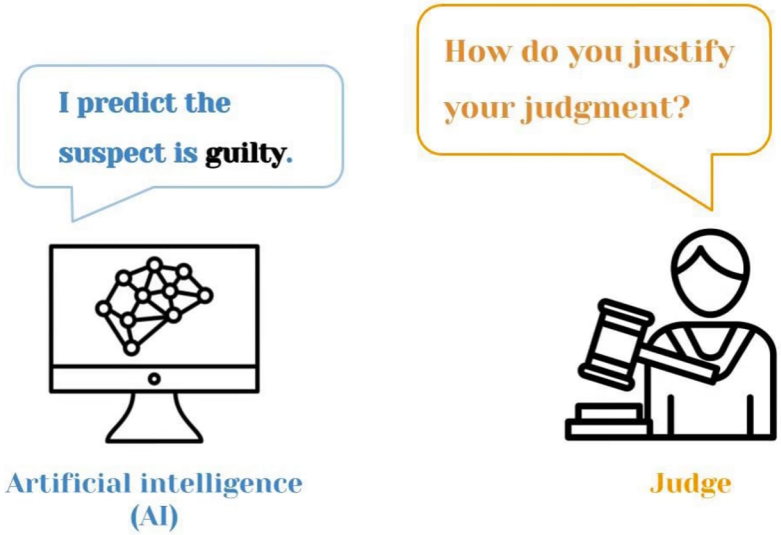
Estimating Example Difficulty using Variance of Gradients

Chirag Agarwal*, Sara Hooker*

Google Research

Interpretability in current models?



Justice system

Medical diagnosis

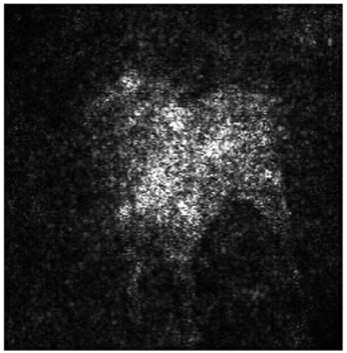
Motivation for Studying Relative Importance

- Automatically surfaces high priority examples for human inspection
- Intra and Inter class understanding
- Model bias

Gradient

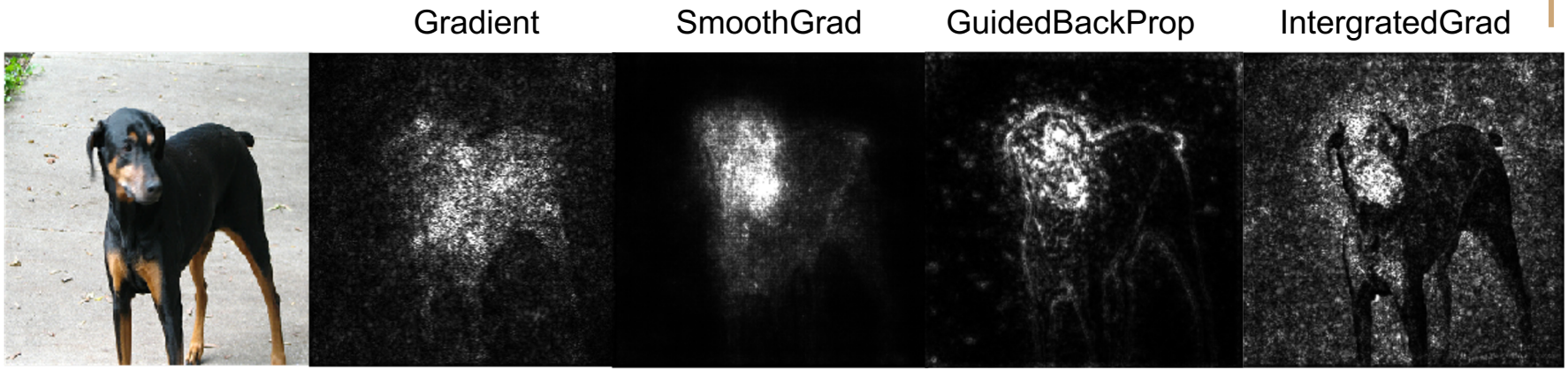
Saliency maps can be noisy and ‘uninterpretable’

