Motivation

Data collection for supervised network learning is prone to random errors and adversarial perturbations. Adversarial examples shatters the usual intuition that the perturbation are artifacts of learning. It has been shown that these examples are non intuitively related to the data distribution. As pointed out by [Ref to be added] adversaries can deliberately corrupt the labels and force the network behavior to fluctuate these are termed as targetted adversarial attacks. The other kind is where the input images are slightly perturbed and added to a clean set.

1. targeted

$$x_{adv} = x - \eta \, .sgn(\nabla_x \max_y logp(y/\mathbf{x})) \tag{1}$$

2. Un-targeted adversarial

$$x_{adv} = x + \eta \, .sgn(\nabla_x logp(y_{target}/\mathbf{x})) \tag{2}$$

As a result we explore statistical characteristics of adversarial examples and their relationship with the clean data

Related Work

1. Transferability and subspace of adversarial examples
    1. The Space of Transferable Adversarial Examples [2]
    2.
2. Geometric properties of decision boundary
    1. Classification regions of deep neural networks [3]
    2. Entropy SGD Biasing Gradient Descent into wide valleys [4]
3. Generating Adversarial perturbations
    1. Adversarial examples in the physical world [5]
    2. Deep Fool [6]
4. Hessian Approximation
    1. Efficient BackProp [7]

Experimental approach

Adversarial examples are shown to be a property of networks being more linear. These examples are transferable to other networks trained on In-domain data.

1. main part of this work we chose to map these perturbations on loss landscape and model their statistical relationship with the clean data.
2. Following which we have also explored targetted attacks by mislabeling data specifically on regression network Pilotnet [reference to be added]

Background

Our Hypothesis is based on the results in the paper 'Classification regions of deep neural networks'. This paper presents results on the geometry of the decision regions around normal images.

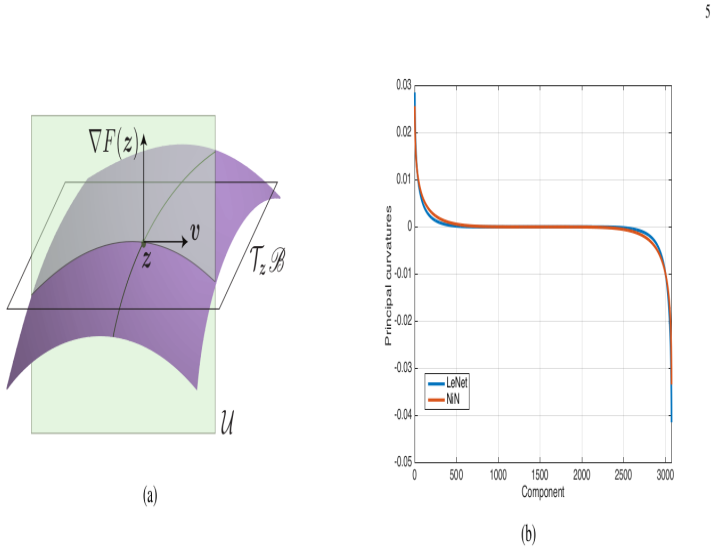


Figure 1: Loss vs Iterations over

The paper points the assymetry in the negative curvatures and natural images share common directions.

Detecting adversarial Images : Initial Hypothesis

'Can the local curvature/landscape around the data point reveal information about how easily an image can be perturbed?.'

The flatness of the eigen spectrum of the original image should reveal details about how easy is it to perturb an image

- Hessian of Loss; $\frac{dE}{dX}$ should reveal information through its Eigen Spectrum.
- flat spectrum implies the neighborhood of the point in space is possibly curved in multiple directions. In other words there are multiple directions in which it can be perturbed.
- Thus the metric should be independent of the chosen method of perturbation.

De-constructing the hypothesis

- The Hessian provides useful information relevant to the local curvature and not the global landscape that the algorithm traverses to place to data just outside the decision region.
- Does KL Divergence of the Hessian spectrum provide enough information about the curvature of the landscape? Do we need more information regarding the curvature?
- What does easily perturbable images mean ? What metric can we chose to correlate with the eigen spectrum?.

Defining easily perturbable image

To correlate with eigen spectrum of hessian we had to learn the local curvature. It generaly depends on below points.

1. Curvature around the point which is usually given by equation below

$$k(v,x) = \frac{vHv^T}{||v||_2^2||\nabla(F(x))||_2} \tag{3}$$

2. The number of possible directions of curvatures around the point x that are close to decision regions.

The assumption was that with change in α algorithm might traverse a different landscape.

- An image if mis classified with high confidence would in general have high cross entropy with its true label
- Images with low KLD appear to traverse different landscape for multiple α. Images with Index 9 and 4 have Low KLD. However they result in misclassification with high confidence.

Cost change ($\nabla(E)$) vs steps ($\nabla(P)$)

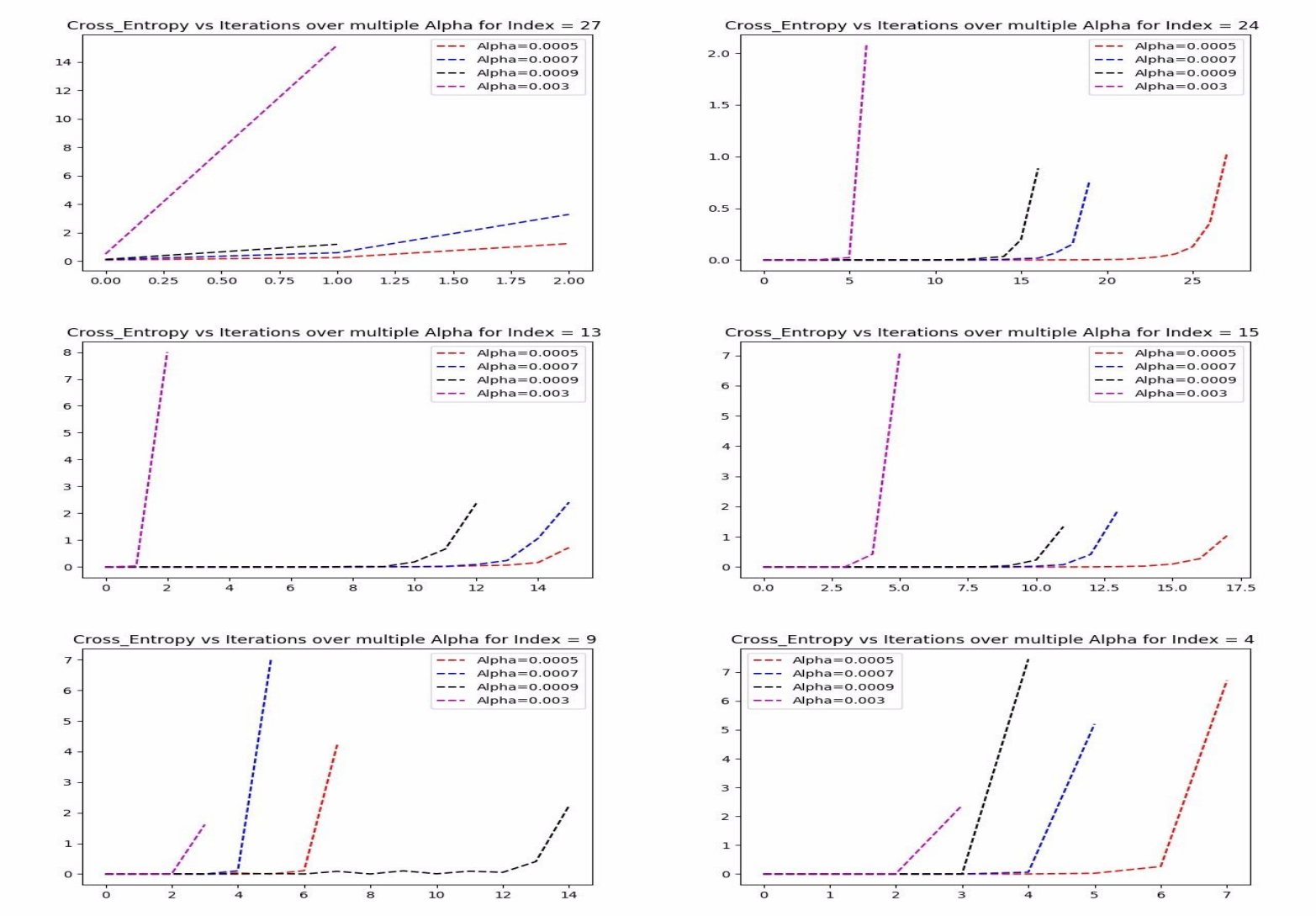


Figure 2: Loss vs Iterations over

- 30 Images from the CIFAR data selected and sorted based on the 'KulbaeckLuber' Divergences of the Histogram.
- They are grouped into 10 Images of each low KL Divergence **Group I** , mid KL Divergence **Group II**, High KL Divergence **Group III**.
- Each of the groups are compared for various metrics

