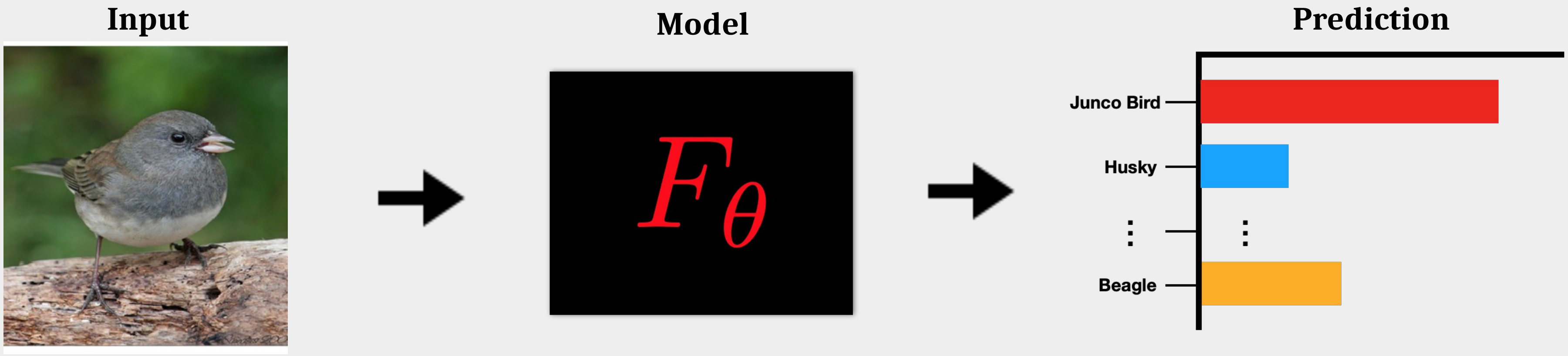
Sanity Checks for Saliency Maps

Overview

- Feature Attribution / Saliency Maps Setup
- Overview of Sanity Checks for Saliency Maps
- Follow-up work
- Parting thoughts / Q&A

Feature Attributions / Saliency Maps



What parts of the input are 'most important' for the model prediction Junco Bird?

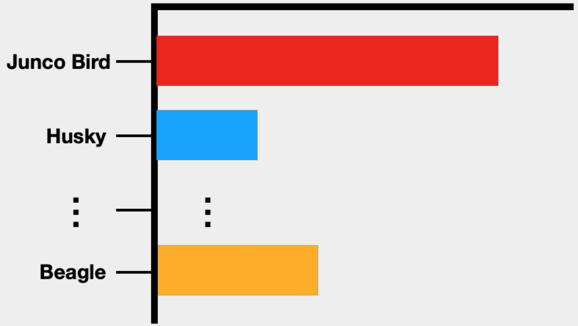
Feature Attributions / Saliency Maps

Input

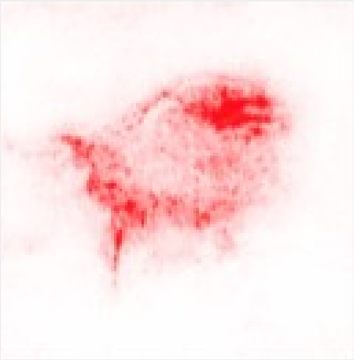
Model

F_θ

Prediction

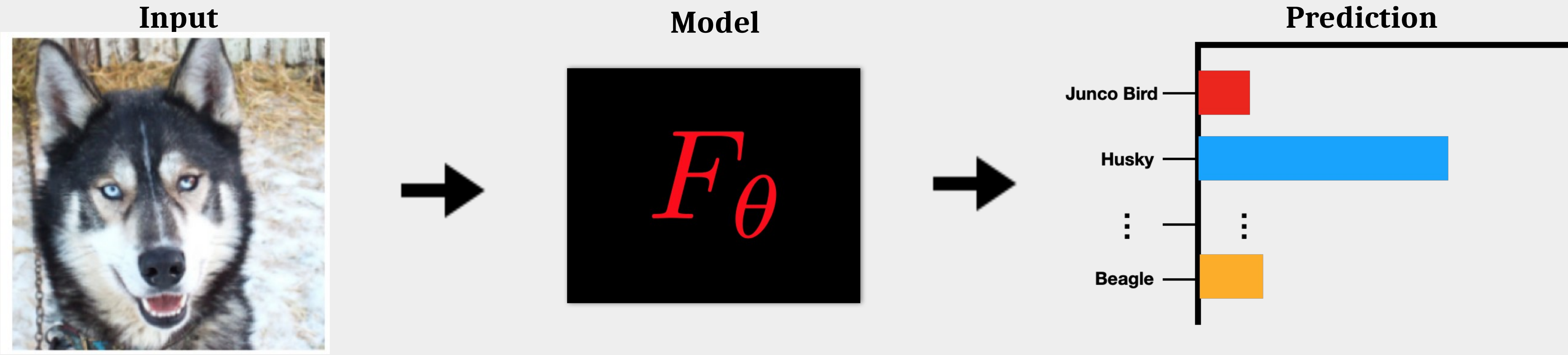


What parts of the input are 'most important' for the model prediction Junco Bird?



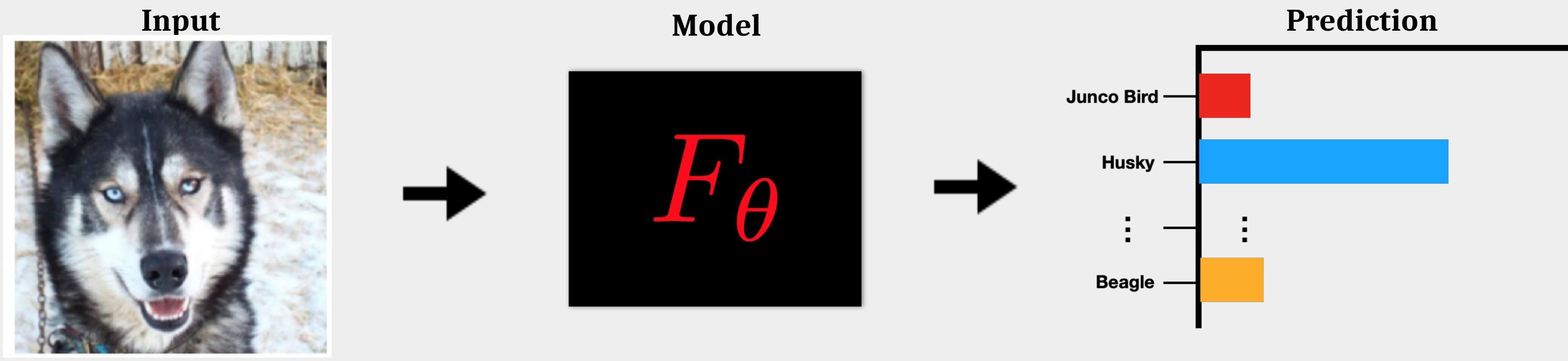
Feature Attribution / Saliency Map

Identifying Shortcuts



What parts of the input are 'most important' for the model prediction Husky?

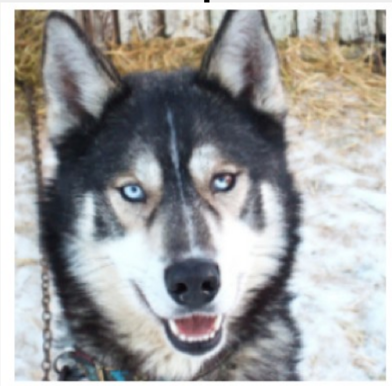
Identifying Shortcuts



What parts of the input are 'most important' for the model prediction Husky?

