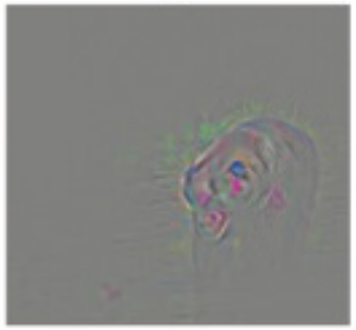
Explaining the Unexplained: A **CL**ass-**E**nhanced **A**ttentive **R**esponse (CLEAR) Approach to Understanding Deep Neural Networks

Devinder Kumar*, Alexander Wong & Graham W. Taylor

UNIVERSITY of GUELPH

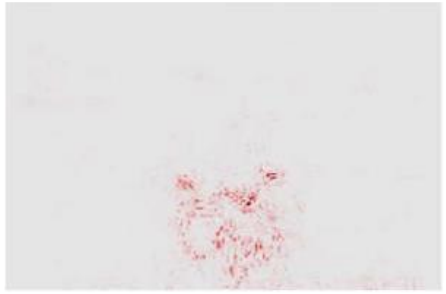
VECTOR INSTITUTE

Current Approaches – Heatmap/Attention based!

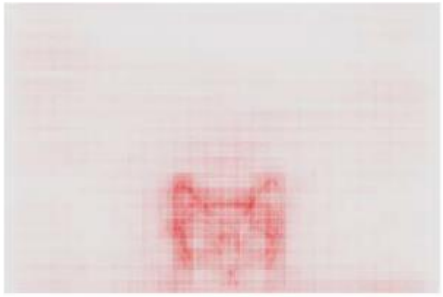



Deconvolution: Zeiler et.al. ECCV 14

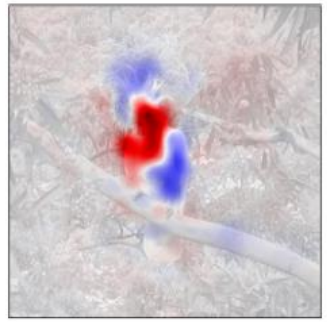

Guided backpropagation: ICLR 2015



Saliency: Simonyan et.al.
CVPR 2013

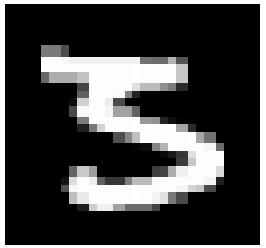
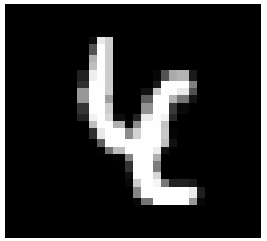
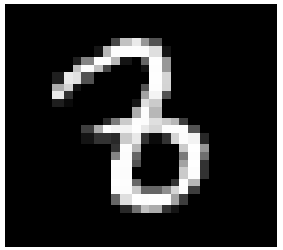
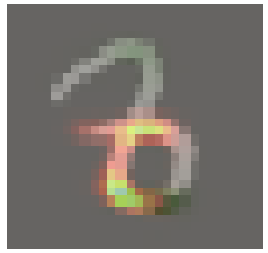


Deep Taylor Decomp.
Montavon et. al.
PR journal 2017

Prediction Difference:
Zintgraf et. al. ICLR 2017

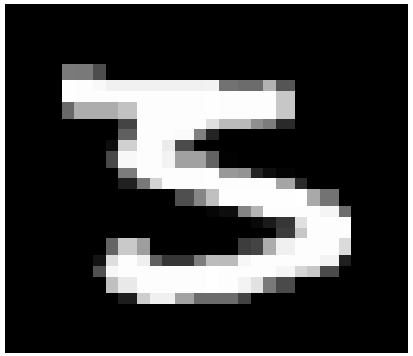
Interpretations

Input	Output	Heatmap	Interpretation
 DNN	3 ✓		Focuses on right areas: Looks correct!
 DNN	2 ✗		Probably focuses on correct part, but why 3?
 DNN	3 ✗		Focuses on wrong part, curve might be two; but why not 3 or 5 or 6?

