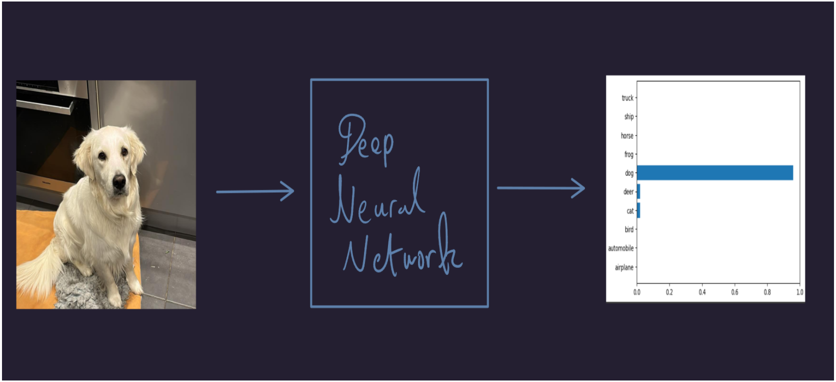
A Benchmark for Interpretability Methods in DNNs
(Google Brain)

Sam Laing

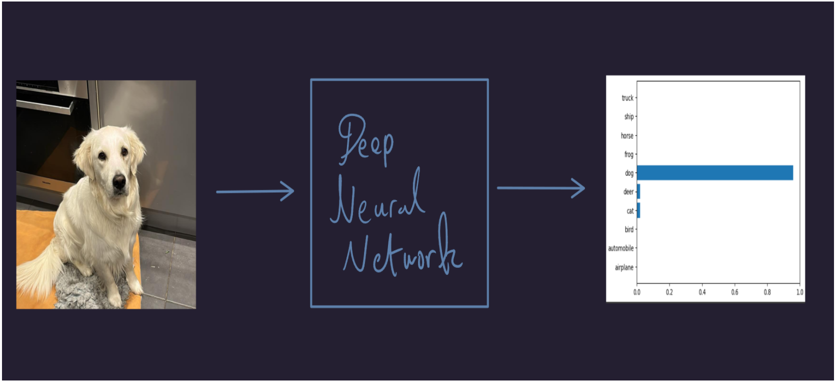
University of Tuebingen

June 26, 2024

A Bit of Background

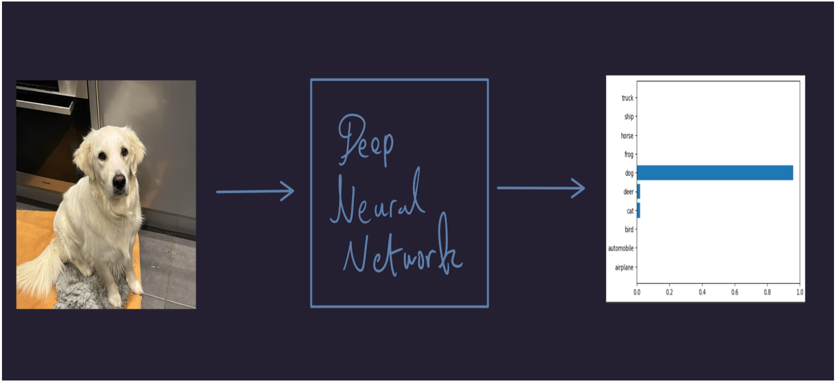


A Bit of Background



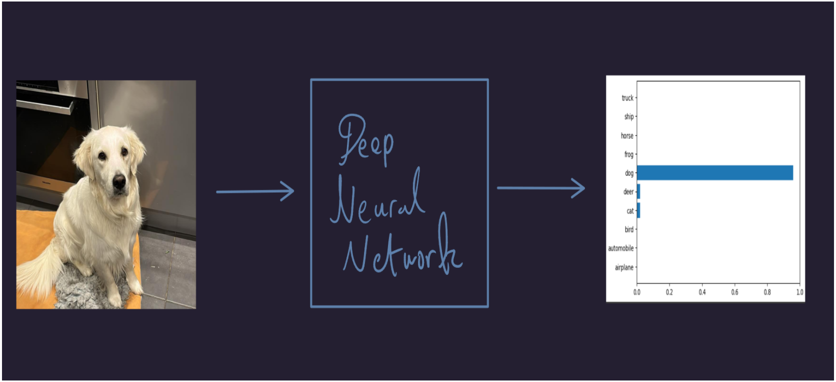
- Deep image classification: "features" = pixels .

