Learning Perturbations to Improve Classification Accuracy

A.I. Specialization Project

Department of Computer Science and Engineering
Indian Institute of Technology, Jodhpur

Presented by:
Qazi Sajid Azam (B16CS026)
Anurag Shah (B16CS034)

Mentor:
Dr. Deepak Mishra

Problem Statement

Standard Classifiers (VGG16, VGG19, RESNet) work great but we don’t know why they work correctly.

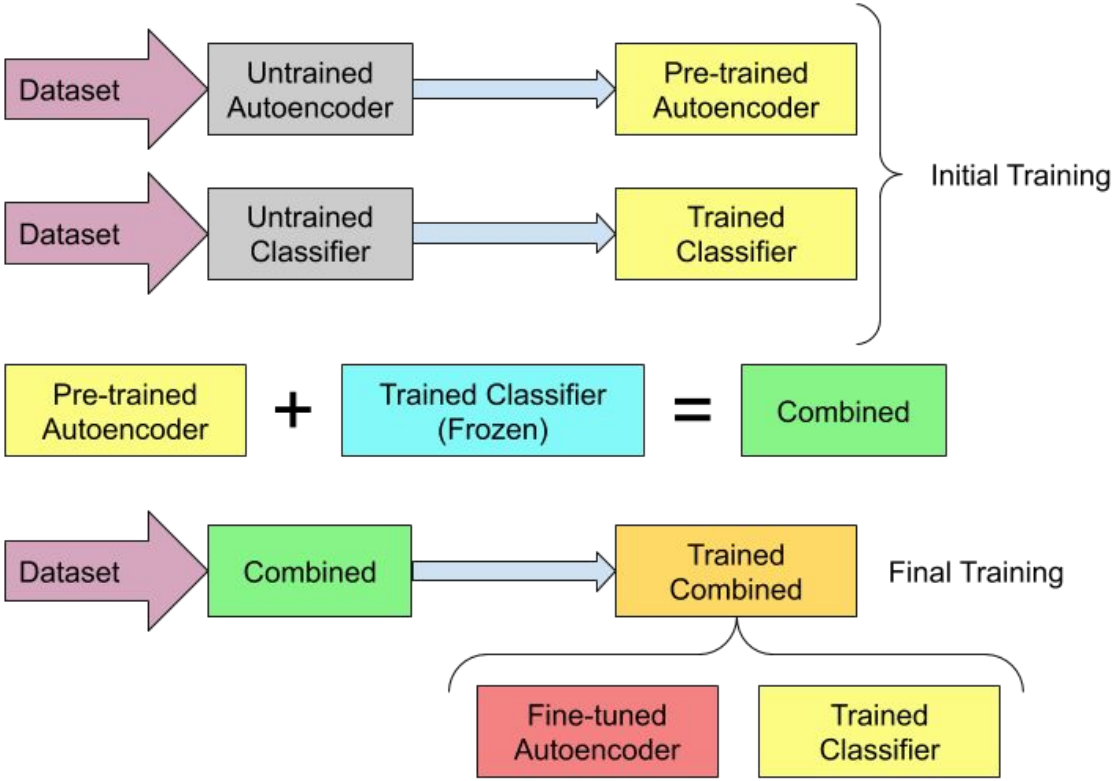
It becomes very important to know the reason or at least have assurance that classifiers are working because of correct reasons especially in Medical field.

We dig into the detail on why classifier work and how we can improve the classification accuracy.