$$y = f(x)$$



$$S = \nabla_x f(x)$$

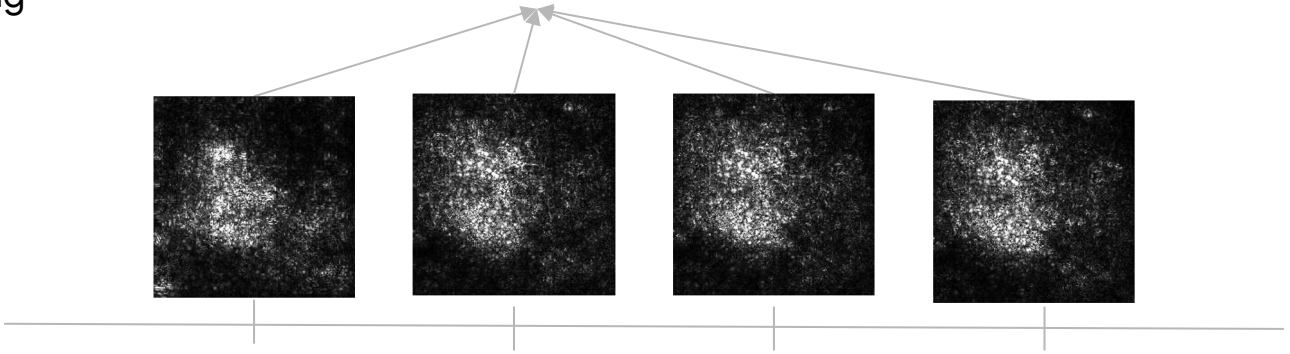
Different variants of Gradients



Variance Of Gradients

$$VOG_i = \sqrt{\left(\frac{1}{K}\sum_{t=1}^{K}\frac{1}{N}\left(\boldsymbol{s_{ti}} - \mu_i\right)^2\right)}$$

Compute the average variance in gradients (VOG) for an image over training



0 epochs

90 epochs

Understand how feature importance forms over the course of training.

Low Variance

High Variance

Low Variance

High Variance

0 epochs

Early-stage training

90 epochs

Late Stage Training

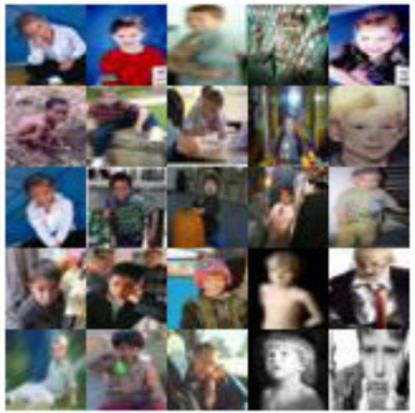
Achille et al. 2017. Critical learning periods in deep neural networks. arXiv preprint arXiv:1711.08856.
Mangalam, K. and Prabhu, V. U. Do deep neural networks learn shallow learnable examples first. 2019.
Jiang et al. 2020.. Exploring the memorization-generalization continuum in deep learning. arXiv preprint arXiv:2002.03206