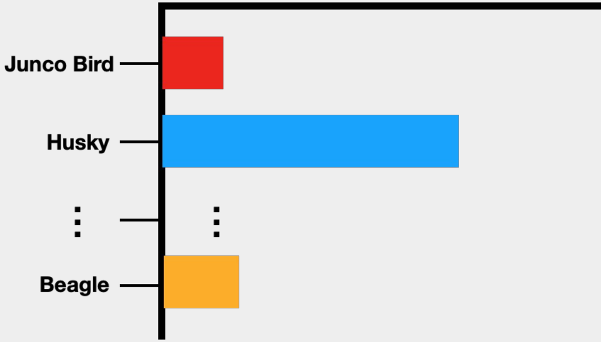
Identifying Shortcuts

Input



Model

Prediction

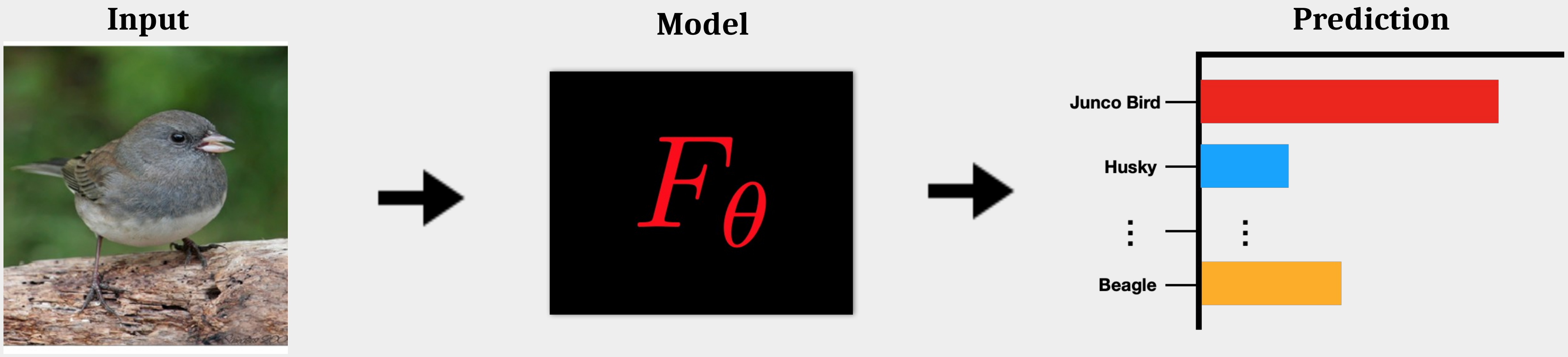


What parts of the input are 'most important' for the model prediction Husky?

Collect additional data to fix the bug.

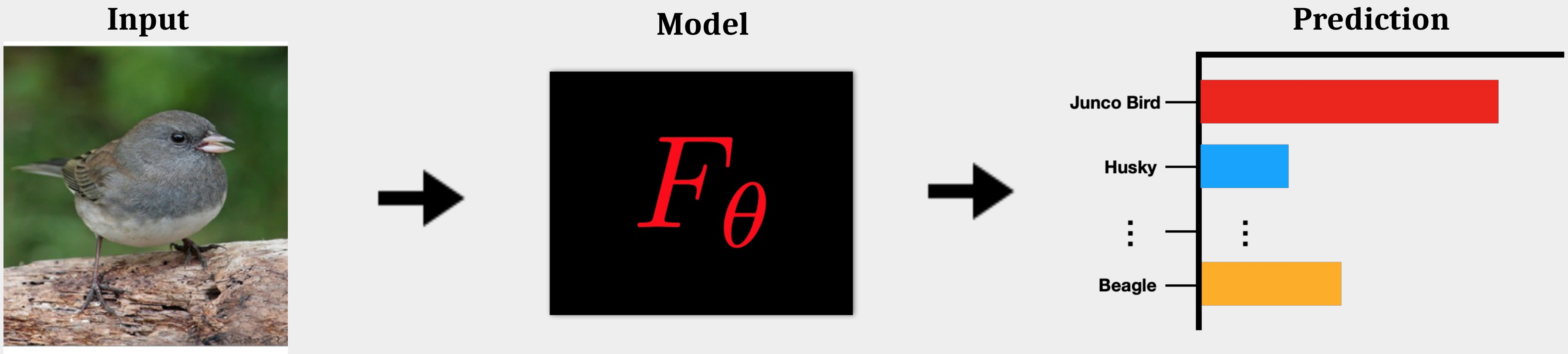
Model relying on snow to identify Huskies.

Feature Attributions / Saliency Maps



Feature attribution method: assigns an output 'relevance' score to each dimension of the input.

Feature Attributions / Saliency Maps



Feature attribution method: assigns an output 'relevance' score to each dimension of the input.

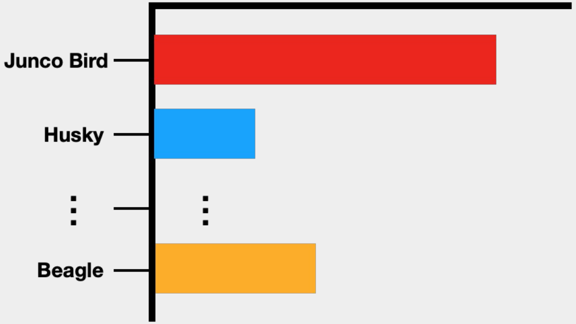
$$F : \mathbb{R}^d \rightarrow \mathbb{R}^c$$ **Model**

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

Input-Gradient / Saliency / Gradient

Model

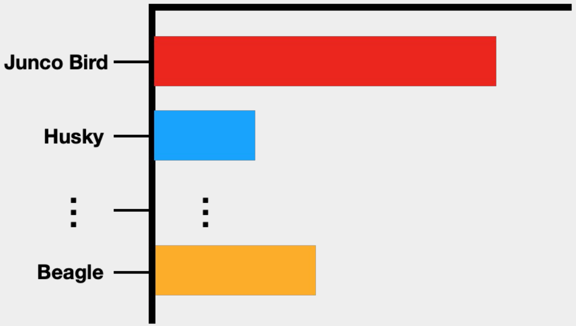
F_θ



Input-Gradient / Saliency / Gradient

Model

$$F_\theta$$



Input-Gradient

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

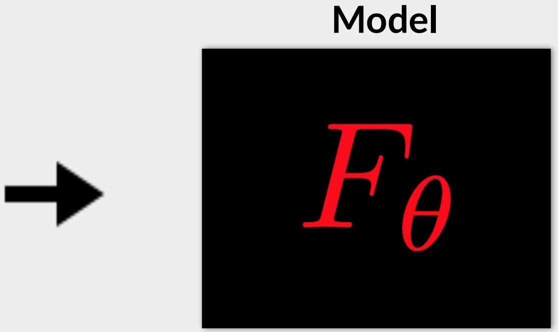
Input

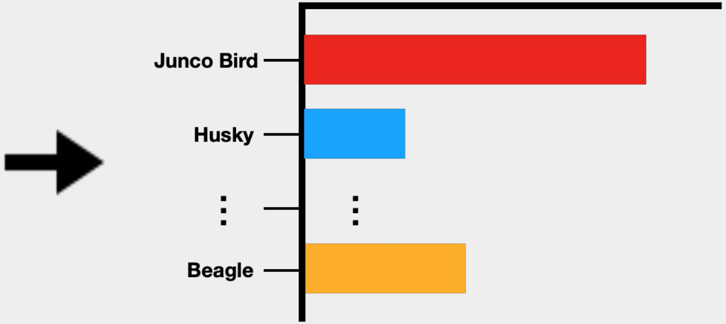
Logit

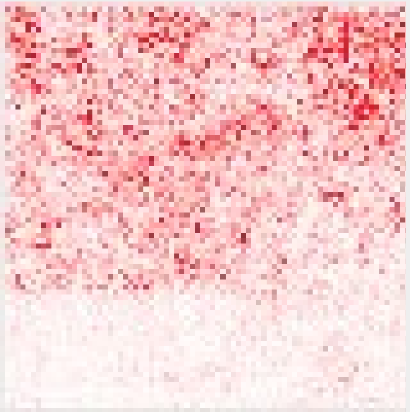
Same dimension as the input.

Baehrens et. al. 2010; Simonyan et. al. 2014

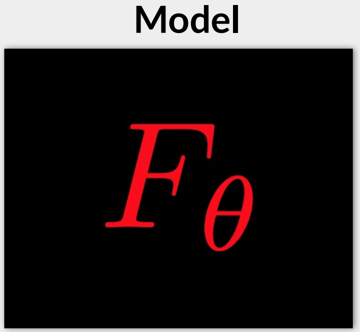
Input-Gradient / Saliency / Gradient

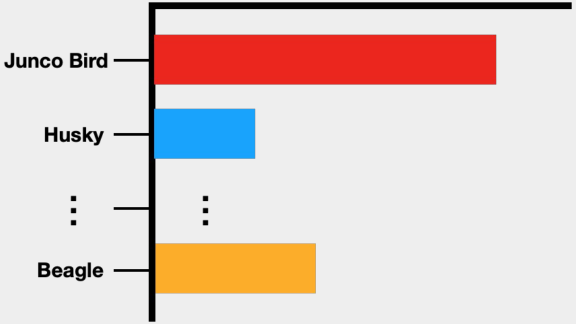






Integrated Gradients





$$(x - \tilde{x}) \times \int_{\alpha=0}^{1} \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x}$$