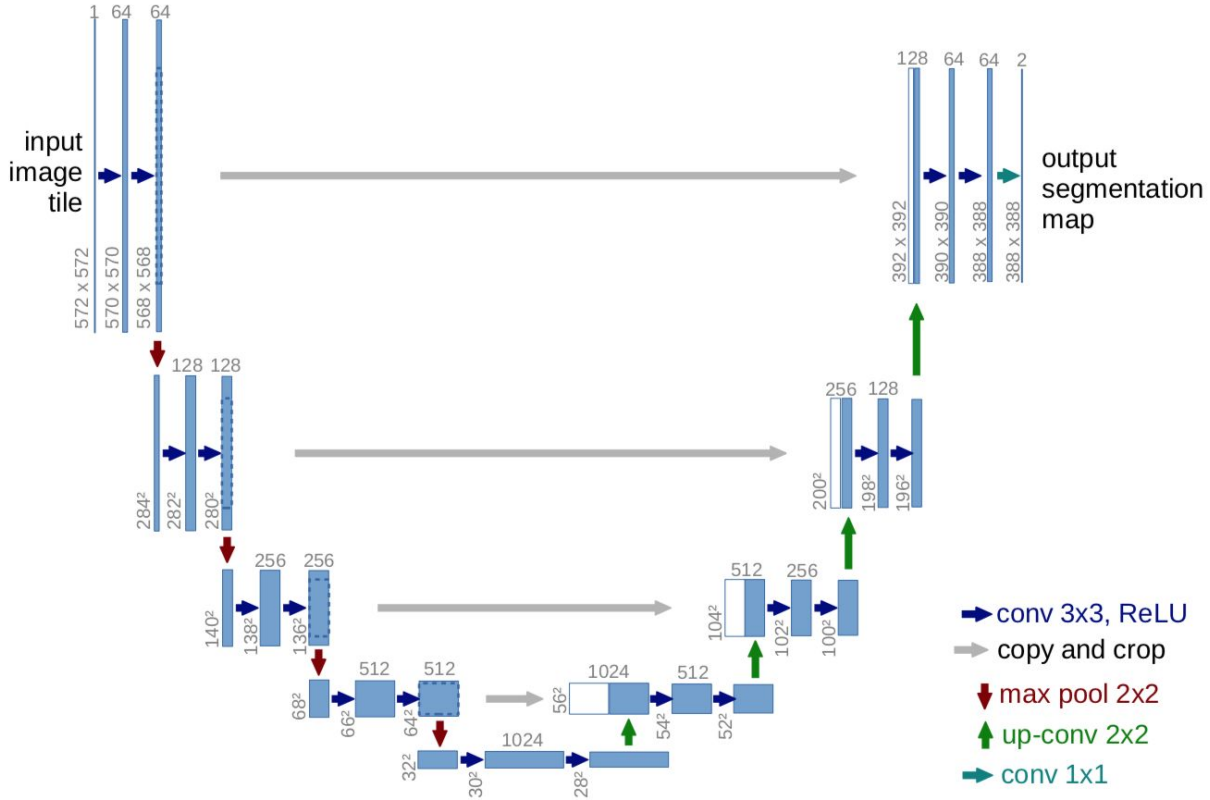
Previous work

- LIME - Local Interpretable Model-Agnostic Explanations
- GRADCAM - Gradient-weighted Class Activation Mapping
- Transferable Recognition-Aware Image Processing
- Uncovering the impact of Background Features on Deep Neural Networks

