A Formal Methods Approach Towards Deep Learning Interpretability

Kristen Kessel, Christopher Lazarus, Javier Sagastuy
Stanford University

Summary

Although deep neural networks have proved to be very successful at classification tasks, their intrinsic complexity makes reasoning about a classification outcome difficult. In recent work [2], statistical methods were introduced as a means to assess the influence of human-intelligible concepts in classification outcomes. We aim to assess and extend such methods by leveraging formal methods for neural network verification.

- **Question:** How important is a concept in classifying image i as label k? e.g. Is the presence of stripes relevant in the classification of an animal as a zebra?
- **Approach #1:** Use TCAV framework to provide statistical guarantees
- **Approach #2:** Use neural network verification methods [1] to provide formal guarantees

Testing with Concept Activation Vectors (TCAV) [2]

- **Idea:** Identify the region in the latent space corresponding to layer ℓ of the network in which a human-intelligible concept (e.g. blue) manifests more intensely with a vector called the Concept Activation Vector (CAV). Measure the relevance of this concept for classification of image i as class k by taking directional derivative of the layer ℓ activations for image i with the CAV.
- **Inputs:**
  - trained classification network
  - set of examples for a user-defined concept C and set of random counterexamples
  - labeled examples for the class k under consideration
- **Outputs:**
  - CAV v_c^ℓ for concept C at layer ℓ
  - TCAV score $S_{C,k}^\ell(\mathbf{x})$ of the sensitivity of the model's prediction of class k to concept C

$$S_{C,k}^\ell(\mathbf{x}) = \nabla h_k^\ell(f_\ell(\mathbf{x})) \cdot \mathbf{v}_C^\ell$$

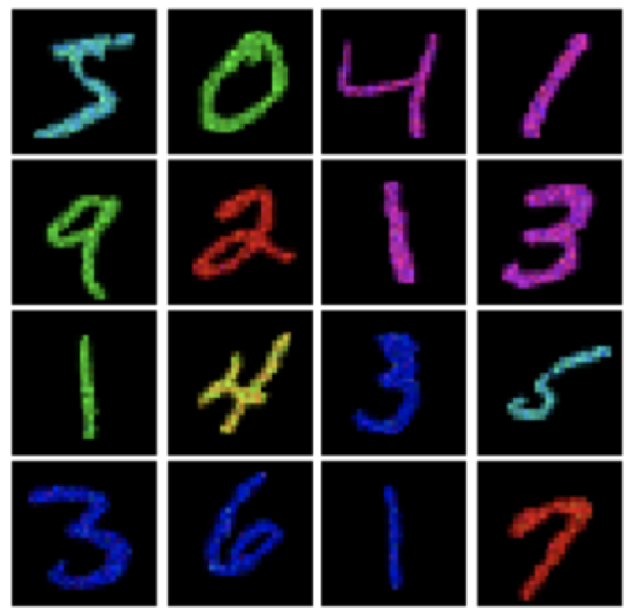
  - p-value testing the hypothesis that concept C is not relevant in classifying images of class k

Neural Network Verification

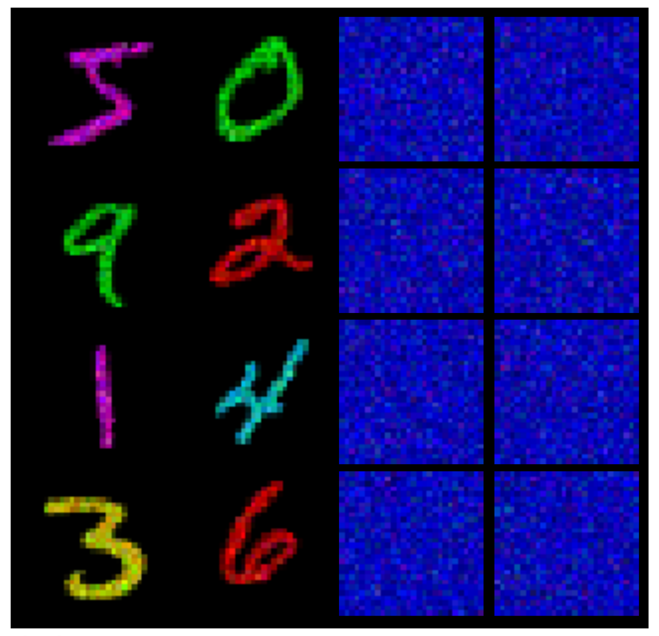
$$\vec{x} \in \mathcal{X} \Rightarrow \vec{y} = \vec{f}(\vec{x}) \in \mathcal{Y}$$

Approach: TCAV + Verification

Custom data sets

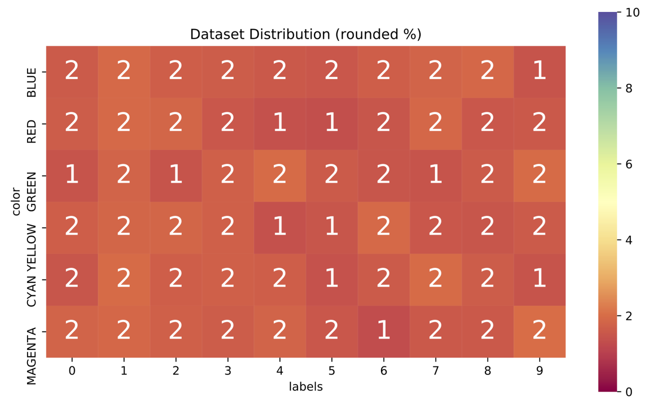


(a) Colorized MNIST training set for classification of hand-written digits

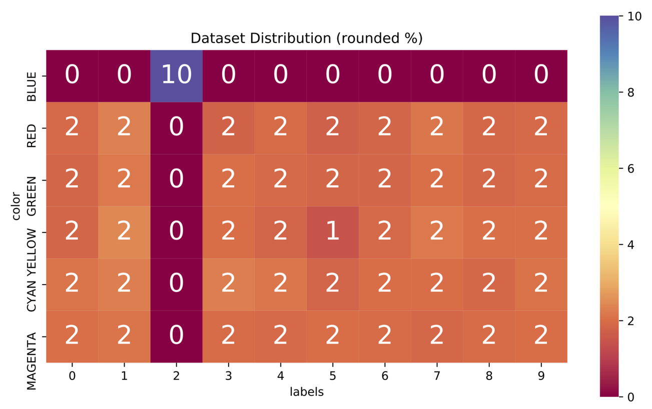


(b) Blue concept training set and non-blue training set to learn CAVs for concept blue

