Concept Distillation: Leveraging Human-Centered Explanations for Model Improvement

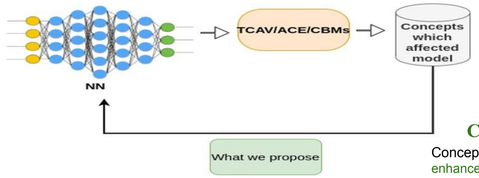
Avani Gupta[1,2] Saurabh Saini[2] P J Narayanan[2]

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

NEURAL INFORMATION PROCESSING SYSTEMS

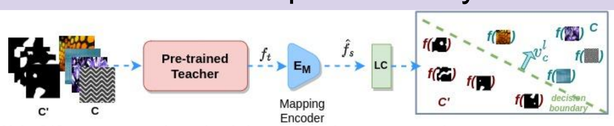
Outline

- Humans think in abstract concepts like color, texture, shapes, etc.
- Human-centered concepts are used by Interpretability methods.
- Can they be used to debias a trained model?
- We introduce a concept loss to ante-hoc finetune a model to make it more sensitive or less sensitive to a concept.
- We also extend concepts to other layers of a model using prototypes.
- We introduce concept distillation to get more informative concepts



Concept Distillation
Concept sensitive finetuning of DNNs
enhance, diminish or add concept

Concept Sensitivity



CAVs [1] are normal to decision boundary separating concepts.
Concept Sensitivity: Nudge a model's gradients towards CAV direction and check the effect on final layer's logit (or loss) prediction to get Sensitivity of trained model for layer l and class sample k.

Key Contributions

Extend CAVs: From post-hoc explanations to ante-hoc model improvement without altering base architecture.
Novel Concept loss: Concept sensitive training of DNNs.
Extend CAV sensitivity calculation to any layer and enhance it by making it more global by employing prototypes.
Concept distillation: Exploit the inherent knowledge of large pretrained models as a teacher in concept definition.
Benchmark results: On standard biased MNIST datasets and introduce a challenging TextureMNIST dataset. Application on severe biases like age.
Application beyond classification: Tackle multi-branch Intrinsic Image Decomposition problem (IID), introducing concepts as priors.

Proposed Concept Loss

Proposed novel concept loss for concept sensitive finetuning of DNNs.
We control the above component to control concept sensitivity.

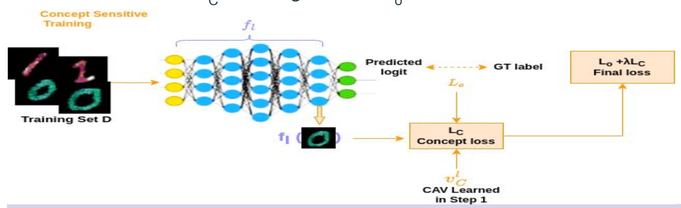
$$L_C(\boldsymbol{x}) = |\cos(\nabla L_p\left(f_l(\boldsymbol{x})\right), \boldsymbol{v}_C^l)|.$$

For C desensitization: make loss gradient perpendicular to cav (align along decision boundary) For sensitization: make it parallel to cav (align along cav direction) use 1 – L_c Final loss = L_c + L_o (conventional ground truth loss)

[1]M42, UAE [2]CVIT, KCIS, IIIT Hyderabad

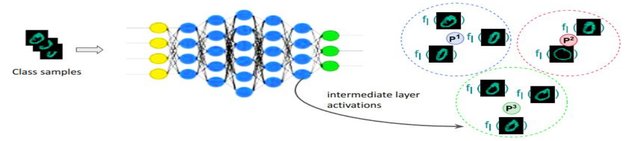
Concept Sensitive Training

1. Choose layer l, Concepts C, learn CAVs.
2. Concept loss L_C to enhance/diminish/add concepts to model.
3. Finetune M with L_C with original loss L_o

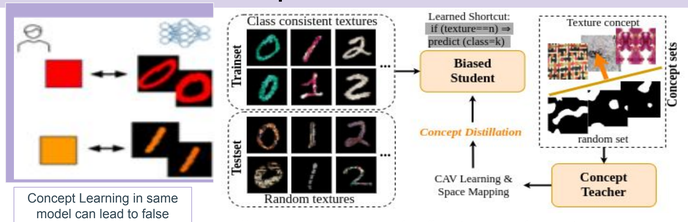


Proto-types for Intermediate Layer Sensitivity

- Kim et al. [1] only calculate sensitivity of final layer wrt any other layer.
- Concepts can exist in any layer.
- We need ANY layer sensitivity wrt intermediate layer outputs. How to calculate intermediate layer sensitivity?
- Class Proto-types as Pseudo-GT labels!
- K-means on class sample activations to get K cluster centers (proto-types). Calculate proto-type loss as avg L2 distance from K class centers. Use it instead of final layer loss for intermediate layer sensitivity.