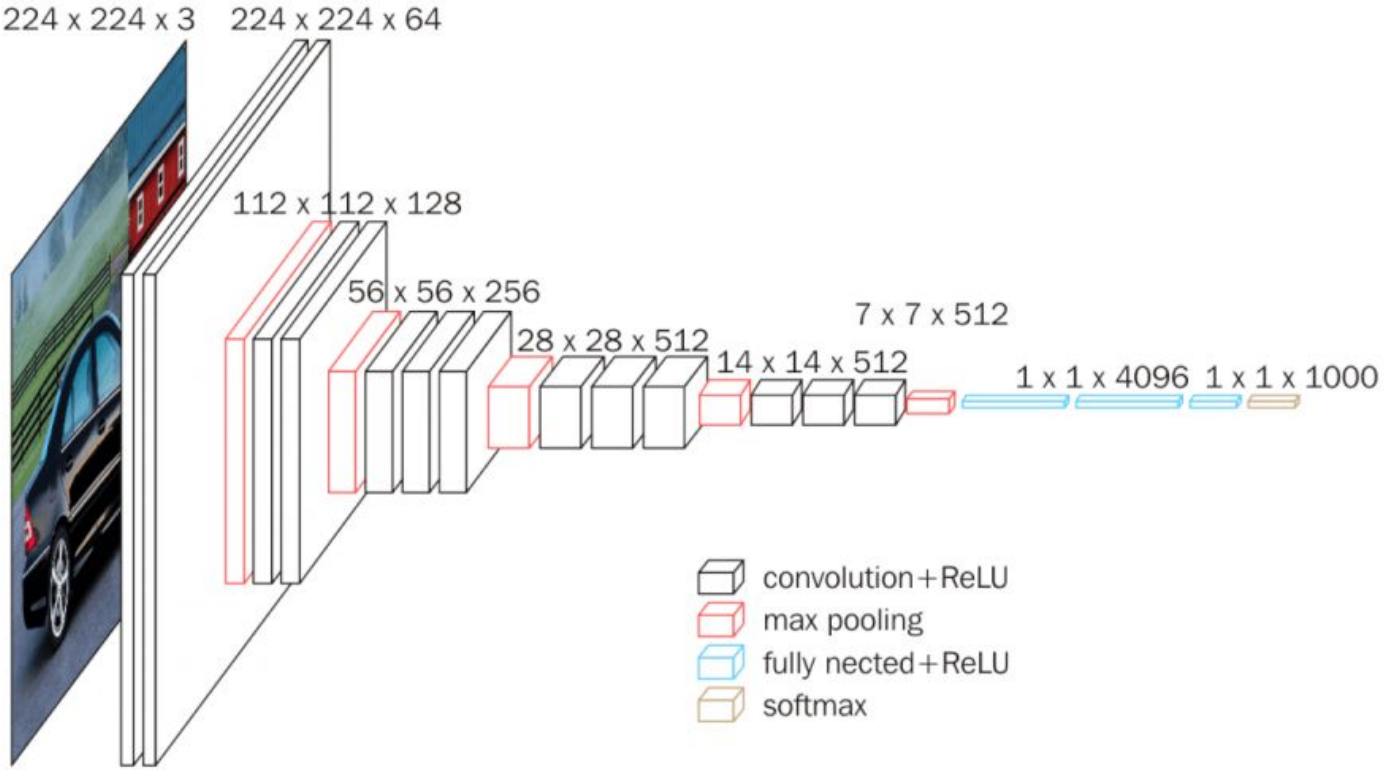
Approach


Dataset
Untrained Autoencoder
Pre-trained Autoencoder
Dataset
Untrained Classifier
Trained Classifier
Initial Training
Pre-trained Autoencoder
+
Trained Classifier (Frozen)
=
Combined
Dataset
Combined
Trained Combined
Final Training
Fine-tuned Autoencoder
Trained Classifier

AutoEncoder - UNet



Classifier - VGG16



Loss Function

- Autoencoder: Initial training to learn to replicate images

$$BCE = -\sum_{i=1}^{C'=2} t_i log(f(s_i)) = -t_1 log(f(s_1)) - (1 - t_1) log(1 - f(s_1))$$