A Bit of Background



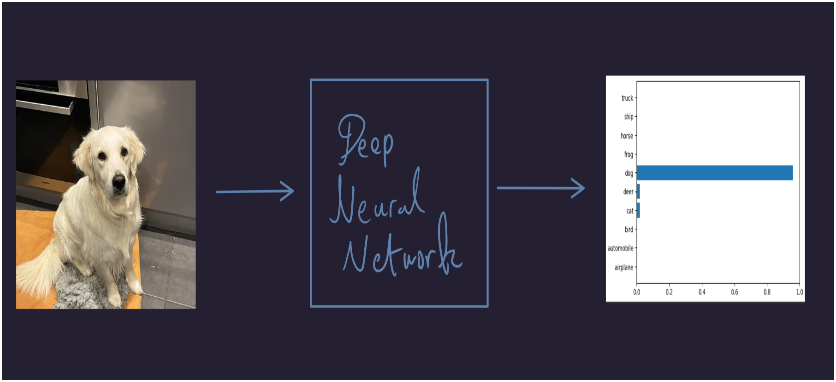
- Deep image classification: "features" = pixels .
- Interpretability methods → help engineer understand their model

A Bit of Background



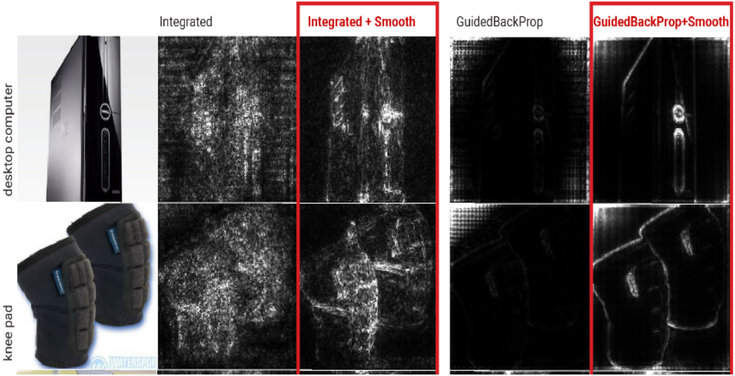
- Deep image classification: "features" = pixels .
- Interpretability methods → help engineer understand their model
- Ostensibly. But are they really doing anything?
- Interpretability method A > Interpretability method B??

A Bit of Background



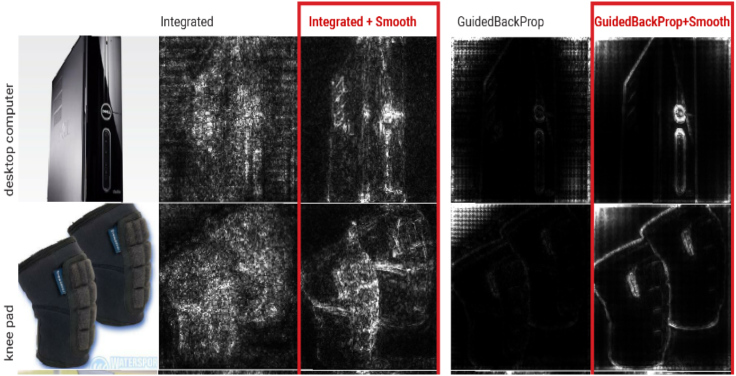
- Deep image classification: "features" = pixels .
- Interpretability methods → help engineer understand their model
- Ostensibly. But are they really doing anything?
- Interpretability method A > Interpretability method B??
- If only there was a benchmarking framework to do this...

Included Interpretability Methods



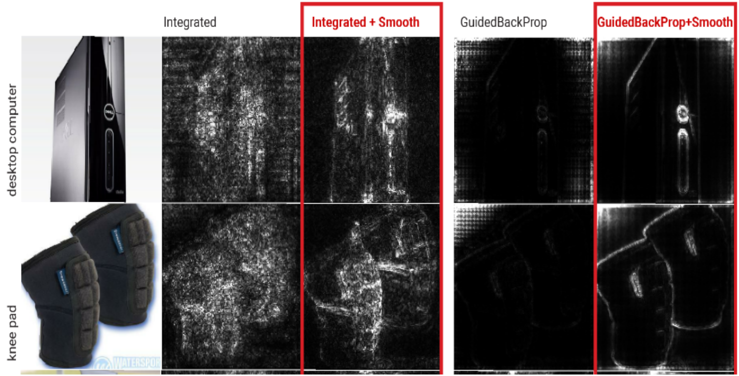
- Gradients (sensitivity heatmaps) $e = \partial_{x_i} f_c(x)$