Path integral: 'sum' of interpolated gradients

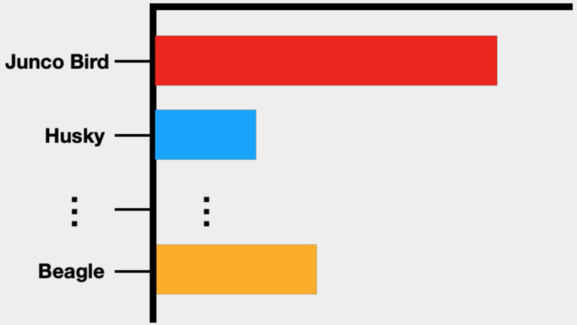
Baseline input

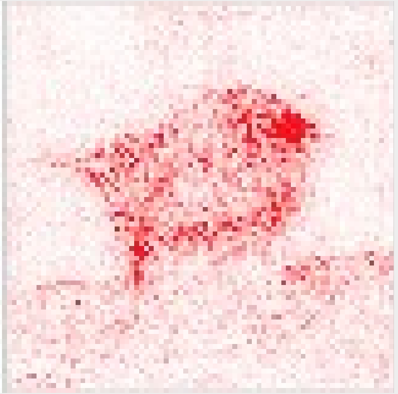
[Sundararajan et. al. 2017]

Integrated Gradients

Model

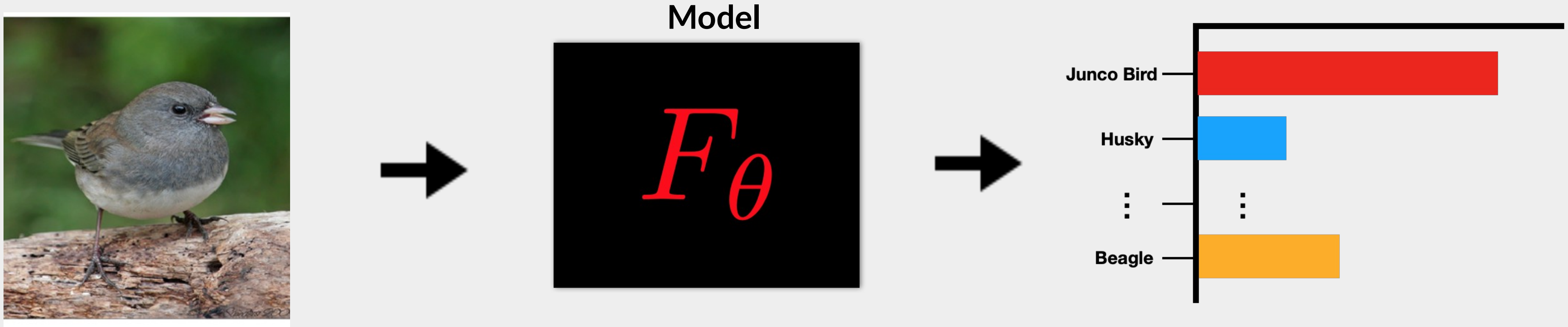
F_θ





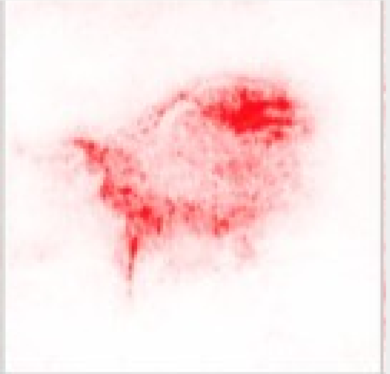
Sundararajan et. al. 2017

SmoothGrad

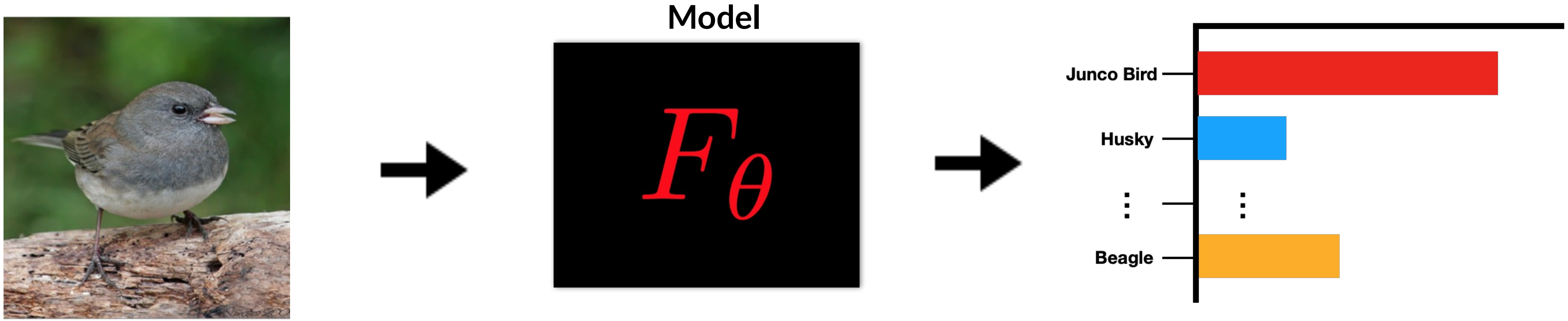


$$\frac{1}{N}\sum_{i}^{N}\nabla_{(x+\epsilon)}F_i(x+\epsilon)$$

Gaussian noise



Guided Backprop: “Modified Backprop”

