Lecture 20: Feature Attribution Methods

Trustworthy Machine Learning

Spring 2024

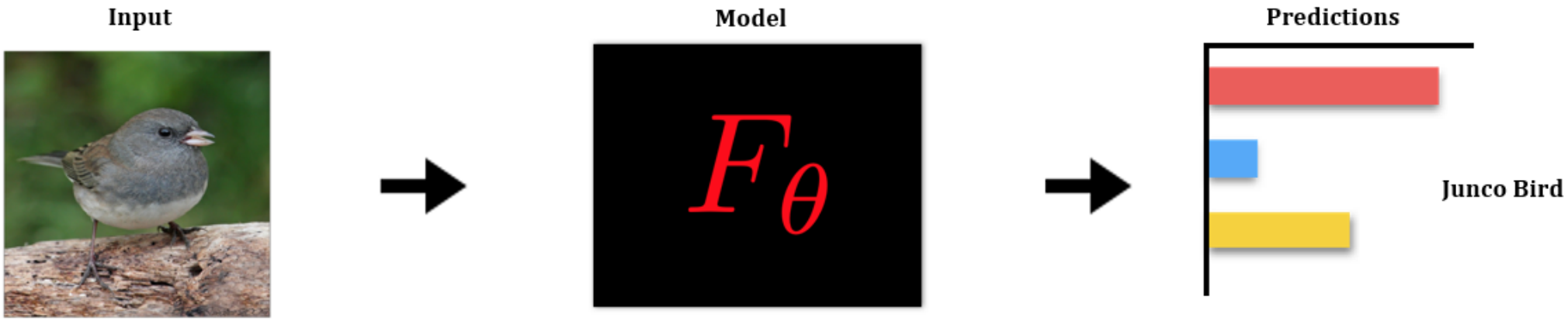
Agenda

- Recap: Feature Attribution Methods
  - Feature attribution problem
  - LIME (Local Interpretable Model-agnostic Explanations) algorithm
  - SHAP method based on cooperative game theory

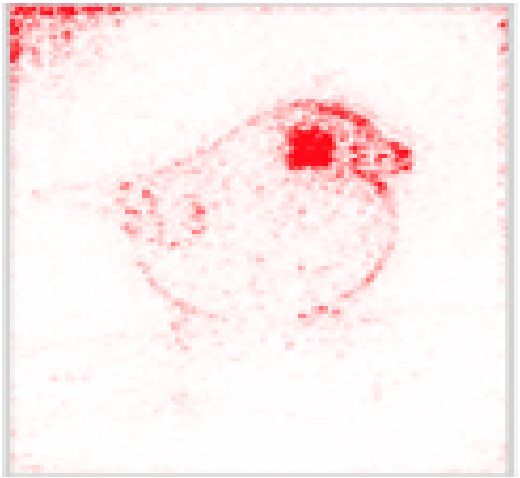
- Today:
  - Saliency Maps
  - Formal guarantees for feature attribution methods (Anton Xue)

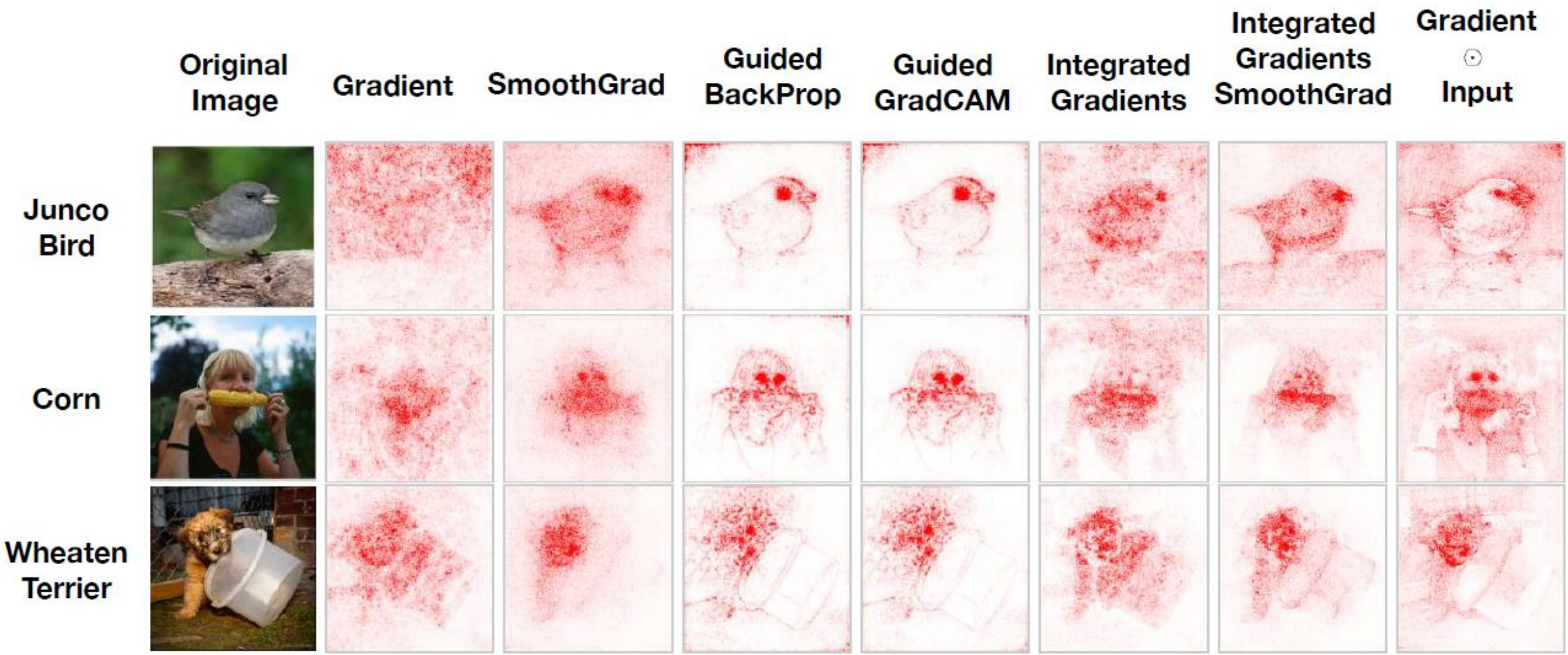
- Resources:
  - Tutorial on “Interpreting ML models” by Hima Lakkaraju
  - SmoothGrad: removing noise by adding noise; Smilkov et al; NeurIPS 2017
  - Stability guarantees for feature attributions with multiplicative smoothing; Xue et al; NeurIPS 2023