- Classifier and Combined model: Learn to recognize image class

$$CCE = -log\left(\frac{e^{s_p}}{\sum_j^C e^{s_j} a}\right)$$

Role of Autoencoder

1. Remove Noise if any in the input image
2. Enhance the input image for better classification

Dataset

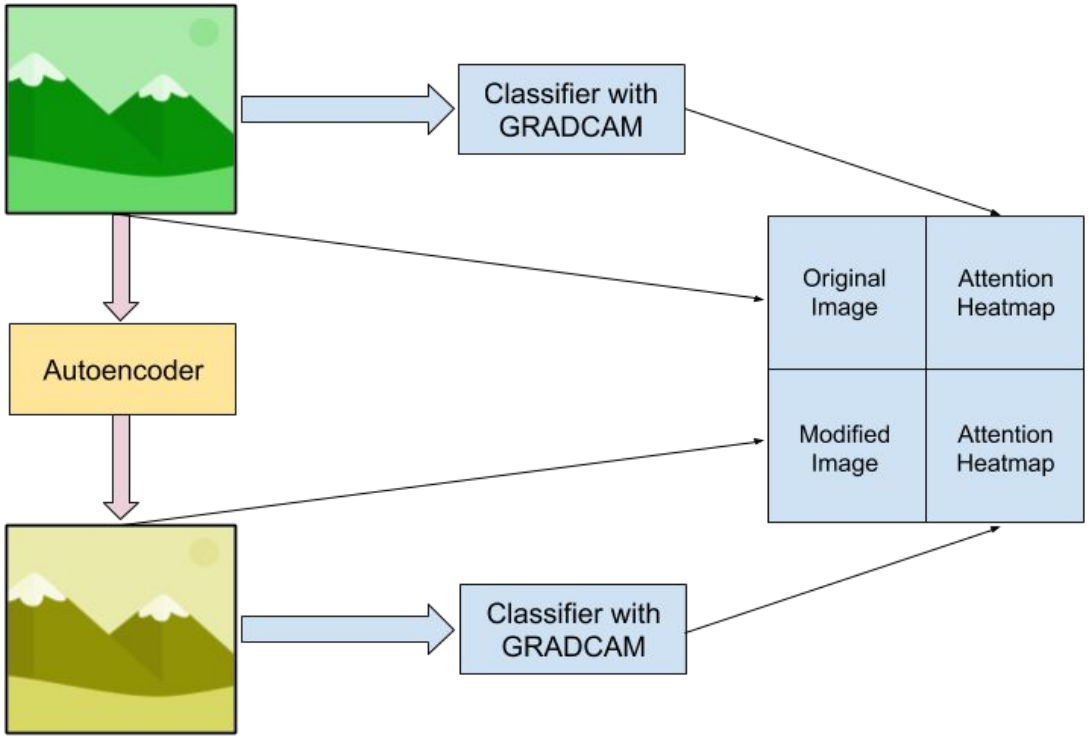
- STL-10: 4010 training images, 990 test images, 96x96 size, 10 classes
- CIFAR-10: 50000 training images, 10000 test images, 32x32 size, 10 classes
- Kaggle Intel Image Classification: 10000 training images, 4000 test images, 150x150 size, 6 classes

Accuracy Comparison

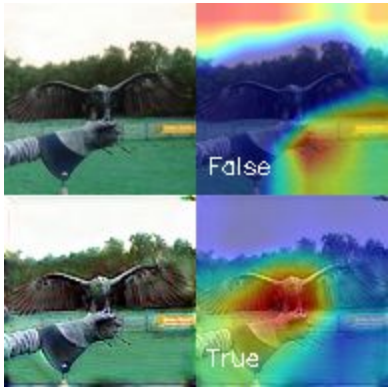
Dataset	Classifier Accuracy (%)	Combined Accuracy (%)
STL-10	82.8	85.5
STL-10 (full)	82.8	96.0
CIFAR-10*	75.7	79.6
Intel*	90.6	93.2
Intel* (full)	90.6	99.0

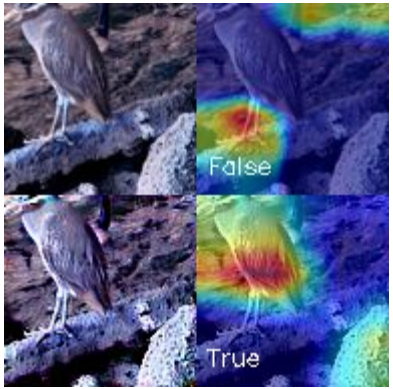
* - Images from these datasets are not analysed. We used these datasets for checking accuracy of combined model only.