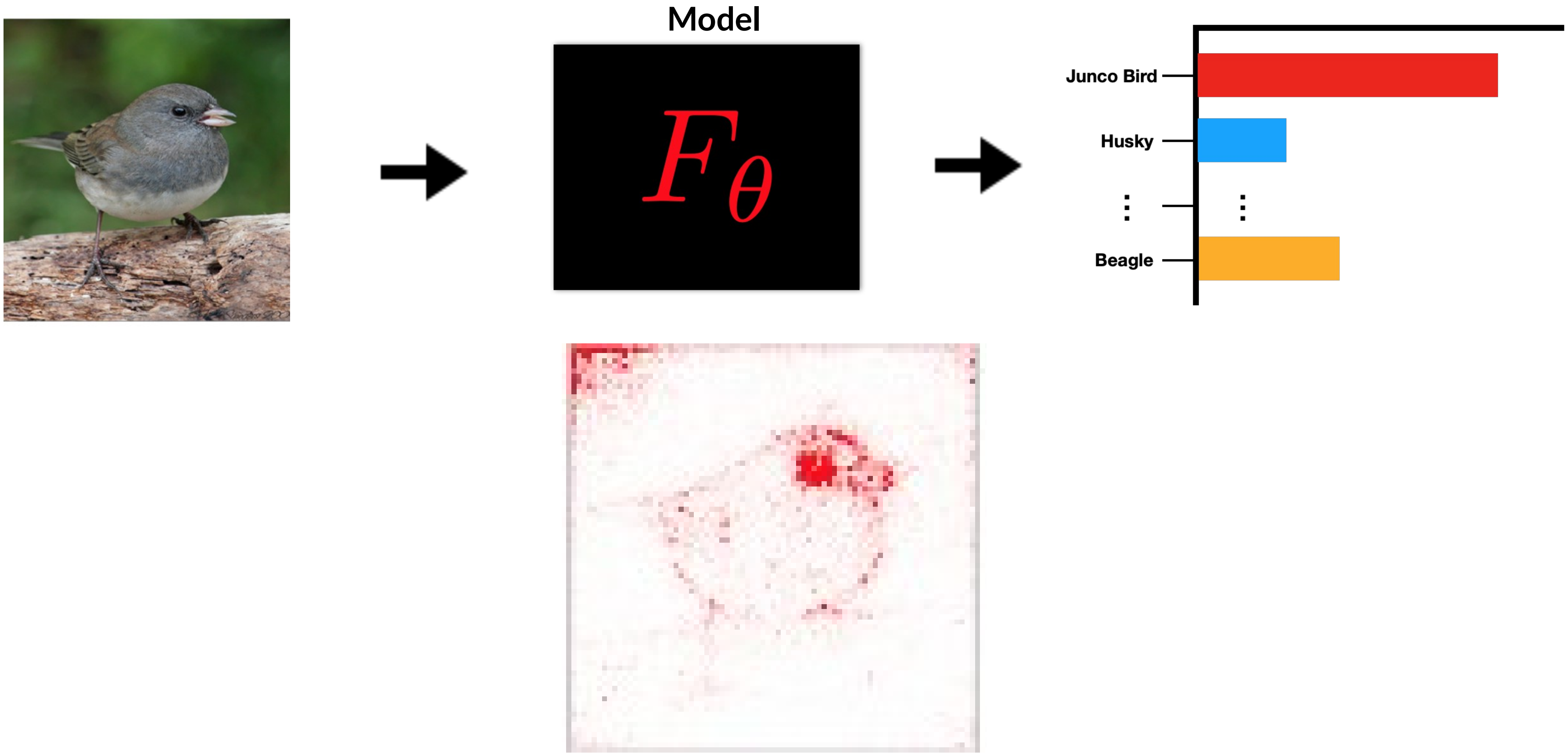


activation: $$f_i^{l+1} = relu(f_i^l) = \max(f_i^l, 0)$$

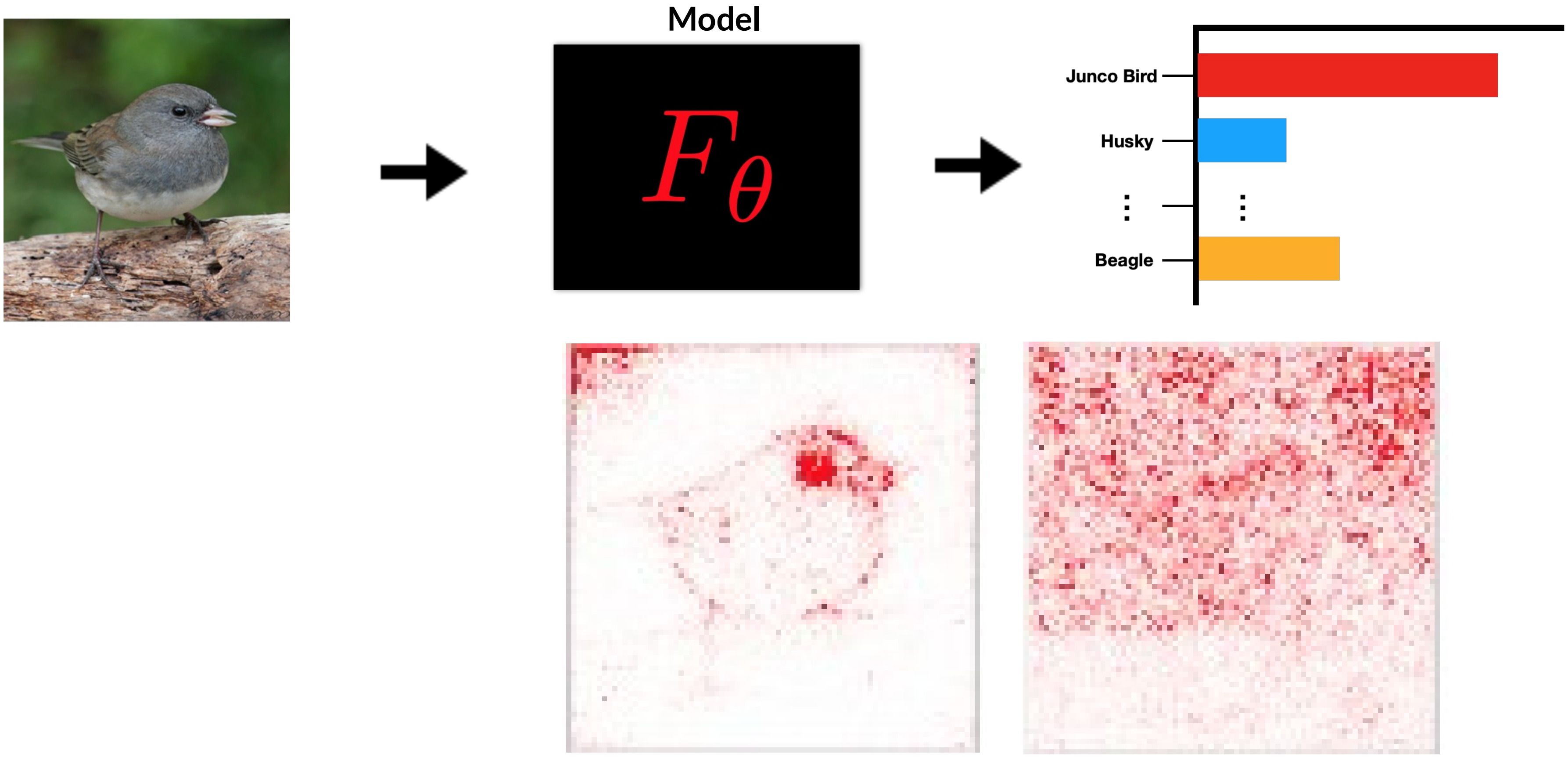
backpropagation: $$R_i^l = (f_i^l > 0) \cdot R_i^{l+1}, \text{ where } R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}$$

guided backpropagation: $$R_i^l = (f_i^l > 0) \cdot \boxed{(R_i^{l+1} > 0)} \cdot R_i^{l+1}$$