Saliency Maps

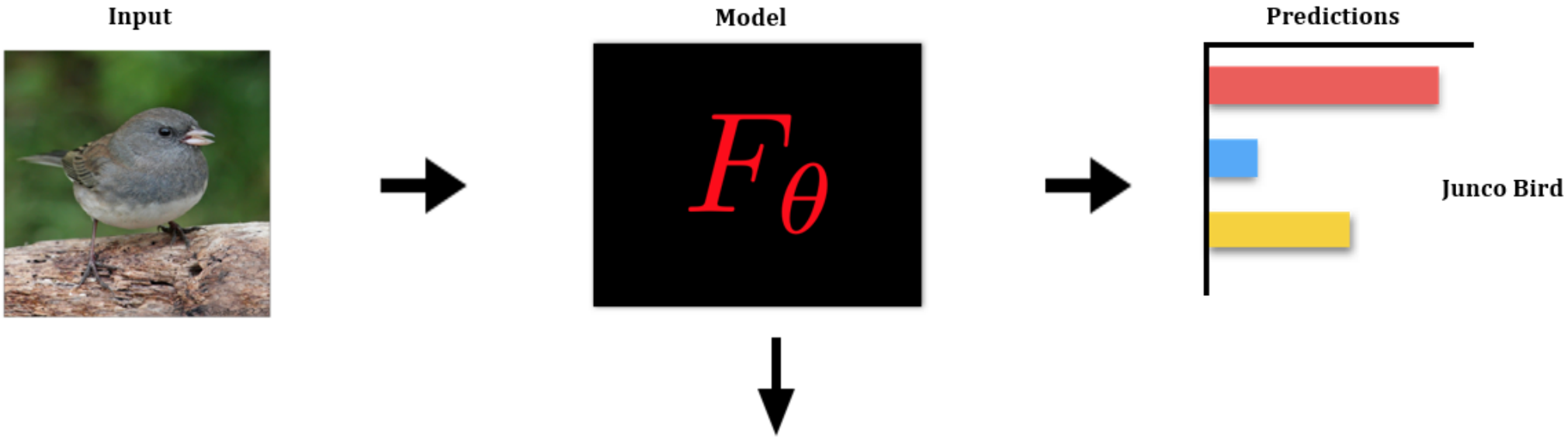


What parts of the input are most relevant for the model's prediction: 'Junco Bird'?




Original Image
Gradient
SmoothGrad
Guided BackProp
Guided GradCAM
Integrated Gradients
Integrated Gradients SmoothGrad
Gradient ⊙ Input
Junco Bird
Corn
Wheaten Terrier