GRADCAM Comparisons: Process

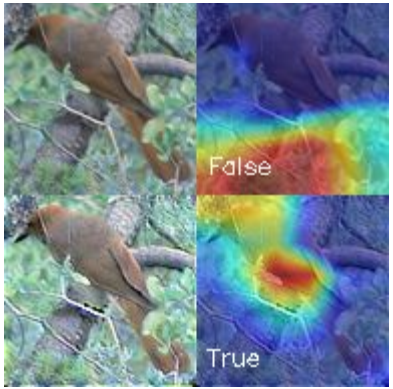


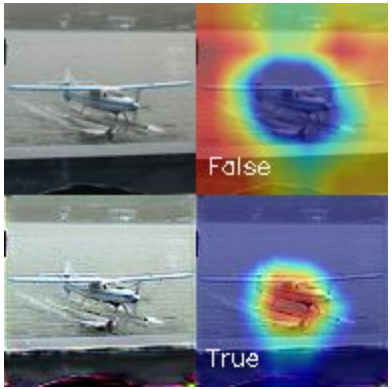
GRADCAM Comparisons: Different





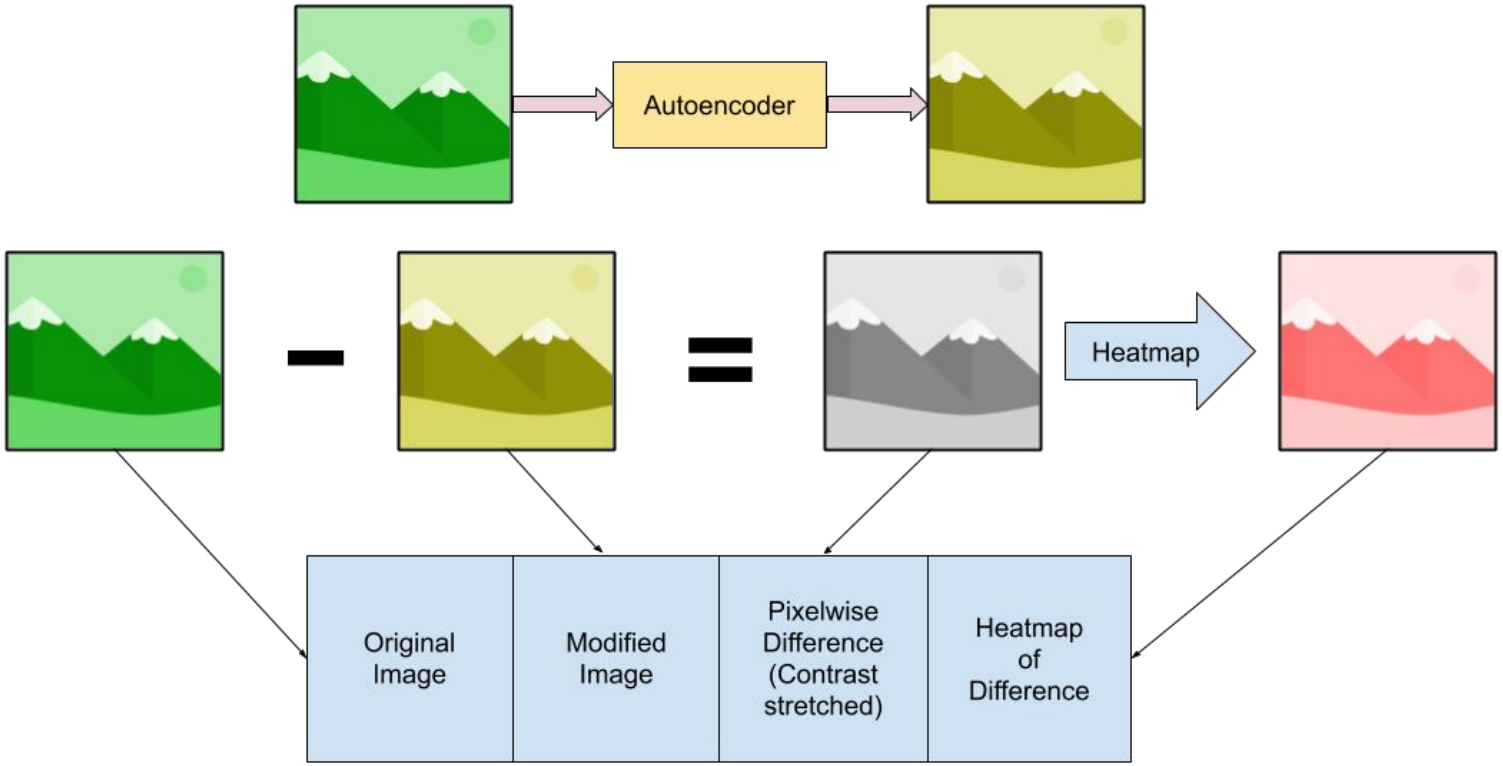