Class Enhanced Attentive Response (CLEAR) Map

Binary Heatmap

CLEAR Map



Class Enhanced Attentive Response (CLEAR)

$$\hat{h}_l = \sum_{k=1}^{K} z_{k,l} * w_{k,l}.$$

Output response at layer l

$$R_l = G_1 U_1 G_2 U_2 .... G_{l-1} U_{l-1} G_l z_l$$

Output response of layer l

$$R(\underline{x}|c) = G_1 U_1 G_2 U_2 .... G_{L-1} U_{L-1} G_L^c z_L.$$

given $c\ (1 \leq c \leq N)$

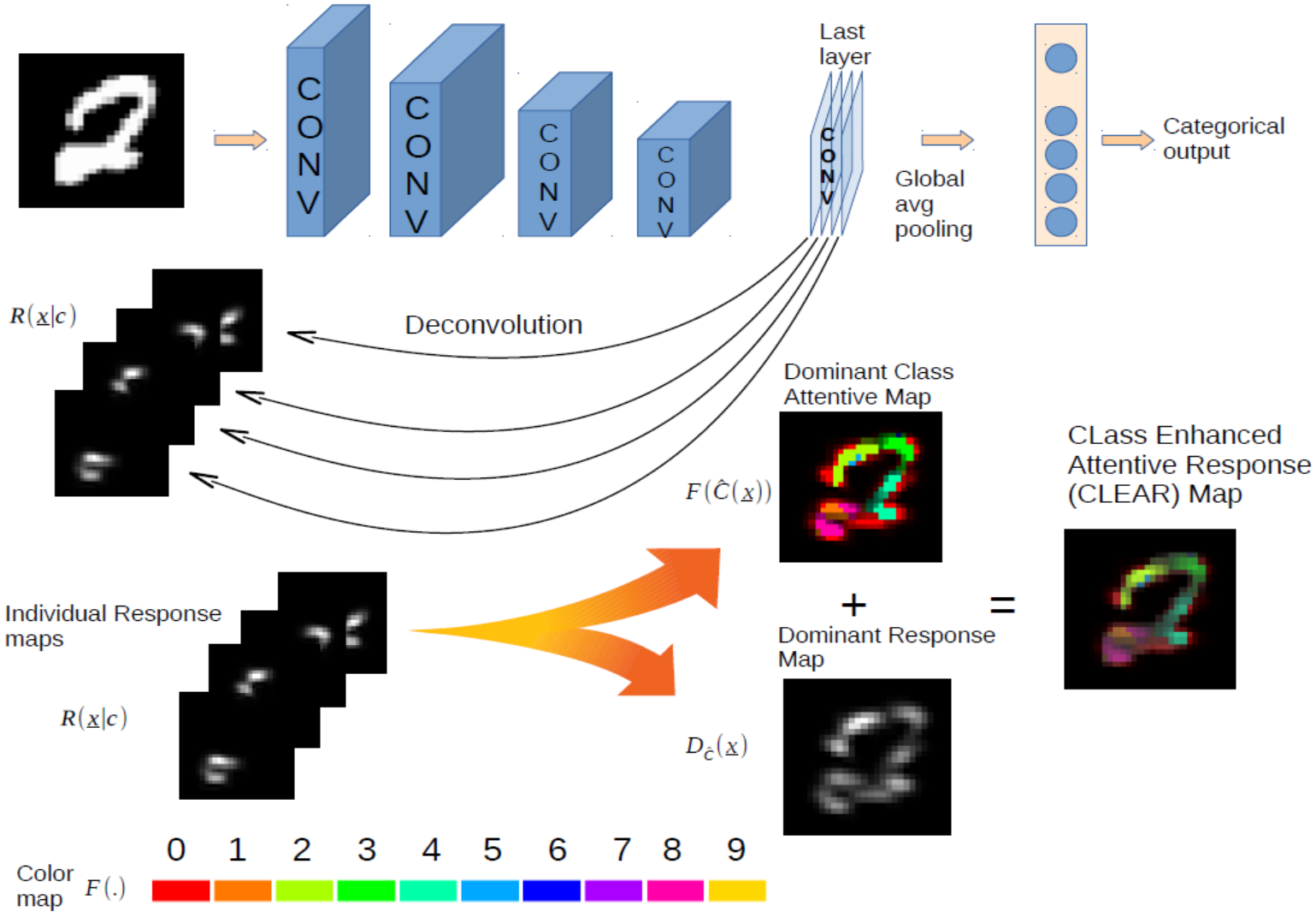
$$\hat{C}(\underline{x}) = \underset{c}{\operatorname{argmax}}\, R(\underline{x}|c).$$