Included Interpretability Methods



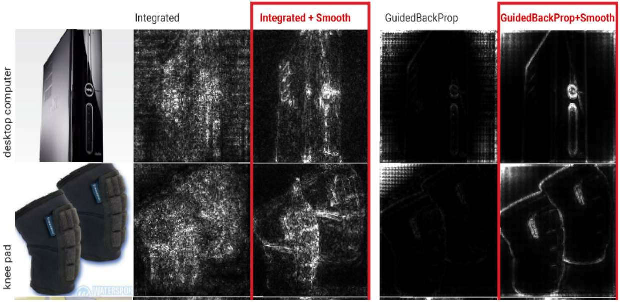
- Gradients (sensitivity heatmaps) $e = \partial_{x_i} f_c(x)$
- Guided Backprop (sort of a tidied up sensitivity map). Keep positives in ReLU

Included Interpretability Methods



- Gradients (sensitivity heatmaps) $e = \partial_{x_i} f_c(x)$
- Guided Backprop (sort of a tidied up sensitivity map). Keep positives in ReLU
- Integrated Gradients

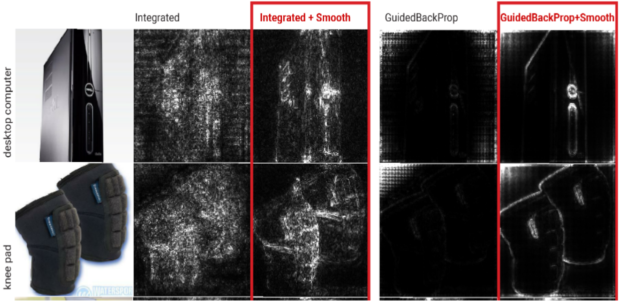
Included Interpretability Methods: Ensembling in a Nutshell



Smilkov's "Smoothgrad" paper (2017):
"Reduce noise by adding noise"

- Inject inputs with Gaussian noise consider mean/variance of outputs

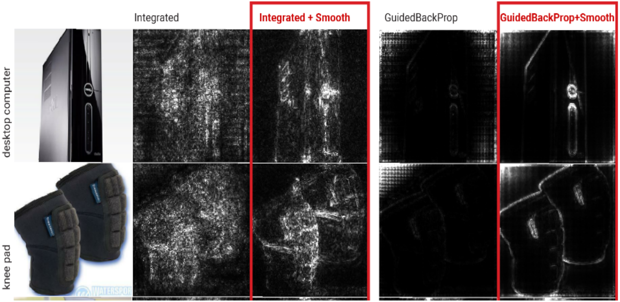
Included Interpretability Methods: Ensembling in a Nutshell



Smilkov's "Smoothgrad" paper (2017):
"Reduce noise by adding noise"

- Inject inputs with Gaussian noise consider mean/variance of outputs
- $\eta_i \sim N(0, \sigma^2 I)$ $i \in \{1, \ldots, J\}$

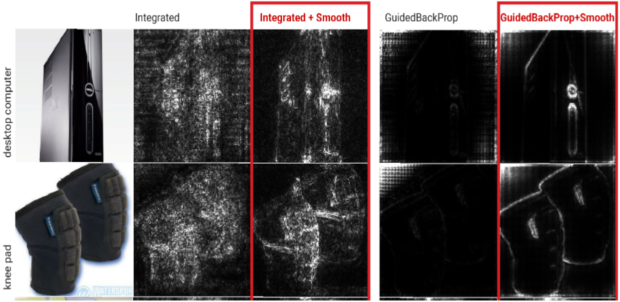
Included Interpretability Methods: Ensembling in a Nutshell



Smilkov's "Smoothgrad" paper (2017):
"Reduce noise by adding noise"

- Inject inputs with Gaussian noise consider mean/variance of outputs
- $\eta_i \sim N(0, \sigma^2 I)$ $i \in \{1, \ldots, J\}$
- SmoothGrad (SG) $e = \sum_{i=1}^{J} f_c(x + \eta_i)$

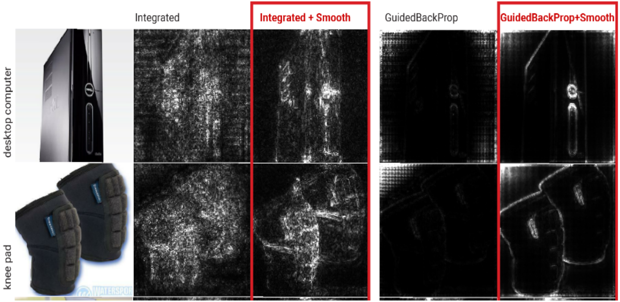
Included Interpretability Methods: Ensembling in a Nutshell



Smilkov's "Smoothgrad" paper (2017): "Reduce noise by adding noise"