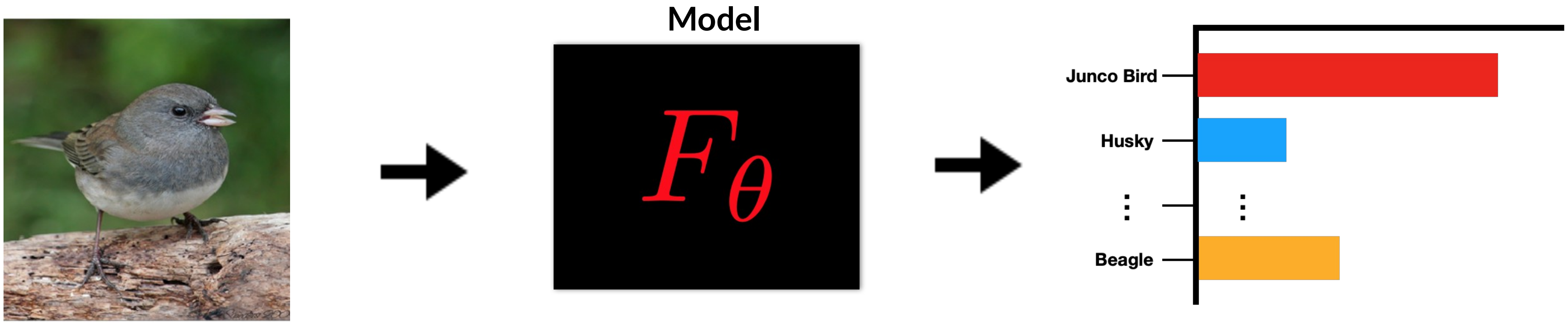
Guided Backprop



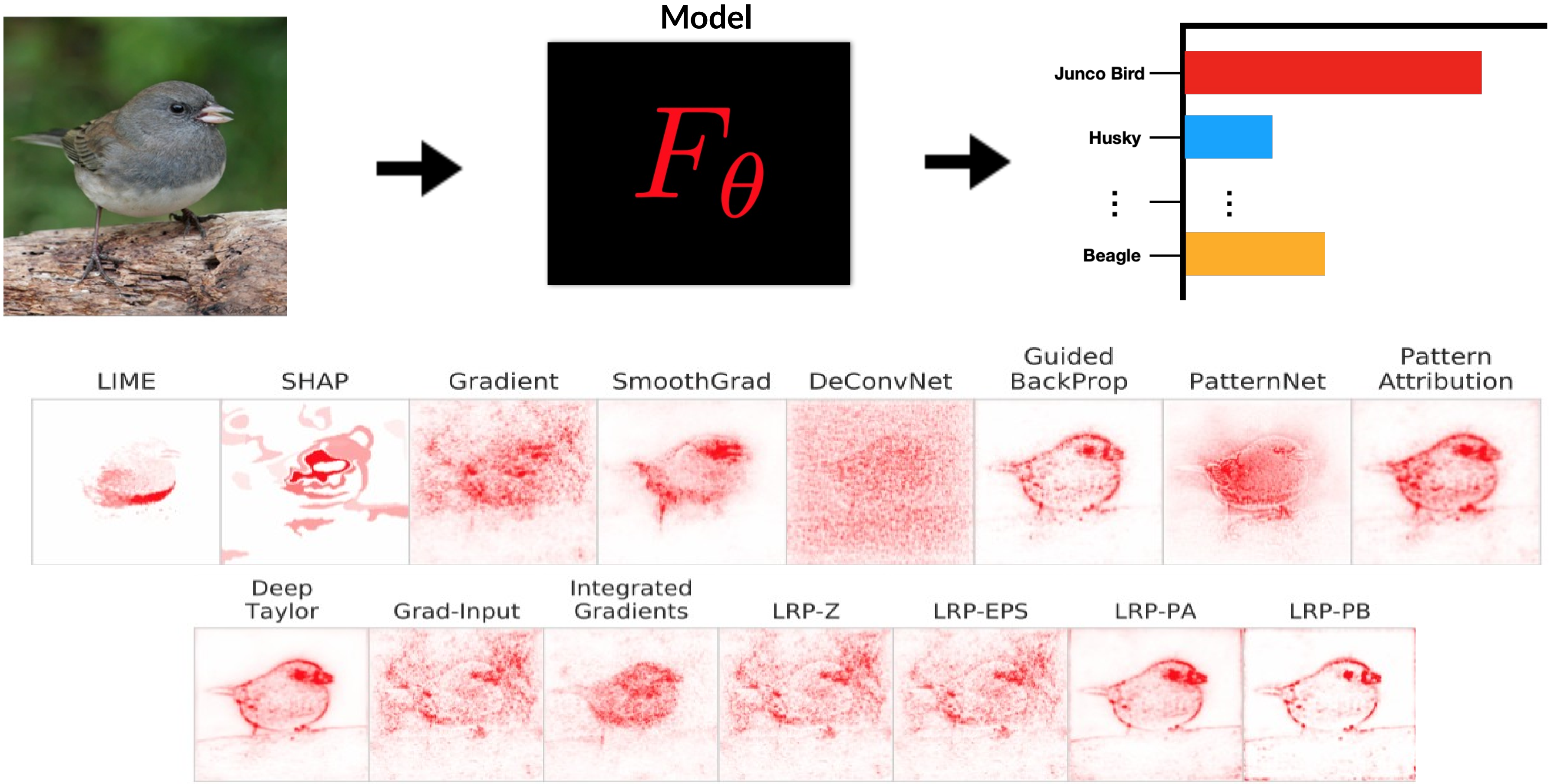
Guided Backprop



Recap



Recap



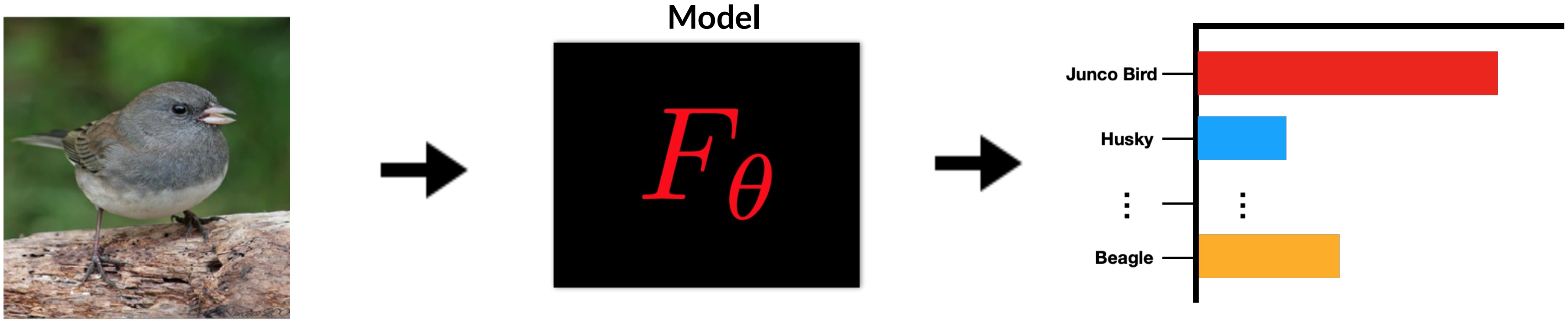
Model
F_θ
Junco Bird
Husky
Beagle
LIME
SHAP
Gradient
SmoothGrad
DeConvNet
Guided BackProp
PatternNet
Pattern Attribution
Deep Taylor
Grad-Input
Integrated Gradients
LRP-Z
LRP-EPS
LRP-PA
LRP-PB

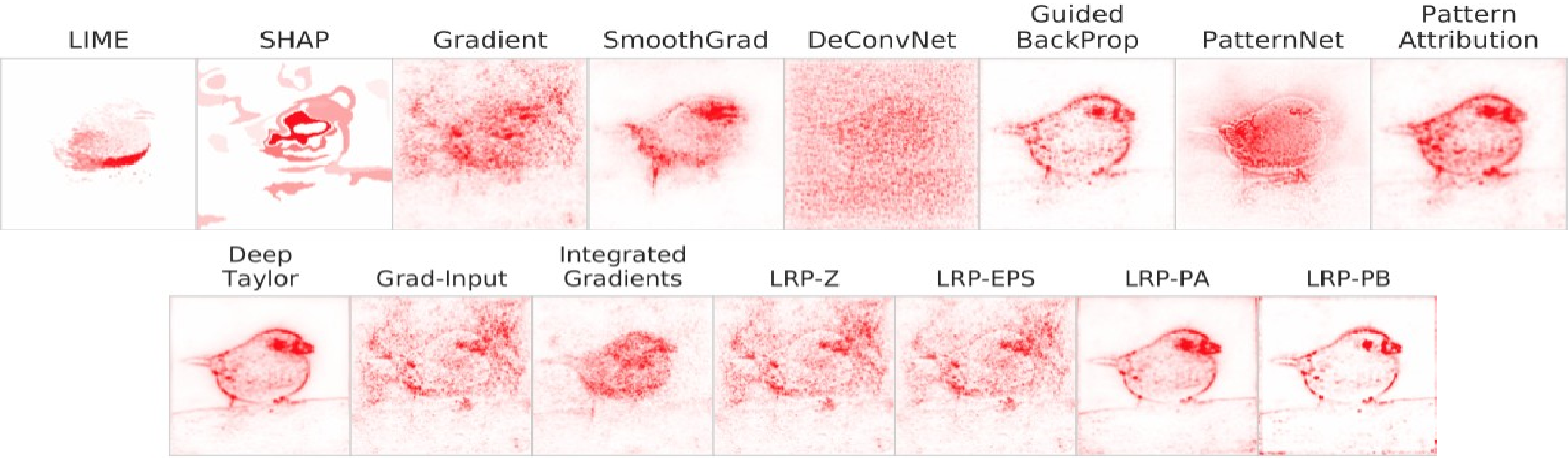
Recap

- **Class Activation Mapping** (Zhou et. al. 2016).
- **Meaningful Perturbation** (Fong et. al. 2017).
- **RISE** (Petsuik et. al. 2018).
- **Extremal Perturbations** (Fong & Patrick 2019).
- **DeepLift** (Shrikumar et. al. 2018).
- **Expected Gradients** (Erion et. al. 2019)
- **Excitation Backprop** (Zhang et. al. 2016)
- **GradCAM** (Selvaraju et. al. 2016)
- **Guided GradCAM** (Selvaraju et. al. 2016)
- **Occlusion** (Zeiler et. al. 2014).
- **Prediction Difference Analysis** (Gu. et. al. 2019).
- **Internal Influence** (Leino et. al. 2018).

See for additional methods: Samek & Montavon et. al. 2020

Recap: which method should you use?





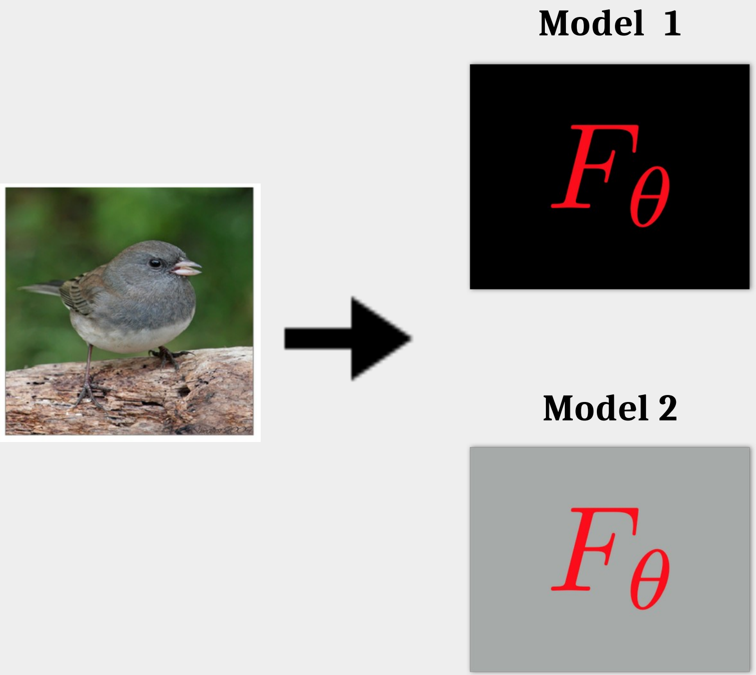
‘Sanity Checks’

Intuitive ‘principles’ that an attribution method should satisfy.