GRADCAM Comparisons: Different





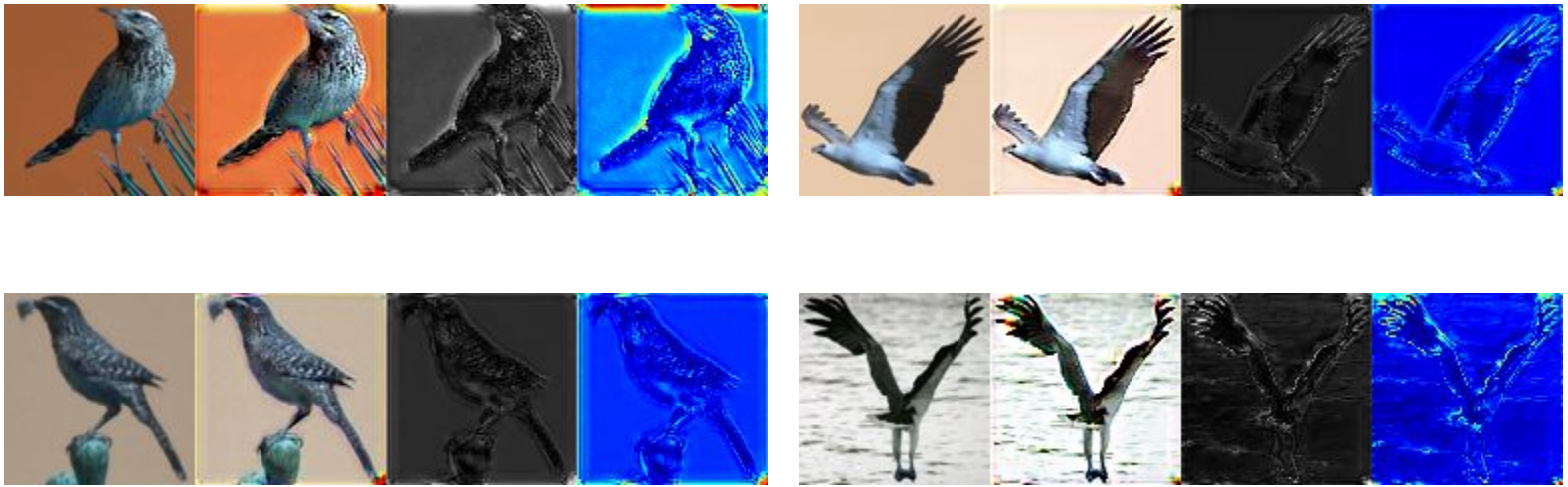
GRADCAM Comparisons: Same

Pixelwise Difference of Images: Process

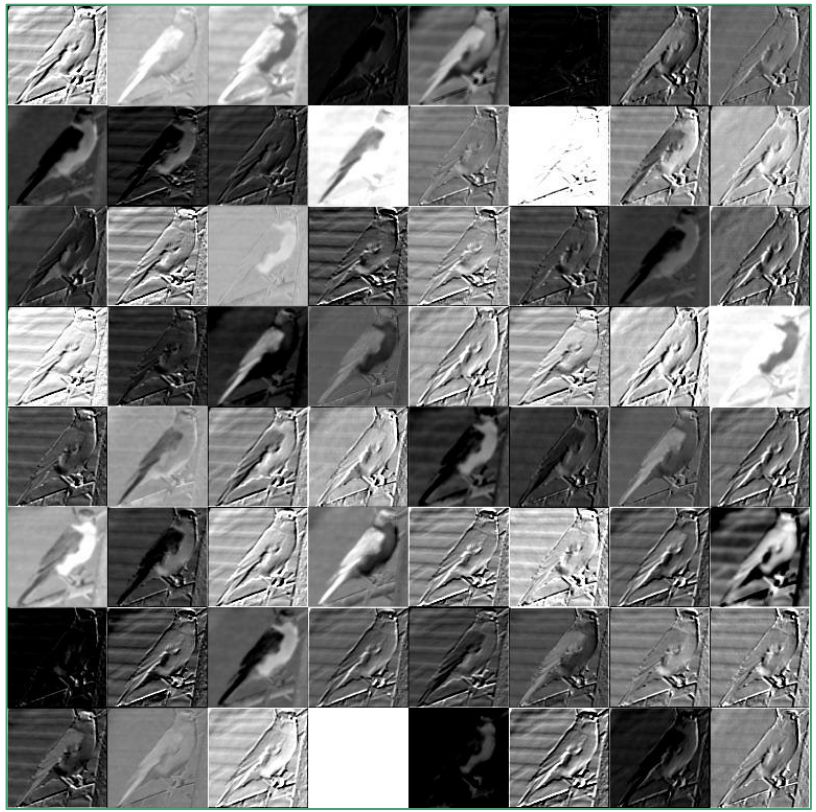