- Inject inputs with Gaussian noise consider mean/variance of outputs
- $\eta_i \sim N(0, \sigma^2 I)$ $i \in \{1, \ldots, J\}$
- SmoothGrad (SG) $e = \sum_{i=1}^{J} f_c(x + \eta_i)$
- VarGrad (VAR) $e = \mathrm{Var}(\{f_c(x + \eta_i)\}_{i=1}^{J})$

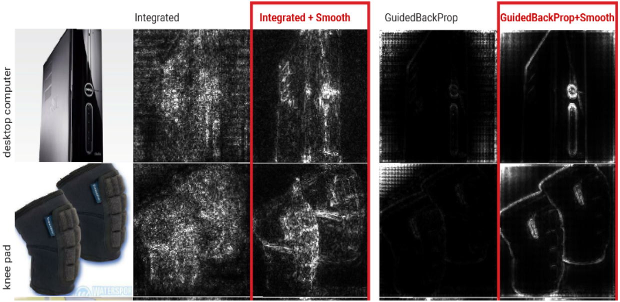
Included Interpretability Methods: Ensembling in a Nutshell



Smilkov's "Smoothgrad" paper (2017): "Reduce noise by adding noise"

- Inject inputs with Gaussian noise consider mean/variance of outputs
- $\eta_i \sim N(0, \sigma^2 I)$ $i \in \{1, \ldots, J\}$
- SmoothGrad (SG) $e = \sum_{i=1}^{J} f_c(x + \eta_i)$
- VarGrad (VAR) $e = \mathrm{Var}(\{f_c(x + \eta_i)\}_{i=1}^{J})$
- SmoothGrad Squared (SG-SQ) $e = \sum_{i=1}^{J} f_c(x + \eta_i)^2$

Included Interpretability Methods: Ensembling in a Nutshell



Smilkov's "Smoothgrad" paper (2017): "Reduce noise by adding noise"

- Inject inputs with Gaussian noise consider mean/variance of outputs
- $\eta_i \sim N(0, \sigma^2 I)$ $i \in \{1, \ldots, J\}$
- SmoothGrad (SG) $e = \sum_{i=1}^{J} f_c(x + \eta_i)$
- VarGrad (VAR) $e = \mathrm{Var}(\{f_c(x + \eta_i)\}_{i=1}^{J})$
- SmoothGrad Squared (SG-SQ) $e = \sum_{i=1}^{J} f_c(x + \eta_i)^2$

$\rightarrow$Apply attribution/interpretability methods to these statistics

ROAR (RemOve And Retrain)

The Idea Behind ROAR

Start with trained classifier f

The Idea Behind ROAR

Start with trained classifier f

$\forall$ method, $\forall$ image $\in$ dataset, sort pixels by ranked importance.

The Idea Behind ROAR

Start with trained classifier f

$\forall$ method, $\forall$ image $\in$ dataset, sort pixels by ranked importance.

So $(e_j)_{j=1}^{D}$ of pixel coordinates $\forall$ image in dataset. $\implies$ $((e_j^{(n)})_{j=1}^{D})_{n=1}^{N}$

The idea behind ROAR

for $j \in \{0, 10, \ldots, 100\}$, replace the top j% ranked pixels with the per channel mean $\forall$ image and retrain.

Proportion: 10%

The idea behind ROAR

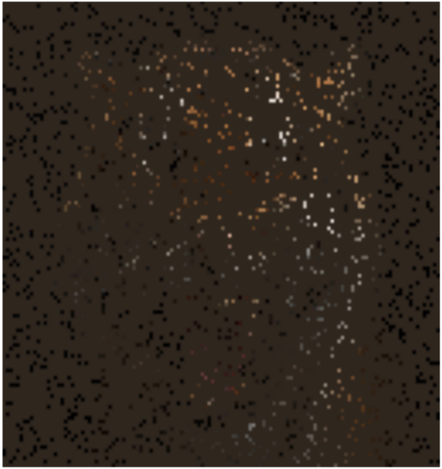
The idea behind ROAR

Proportion: 50%

The idea behind ROAR

The idea behind ROAR

Proportion: 90%



The idea behind ROAR

- Effect of having dropped the "most informative pixels" as determined by each interpretability method.
- Investigate how much their removal from the training process effects accuracy.
- Also a no-retraining variant

