# Interpretation of Image Related Deep Learning Model

Yunhao Zhang
Department of Informatics
New Jersey Institute of Technology & Urban Data Science Lab, Texas A&M University

Email: yz862@njit.edu

# Index

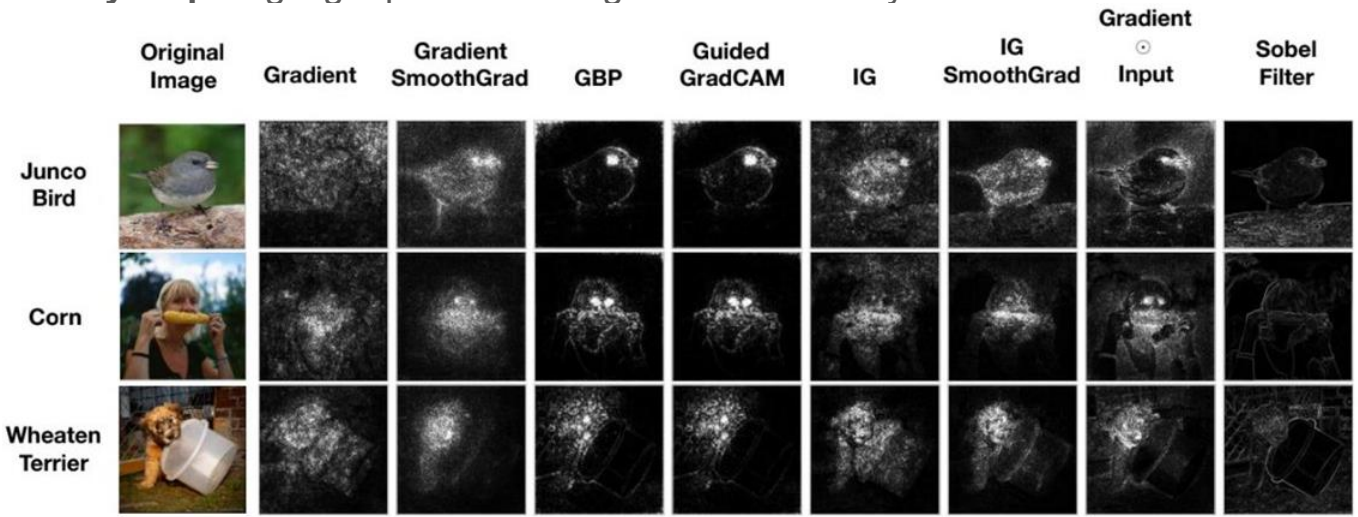
1. Research Background
2. Replicated Model(Others Work)
3. Further Extension(Our Work)
4. Replication Tool

# Research Background

# Background: Saliency Map

**Summary**: Providing clues why Deep Learning models generate a such result(prediction)?

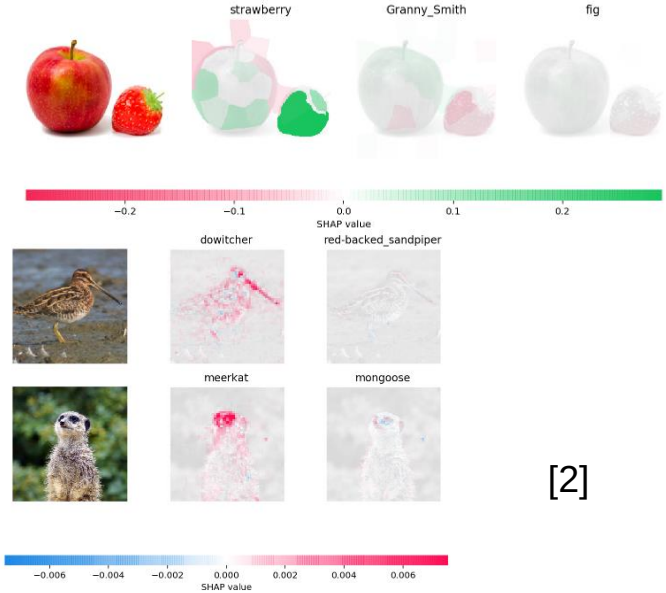
**Saliency Map**: Highlight pixels on image to indicate they cause the models' decision.




[1]

1.Adebayo, Julius, et al. "Sanity checks for saliency maps." *Advances in Neural Information Processing Systems*. 2018.