- **‘Model’ Faithfulness**: is the ‘explanation’ sensitive to model parameters?
  - Test: change the model weights and measure corresponding change in explanation.
  - Operationalize by reinitialization of model weights.
- **Data Faithfulness**: is the attribution sensitive to training data?
  - Test: change training label and measure corresponding change in explanation.
  - Operationalize by randomization labelling in training data.

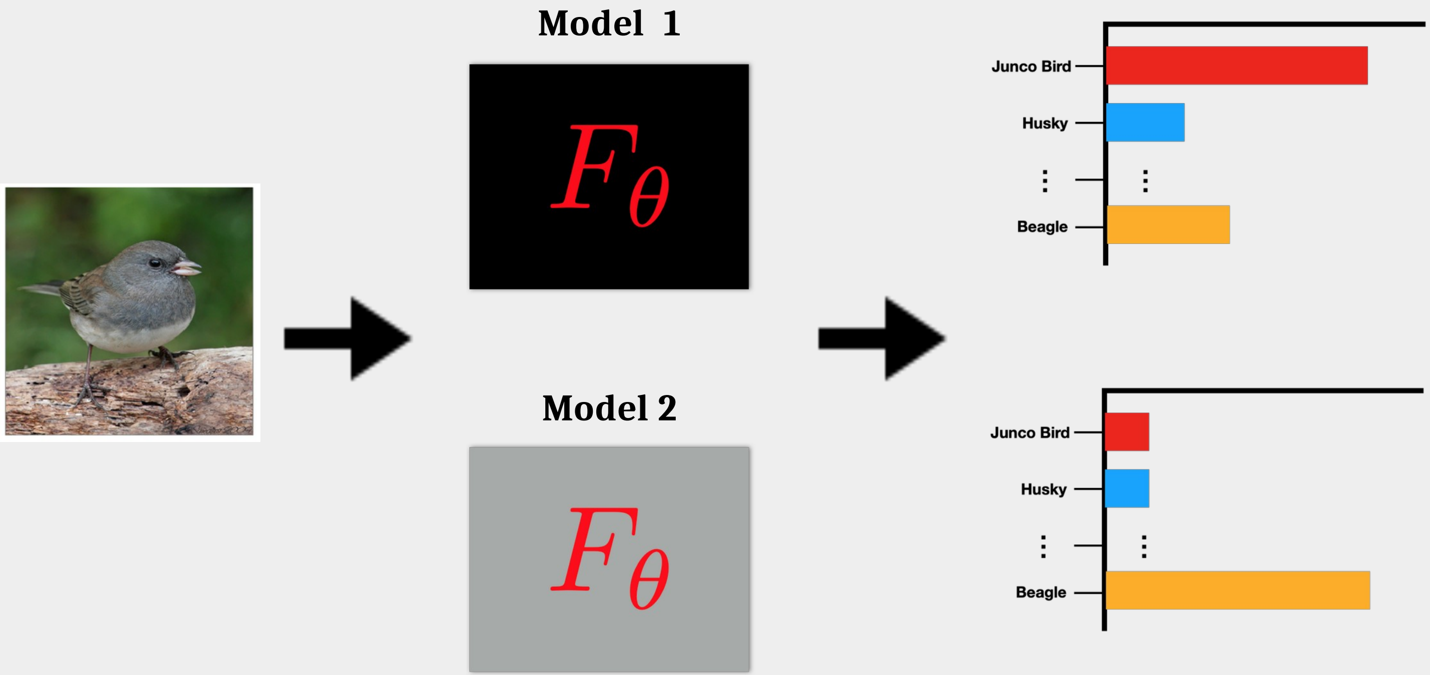
Sensitivity to model parameters

If the parameter settings change of model changes the saliency map should change.



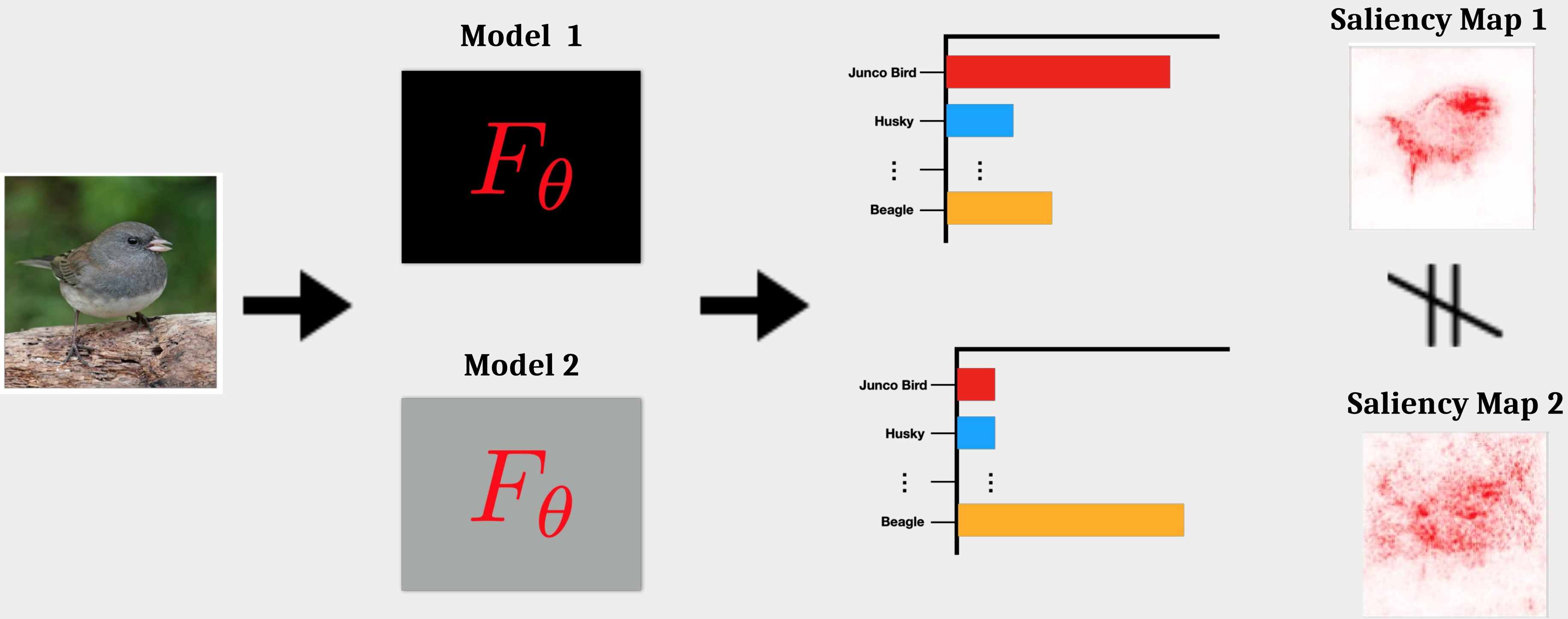
Sensitivity to model parameters

If the parameter settings change of model changes the saliency map should change.