To Retrain Or not To Retrain

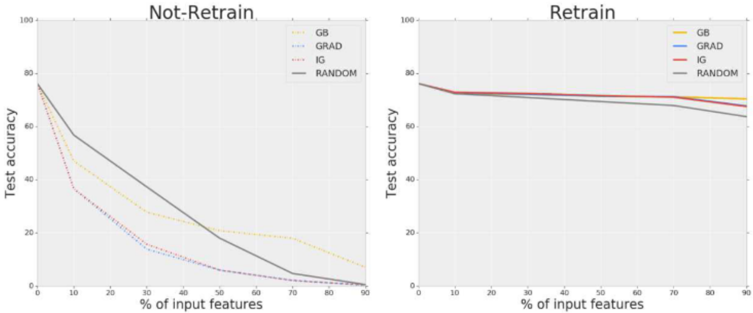
- No retraining $\implies p_{\text{ train}} \neq p_{\text{ test}} \ldots \rightarrow \leftarrow$ in ML

To Retrain Or not To Retrain

- No retraining $\implies p_{\text{train}} \neq p_{\text{test}} \ldots \rightarrow\leftarrow$ in ML
- Paper therefore argues it is necessary

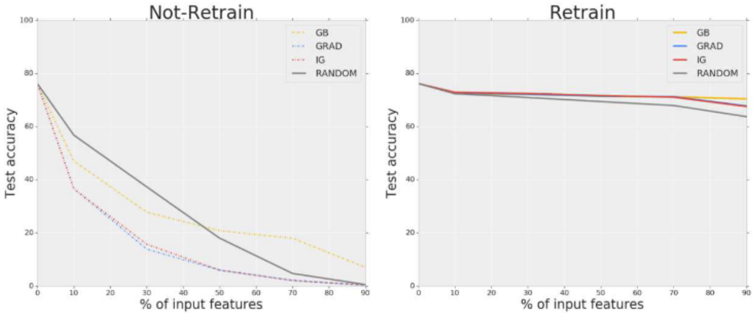
To Retrain Or not To Retrain

- No retraining $\implies p_{\text{train}} \neq p_{\text{test}} \ldots \rightarrow \leftarrow$ in ML
- Paper therefore argues it is necessary



To Retrain Or not To Retrain

- No retraining $\implies p_{\text{ train}} \neq p_{\text{ test}} \ldots \rightarrow \leftarrow$ in ML
- Paper therefore argues it is necessary



- Paper does some validation with synthetic data but unconvincing.

An Outline of the Experiment

Really just a refinement of above.

- ResNet50 classifier: Imagenet, Birdsnap and Food 101

An Outline of the Experiment

Really just a refinement of above.

- ResNet50 classifier: Imagenet, Birdsnap and Food 101
- Random pixel selection and Sobel Edge filter benchmarks.

An Outline of the Experiment

Really just a refinement of above.

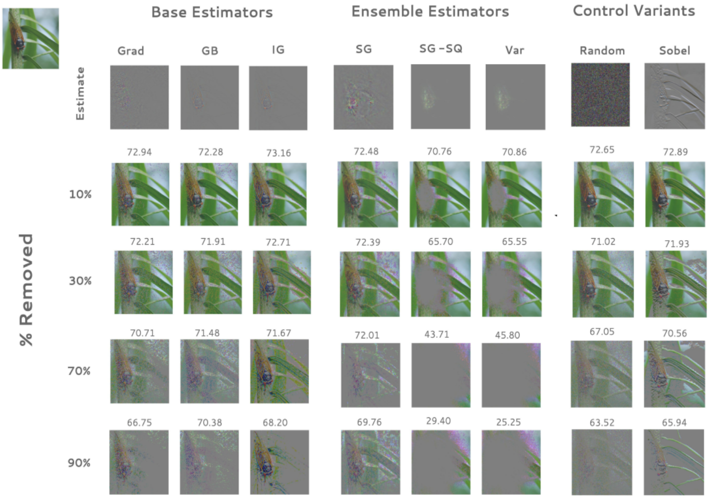
- ResNet50 classifier: Imagenet, Birdsnap and Food 101
- Random pixel selection and Sobel Edge filter benchmarks.
- New train and test sets $\forall\, j \in \{0, 10, 30, 50, 70, 90\}$

An Outline of the Experiment

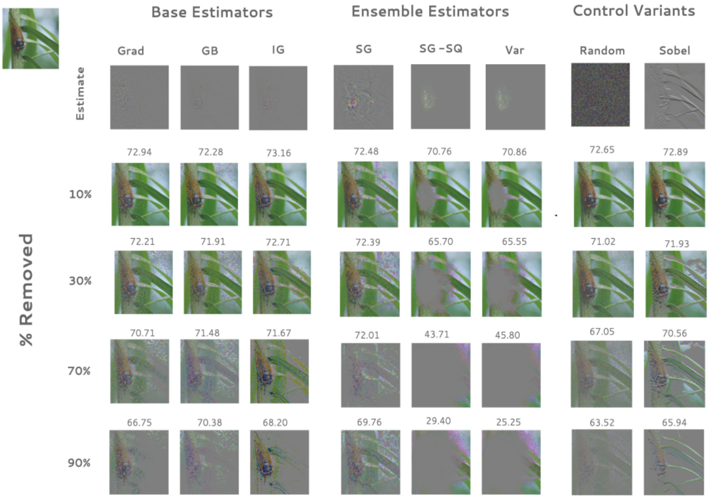
Really just a refinement of above.

