Distorted Representation Space Characterization Through Backpropagated Gradients

Gukyeong Kwon, Mohit Prabhushankar*, Dogancan Temel, and Ghassan AlRegib (*: equal contribution)*
Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
{gukyeong.kwon,mohit.p,cantemel,alregib}@gatech.edu

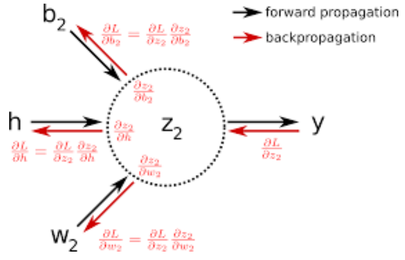
ICIP 2019

https://ghassanalregib.info/

Motivation

- Distortions in the visual domain creates a distributional shift between train and test images for neural networks.
- Existing approaches overcome this distributional shift by utilizing weight retraining and domain adaptation techniques both of which require large amount of distorted data.
- However, in some applications like Image Quality Assessment (IQA) and Out-of-distribution (OOD) detection, we do not have access to distorted data.
- In this paper, we propose a technique to characterize distorted representation spaces in an unsupervised manner.

Gradients in Neural Networks



y : Output
z_2: Hidden Unit
b_2, w_2: Bias and Weight
h : Input
L : Loss

Forward Propagation :

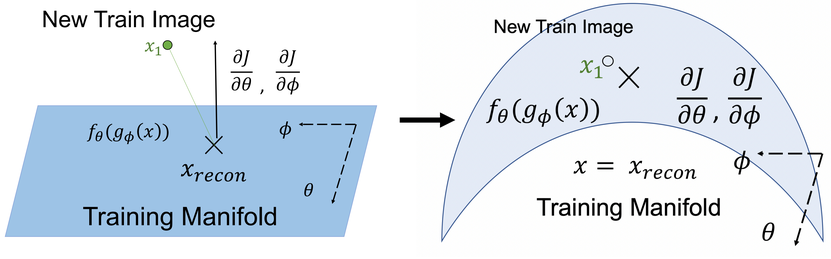
$$y = z_2 = sigmoid(w_2 \times h + b_2)$$

Backpropagation on weights :

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial w_2}$$

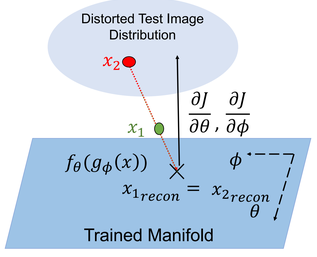
Geometric Interpretation of Gradients

Gradients represent required changes in the training manifold to incorporate new train images.



Existing Features for Distortion Characterization

x_1(pristine) and x_2 (distorted) are projected onto trained manifold to obtain $x_{1_{recon}}, x_{2_{recon}}$

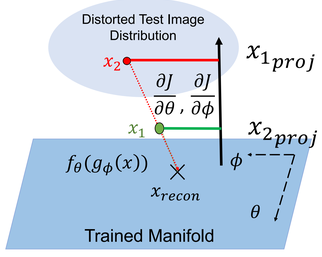


Feature = Difference($x_{1_{recon}}, x_{2_{recon}}$)

Distortion not Characterized!

Backpropagated Gradients as Features

x_1(pristine) and x_2 (distorted) are projected onto gradient space to obtain $x_{1_{proj}}, x_{2_{proj}}$



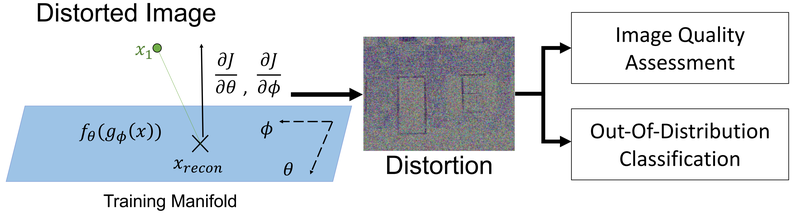
Feature = Difference($x_{1_{proj}}, x_{2_{proj}}$)

Distortion Characterized!

Gradient features :

- By definition, characterizes changes from trained pristine manifold
- Unsupervised technique
- Effective directional measures

Gradients for Characterizing Distortions



- Image Quality Assessment : To objectively estimate the perceptual quality of images.
- Out-of-Distribution Classification: To classify challenge-free images as in-distribution and challenge images as out-of-distribution

Experimental Results

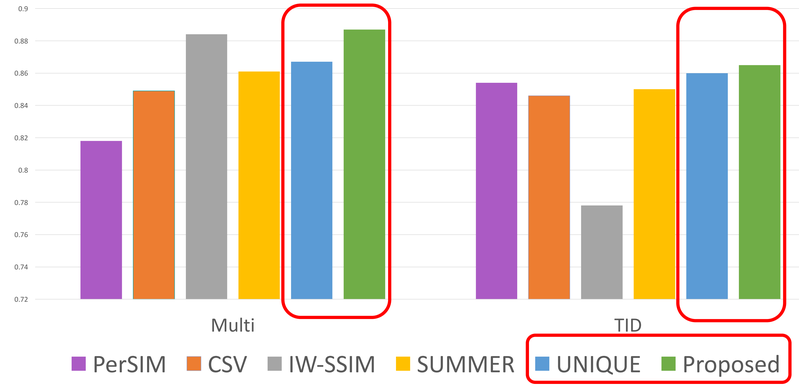
Out-of-Distribution Classification

- DE (Decolorization), CE (Codec error), NO (Noise), LB (Lens Blur), DL (Dirty Lens), RA (Rain)

Method	Accuracy					
	Non-Blur Types			Blur Types		
	DE	CE	NO	LB	DL	RA
VAE-A (z_g)	56.51	62.01	84.72	93.49	93.70	94.10
VAE-R ($\frac{\partial L}{\partial \phi}$)	**70.24**	60.64	68.59	89.08	90.83	92.57
VAE-L ($\frac{\partial L}{\partial \theta}$)	56.18	63.94	83.04	94.51	94.08	94.74
Proposed	62.67	**64.76**	**87.57**	**95.26**	**95.71**	**95.85**

Image Quality Assessment

Spearman Correlation Coefficient on TID2013 and Multi-LIVE dataset



Useful Links