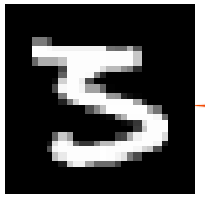
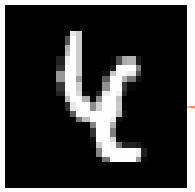
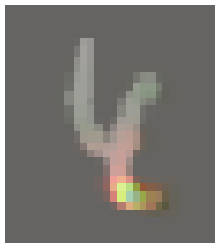
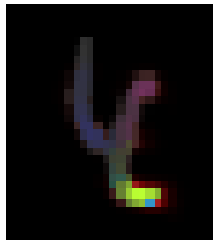
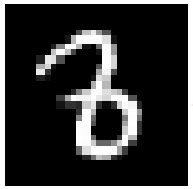
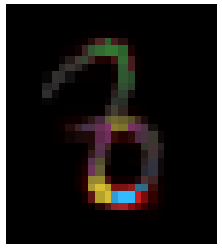
Dominant Class Response

$$D_{\hat{C}}(\underline{x}) = R(\underline{x}|\hat{C}).$$

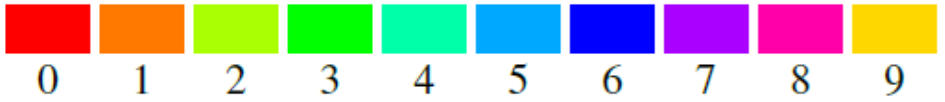
Dominant Response

Class Enhanced Attentive Response (CLEAR) Map

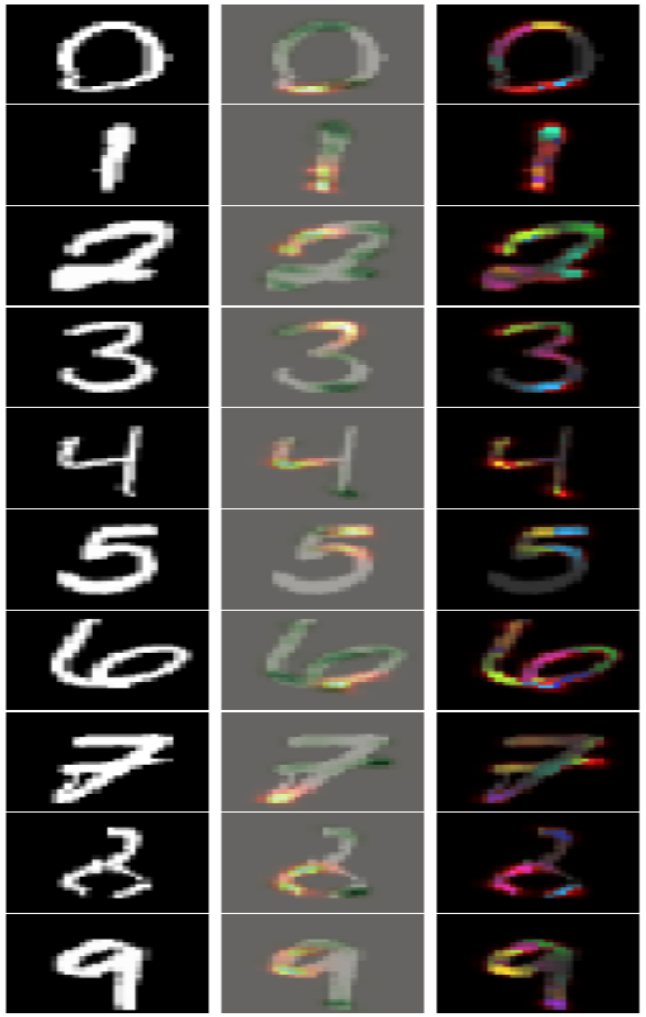


Input	Output	Heatmap	Interpretation	CLEAR map	Interpretation
	DNN → 3 ✓		Focuses on right areas: Looks correct!		Major part of the positive focus is of 3
	DNN → 2 ✗		Focuses on wrong part, curve might be two; but why not 3 or 5 or 6?		Major part of the positive focus represents 2
	DNN → 3 ✗		Probably focuses on correct part, but why 3?		Major part of negative focus is 3; higher activation than any other class