- ResNet50 classifier: Imagenet, Birdsnap and Food 101
- Random pixel selection and Sobel Edge filter benchmarks.
- New train and test sets $\forall\, j \in \{0, 10, 30, 50, 70, 90\}$
- Each model retrained 5 times $\forall$ method (DNN training is noisy)

ROAR in Action

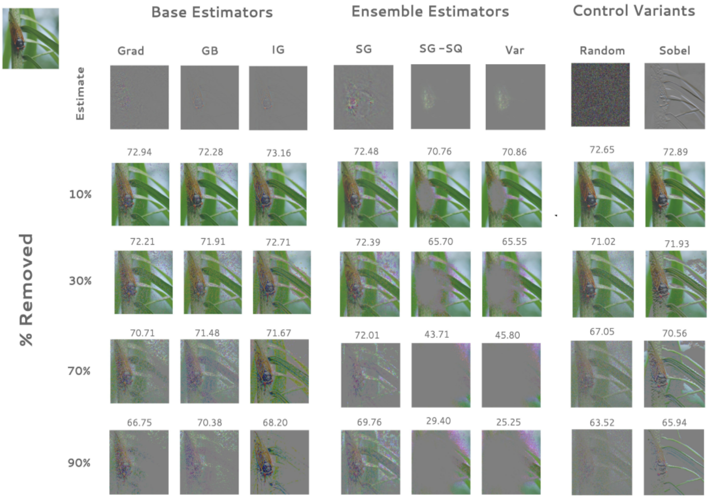


ROAR in Action



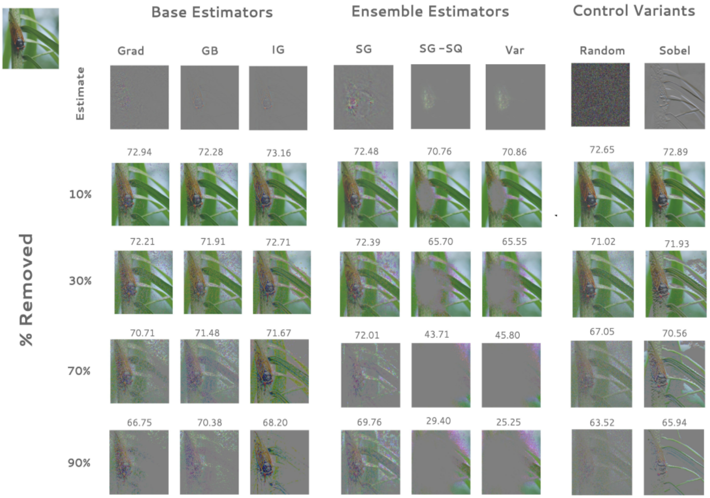
Initial Imagenet accuracy: ~ 78 % ...

ROAR in Action



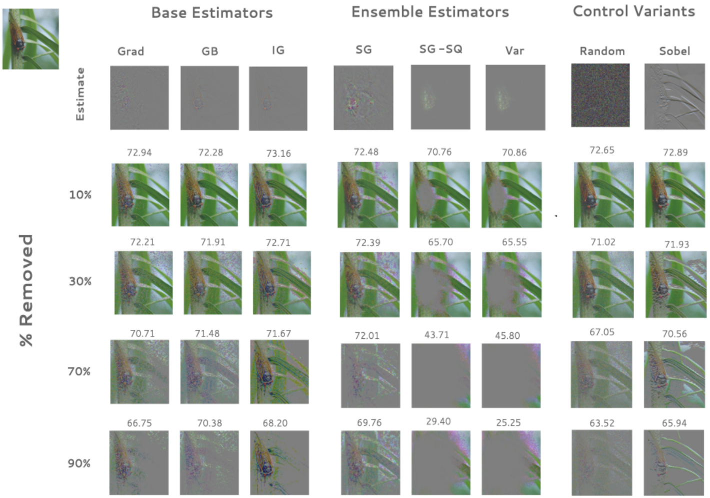
Initial Imagenet accuracy: ~ 78 % ... Paper is from 2018!