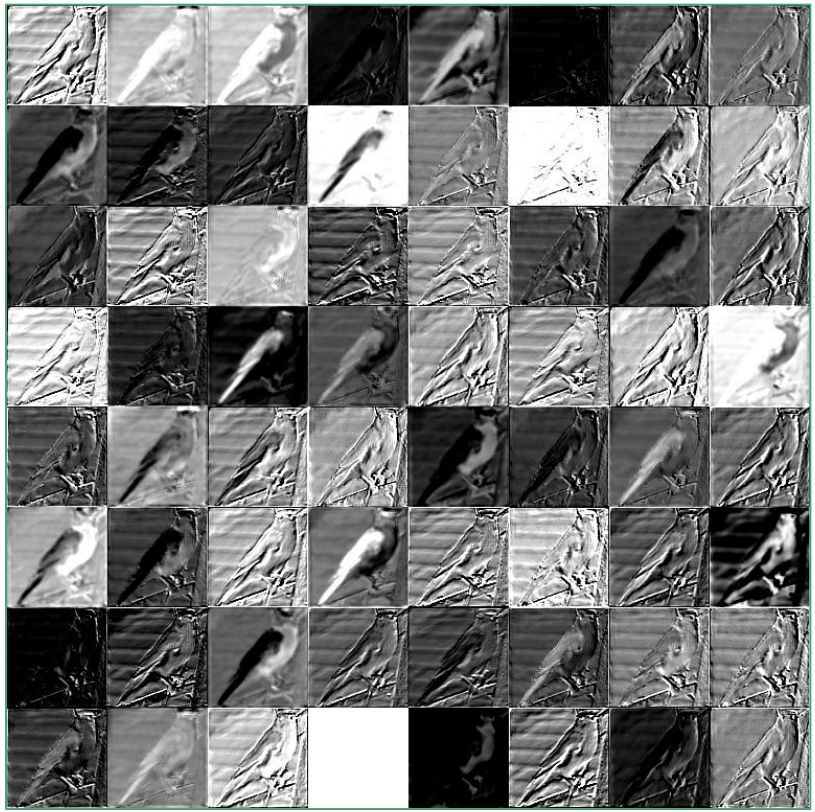
Pixel-wise Difference of Images

Pixel-wise Difference of Images



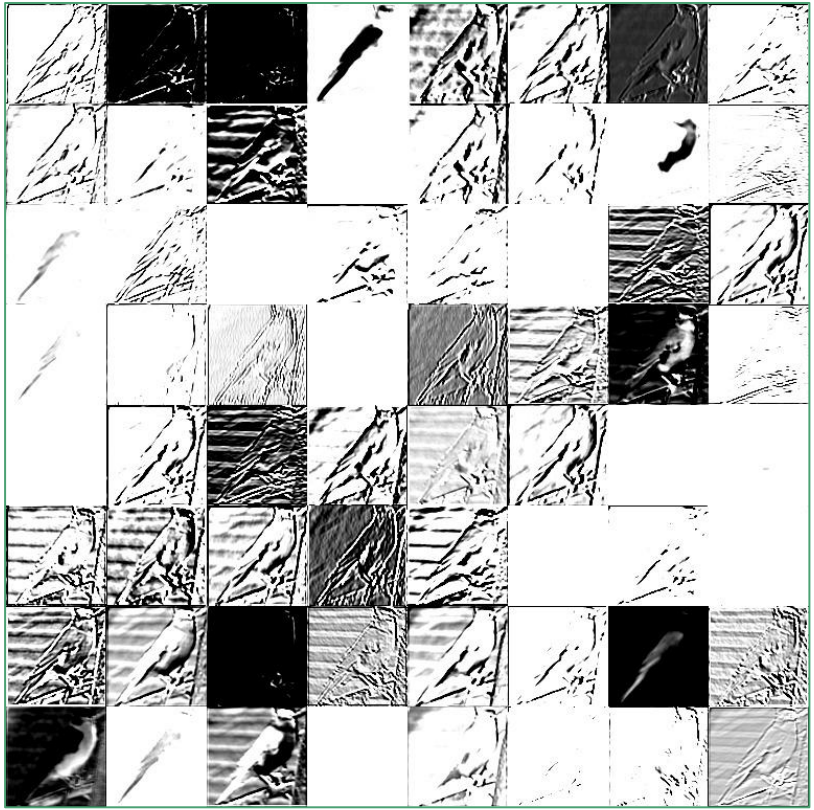
Internal Representations: Conv1 (old | new)