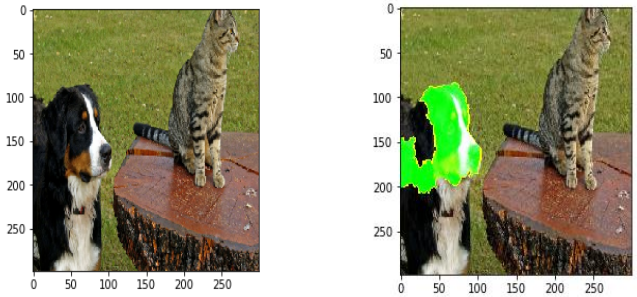
# Other Saliency Map Applications: LIME, SHAP




[2]

```
images = transform_img_fn(['dogs.jpg'])
# I'm dividing by 2 and adding 0.5 because of how this Inception represents images
plt.imshow(images[0] / 2 + 0.5)
preds = predict_fn(images)
for x in preds.argsort()[0][-5:]:
    print x, names[x], preds[0,x]
```

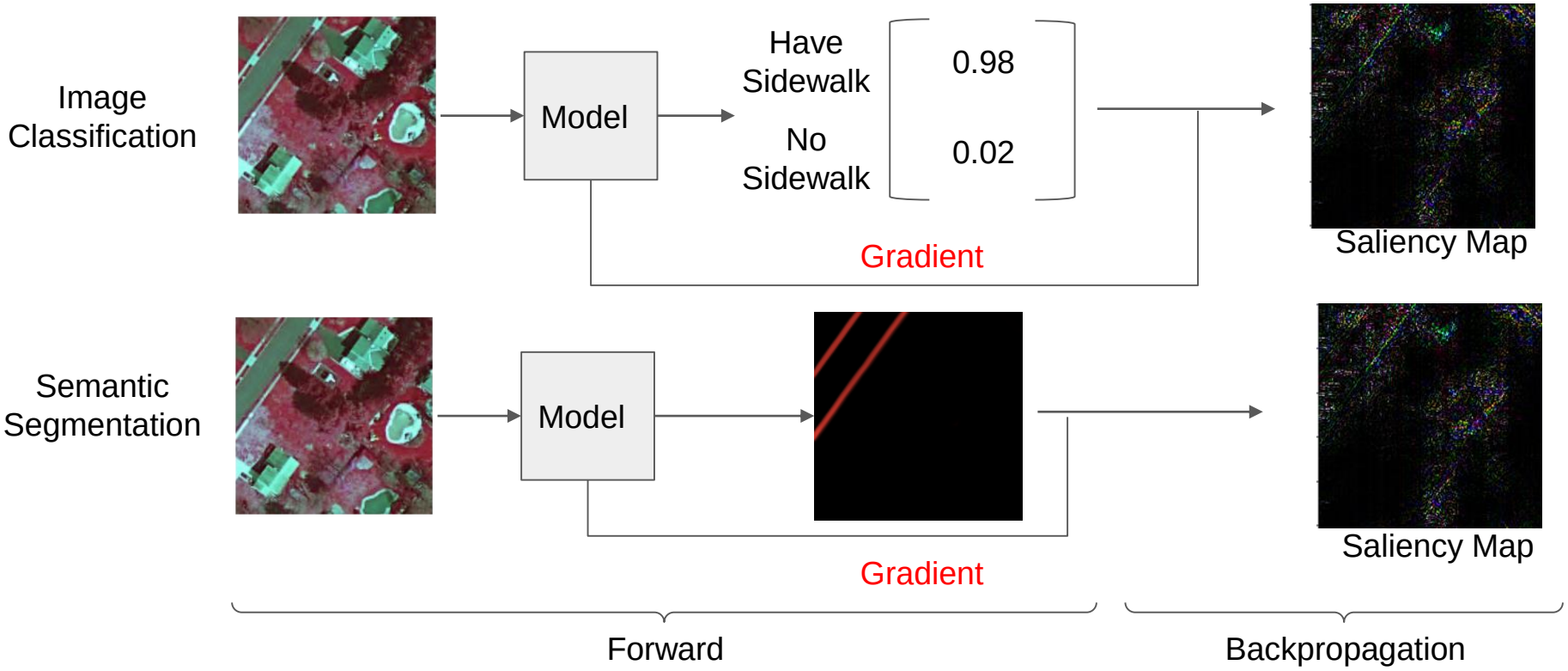
```
286 Egyptian cat 0.000892741
242 EntleBucher 0.0163564
239 Greater Swiss Mountain dog 0.0171362
241 Appenzeller 0.0393639
240 Bernese mountain dog 0.829222
```