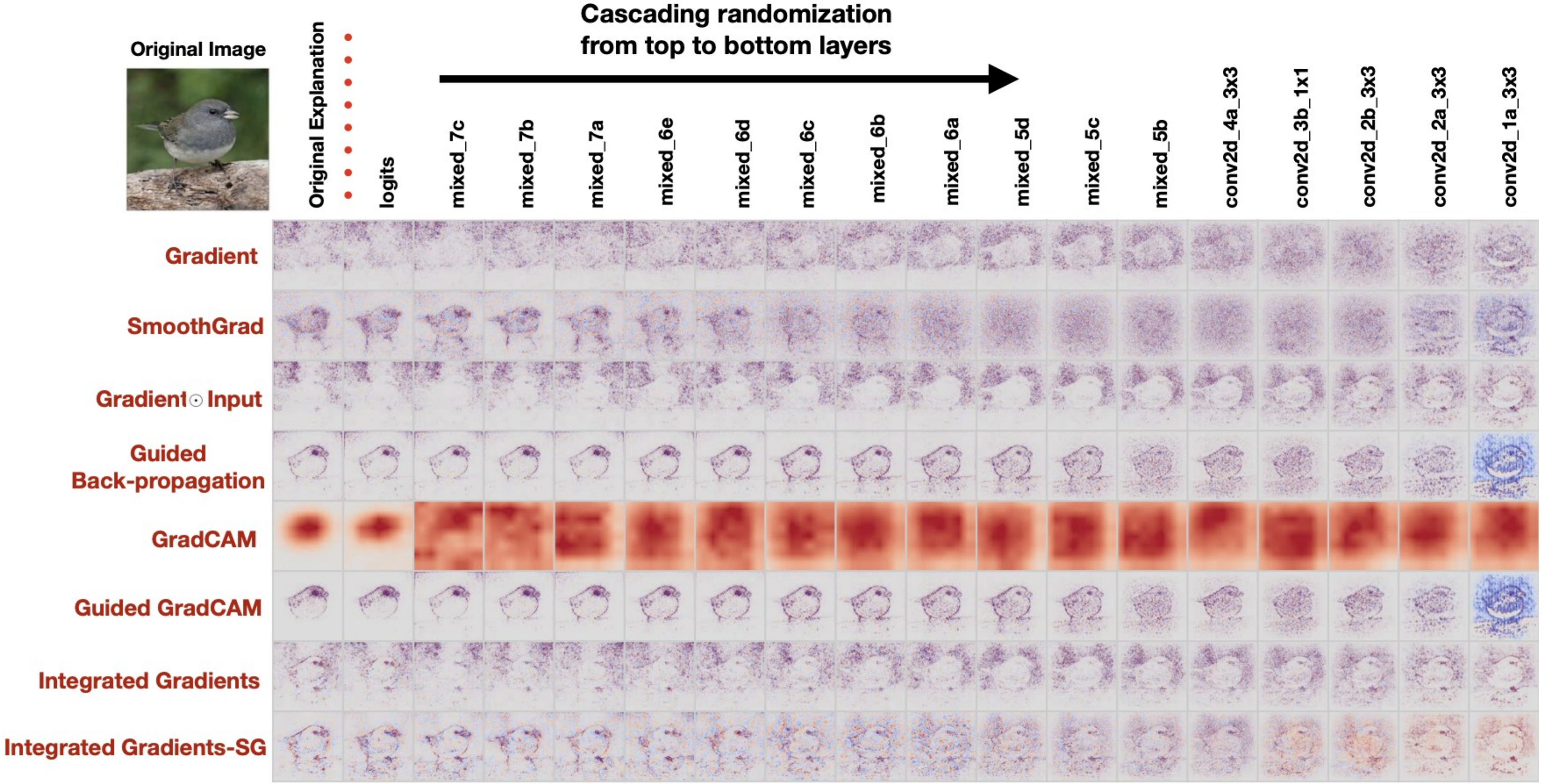


Sensitivity to model parameters

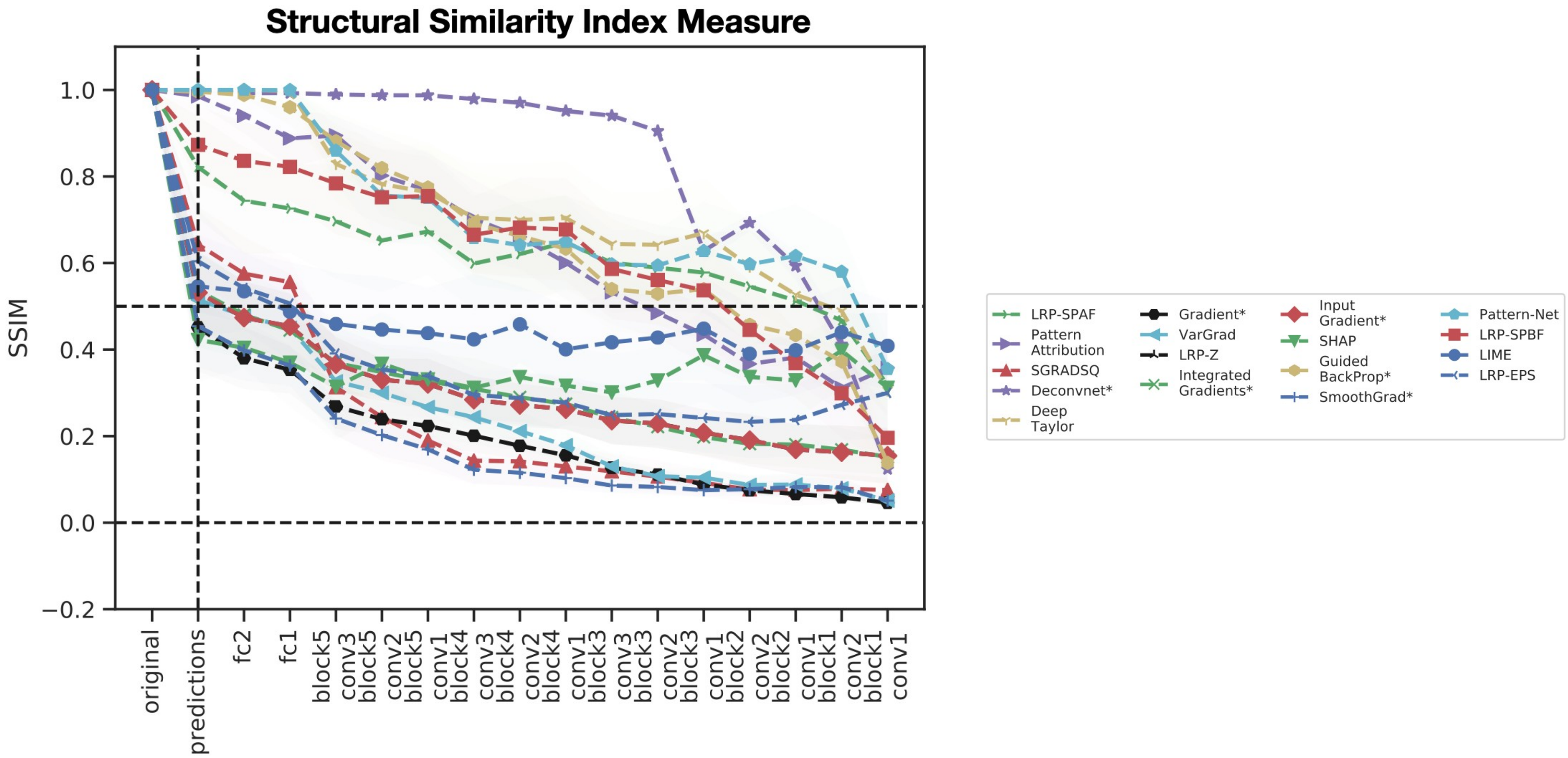
If the parameter settings change of model changes the saliency map should change.