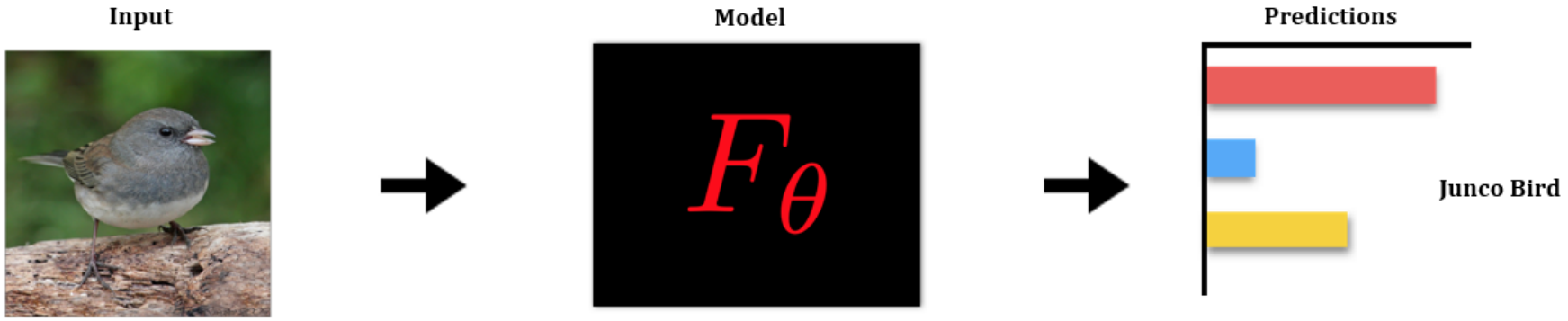
Saliency Maps for Deep Neural Networks



$$F : \mathbb{R}^d \rightarrow \mathbb{R}^c$$ **Model**

$$F_i : \mathbb{R}^d \rightarrow \mathbb{R}$$ **class specific logit**

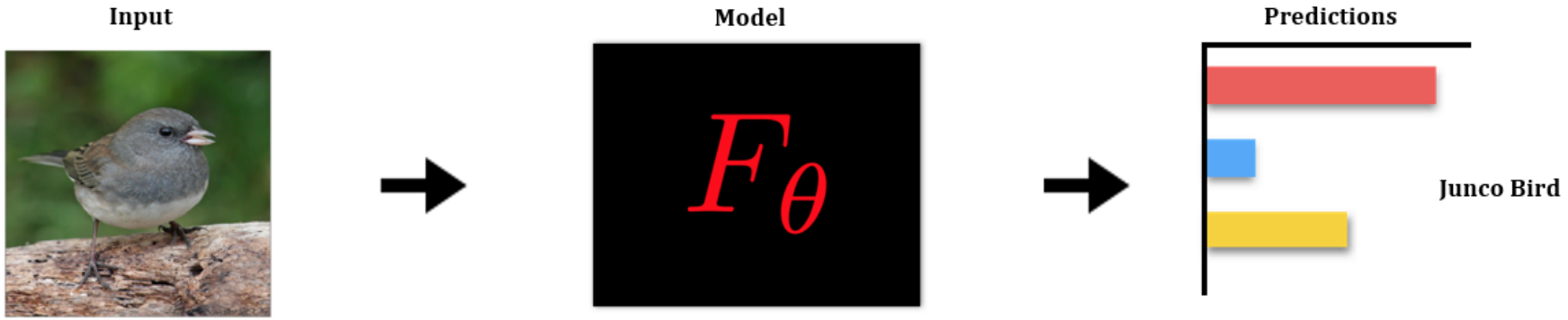
Input Gradient



What's the contribution of a pixel in input image to the prediction "Junco bird"