[3]

2. https://github.com/marcotcr/lime
3. https://github.com/slundberg/shap

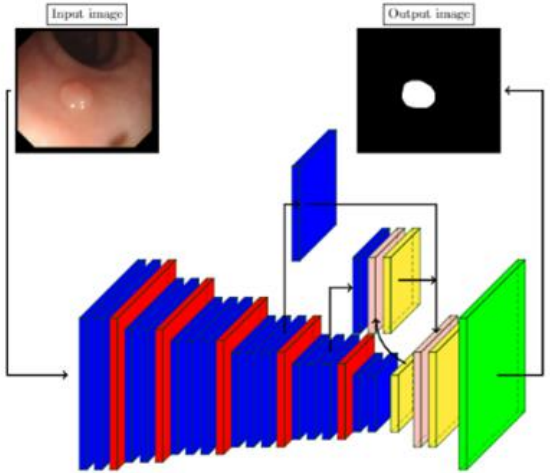
# Pipeline: Get Saliency Map



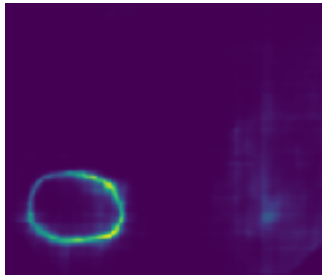
# Replicated Model

# EFCN: Guided Backpropagation Saliency Map

**Objective**: Using their own model to perform Semantic Segmentation task(prediction)  and applying Saliency Map method to interpret prediction.




Model Architecture[4]



Saliency Map: Guided Backpropagation[4]

| Model | # P(M) | IoU B | IoU P | IoU M | Acc M |
|---|---|---|---|---|---|
| SDEM [4] | << 1 | 0.739 | 0.221 | 0.480 | 0.756 |
| FCN-8 [2] | 134.5 | **0.946** | 0.509 | 0.728 | **0.949** |
| ESegNet | 29.5 | 0.933 | 0.522 | 0.728 | 0.935 |
| EFCN-8 | 134.5 | **0.946** | **0.587** | **0.767** | **0.949** |

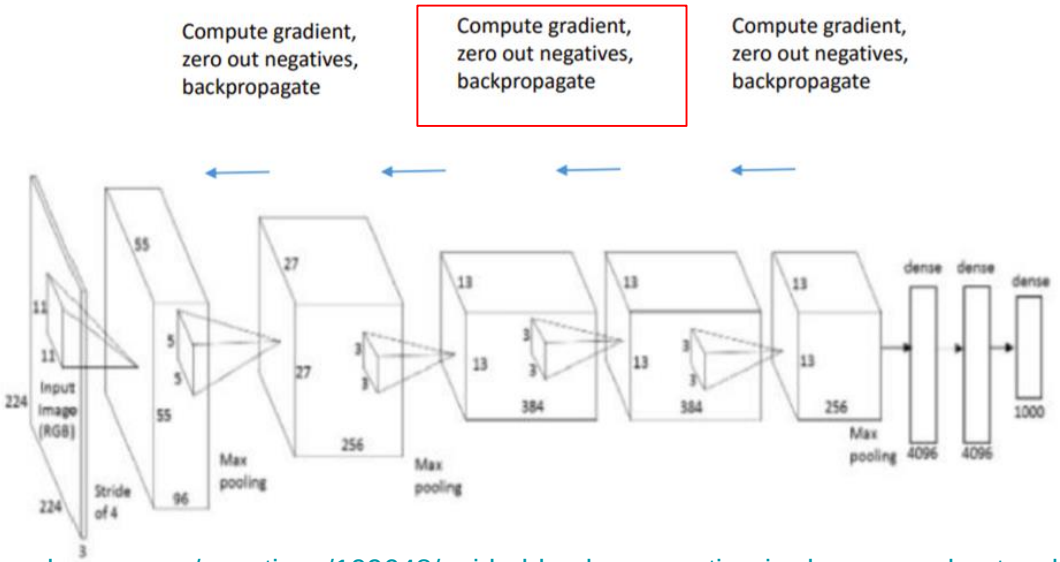
Result Comparison[4]

4. Wickstrøm, Kristoffer, Michael Kampffmeyer, and Robert Jenssen. "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps." *Medical Image Analysis* 60 (2020): 101619.

# EFCN: Guided Backpropagation

**Dependence**: Python, Pytorch[6]

**Guided Backpropagation Details**:

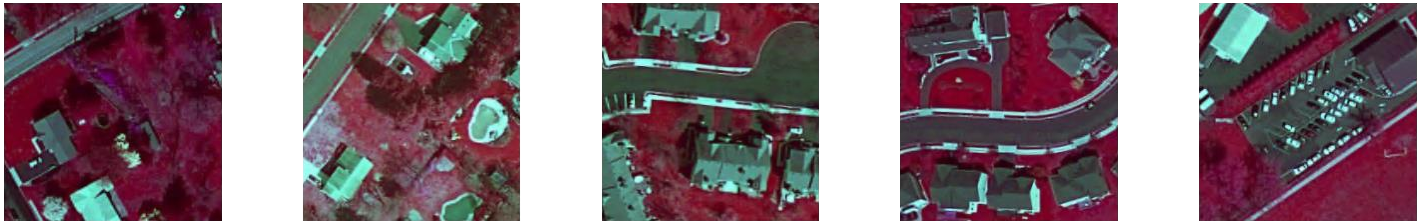



[5]

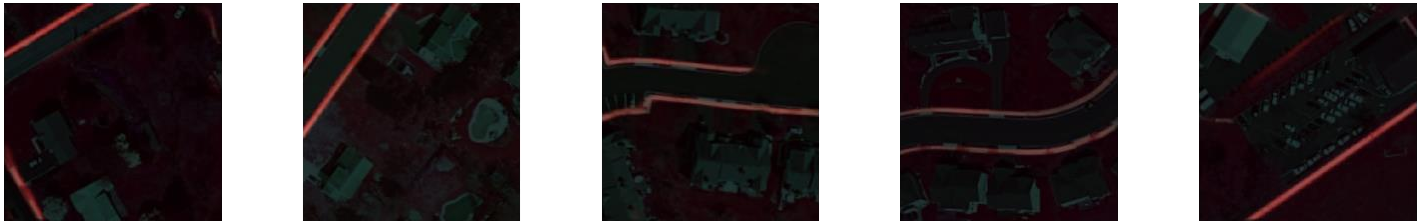
5. https://cs.stackexchange.com/questions/109648/guided-backpropagation-in-deep-neural-networks
6. https://pytorch.org/

# Result EFCN

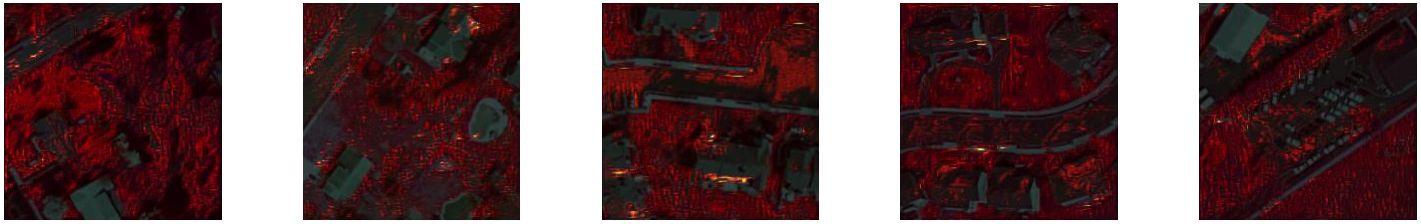
Image



Prediction



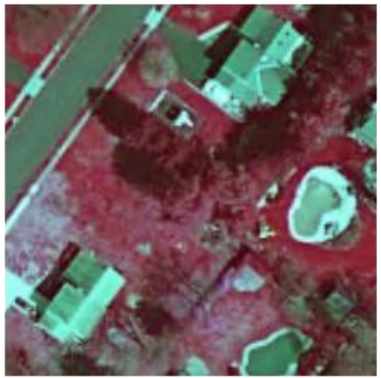
Guided Saliency



# Further Extension

# Further Consideration: Reliability

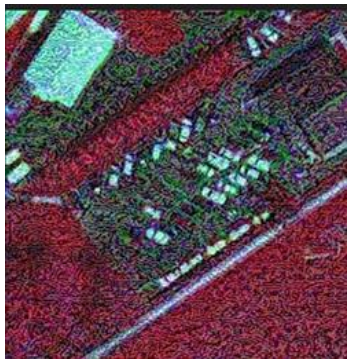
1. Can we get saliency from a part(certain region) of image?
2. Is this method robust?