KLD vs MSE plot for CIFAR on ResNet model

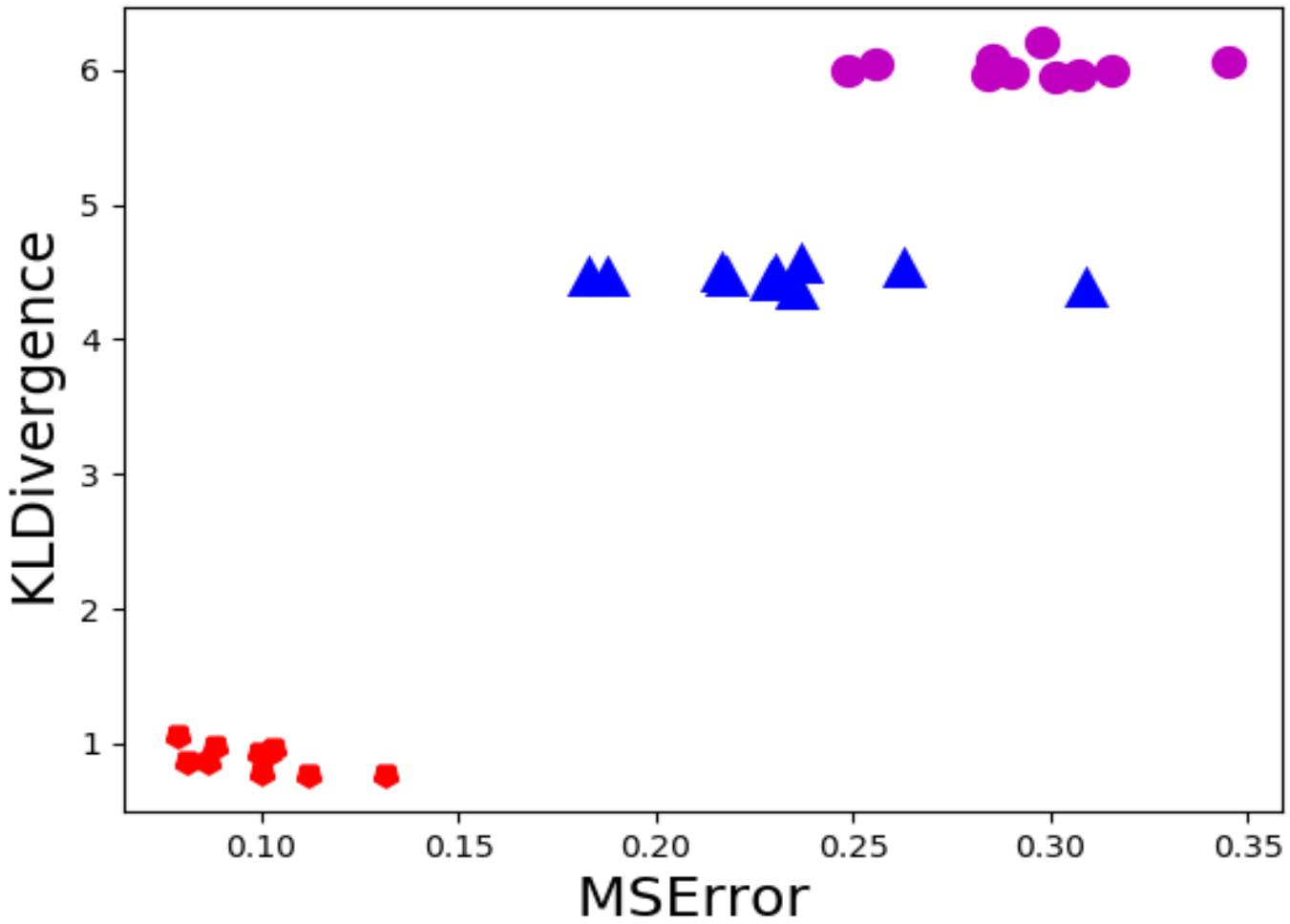


Figure 3: KLD vs MSE

Confidence of misclassification for multiple α

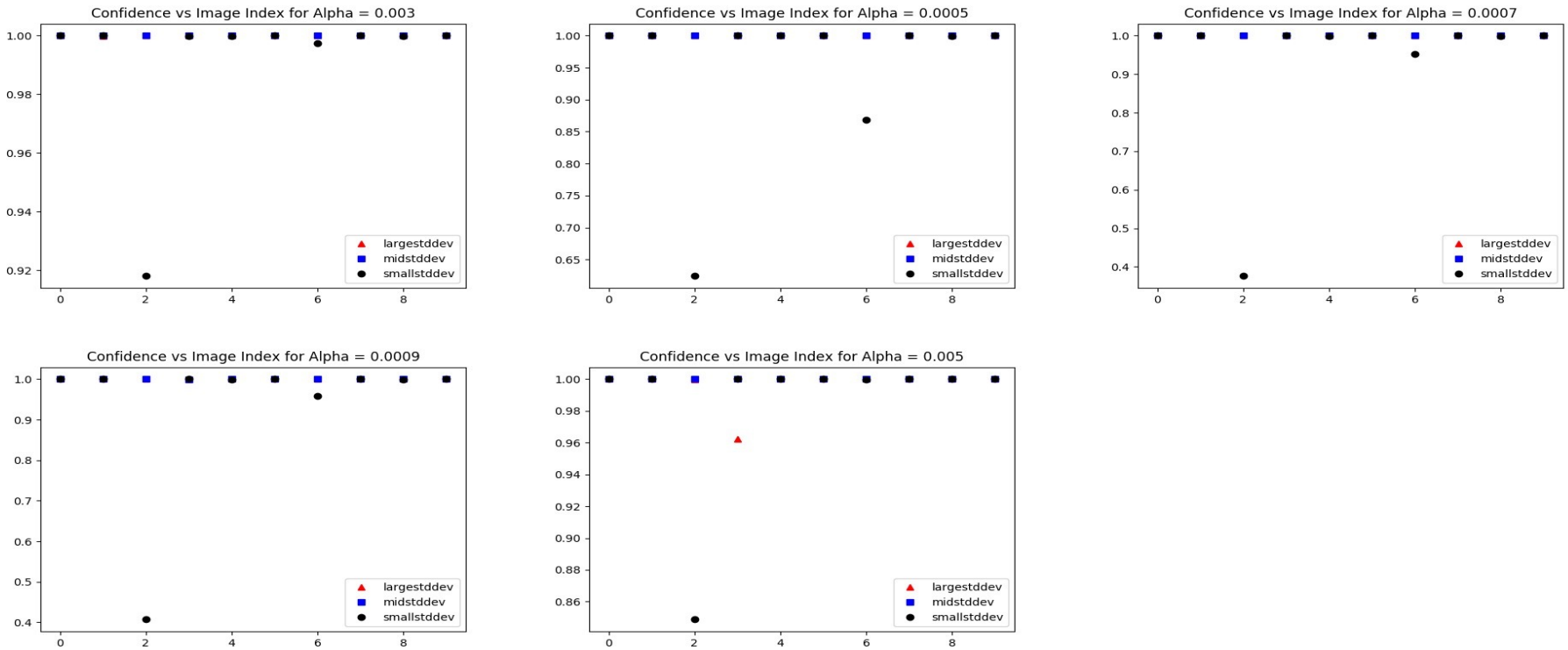


Figure 4: Confidence of Misclassified Images vs Image Index

Perturbed Image for VGG

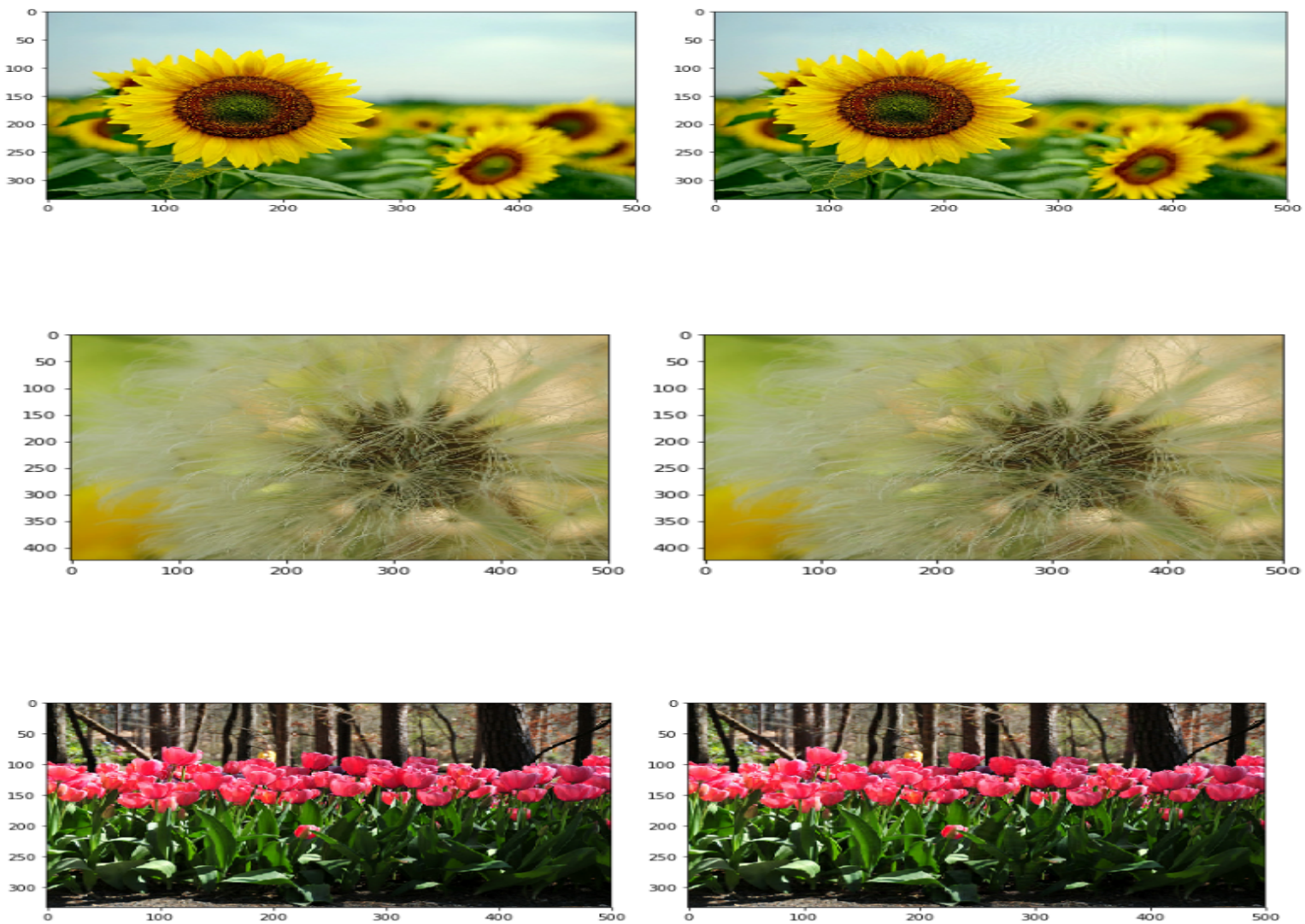


Figure 5: 3 Images: Original and Perturbed (High KLD , Mid KLD, Low KLD)

KI Divergence vs MSE plots (Imagenet on VGG)

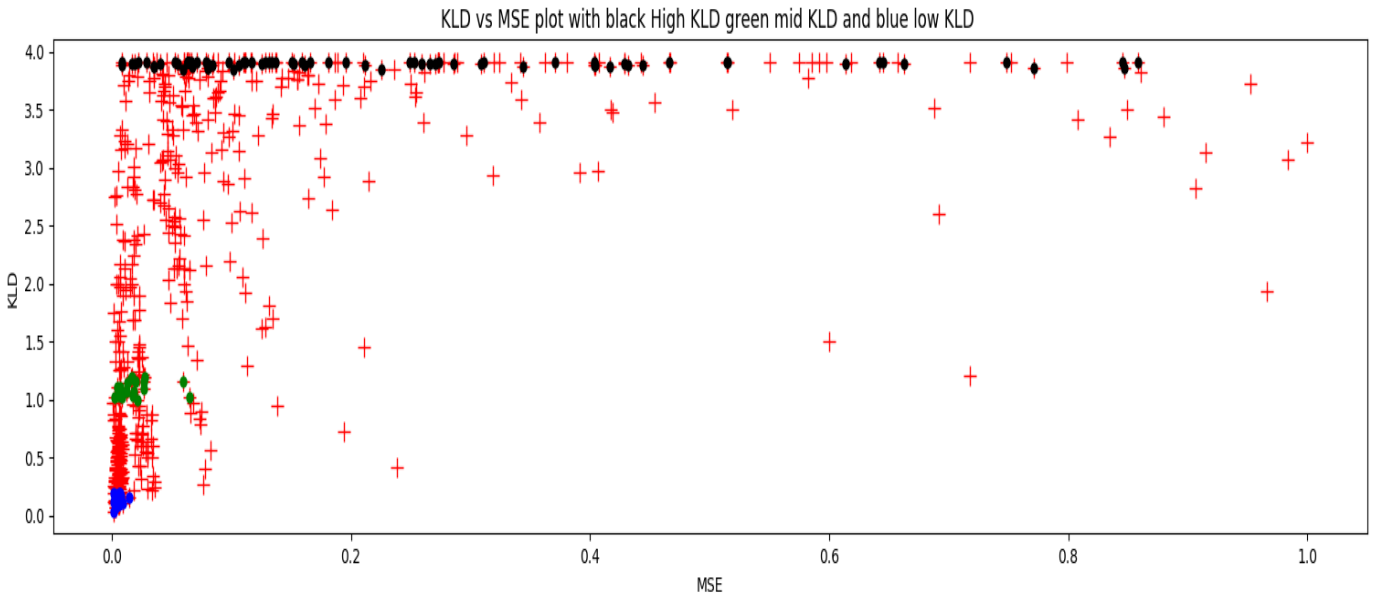


Figure 6: Black: points with Top 20 KLD , green : Mid 20 KLD , blue low 20 KLD