Solution: Compute the gradient of the output logit w.r.t. the pixel

Input Gradient



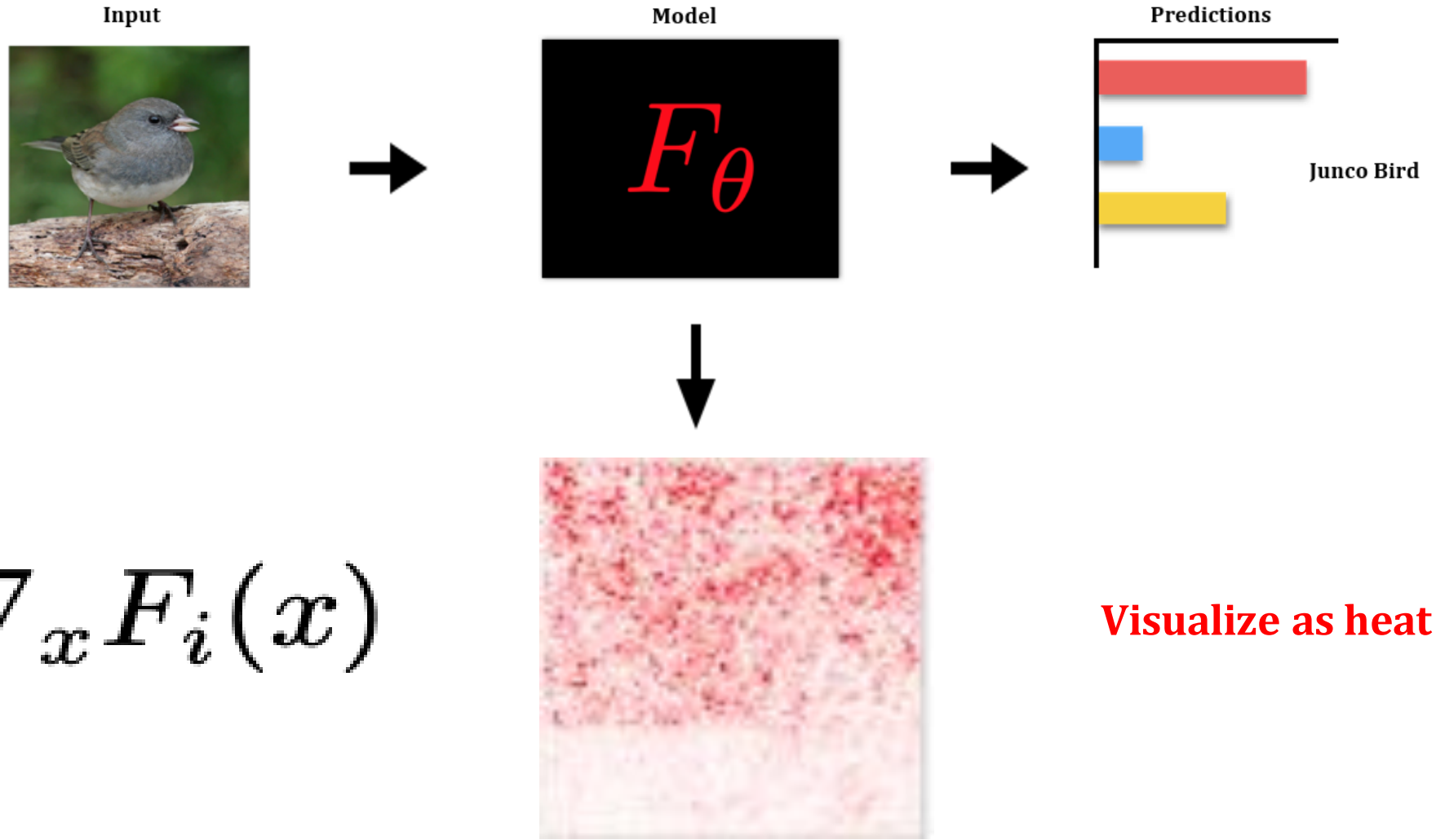
$$\nabla_x F_i(x) \quad \in \mathbb{R}^d$$