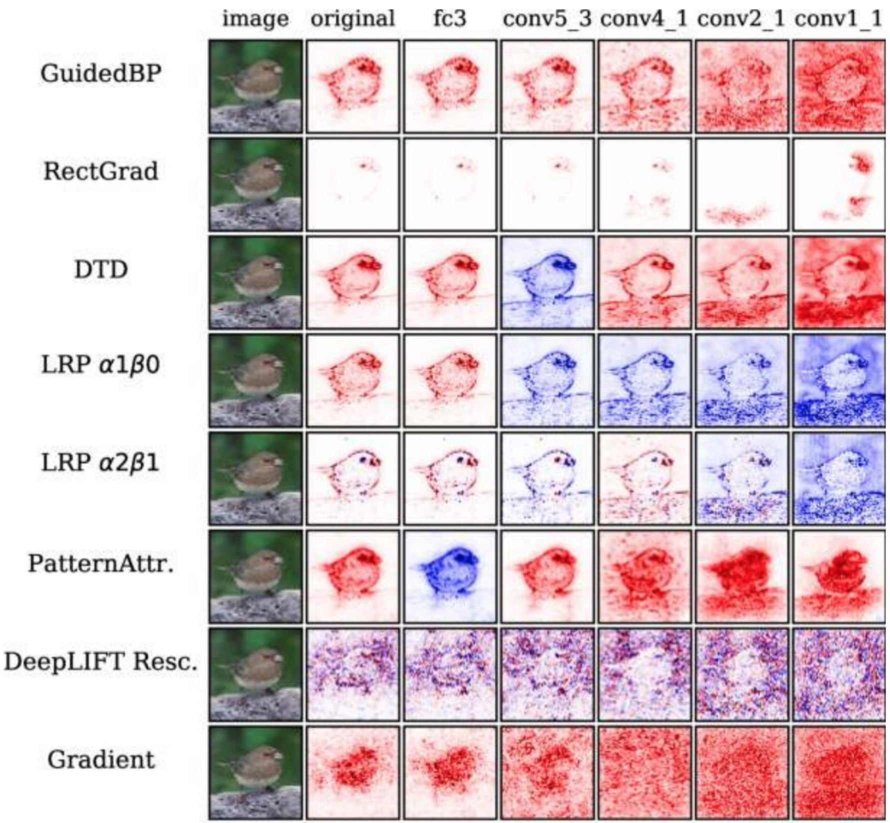
Sensitivity to model parameters



Sensitivity to model parameters

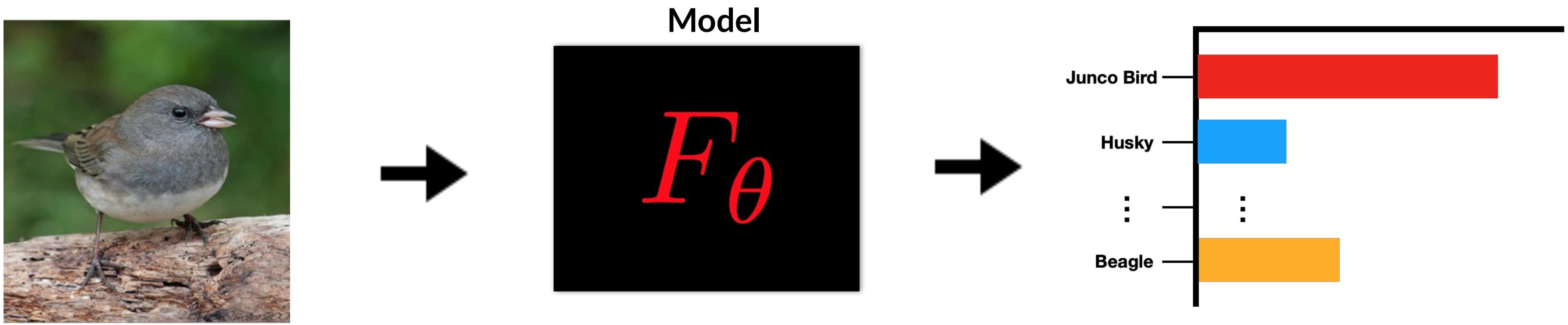


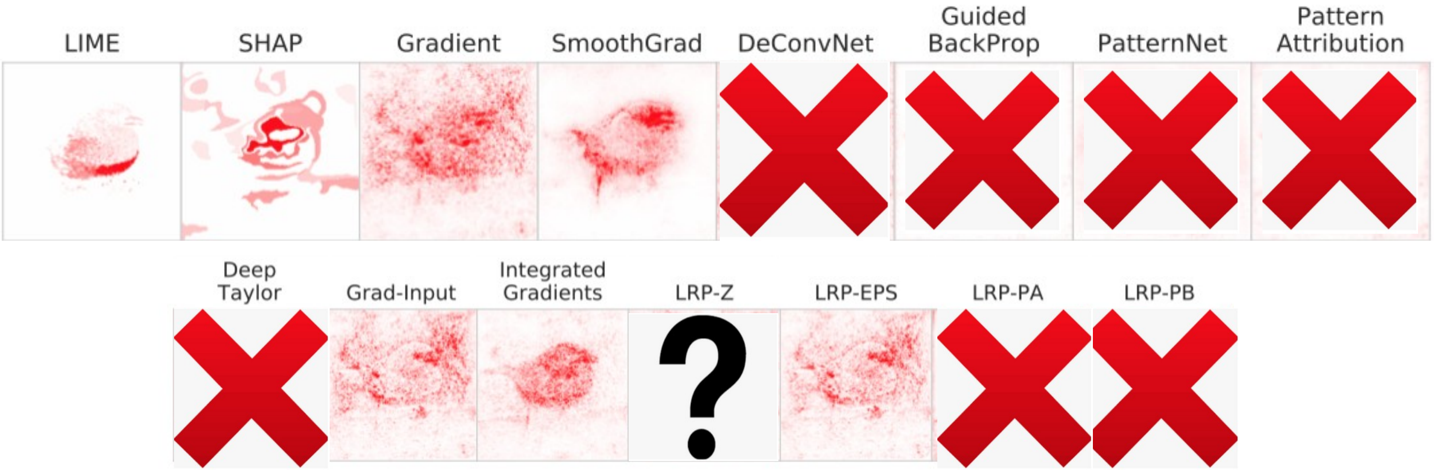
Modified BackProp Approaches



These modified backprop methods converge to a rank-1 matrix! This is because the product of a sequence of non-negative matrices (non-orthogonal columns) converges to a rank-1 matrix (*Theorem 1 in Sixt et. al. 2020*).

Recap: which method should you use?





Some Takeaways

- Identified certain classes of feature attribution methods that are invariant to higher layer weights.
- ‘Sanity Checks’ are actually ‘weak’ requirements, i.e., does not tell you whether a method is effective.

Some objections

Causal reframing suggests that sanity checks results might be task specific.

Revisiting Sanity Checks for Saliency Maps

Gal Yona
Weizmann Institute of Science

Daniel Greenfeld
Jether Energy Research

On the Relationship Between Explanation and Prediction: A Causal View

Amir-Hossein Karimi[1 2 3] **Krikamol Muandet**[4] **Simon Kornblith**[3] **Bernhard Schölkopf**[1] **Been Kim**[3]

Some objections

Where you choose to perform randomization matters, and perhaps the weight randomization is not the best approach.

Shortcomings of Top-Down Randomization-Based Sanity Checks for Evaluations of Deep Neural Network Explanations

Alexander Binder[1,2][0000-0001-9605-6209], Leander Weber[3], Sebastian Lapuschkin[3][0000-0002-0762-7258], Grégoire Montavon[4,5][0000-0001-7243-6186], Klaus-Robert Müller[5,6,7,8], and Wojciech Samek[3,5,6][0000-0002-6283-3265]

More recent observations: Spurious

Beyond faithfulness, it is unclear whether these feature attribution methods are effective for model debugging.

Do Feature Attribution Methods Correctly Attribute Features?

Yilun Zhou[1], Serena Booth[1], Marco Tulio Ribeiro[2], Julie Shah[1]

[1]MIT CSAIL, [2]Microsoft Research

[1]{yilun, serenabooth, julie_a_shah}@csail.mit.edu, [2]marcotcr@microsoft.com

More recent observations

Do Input Gradients Highlight Discriminative Features?

Harshay Shah* Microsoft Research India harshay@google.com

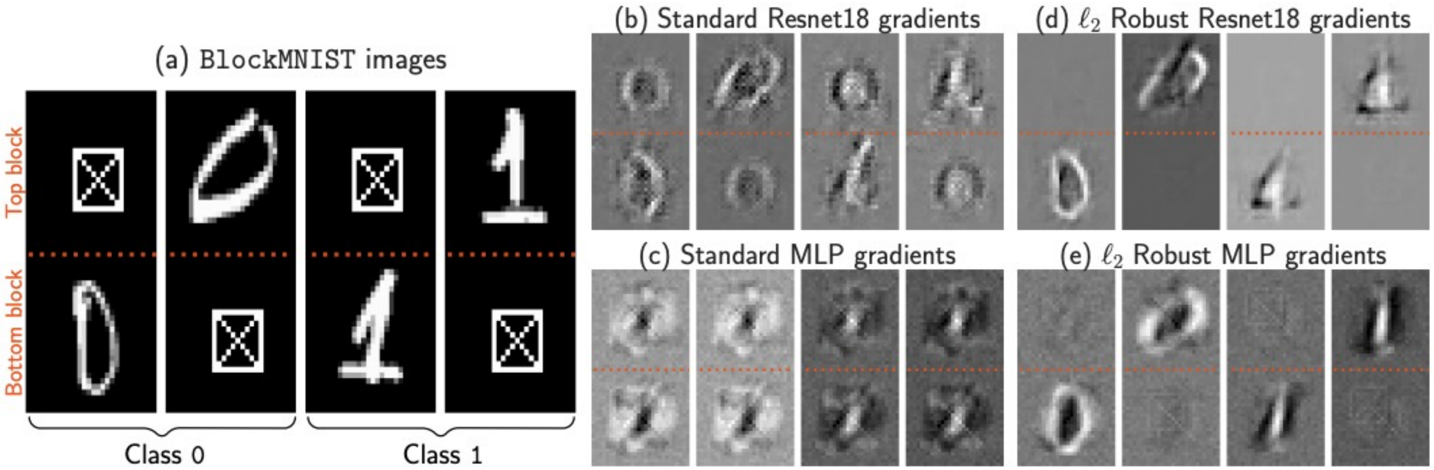
Prateek Jain* Microsoft Research India prajain@google.com

Praneeth Netrapalli* Microsoft Research India pnetrapalli@google.com

RETHINKING THE ROLE OF GRADIENT-BASED ATTRIBUTION METHODS FOR MODEL INTERPRETABILITY

Suraj Srinivas Idiap Research Institute & EPFL suraj.srinivas@idiap.ch

François Fleuret University of Geneva francois.fleuret@unige.ch



Parting Thoughts

Feature attribution is still important for applications, however, additional is needed to characterize the properties of DNN model training that will result in 'gradients' that capture discriminative signals.