ROAR in Action



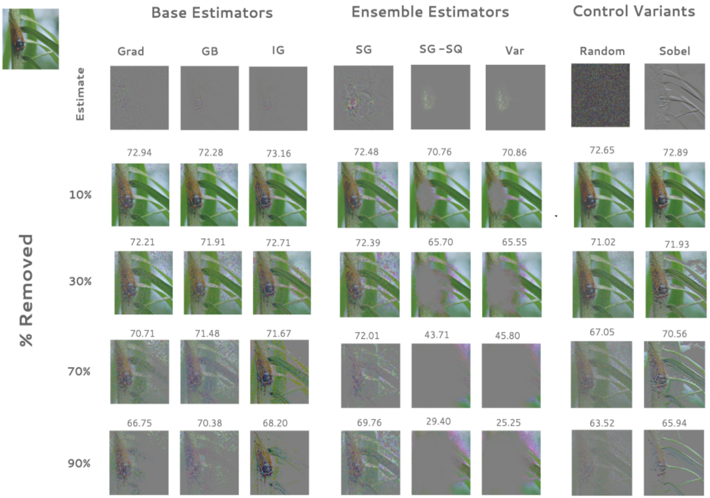
On average Replacing many pixels $\implies$ big decrease of predictive power!

Results



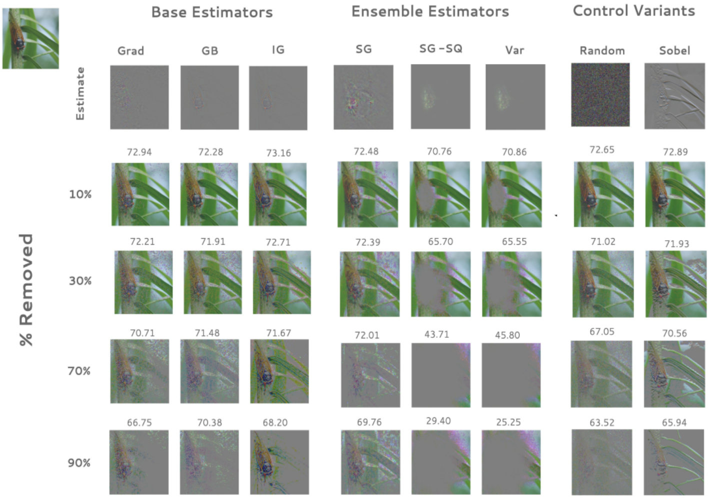
ImageNet, 90% pixels randomly removed... still 63.52% accuracy relative to the original 78.68%

Results



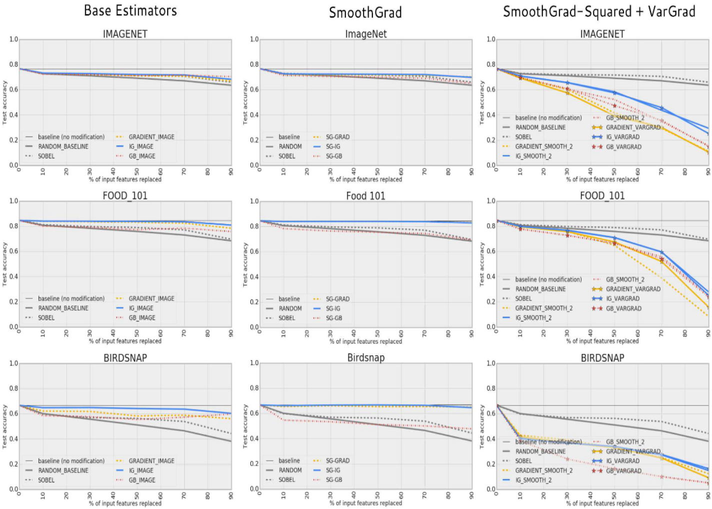
ImageNet, 90% pixels randomly removed... still 63.52% accuracy relative to the original 78.68%

Results



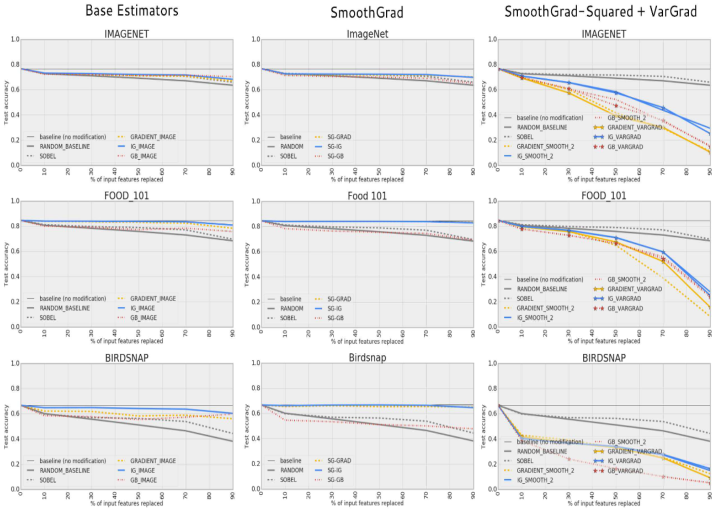
According to the paper, SG-SQ and VarGrad are the real heros

Plots



SG - SQ and VarGrad always outperformed...