Input

Logit

Same dimension as the input

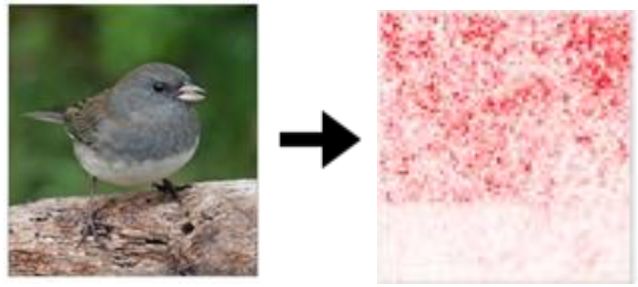
Input Gradient



$$\nabla_x F_i(x)$$

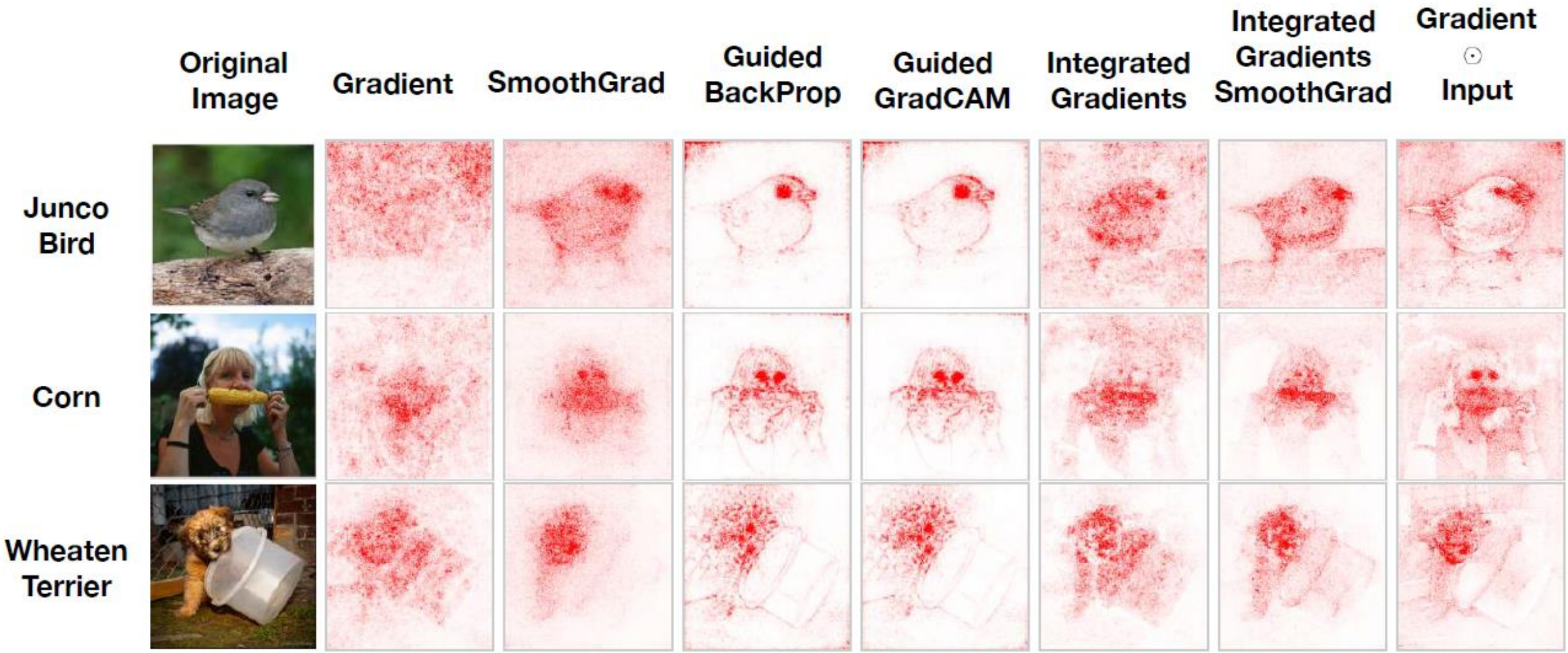
Visualize as heat map

Beyond Input Gradients

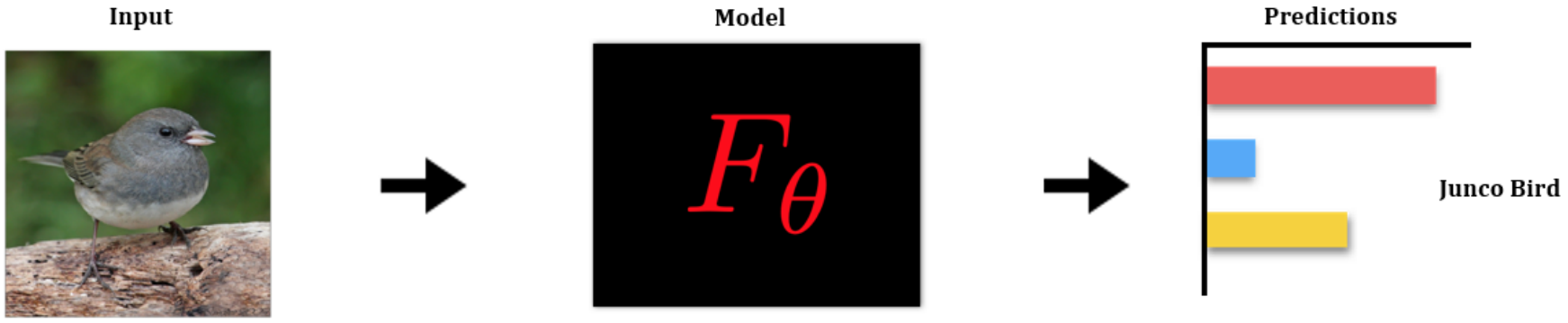


Input gradients capture sensitivity to individual pixels/features well, but …
Raw gradients are visually noisy

Many improvements proposed in literature


Original Image
Gradient
SmoothGrad
Guided BackProp
Guided GradCAM
Integrated Gradients
Integrated Gradients SmoothGrad
Gradient ⊙ Input
Junco Bird
Corn
Wheaten Terrier