The pattern which appeared in CIFAR does not show in VGG. Instead 'High KLD' images appear to have more spread out MSE

Histograms for Imagenet on VGG (Uses Hessian Approximation)

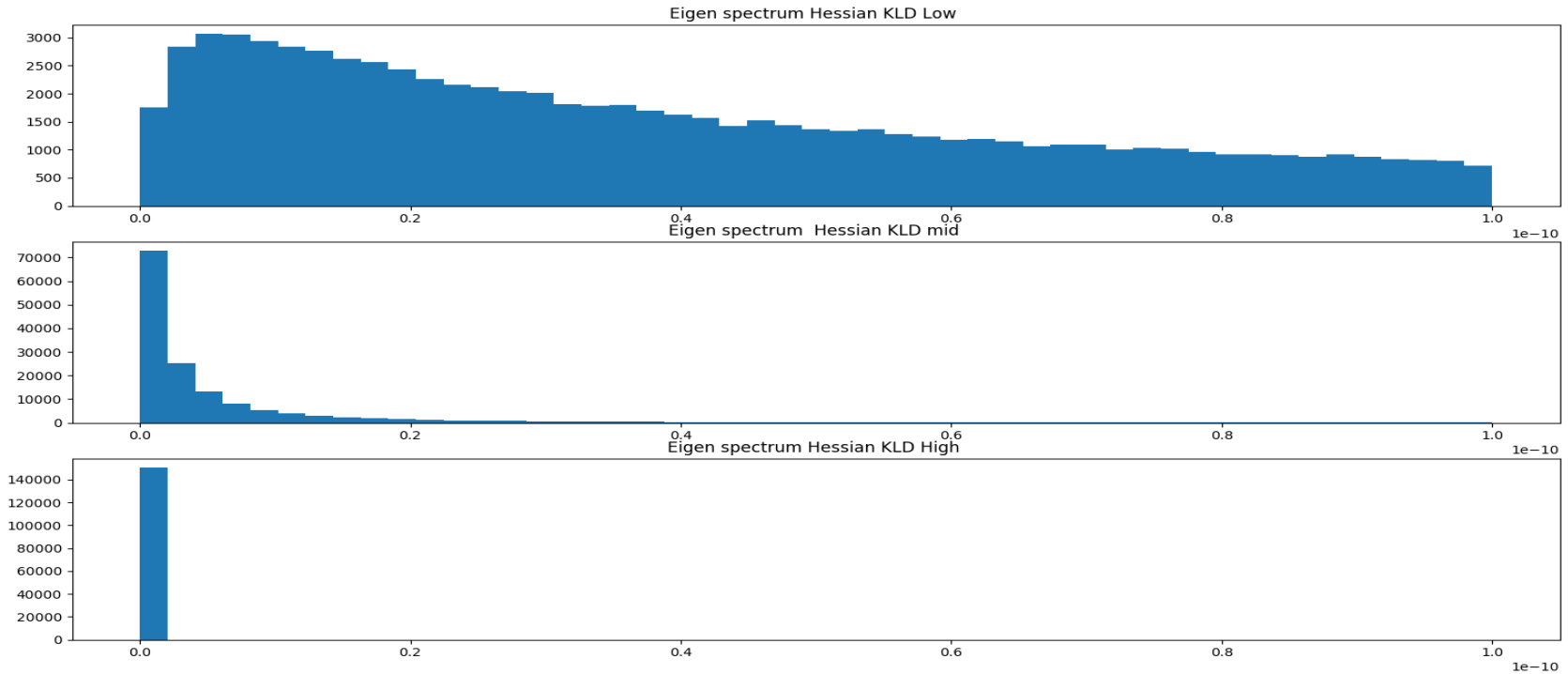


Figure 7: Top:Eigen Spectrum Low KLD , Middle Eigen Spectrum Med KLD , Bottom Eigen Spectrum high KLD

KLD of perturbed images for every iteration

Plot below shows the variation of KLD for every perturbation. Changes in color are used to represent the Image label flips that occur. ' Note that the label flips for high and mid KLD appear at low KLD points of the landscape'.

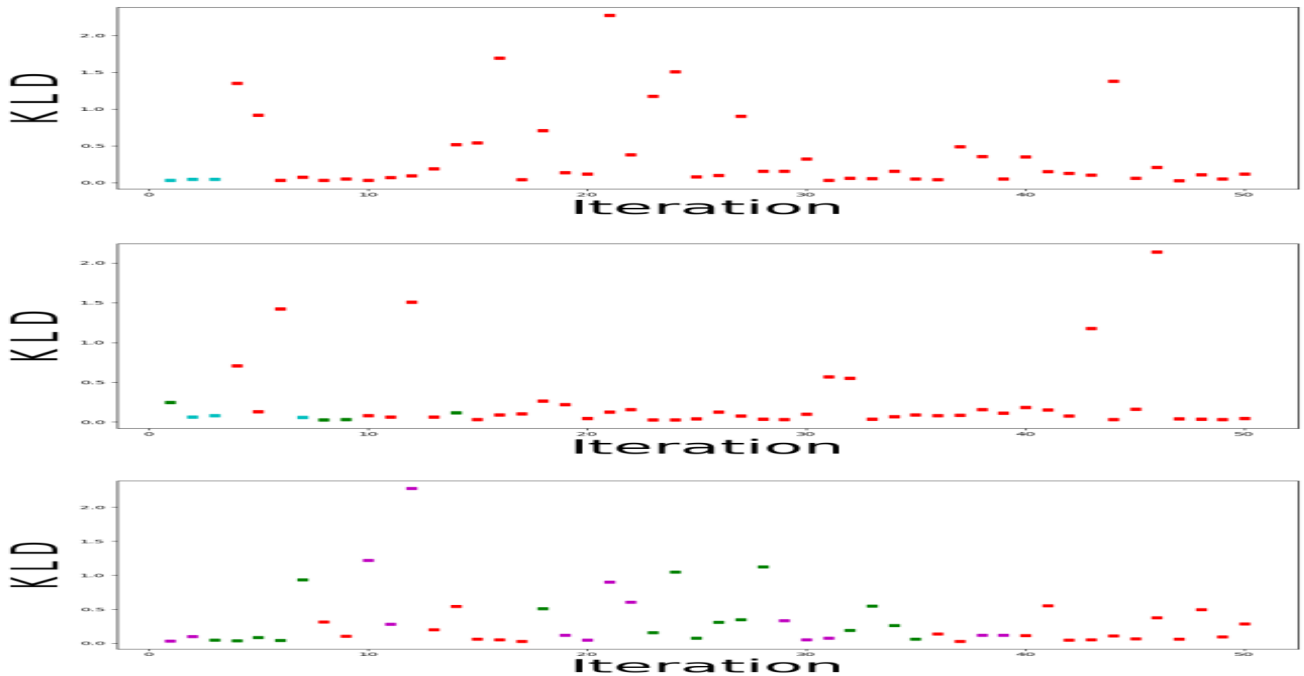


Figure 8: Top: Images with low KLD , Mid Image with mid KLD , Bottom Image with high KLD

Plots on (i) MSE (ii) Number of Steps

The plots below seem consistent with what was seen in previous slide. Image with low KLD appeared to switch labels at high KLD regions and switch more often. Below it is seen that the number of steps to misclassify is also high along with the MSE of the misclassified image. Both mid and high KLD images points show similar behavior

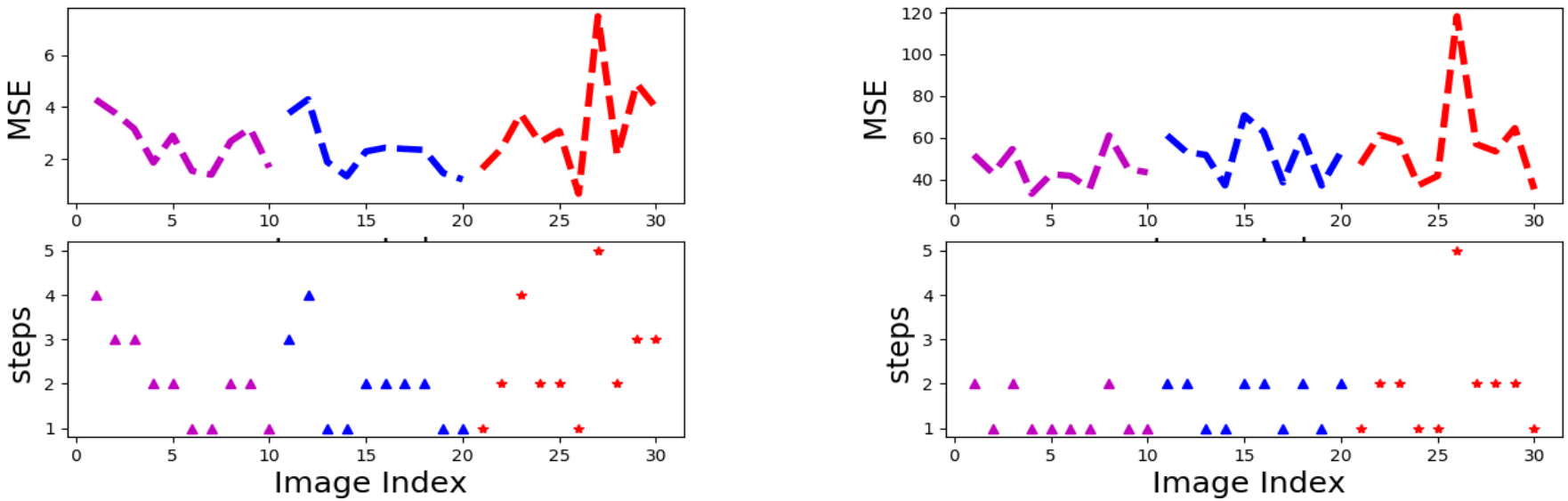


Figure 9: Left: $\alpha = 0.1$, Right: $\alpha = 0.5$, *HighKLD*, *MidKLD* , *LowKLD*

Plots on (i) MSE (ii) Number of Steps

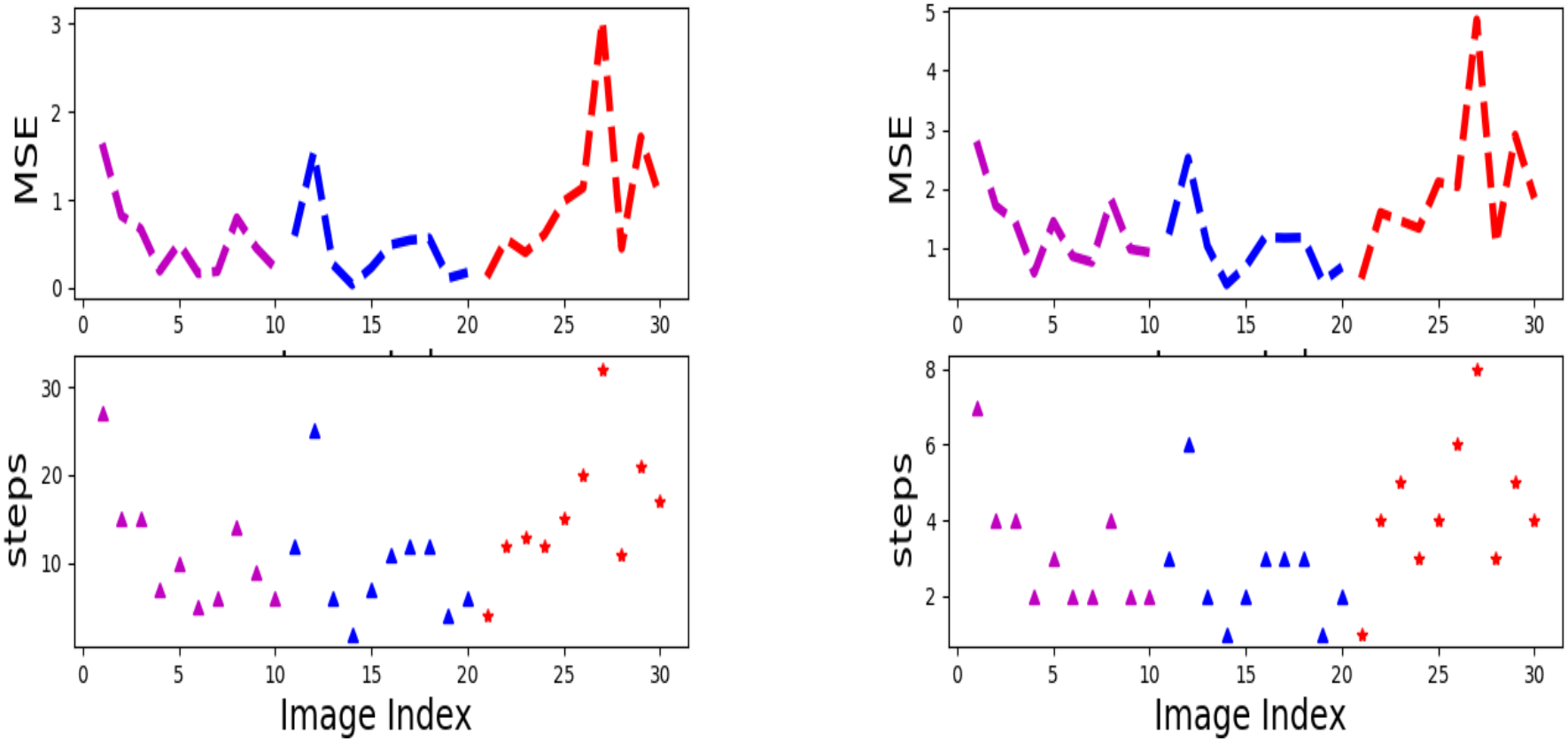


Figure 10: Left:$\alpha = 1.0$, Right: $\alpha = 10$. *HighKLD*, *MidKLD* , *LowKLD*

Confidence of misclassification

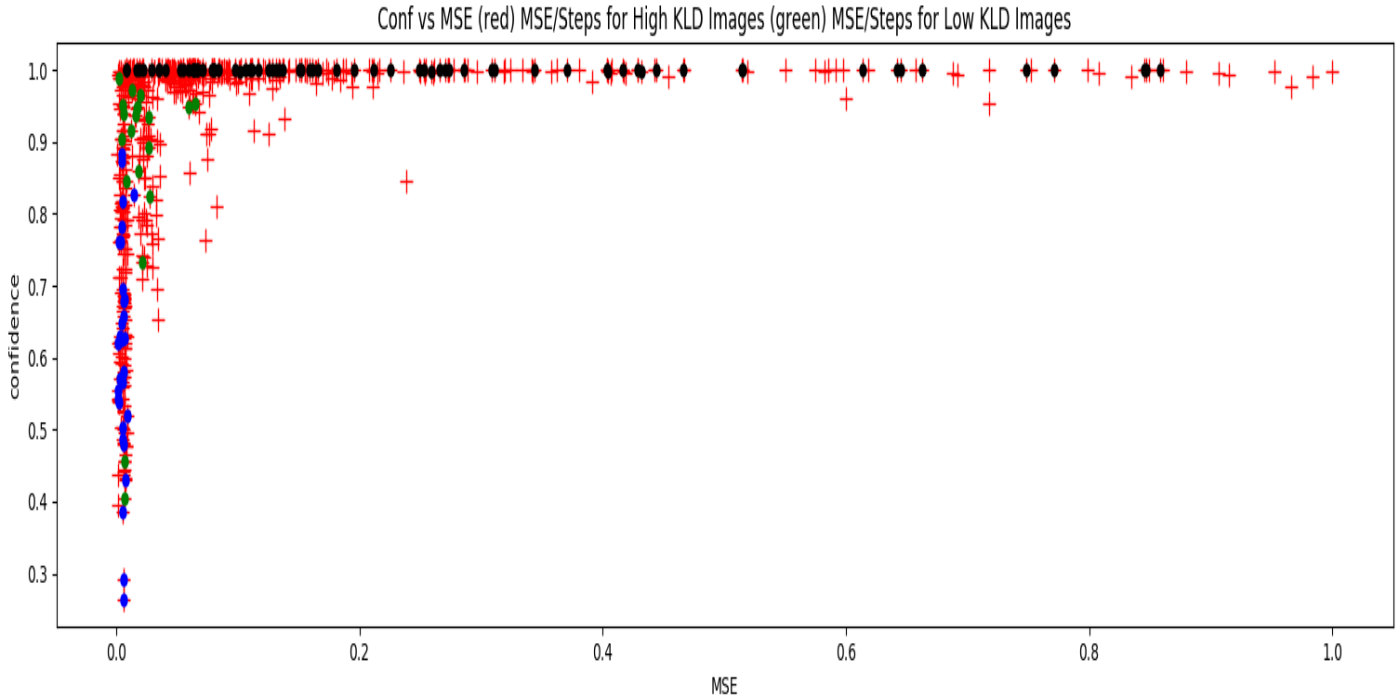


Figure 11: Blue : Low KLD , Black : High KLD , green Mid KLD

Imagenet on Resnet: Peturbed Image

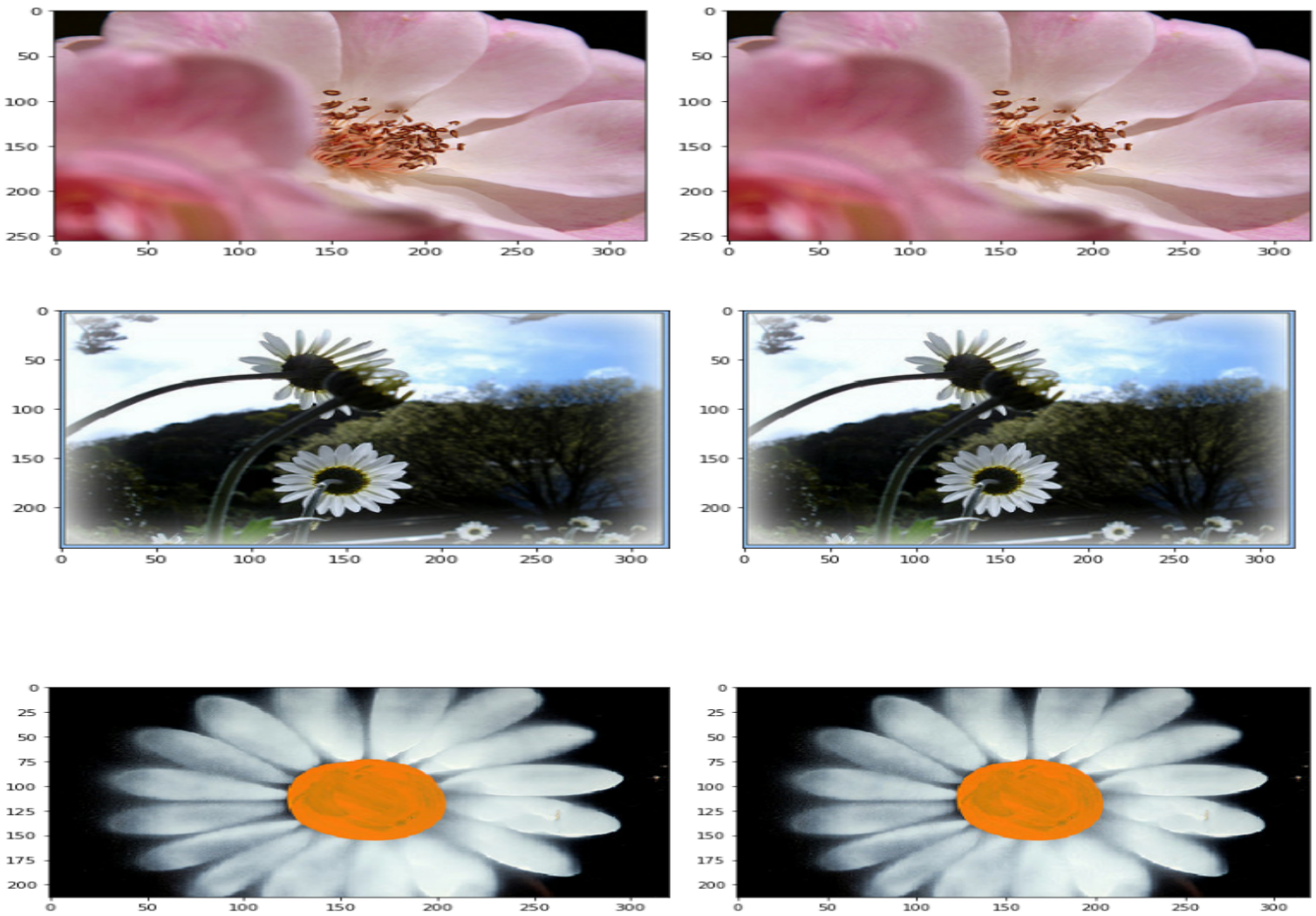


Figure 12: 3 Images: Original and Perturbed (High KLD , Mid KLD, Low KLD)

Imagenet on Resnet: KLD Vs MSE(Uses Approximation)

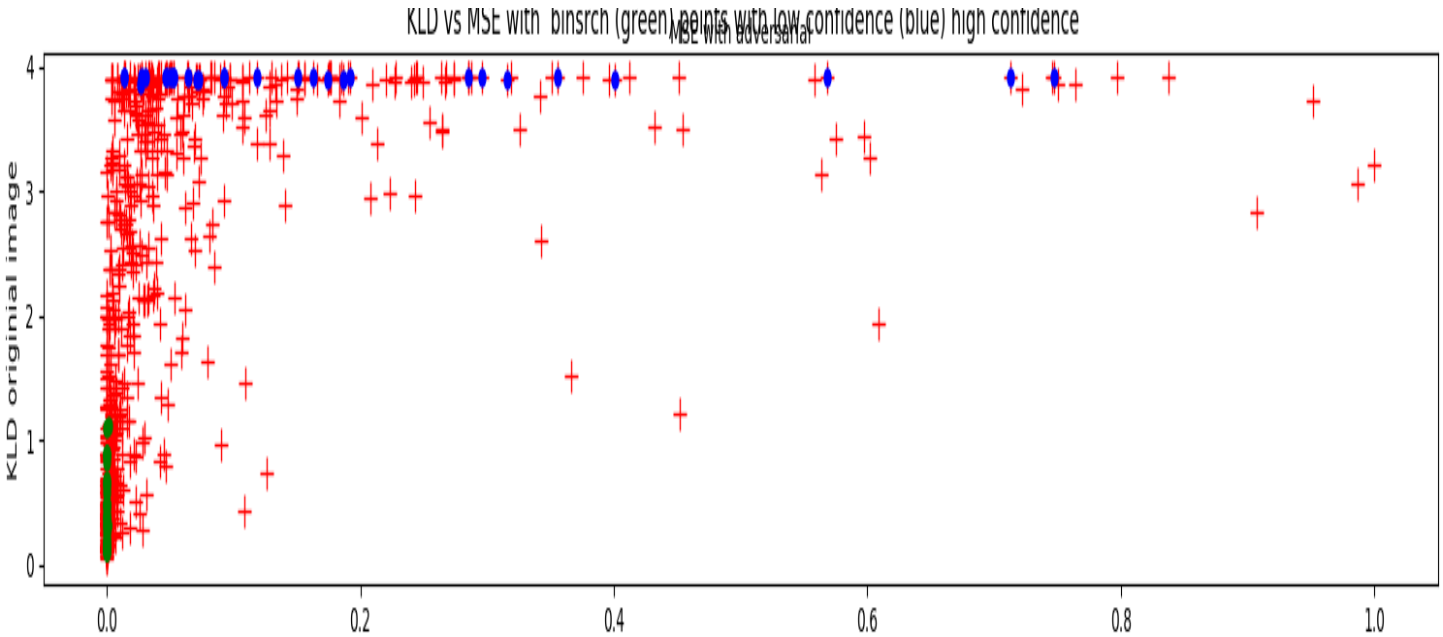


Figure 13: Green: Low KLD , Blue L high KLD

Histogram for Imagenet Resnet(Uses Approximation)

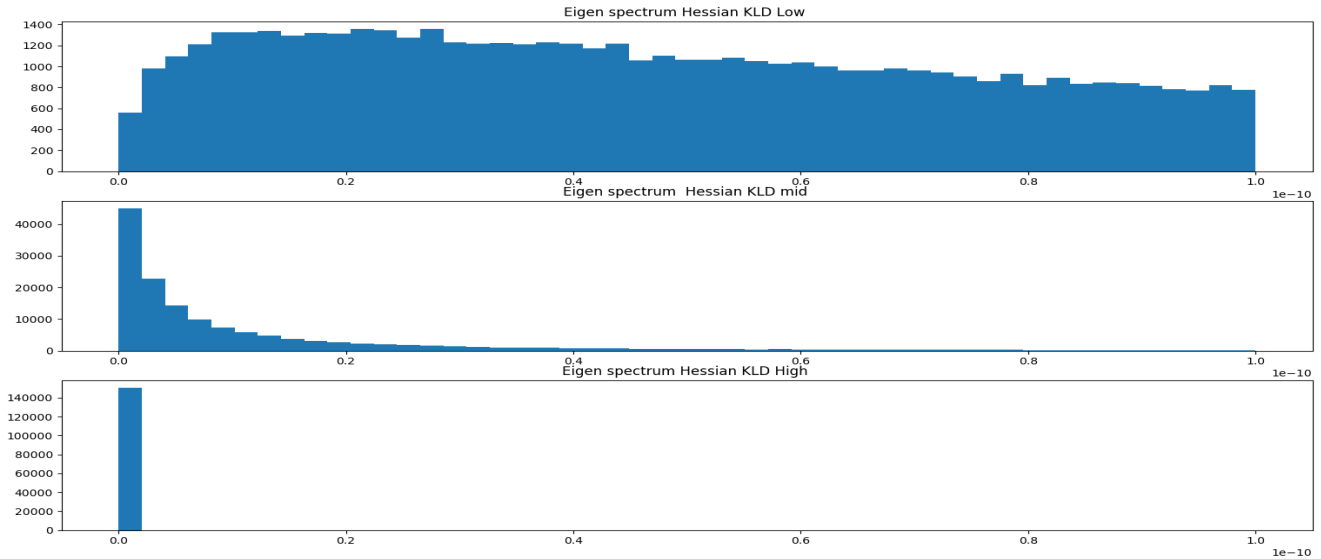


Figure 14: Top:Eigen Spectrum Low KLD , Middle Eigen Spectrum Med KLD , Bottom Eigen Spectrum high KLD

Imagenet on Resnet : KLD of perturbed images for every iteration

Plot below shows the variation of KLD for every perturbation. Changes in color are used to represent the Image label flips that occur.

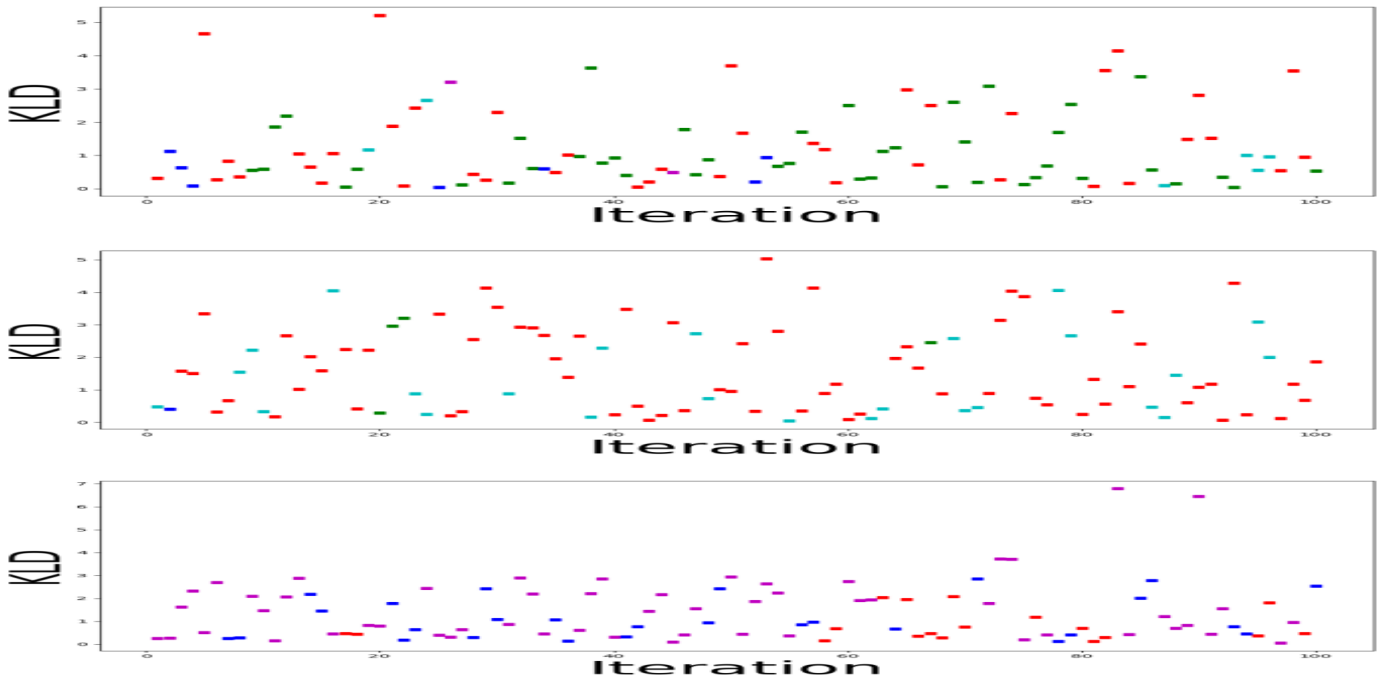
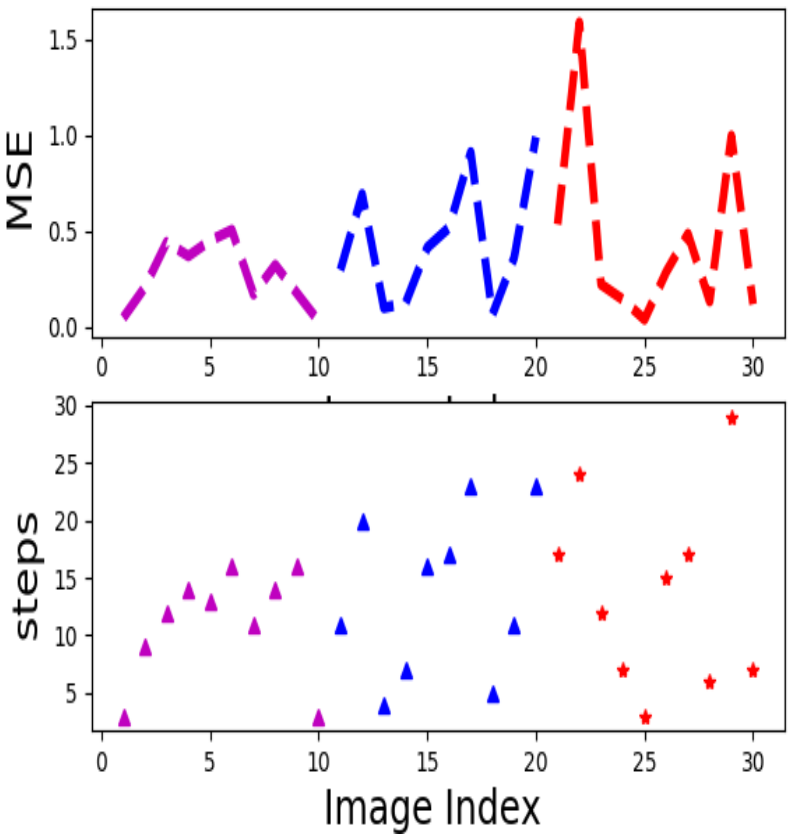


Figure 15: Top: Images with low KLD , Mid Image with mid KLD , Bottom Image with high KLD

Imagenet on Resnet: Plots on (i) MSE (ii) Number of Steps



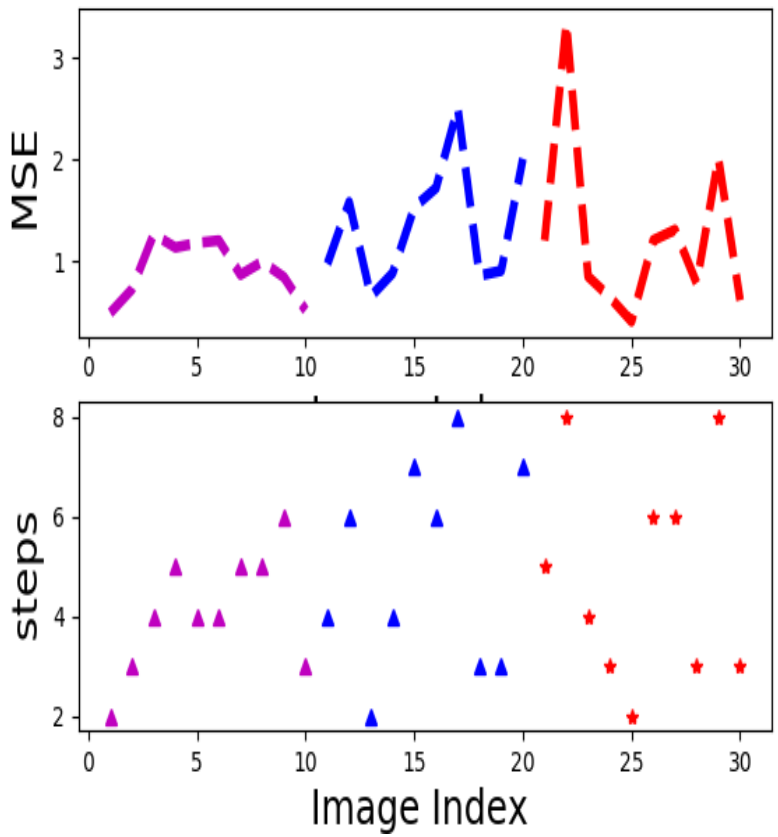
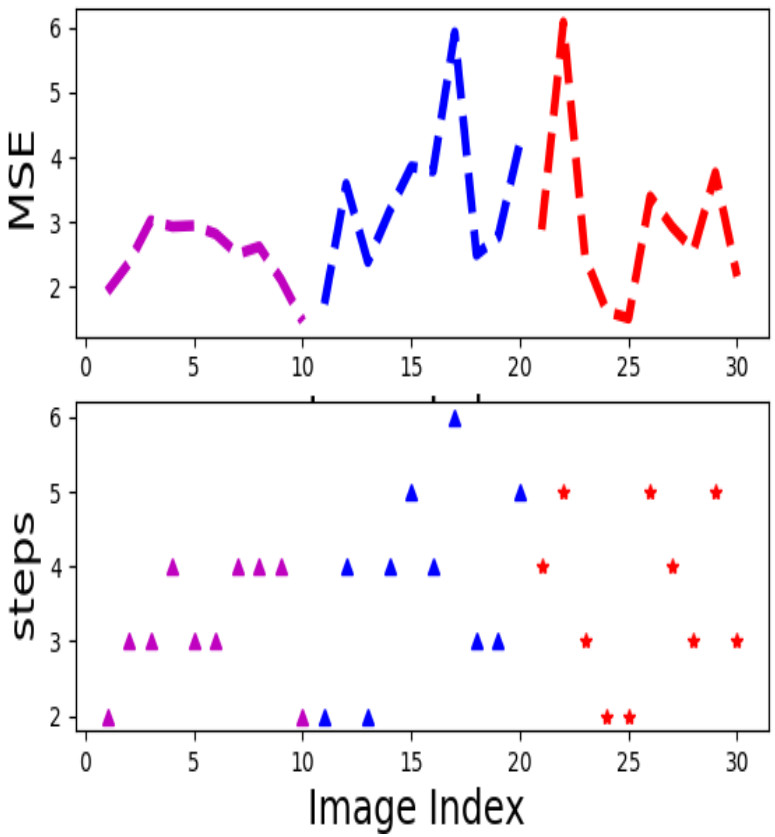


Figure 16: Left:$\alpha = 0.1$, Right: $\alpha = 0.5$, *HighKLD*, *MidKLD* , *LowKLD*

Imagenet on Resnet: Plots on (i) MSE (ii) Number of Steps



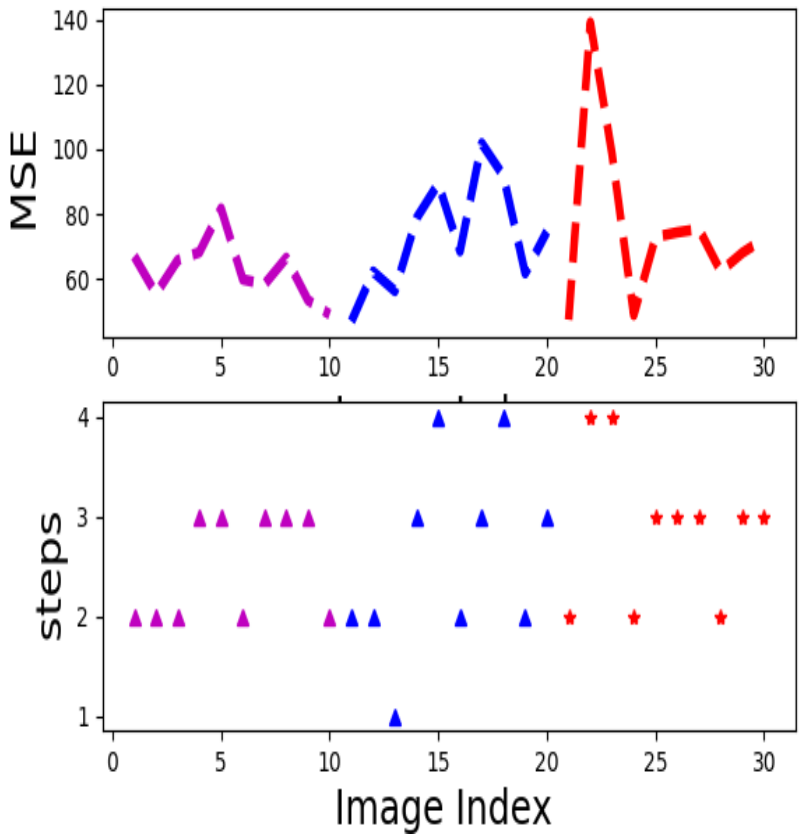


Figure 17: Left:$\alpha = 1$, Right: $\alpha = 10$, *HighKLD*, *MidKLD* , *LowKLD*

Binary Search to find closest points to decision boundary

- Another method was to find the closest data point on the decision boundary for every image in the test set
- perform a search between adversarial and original image until the label flips.
- Now pick the point closest to the decision boundary and plot the KLD and MSE .