Maybe talk about classe sand support vector LALALALALAL



(c) label 1



(d) label 2

2 Figures side by side

$$\mathcal{L}_{LP} = -\mathbb{E}_{Z\sim q(Z|A,X)}[A_{ij}\log\tilde{A}_{ij} + (1-A_{ij})\log(1-\tilde{A}_{i,j})] + \mathrm{KL}(q(Z|A,X)||p(Z)).$$

2. Edge hallucination produces $\hat{A}$:

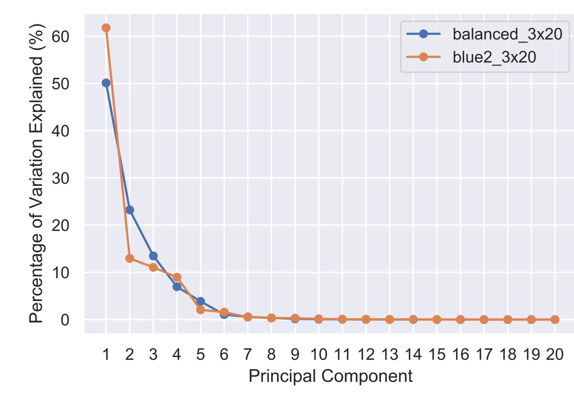
- top$K$ (K hyper-parameter)
- sampling using gumbel softmax [4] trick (allows gradients to flow)

3. Node classification

- $\hat{y} = \mathrm{GCN}(\hat{A}, X)$

Results

Need to talk about significant CAVs Then talk about the avenues and boulevards.



(a) label 1

model	layer	TCAV Score	signif cant
balanced_5x50	fc1	0.15 ± 0.10	yes
	fc4	0.14 ± 0.13	yes
blue2_5x50	fc1	1.00 ± 0.00	yes
	fc4	1.00 ± 0.00	yes
balanced_3x50	fc1	0.14 ± 0.13	yes
	fc2	0.08 ± 0.05	yes
blue2_3x50	fc1	0.80 ± 0.08	yes
	fc2	0.78 ± 0.11	yes
balanced_3x20	fc1	0.20 ± 0.06	yes
	fc2	0.14 ± 0.05	yes
blue2_3x20	fc1	1.00 ± 0.01	yes
	fc2	0.98 ± 0.01	yes

(b) label 2

2 Figures side by side

Node classification results

#	network	in/out sets	algorithm	result
1	blue2_5x50	$\mathcal{X}_1/\mathcal{Y}_{+\text{PC1,fc4}}$	NSVerify	violated
	bal2_5x50	$\mathcal{X}_1/\mathcal{Y}_{+\text{PC1,fc4}}$	NSVerify	violated
1.1	blue2_5x50	$\mathcal{X}_1/\mathcal{Y}_{+\text{mean,fc4}}$	NSVerify	violated
	bal2_5x50	$\mathcal{X}_1/\mathcal{Y}_{+\text{mean,fc4}}$	NSVerify	violated
2	blue2_3x20	$\mathcal{X}_1/\mathcal{Y}_{+\text{PC1,fc2}}$	Reluplex	violated
	bal2_3x20	$\mathcal{X}_1/\mathcal{Y}_{+\text{PC1,fc2}}$	Reluplex	violated
2.1	blue2_3x20	$\mathcal{X}_1/\mathcal{Y}_{+\text{mean,fc2}}$	Reluplex	violated
	bal2_3x20	$\mathcal{X}_1/\mathcal{Y}_{+\text{mean,fc2}}$	Reluplex	violated
3	blue2_5x50	$\mathcal{X}_{2,5}/\mathcal{Y}_{+\text{PC1,fc4}}$	NSVerify	unknown
	bal2_5x50	$\mathcal{X}_{2,5}/\mathcal{Y}_{+\text{PC1,fc4}}$	NSVerify	unknown
3.1	blue2_3x20	$\mathcal{X}_{2,5}/\mathcal{Y}_{+\text{PC1,fc2}}$	NSVerify	holds
	bal2_3x20	$\mathcal{X}_{2,5}/\mathcal{Y}_{+\text{PC1,fc2}}$	NSVerify	violated

Table: Results of formal verification experiments for various networks, input and output sets, and algorithms. If the result is violated, this indicates that $\vec{x} \in \mathcal{X} \not\Rightarrow \vec{y} = \vec{f}(\vec{x}) \in \mathcal{Y}$.

- Try to salvage somethibg
- Nothing worked :)

References

[1] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In International Conference on Computer Aided Verification, pages 97–117. Springer, 2017
[2] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In ICML, 2018
[3] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann. lecun. com/exdb/mnist, 2:18, 2010
[4] C. Liu, T. Arnon, C. Lazarus, C. Barrett, and M. J. Kochenderfer. Algorithms for verifying deep neural networks. CoRR, abs/1903.06758, 2019

{kkessel, clazarus, jvrsgsty}@stanford.edu