SmoothGrad



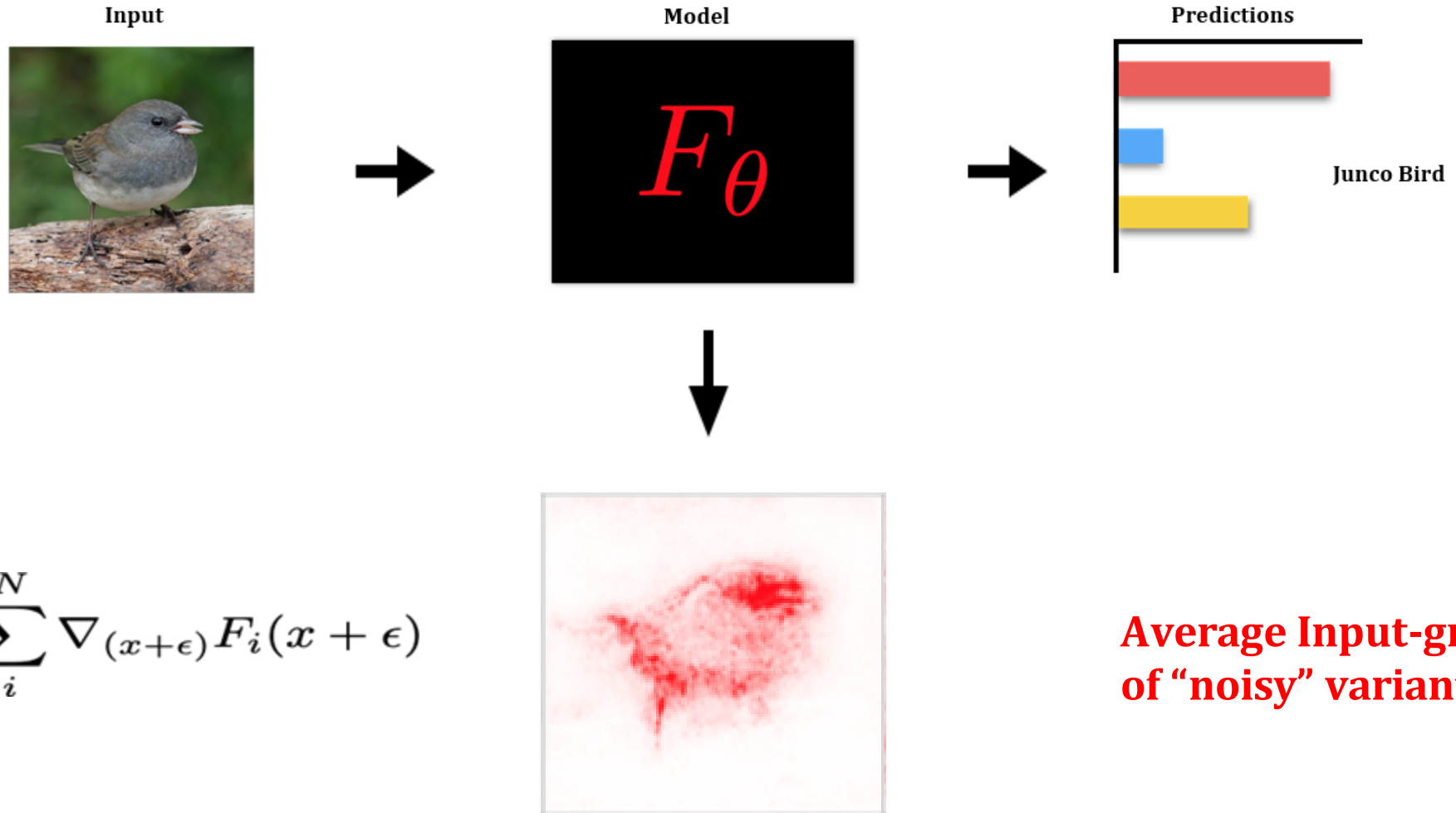
SmoothGrad

$$\frac{1}{N}\sum_{i}^{N}\nabla_{(x+\epsilon)}F_i(x+\epsilon)$$

Gaussian noise

- Construct N perturbations of input x
- By adding noise ε sampled from Gaussian distribution with standard deviation σ
- For each variant compute input gradient
- Take average