Grassland + Impervious Surface + Sidewalk

# Perturbation: FGSM Attack

Pertibarted

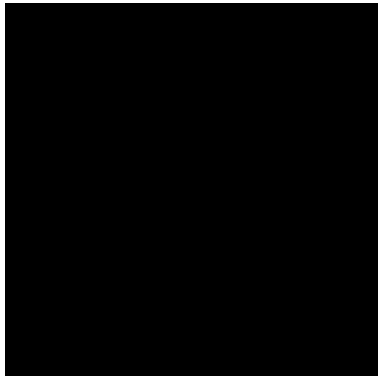


Epsilon

Noise

Original

Pertibarted Pred

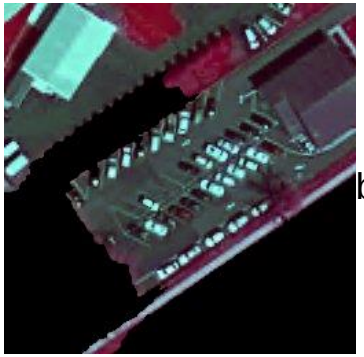


Desired Change

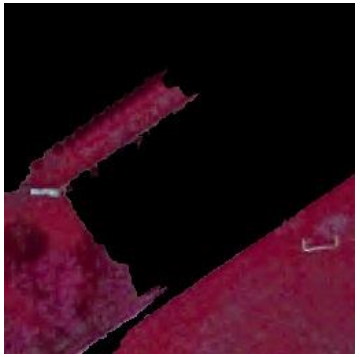
Original Pred

# Perturbation: Mask Out Certain Region

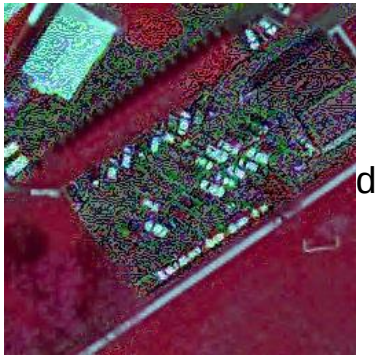
a



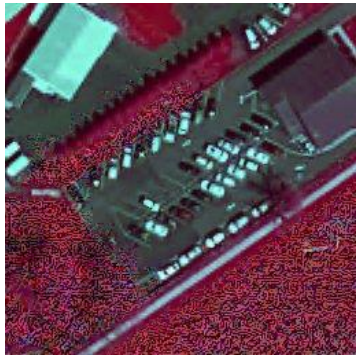
b



c



d



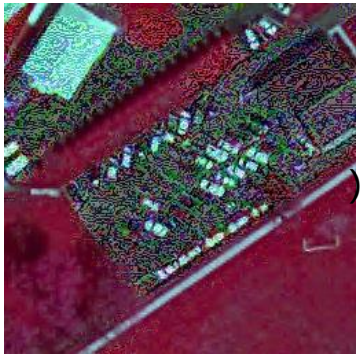
# Perturbation: Get Sailency on Desired Region