Plots



SG - SQ and VarGrad always outperformed...

But best method to wrap around changed

My Thoughts (Dummy Disclaimer!)

- Why VAR and SG-SQ over Vanilla SG?

My Thoughts (Dummy Disclaimer!)

- Why VAR and SG-SQ over Vanilla SG?
- Variance captures more of the distribution's properties than mean?

My Thoughts (Dummy Disclaimer!)

- Why VAR and SG-SQ over Vanilla SG?
- Variance captures more of the distribution's properties than mean?
- SG-SQ something of a middleground as unnormalized

My Thoughts (Dummy Disclaimer!)

- Why VAR and SG-SQ over Vanilla SG?
- Variance captures more of the distribution's properties than mean?
- SG-SQ something of a middleground as unnormalized

$$\text{SG-SQ: } e_2 := \sum_{i=1}^{J} f_{c_i}(x+\eta_i)^2 \quad , \quad \text{SG: } e := \sum_{i=1}^{J} f_{c_i}(x+\eta_i)$$

$$\implies \partial_{x_d} e_2 = 2e.\partial_{x_d} e \quad \forall d$$

Gradient of e_2 might encode more info since mean explicitely there

My Thoughts (Dummy Disclaimer!)

- Why VAR and SG-SQ over Vanilla SG?
- Variance captures more of the distribution's properties than mean?
- SG-SQ something of a middleground as unnormalized

$$\text{SG-SQ: } e_2 := \sum_{i=1}^{J} f_{c_i}(x+\eta_i)^2 \quad , \quad \text{SG: } e := \sum_{i=1}^{J} f_{c_i}(x+\eta_i)$$

$$\implies \partial_{x_d} e_2 = 2e.\partial_{x_d} e \quad \forall d$$

Gradient of e_2 might encode more info since mean explicitely there

A Few Possible Issues in the Approach

- Replace the top j pixels the mean.

A Few Possible Issues in the Approach

- Replace the top j pixels the mean.
- Still conveys possibly useful information

A Few Possible Issues in the Approach

- Replace the top j pixels the mean.
- Still conveys possibly useful information
- "Missingness" not well defined for images

A Few Possible Issues in the Approach

- Replace the top j pixels the mean.
- Still conveys possibly useful information
- "Missingness" not well defined for images

Another possible Issue

- Cost!
- More compute needed for big datasets

Another possible Issue

- Cost!
- More compute needed for big datasets
- Retraining a large image classifier many times may be unfeasible

Another possible Issue

- Cost!
- More compute needed for big datasets
- Retraining a large image classifier many times may be unfeasible
- Best depends on dataset
- Without retraining, you run into theoretical violations of ML!

Why I chose this paper

Why I chose this paper

So many interpretability methods boast high performance: How to pick the right one?
Hard to find strong quantitative statements about explainability accuracy.

Why I chose this paper

So many interpretability methods boast high performance: How to pick the right one?
Hard to find strong quantitative statements about explainability accuracy.

Did I like it?

Why I chose this paper

So many interpretability methods boast high performance: How to pick the right one?
Hard to find strong quantitative statements about explainability accuracy.

Did I like it?

Yes and No!

Why I chose this paper

So many interpretability methods boast high performance: How to pick the right one?
Hard to find strong quantitative statements about explainability accuracy.

Did I like it?

Yes and No!
Quantitative framework and discovery are nice.

Why I chose this paper

So many interpretability methods boast high performance: How to pick the right one?
Hard to find strong quantitative statements about explainability accuracy.

Did I like it?

Yes and No!
Quantitative framework and discovery are nice.
Some of the details might need refinement

Why I chose this paper

So many interpretability methods boast high performance: How to pick the right one?
Hard to find strong quantitative statements about explainability accuracy.

Did I like it?

Yes and No!
Quantitative framework and discovery are nice.
Some of the details might need refinement
Slightly limited number of techniques compared
Paper used unclear notation and ommitted details at times

Questions?