SmoothGrad



$$\frac{1}{N}\sum_{i}^{N}\nabla_{(x+\epsilon)}F_i(x+\epsilon)$$

Average Input-gradients of "noisy" variants

SmoothGrad: Effect of noise

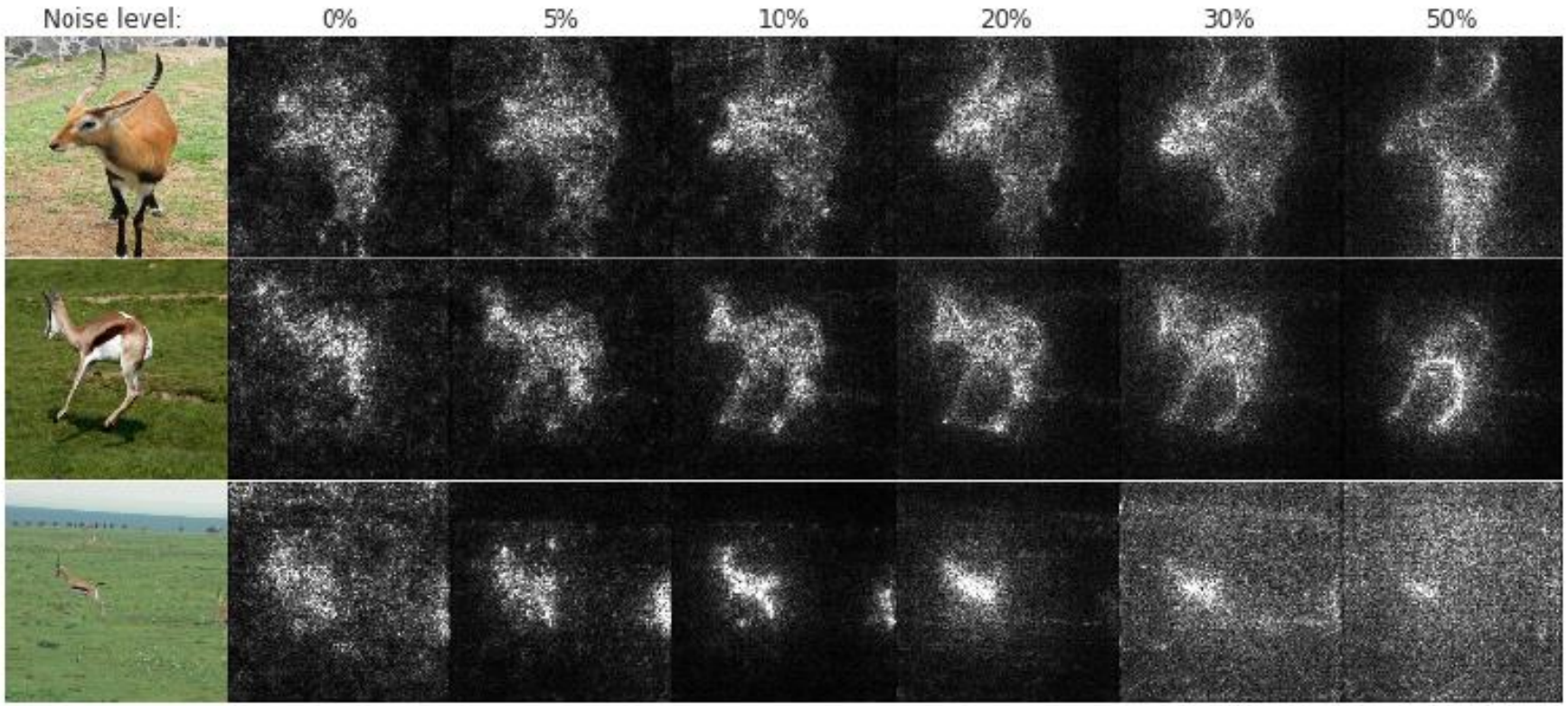
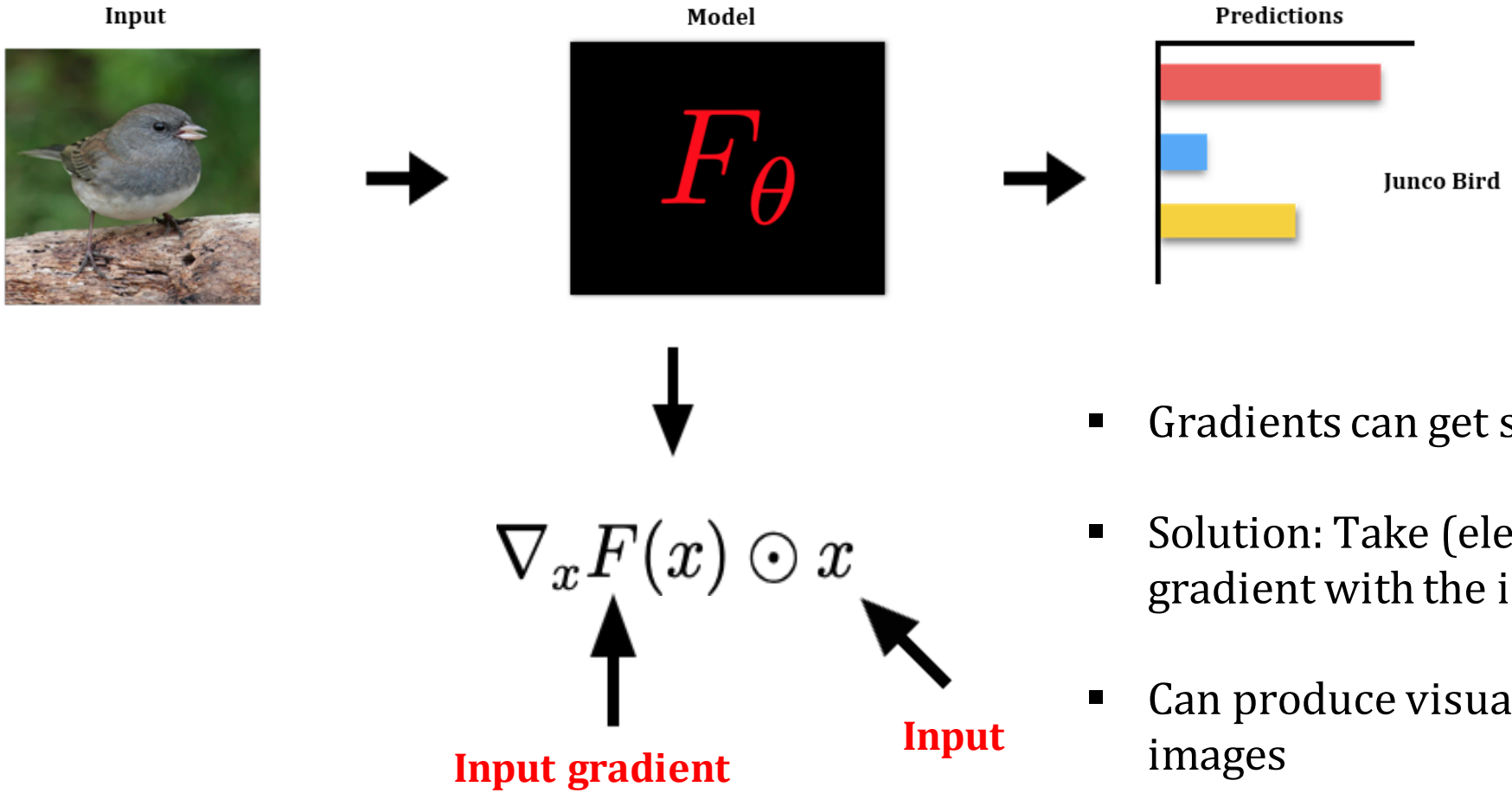