Assume S(...) is a function to get saliency map……

Original Image

S( ) — S( ) =

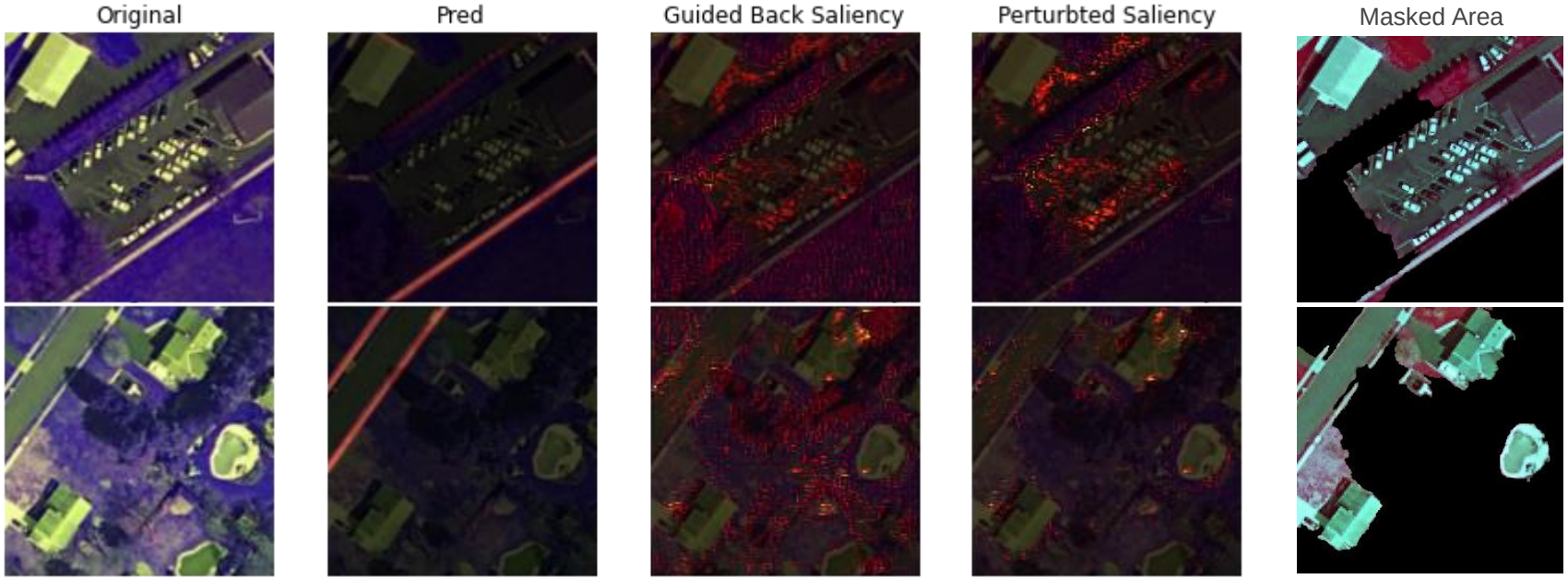
Original Image with masked area



Saliency Map Which should only has highlighted pixels on masked area

S(Grassland) + S(impervious surface) + S(Sidewalk) — S(Grassland) + S(Sidewalk) = S(impervious surface)

# Perturbation: Result



# Replication Tool

# Google Colab

**Merit**: Providing a remote environment for developing and debugging.

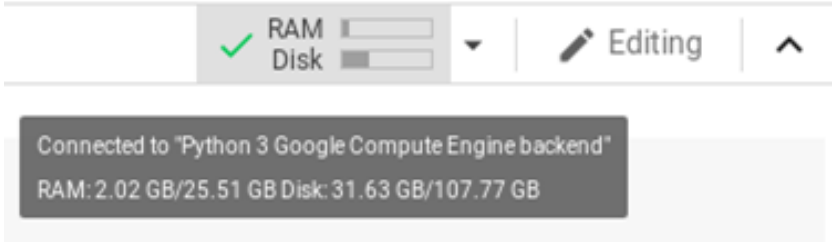
```
[18]  1  if args.cuda:
      2      print('cuda')
      3      model = torch.load(args.model).cuda()
      4  else:
      5      model = torch.load(args.model, map_location=torch.device('cpu')).cpu()
      6      print('cpu')
      7
      8  # Compute guided smooth gradient
      9  guided_smooth_grad = GuidedBackpropSmoothGrad(
     10      pretrained_model=model,
     11      cuda=args.cuda,
     12      n_samples=1,
     13      magnitude=True)
     14
```

```
cpu
/usr/local/lib/python3.6/dist-packages/torch/serialization.py:657: SourceChangeW
  warnings.warn(msg, SourceChangeWarning)
```

```
[22]  1  count = 0
      2  for gra in gradients:
      3
      4    print(files[inter_index[count]])
      5    mat = gra[0]
      6    mat = np.transpose(mat, (1,2, 0))
      7    mat[mat<=0] = 0
      8    mat = np.sum(mat, axis=2)
      9    plt.imshow(mat, cmap='hot', )
     10    plt.axis("off")
     11    plt.show()
     12    count += 1
     13
```

# Google Colab

**Merit**: Offering high capacity of RAM and GPU




```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.36.06    Driver Version: 418.67       CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla P100-PCIE...  Off  | 00000000:00:04.0 Off |                    0 |
| N/A   41C    P0    28W / 250W |      0MiB / 16280MiB |      0%      Default |
|                               |                      |                 ERR! |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

# Reference


1. Adebayo, Julius, et al. "Sanity checks for saliency maps." *Advances in Neural Information Processing Systems*. 2018.
2. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems*. 2017. https://github.com/marcotcr/lime
3. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016. https://github.com/slundberg/shap
4. Wickstrøm, Kristoffer, Michael Kampffmeyer, and Robert Jenssen. "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps." *Medical Image Analysis* 60 (2020): 101619.
5. https://cs.stackexchange.com/questions/109648/guided-backpropagation-in-deep-neural-networks
6. https://pytorch.org/