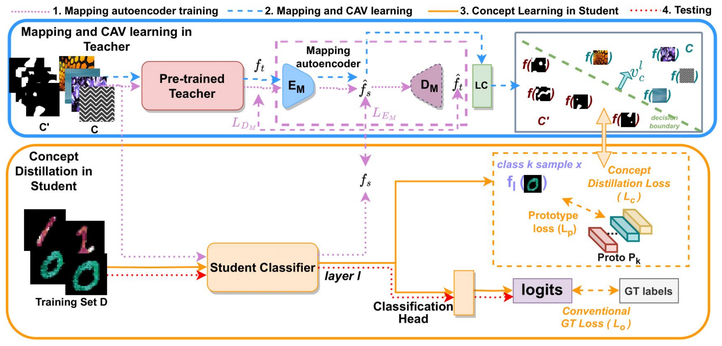


Concepts from Teacher



Concept Learning in same model can lead to false concept associations.

We propose using a teacher for efficient learning of concepts as shown.



Concept Sensitive Finetuning by our proposed concept loss.

Project Page: **https://avani17101.github.io/Concept-Distilllation/**

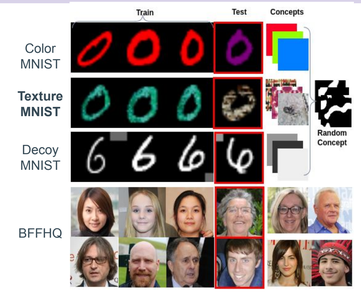
Applications in Debiasing

Teacher?	Prototype?	Accuracy
✗	✗	9.96
✗	✓	26.97
✓	✗	30.94
✓	✓	**50.93**

Different Components of Our Method.

Dataset	Concept	Base Model	Ours
ColorMNIST	Color	0.52	**0.21**
DecoyMNIST	Spatial patches	0.57	**0.45**
TextureMNIST	Textures	0.68	**0.43**
BFFHQ	Age	0.78	**0.13**

Reduced TCAV scores of bias concept.



Dataset	Bias	Base	CDEP[55]	RRR[56]	EG[17]	Ours w/o Teacher	Ours	Ours+L
ColorMNIST	Digit color	0.1	31.0	0.1	10.0	26.97	41.83	**50.93**$\pm$1.42
DecoyMNIST	Spatial patches	52.84	97.2	**99.0**	97.8	87.49	98.58	**98.98**$\pm$0.20
TextureMNIST	Digit textures	11.23	10.18	11.35	10.43	38.72	48.82	**56.57**$\pm$0.79

Improvements over Zero-shot Interpretability based baselines.

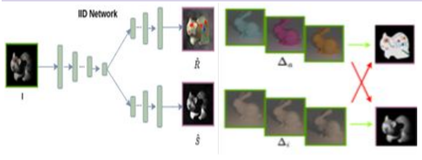
Dataset	Bias	Base	EnD [69]	DFA [43]	Ours w/o Teacher	Ours
BFFHQ	Age	56.87$\pm$2.69	56.87$\pm$1.42	61.27$\pm$3.26	59.4	**63**$\pm$0.79

Few Shot baselines comparison: Our method not limited to concept sets but can work with bias-conflicting samples too.

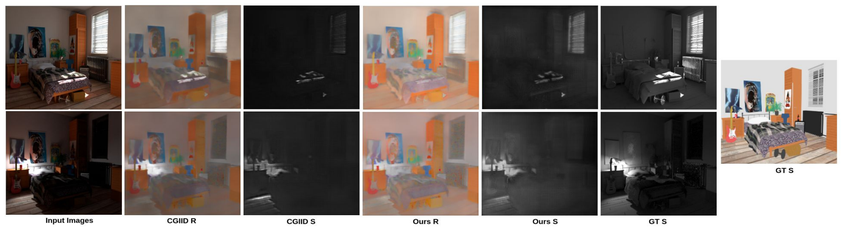
	ColorMNIST Trained			TextureMNIST Trained		
Test Dataset	Base	CDEP[55]	Ours+L	Base	CDEP[55]	Ours+L
Invert color	0.00	23.38	**50.93**	11.35	10.18	**45.36**
Random color	16.63	37.40	**46.62**	11.35	10.18	**64.96**
Random texture	15.76	28.66	**32.30**	11.35	10.18	**56.57**
Pixel-hard	15.87	33.11	**38.88**	11.35	10.18	**61.29**

Generalizability across different test-sets.

Prior Knowledge Induction



Model	MSE R ↓	MSE S ↓	SSIM R ↑	SSIM S ↑
CGIID [47]	0.066	**0.027**	0.536	0.581
CGIID++	0.080	0.032	0.520	0.552
Ours (R only)	**0.052**	**0.027**	**0.54**	0.581
Ours (R & S)	0.059	0.028	0.538	**0.586**



References

[1] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). ICML, 2018
[17] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nature machine intelligence, 3 (7):620–631, 2021
[43] J. Lee, E. Kim, J. Lee, J. Lee, and J. Choo. Learning debiased representation via disentangled feature augmentation. Neurips, 21.
[47] Z. Li and N. Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. ECCV, 2018.
[55] L. Rieger, C. Singh, W. Murdoch, and B. Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. PMLR, 2020.
[56] d F. Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations.