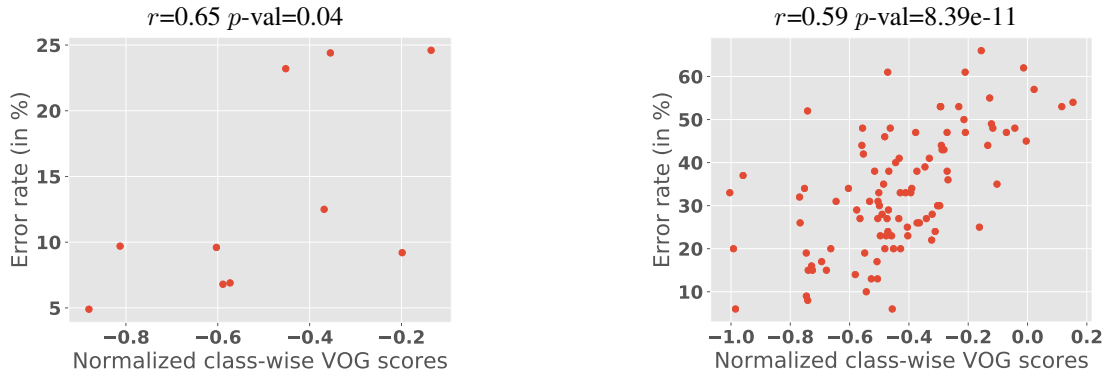
# Supplementary Materials: Estimating Example Difficulty using Variance of Gradients

## A. Toy Experiment

We generate the clusters for classification using scikit-learn and use a 90-10% split for dividing the dataset into train and test set [1]. We train a linear Multiple Layer Perceptron network with a hidden layer of 10 neurons using Stochastic Gradient Descent optimizer for 15 epochs. We divided the training process into three epoch stages: (1) $Early$ [0, 5), (2) $Middle$ [5, 10), and (3) $Late$ stage [10, 15). The trained model achieves a 0% test set error using a linear boundary (Fig. 1a).

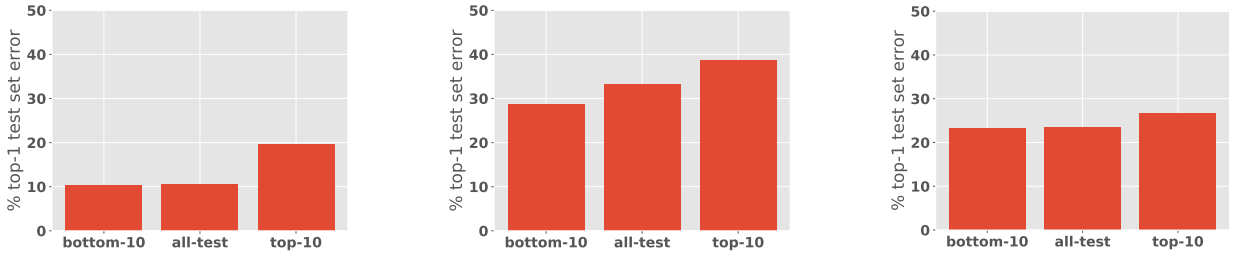
## B. Class Level Error Metrics and VoG

Here, we explore whether VoG is able to capture class level differences in difficulty. We compute VoG scores for each image in the test set of Cifar-10 and Cifar-100 (both test sets have $10,000$ images). In Fig. 9, we plot the average absolute VoG score for each class against the false negative rate for each class. We find that there is a positive, albeit weak, correlation between the two, classes with higher VoG scores have higher mis-classification error rate. The correlation between these metrics is $0.65$ and $0.59$ for Cifar-10 and Cifar-100 respectively. Given that VoG is computed on a per-example level, we find it interesting that the aggregate average of VoG is able to capture class level differences in difficulty.




(a) Cifar-10 (b) Cifar-100

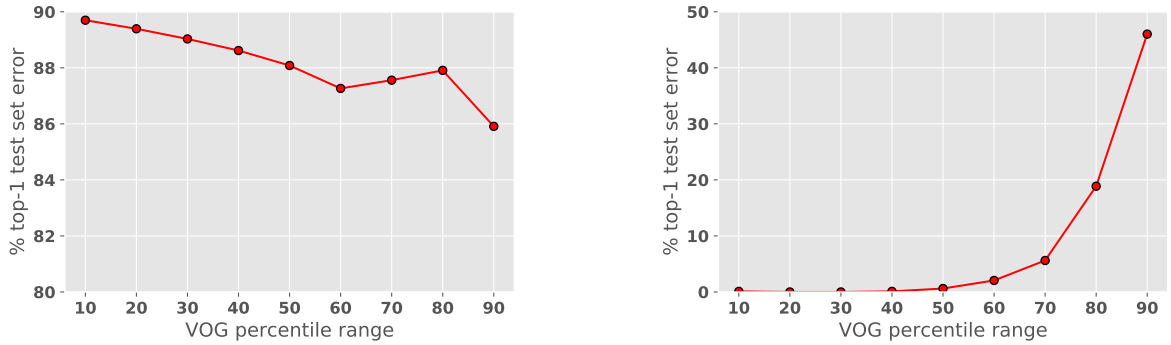
Figure 9. Plot of error rate (y-axis) against normalized class VoG scores for all classes (x-axis). There is a statistically significant positive correlation between class level error metrics and average VoG score (alpha set at 0.05).




(a) Cifar-10 (b) Cifar-100 (c) ImageNet

Figure 10. Bar plots showing the mean top-1 error rate (in %) for three group of samples from (1) the subset of the test set with the bottom 10th percentile of VoG scores, (2) the complete testing dataset, and (3) the subset of the test set with the top 10th percentile of VoG scores.

[1] Code and datasets available at https://github.com/chirag126/VOG.git




(a) Early-stage training (b) Late-stage training

Figure 11. The mean top-1 test set error (y-axis) for the exemplars thresholded by VoG score percentile (x-axis) in Cifar-10 testing set. The Early (a) and Late (b) stage VoG analysis shows inverse behavior where the role of VoG flips as the training progresses.

## C. Statistical Significance of Memorization Experiments

The two-sample $t$-test produces a $p$-value that can be used to decide whether there is evidence of a significant difference between the two distributions of VoG scores. The $p$-value represents the probability that the difference between the sample means is large, *i.e*. smaller the $p$-value, stronger is the evidence that the two populations have different means.

**Null Hypothesis**: $\mu_1 = \mu_2$ **Alternative Hypothesis**: $\mu_1 \neq \mu_2$

If the $p$-value is less than your significance level ($\alpha = 0.05$ in this experiment), you can reject the null hypothesis, *i.e*. the difference between the two means is statistically significant. The details for the individual $t$-tests for Cifar-10 and Cifar-100 are given below:

**Cifar-10:** The statistics for the samples in the correct and shuffled labels are:
Corrected labels: $\mu_1 = 0.62; \sigma_1 = 0.54; N_1 = 40000$
Shuffled labels: $\mu_2 = 0.85; \sigma_2 = 0.75; N_2 = 10000$
Result: $p$-value is $< 0.001$ — Reject Null Hypothesis (the two populations have different VoG means)

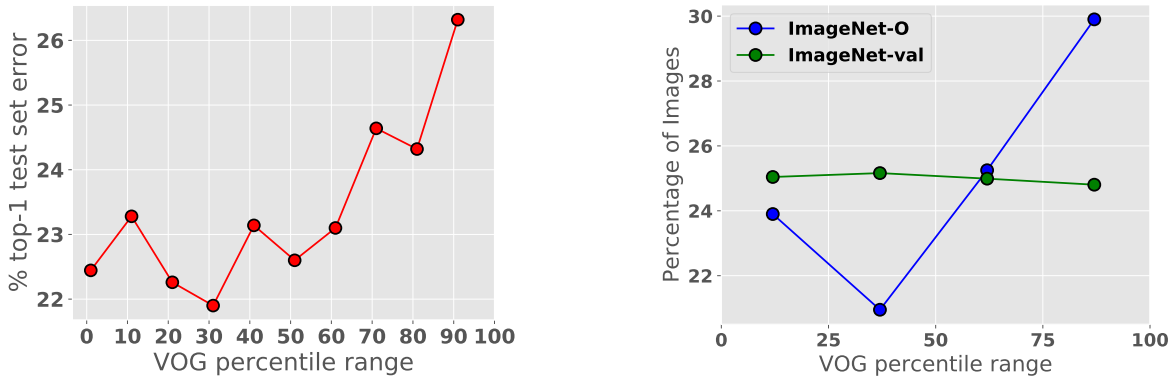
**Cifar-100:** The statistics for the samples in the correct and shuffled labels are:
Corrected labels: $\mu_1 = 0.54; \sigma_1 = 0.46; N_1 = 40000$
Shuffled labels: $\mu_2 = 0.82; \sigma_2 = 0.71; N_2 = 10000$
Result: $p$-value is $< 0.001$ — Reject Null Hypothesis (the two populations have different VoG means)

## D. Early training dynamics of Deep Neural Networks

Following Sec. 3, we plot the relationship between VoG and error rate of the testing dataset for Cifar-10 and Cifar-100. As in ImageNet, we observe a *flipping* trend between the early and late stages for both datasets (Figs. 8,11). We find that for easier datasets like Cifar, this point is only seen on using a lower learning rate (1e-3 in our experiments) for the early training stages.

## E. Detection of Distribution Shifts

We consider ImageNet-O [28], an open source curated out-of-distribution (OoD) dataset designed to fool classifiers. ImageNet-O consists of images that are not included in the original 1000 ImageNet classes. These images were selected with the goal of producing high confidence incorrect ImageNet-1K predictions of labels from within the training distribution. We are interested in understanding if VoG can correctly rank ImageNet-O examples as being atypical or OoD and expect to observe that ImageNet-O examples would be over-represented in top percentiles of VoG scores. In Fig. 12b, we observe that the percentage of ImageNet-O images are relatively over-represented at high levels of VoG, with 30% of all images in the top-25th percentile vs 24% in the bottom 25th percentile.




(a) (b)

Figure 12. **Left:** VoG is a valuable unsupervised tool as it can be computed using either the predicted/true label. We observe that misclassification increases with an increase in VoG scores. Across ImageNet, we observe that VoG calculated for the predicted labels follows the same trend as Fig. 7, where the top-10 percentile VoG scores have the highest error rate. **Right:** Number of ImageNet-O images across different VoG percentiles. We find that higher percentiles of VoG are significantly more likely to over-index on these OoD images.

Table 2. Number of images for each of the OoD dataset used in our OoD detection experiments.

| DATASET | DATASET SIZE |
|---|---|
| CIFAR-100 | 10000 |
| GAUSSIAN | 10000 |
| iSUN | 8920 |
| TINY-IMAGENET-RESIZE | 9810 |
| LSUN-RESIZE | 10000 |

## F. Out-of-Distribution Detection (OoD) Datasets and Model Architectures

Here, we carry out additional experiments to measure the effectiveness of VoG to detect OoD data. We run experiments using three DNN architectures: ResNet-18 [25], DenseNet [34] and WideResNet [71], and benchmark against Maximum Softmax Probability (MSP) [26], which is widely considered a strong baseline in OoD detection [26, 27] We follow the setup in [26] by setting all test set examples in CIFAR-10 as in-distribution (positive). For OoD examples (negative), we benchmark across four datasets: CIFAR-100, iSUN [48], TinyImageNet (Resize) [48], LSUN (Resize) [48], and Gaussian Noise. The Gaussian dataset was generated as described in [48], with $\mathcal{N}(0.5, 1)$. For the various ablations, the size of the OoD dataset can be seen in Table 2.

**Findings.** From Table 3, we observe that VoG is a valuable ranking for OoD detection and improves upon state-of-the-art uncertainty measures for many different tasks. On average, VoG outperforms MSP by large margins with a mean gain of 2.62% in AUROC, 2.33% in AUPR/In, and 2.47% in AUPR/Out across all three architectures and five datasets.

Table 3. Baseline comparison between VoG and Max Softmax Probability (MSP) for different models trained on Cifar-10. VoG is able to detect, both, In- and Out-Of-Distribution (OoD) samples with higher precision across different real-world datasets. For each row, values in **bold** represents superior performance.

| MODEL | IN- / OUT-OF-DISTRIBUTION | METRICS | AUROC /BASE | AUPR IN /BASE | AUPR OUT/BASE |
|---|---|---|---|---|---|
| W-RN-28-10 | C-10/C-100 | MSP | 80.9/50 | 83.4/50 | 75.4/50 |
| | | VOG | **89**/50 | **90.5**/50 | **87.3**/50 |
| | C-10/GAUSSIAN | MSP | 78.1/50 | 84.6/50 | 66.4/50 |
| | | VOG | **88.2**/50 | **91.6**/50 | **80.6**/50 |
| | C-10/ISUN | MSP | 87.8/50 | 90.7/52.8 | 82.9/47.2 |
| | | VOG | **93.3**/50 | **95.3**/52.8 | **89.4**/47.2 |
| | C-10/TINY-IMAGENET-RESIZE | MSP | 88.4/50 | 91/50.5 | 83.4/49.5 |
| | | VOG | **92.8**/50 | **94.3**/50.5 | **89.9**/49.5 |
| | C-10/LSUN-RESIZE | MSP | 90.4/50 | 92.7/50 | 86.6/50 |
| | | VOG | **93.5**/50 | **94.9**/50 | **90.8**/50 |
| RESNET-18 | C-10/C-100 | MSP | 86.8/50 | 89.7/50 | 82.3/50 |
| | | VOG | **87.6**/50 | **90**/50 | **84**/50 |
| | C-10/GAUSSIAN | MSP | **92.7**/50 | **95.1**/50 | **88.2**/50 |
| | | VOG | 85.1/50 | 90.6/50 | 73/50 |
| | C-10/ISUN | MSP | 85.5/50 | 89/52.8 | 79.9/47.2 |
| | | VOG | **92.3**/50 | **94.2**/52.8 | **89.3**/47.2 |
| | C-10/TINY-IMAGENET-RESIZE | MSP | 84.7/50 | 87.4/50.5 | 79.8/49.5 |
| | | VOG | **91.6**/50 | **93.1**/50.5 | **89.5**/49.5 |
| | C-10/LSUN-RESIZE | MSP | 84.3/50 | 86.4/50 | 80/50 |
| | | VOG | **92.3**/50 | **93.6**/50 | **90.4**/50 |
| DENSENET-BC | C-10/C-100 | MSP | 91.4/50 | 93.1/50 | 88.5/50 |
| | | VOG | **93.1**/50 | **94.3**/50 | **91**/50 |
| | C-10/GAUSSIAN | MSP | **95.8**/50 | **97.3**/50 | **92.7**/50 |
| | | VOG | 88.2/50 | 93.4/50 | 74.3/50 |
| | C-10/ISUN | MSP | **92.8**/50 | **95**/52.8 | **88.9**/47.2 |
| | | VOG | 92.5/50 | 94.9/52.8 | 86.5/47.2 |
| | C-10/TINY-IMAGENET-RESIZE | MSP | **91.3**/50 | **93.1**/50.5 | **88.2**/49.5 |
| | | VOG | 90.6/50 | 92.6/50.5 | 86.1/49.5 |
| | C-10/LSUN-RESIZE | MSP | 92.9/50 | 94.7/50 | **90**/50 |
| | | VOG | **93**/50 | **94.9**/50 | 88.2/50 |