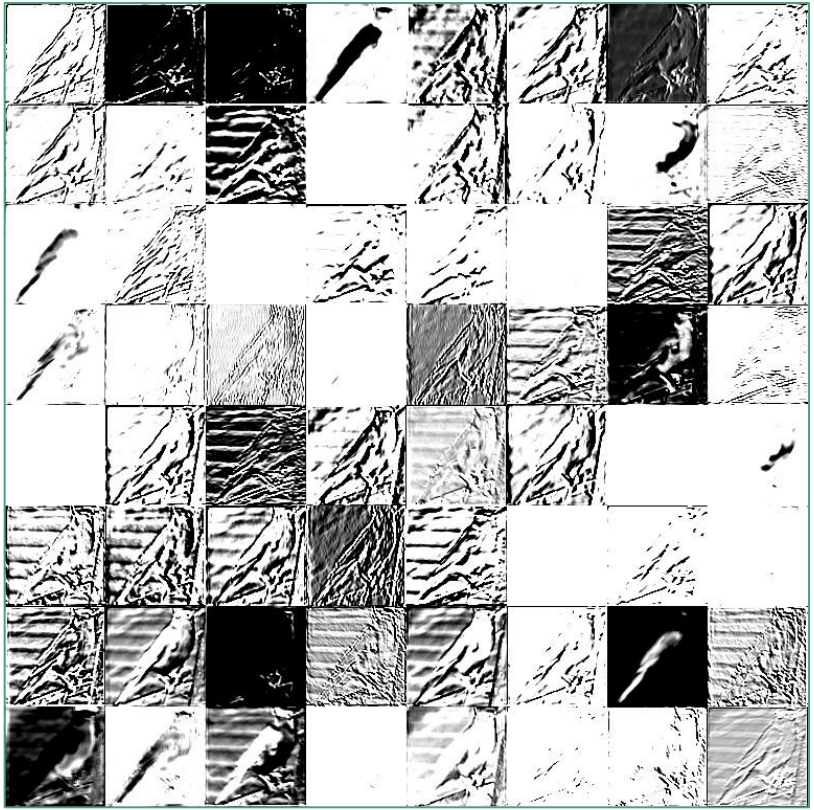




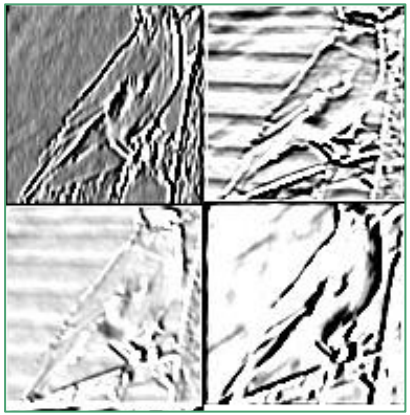
Internal Representations Close-Up: Conv1 (old | new)

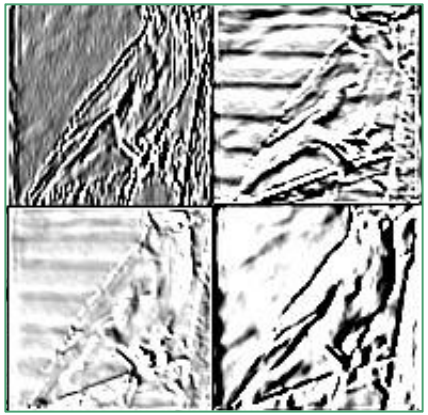
Internal Representations: Conv2 (old | new)





Internal Representations Close-Up: Conv2 (old | new)





Conclusions

1. There are some features that a classifier looks for when classifying an image
2. It is expected that a classifier can classify an image more easily if such features are more prominent
3. We employ an image transformation network which learns which features are important for classifier, and enhances those
4. The processed images are shown to get better accuracy on the classifier compared to the original images on the same classifier
5. By comparing the original images with the processed images, we can get insight about what characteristics the classifier is looking for in the image

Thank You