :D]

To Summarize

- **DNNs are not that Opaque Anymore**

  Some tasks have received more attention than others

To Summarize

- **DNNs are not that Opaque Anymore**

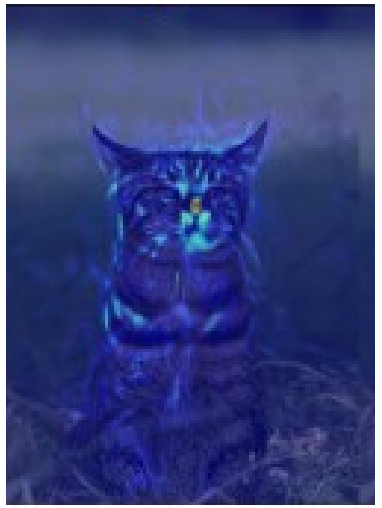
  Some tasks have received more attention than others

- **Post-hoc Model Interpretation & Explanation**

  Some methods have different requirements

  Generated explanations are not always reliable → sanit checks

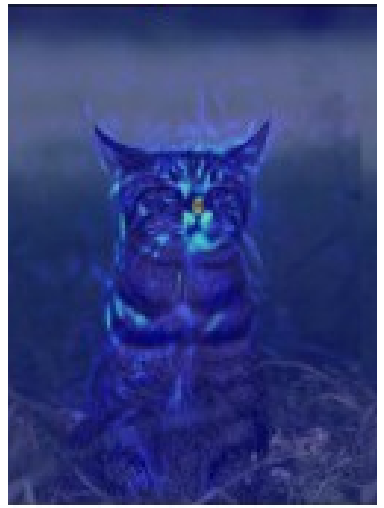
  Features used by the machine and humans may differ



To Summarize

- **DNNs are not that Opaque Anymore**

  Some tasks have received more attention than others



- **Post-hoc Model Interpretation & Explanation**

  Some methods have different requirements

  Generated explanations are not always reliable → sanit checks

  Features used by the machine and humans may differ

- **Interpretable-by-design Models**

  Recent trend of intelligible AI.

  Avoid creating opaque-boxes from the beginning.

Lots of energy and lots of success in your exams

Questions?

References

Post-hoc Model Interpretation

- D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, **Network dissection: Quantifying interpretability of deep visual representations,** CVPR, 2017.
- H. Cunningham, A. Ewart, L. Riggs, R. Huben, L. Sharkey. **Sparse Autoencoders Find Highly Interpretable Features in Language Models**. ICLR 2024.
- K. Simonyan, A. Vedaldi, A. Zisserman, **Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps**, arXiv:1312.6034, 2013.
- V. Escorcia, J. Niebles, B. Ghanem. **On the Relationship Between Visual Attributes and Convolutional Networks**. CVPR 2015.
- J. Oramas, K. Wang, T. Tuytelaars. **Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks**. ICLR 2019.
- R. Fong and A. Vedaldi. **Net2Vec: Quantifying and Explaining How Concepts are Encoded by Filters in Deep Neural Networks**. CVPR 2018
- C. Olah, L. Schubert. Exploring Neural Networks with Activation Atlases. Distill 2019.
- J. Mu & J. Andreas. **Compositional Explanations of Neurons**. NeurIPS 2020.

References

Post-hoc Model Explanation

- M. Zeiler and R. Fergus, **Visualizing and understanding convolutional networks,** European Conference on Computer Vision (ECCV), 2014.
- S. Bach, A. Binder, G. Montavon, F. Klauschen, KR. Müller, W. Samek, **On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation**, in PLOS ONE 10 (7), 2015.
- F. Grün, C. Rupprecht, N. Navab, F. Tombari, **A taxonomy and library for visualizing learned features inconvolutional neural networks**, in International Conference on Machine Learning (ICML) Workshops, 2016
- R. Fong & A. Vedaldi. **Interpretable Explanations of Black Boxes by Meaningful Perturbation**, ICCV 2017.
- V. Petsiuk, A. Das, K. Saenko BMVC 2018.
- R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, **Grad-CAM: Visual explanations from deep networks via gradient-based localization**, ICCV 2017.
- P. Lipton. **Contrastive explanation**. Royal Institute of Philosophy Supplement, 1990.
- A. Dhurandhar, P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam, P. Das. **Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives**. NeurIPS 2018.
- A. Feghahati, C. Shelton, M. Pazzani, K. Tang. **CDeepEx: Contrastive Deep Explanations**. ECAI 2020

References

Evaluation

- Nie et al. **A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations**. ICML 2018
- J. Adebayo, J. Gilmer, M. Muelly , I. Goodfellow , M. Hardt, B. Kim. **Sanity Checks for Saliency Maps**. NeurIPS 2018
- L. Arras, A. Osman, W. Samek, **Ground Truth Evaluation of Neural Network Explanations with CLEVR-XAI**, in arXiv:2003.07258, 2020.
- H. Behzadi-Khormouji & J. Oramas. **A Protocol for Evaluating Model Interpretation Methods from Visual Explanations**. WACV 2023.
- Kindermans et al. **The (Un)reliability of saliency methods.** NeurIPS workshop 2017
- W. Samek, A. Binder, G. Montavon, S. Bach, KR. Müller. **Evaluating the Visualization of What a Deep Neural Network has Learned**. TNNLS 2017.
- S. Lapuschkin , S. Wäldchen, A. Binder, G. Montavon, W. Samek, KR. Müller. **Unmasking Clever Hans Predictors and Assessing What Machines Really Learn**, Nature Communications, 2019.
- Nie et al. **A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations**. ICML 2018
- S. Yang, B. Kim, **BIM: Towards Quantitative Evaluation ofInterpretability Methods with Ground Truth**, arXiv:1907.09701, 2019.
- Vandersmissen, **Oramas. On the Coherency of Quantitative Evaluation of Visual Explanations**. CVIU 2023.

References

Interpretable-by-design Models

- Q. Zhang, Y. Wu, S. Zhu. **Interpretable Convolutional Neural Networks**, CVPR 2018.
- C. Chen, O. Li, C. Tao, A. Barnett, J. Su, C. Rudin. **This Looks Like That: Deep Learning for Interpretable Image Recognition**. NeuriPS 2019.
- P. Wei Koh, T. Nguyen, Y. Tang, S. Mussmann, E. Pierson, Been Kim, P. Liang. **Concept Bottleneck Models**. ICML 2020.
- Z. Chen, Y. Bei, C. Rudin. **Concept Whitening for Interpretable Image Recognition**, Nature Machine Intelligence 2020.

Human-Understandable Explanations

- K. Wang, J. Oramas, T. Tuytelaars. **Towards Human-Understandable Visual Explanations: Imperceptible Cues Can Better Be Removed**, arXiv: 2104:07954, 2021.

University of Antwerp

Artificial Neural Networks

[2500WETANN]

José Oramas