VOG computes a relative ranking of each class

Higher VOG score samples have cluttered backgrounds

Lowest VOG

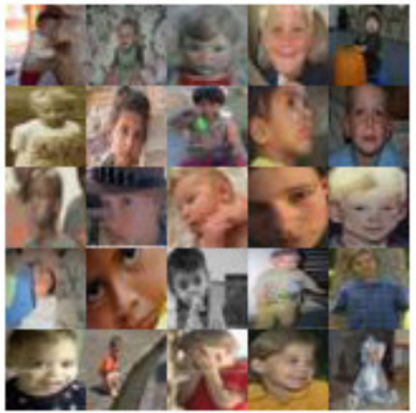
Highest VOG



Early-stage training

Lowest VOG

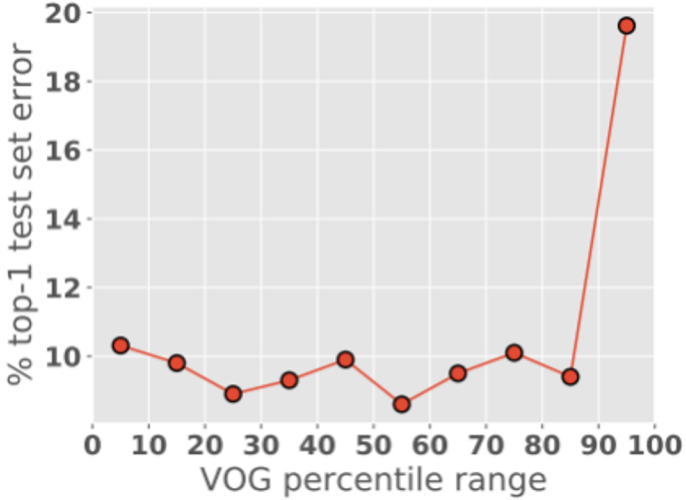
Highest VOG



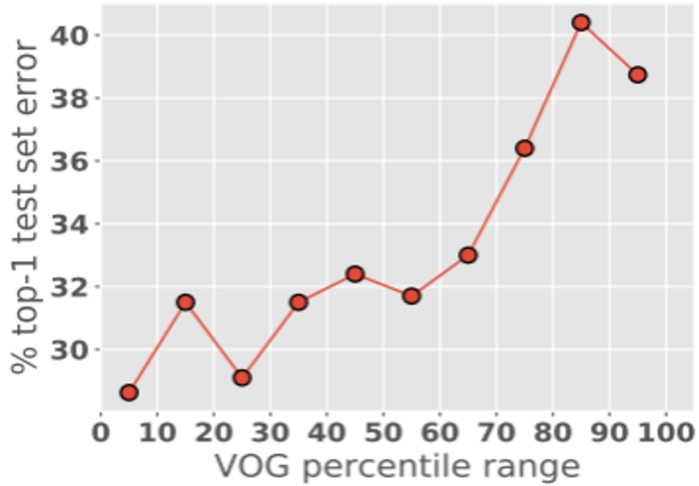
Late-stage training

VOG effectively discriminates easy & challenging examples

Mis-classification increases with an increase in VOG scores



CIFAR-10



CIFAR-100

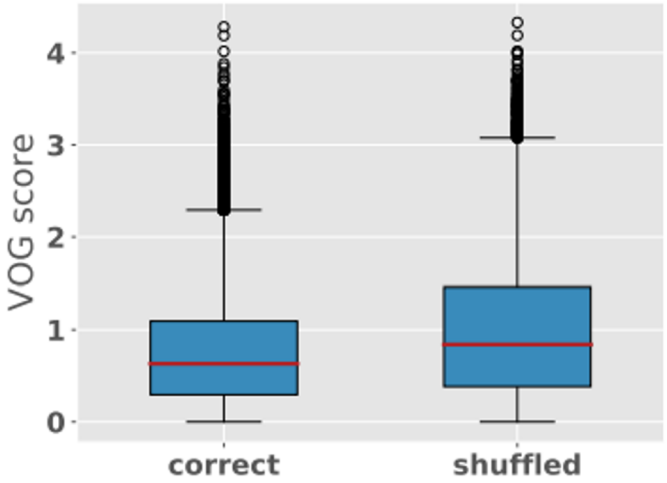
Memorization experiment

- Overparameterized networks achieve zero training error by memorizing examples*
- Replace 20% of all labels in the training set with random shuffled labels
- Re-train the model from random initialization and compute VOG scores

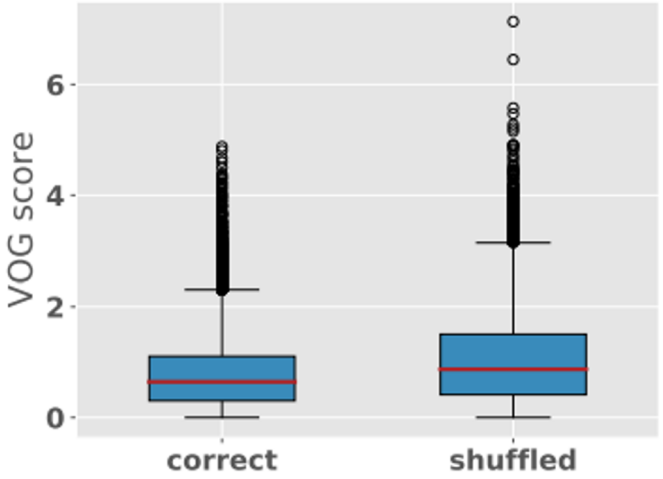
*Zhang et al. Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530, 2016.

VOG distribution

Higher mean (red line) and spread for shuffled data



CIFAR-10



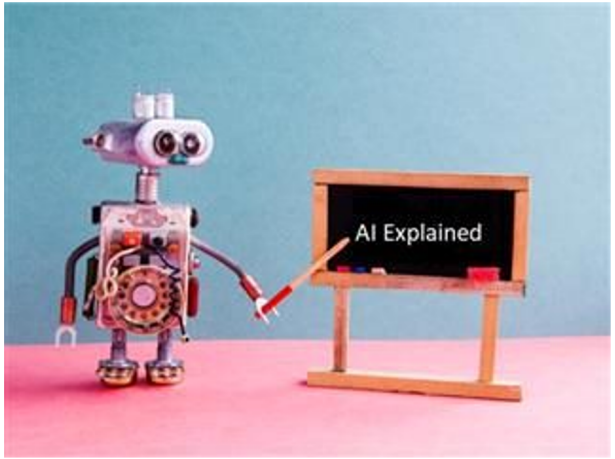
CIFAR-100

Future directions

- Use VOG score to analyze complex datasets
- Investigate various dataset and model bias
- Leverage VOG to aid in curriculum learning

Final takeaways

- An interpretability tool for ranking training and test examples
- No need for modifying the architecture
- VOG aids in clustering images with distinct visual properties



Paper: https://arxiv.org/pdf/2008.11600.pdf

chiragagarwall12@gmail.com
@_cagarwal
shooker@google.com
@sarahookr