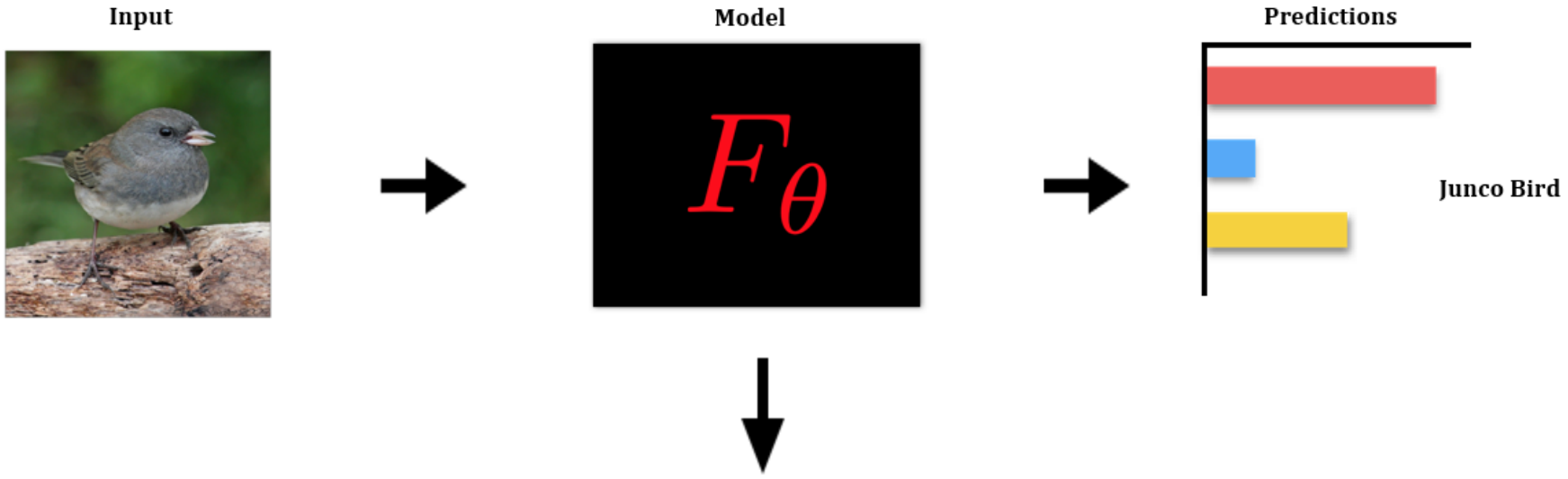
Figure 3. Effect of noise level (columns) on our method for 5 images of the gazelle class in ImageNet (rows). Each sensitivity map is obtained by applying Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the input pixels for 50 samples, and averaging them. The noise level corresponds to $\sigma/(x_{max} - x_{min})$.

Gradient-Input

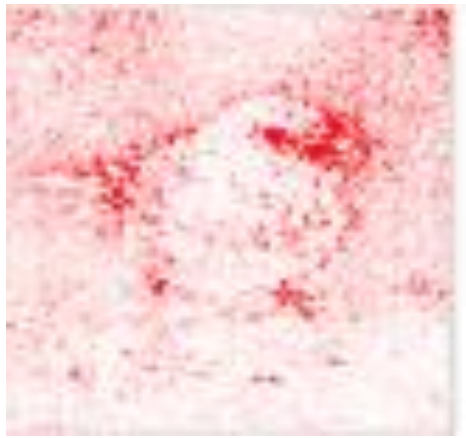


- Gradients can get saturated
- Solution: Take (element-wise) product of the gradient with the input itself
- Can produce visually simpler/sharper images
- Can be used in conjunction with any method

Gradient-Input



$$\nabla_x F(x) \odot x$$



Element-wise product of input-gradient and input

Recap: Feature Attribution Methods

- Feature Attribution Methods
  - LIME (Local Interpretable Model-agnostic Explanations) algorithm
  - SHAP methods based on cooperative game theory
  - Saliency Maps (different versions)

- Reference for application in radiology:

  On the interpretability of AI in Radiology: Challenges and opportunities

  Reyes et al, Radiology: Artificial Intelligence, 2020

- Next: Can we get some guarantees regarding output of an explanation method ?