Findings on binary search

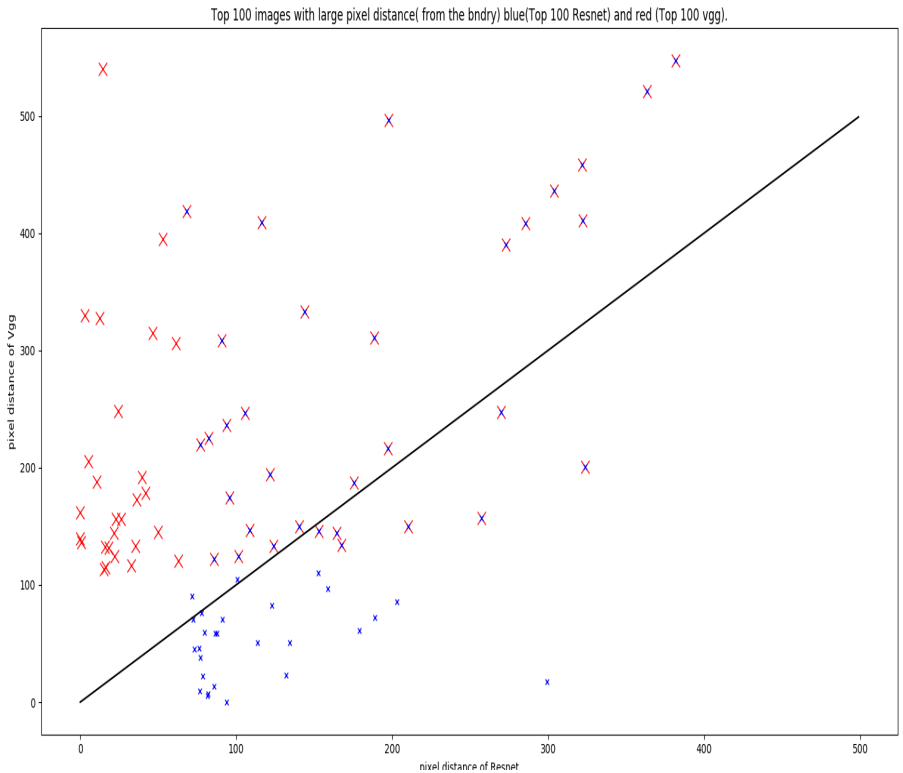


Figure 18: Red Labels are Images with high pixel distance on Resnet , Blue with VGG

We can note that most of the images are overlapping. However images with lower MSE are not transferable

High KLD ; Exact vs Approximate KL-Div Comparison

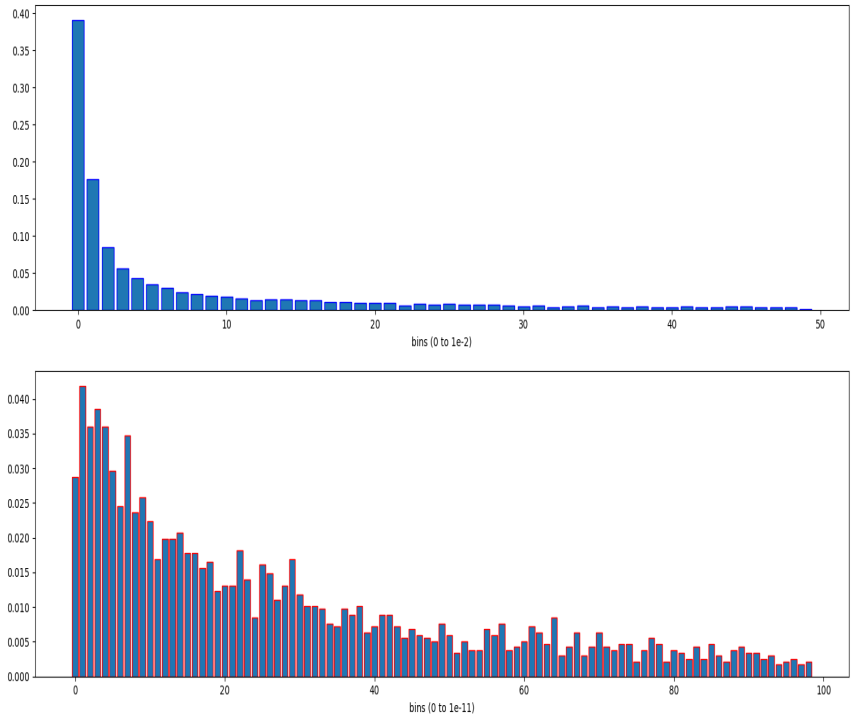


Figure 19: Top Exact : 0.713707614153 , Bottom Approx: 1.15163415619

Low KLD ; Exact vs Approximate KL-Div Comparison

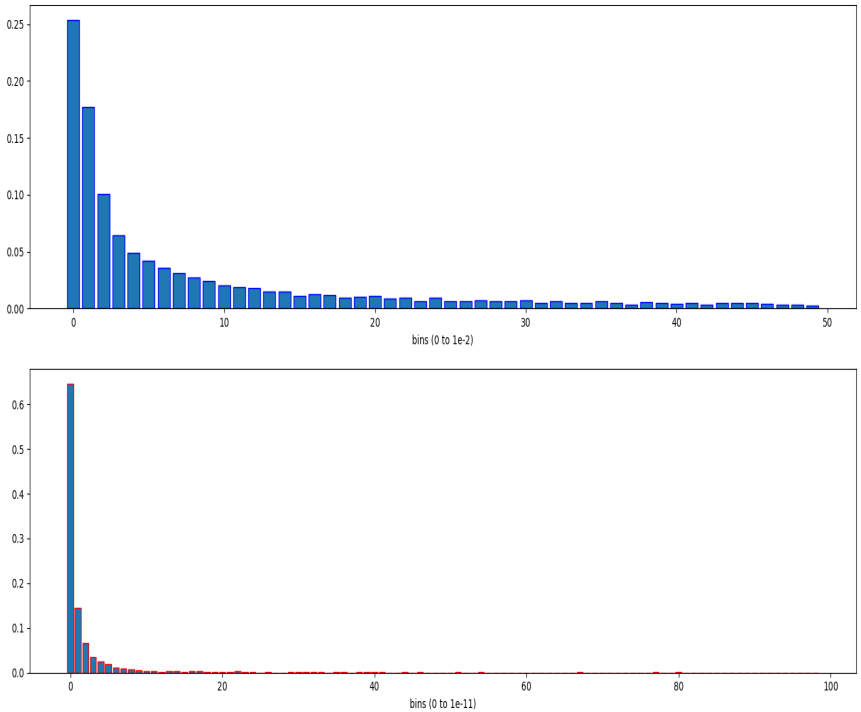


Figure 20: Top Exact : 0.677984972973 , Bottom Approx : 0.807335477142

Important findings on using KLD

1. $KL(p||q)$: If we are trying to approximate intractable distribution p with a tractable distribution q it works except for the case I describe below
2. Cause histograms for all images are different in scale bin sizes have to be varied and the range as well to view the histogram.
3. Even if one makes sure that $sum(q) = 1$ by scaling the values there is another condition of KLD that might not be satisfied. For a fixed bin size if we compute the KLD over all the test images some test images show 0 values in certain bins (peaky'ness in the distribution) this results in KLD being negative.

Hence we planned to re implement all of these for Jenson Shanon Divergence instead of KLD. But except for mathematical accuracy it did not provide any new insights

Conclusion

- The figures and conclusions in the report are based on the learnings from the CIFAR on ResNet, ImageNet on VGG & on ResNet .
- We cannot conclude whether an image can be easily perturbed or not based on the MSE and KLD stats of the single Image alone. But can only provide a probable measure
- we noticed that most of the adversarial images do lie close to decision boundary but there are some points that lie in flat loss landscapes.
- One can provide a `Expected measure of steps to misclassification` if we know the KL Divergence of the Image. But cannot conclude with absolute certainty about its closeness to decision boundary.

- We also 'Mean steps to mis-classify' a metric however we need a measure independent of the algorithm chosen.
- From binsearch we see that the most of the adversarial examples generated for Resnet with high MSE on flatter landscapes can misclassify on VGG as well . They are thus transferable.
- This statistic however cannot explain why for some images the minimum perturbation needed to flip the label places the data point on a flatter landscape.