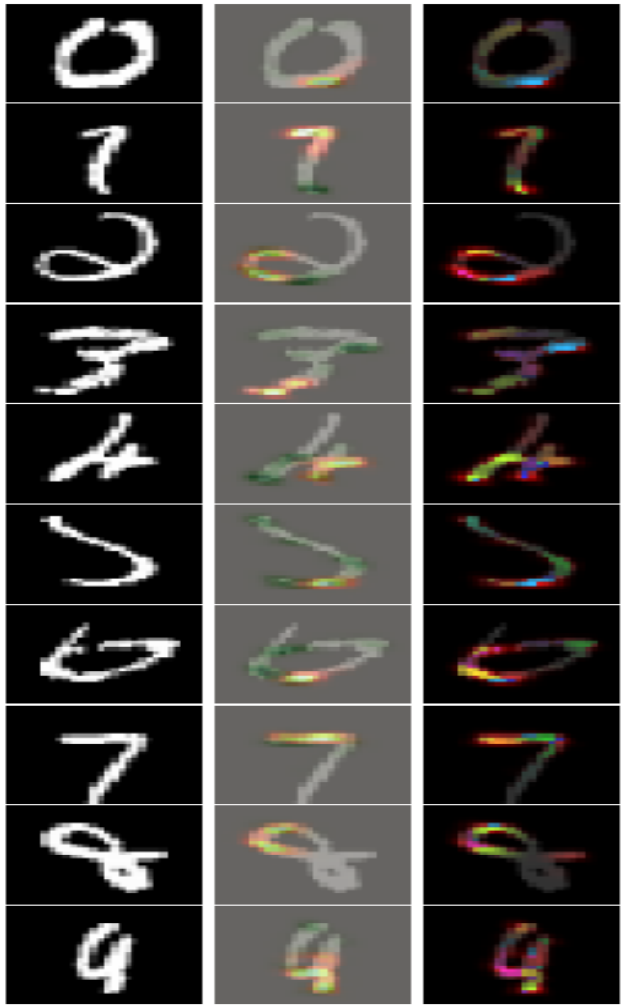
MNIST RESULTS



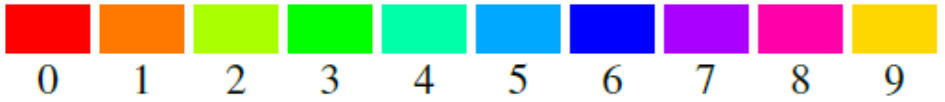
Correctly Classified



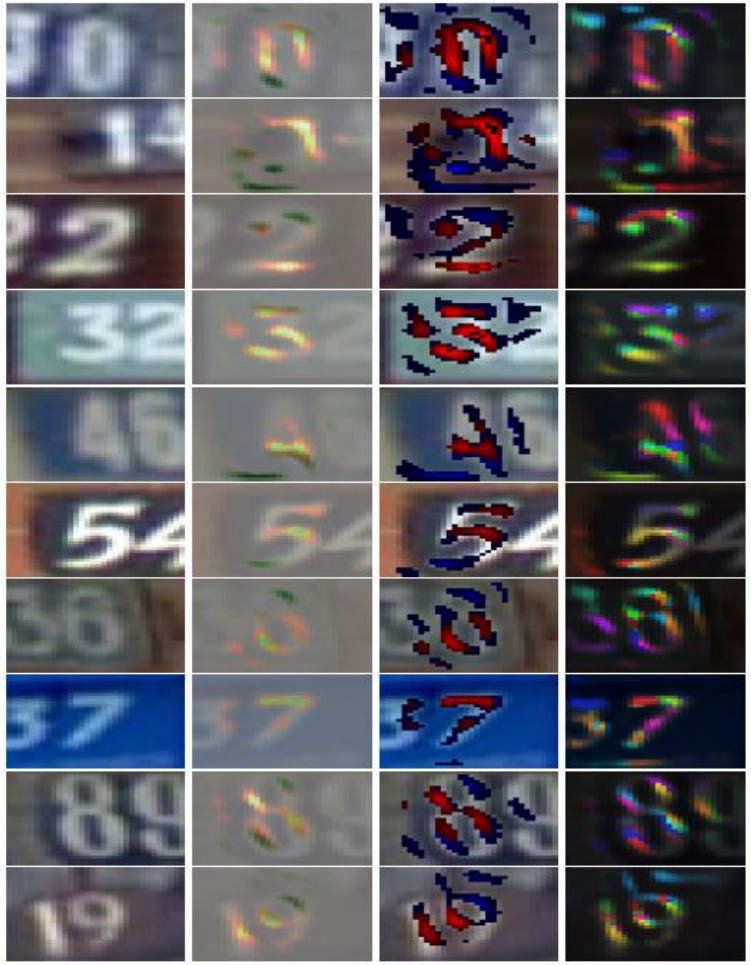
Wrongly Classified



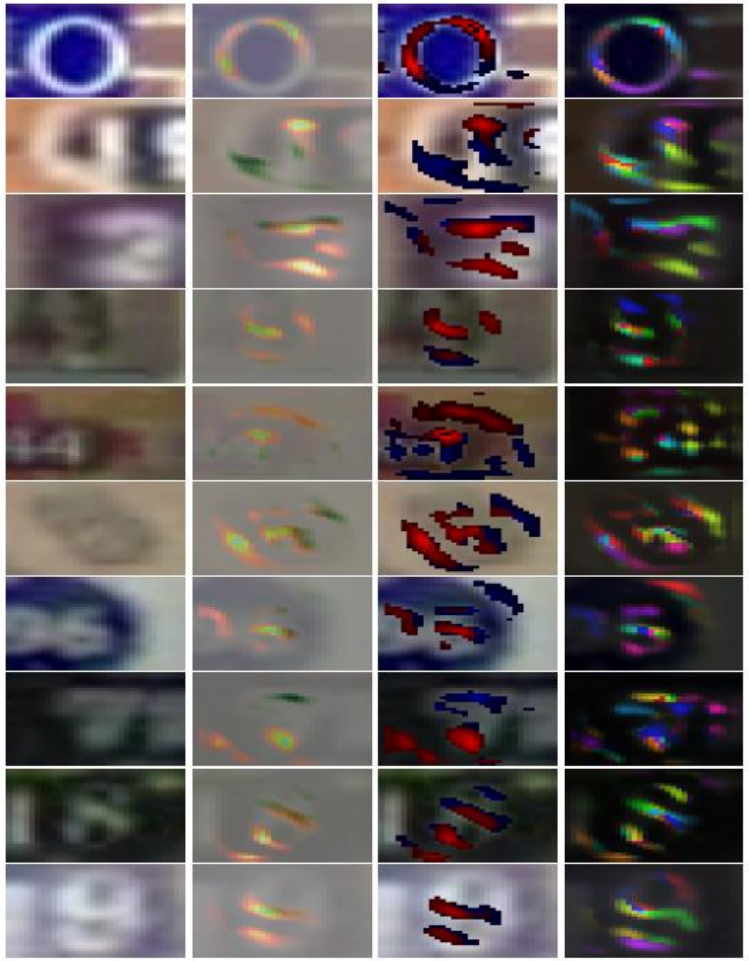
SVHN RESULTS



Correctly Classified



Wrongly Classified

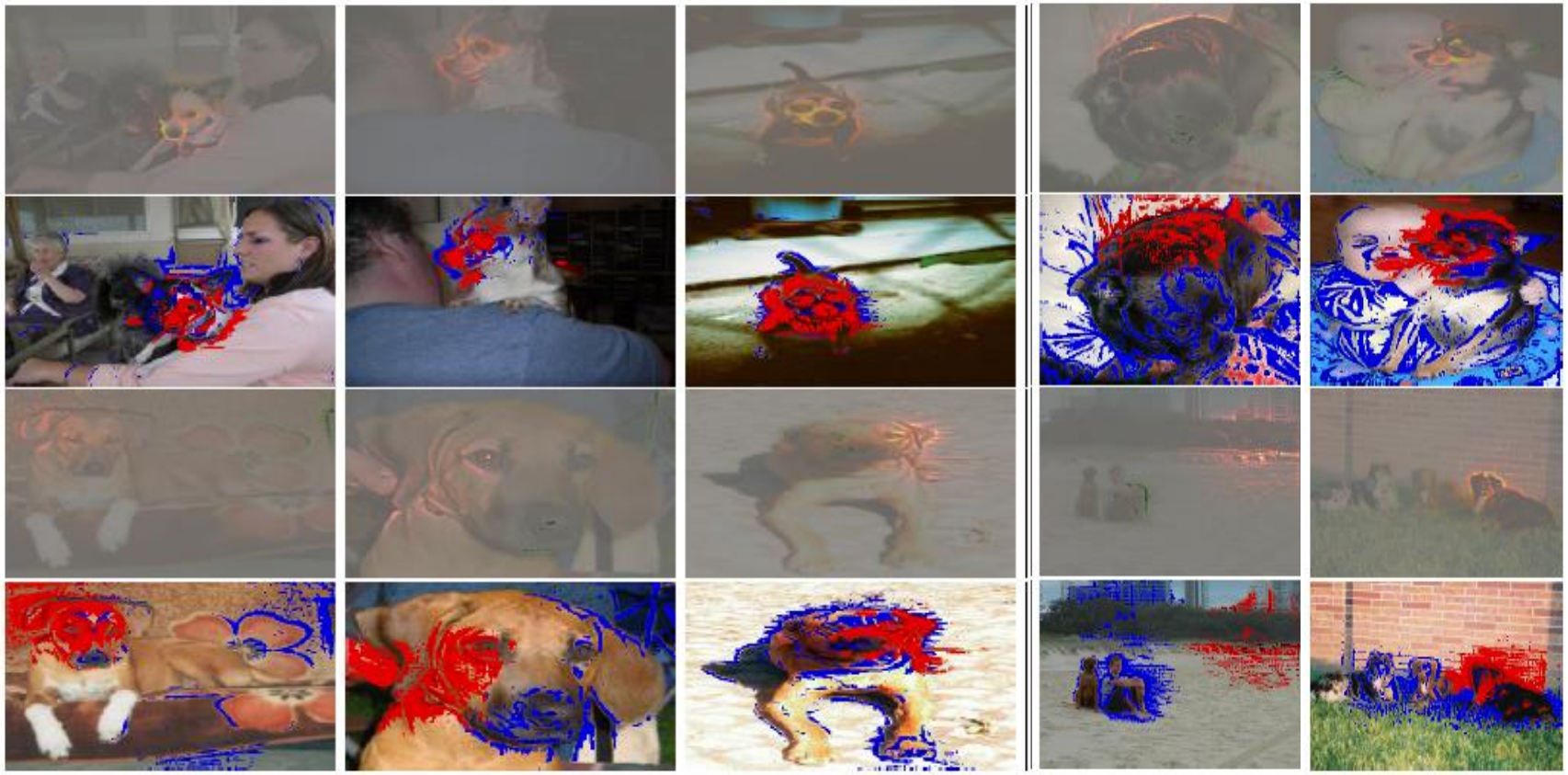


MNIST & SVHN RESULTS

Accuracy(%)	MNIST	SVHN
Full image	99.26	92.60
with only strong features	79.89	69.12
without strong features	43.45	54.46

Stanford Dog Dataset Results

Stanford Dog Dataset Results



Stanford Dog Dataset Results

Conclusion

- Sparsity in the individual response maps from the last layer kernels : same pattern for all datasets considered.
- Evidence for classes tend to come from very specific localized regions.
- **CLEAR maps** enable the visualization of not only the areas of interest that predominantly influence the decision-making process, but also the **degree of influence** as well as the **dominant class** of influence in these areas.
- Showed efficacy of CLEAR maps both quantitatively and qualitatively.

Thank You!

devinder.kumar@uwaterloo.ca

http://devinderkumar.com