Learning Bottleneck Concepts in Image Classification

Bowen Wang, Liangzhi Li, Yuta Nakashima, Hajime Nagahara

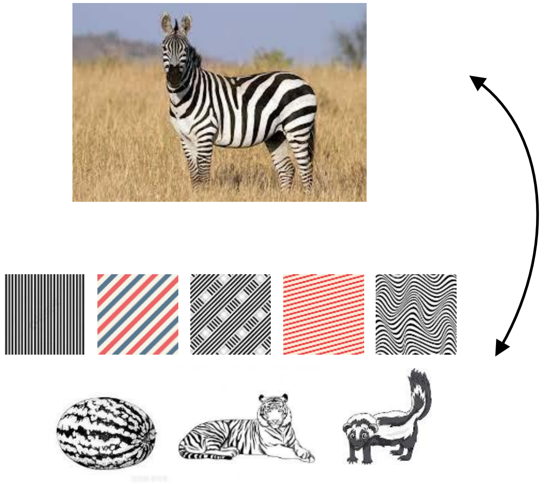
Osaka University, Institute for Datability Science

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

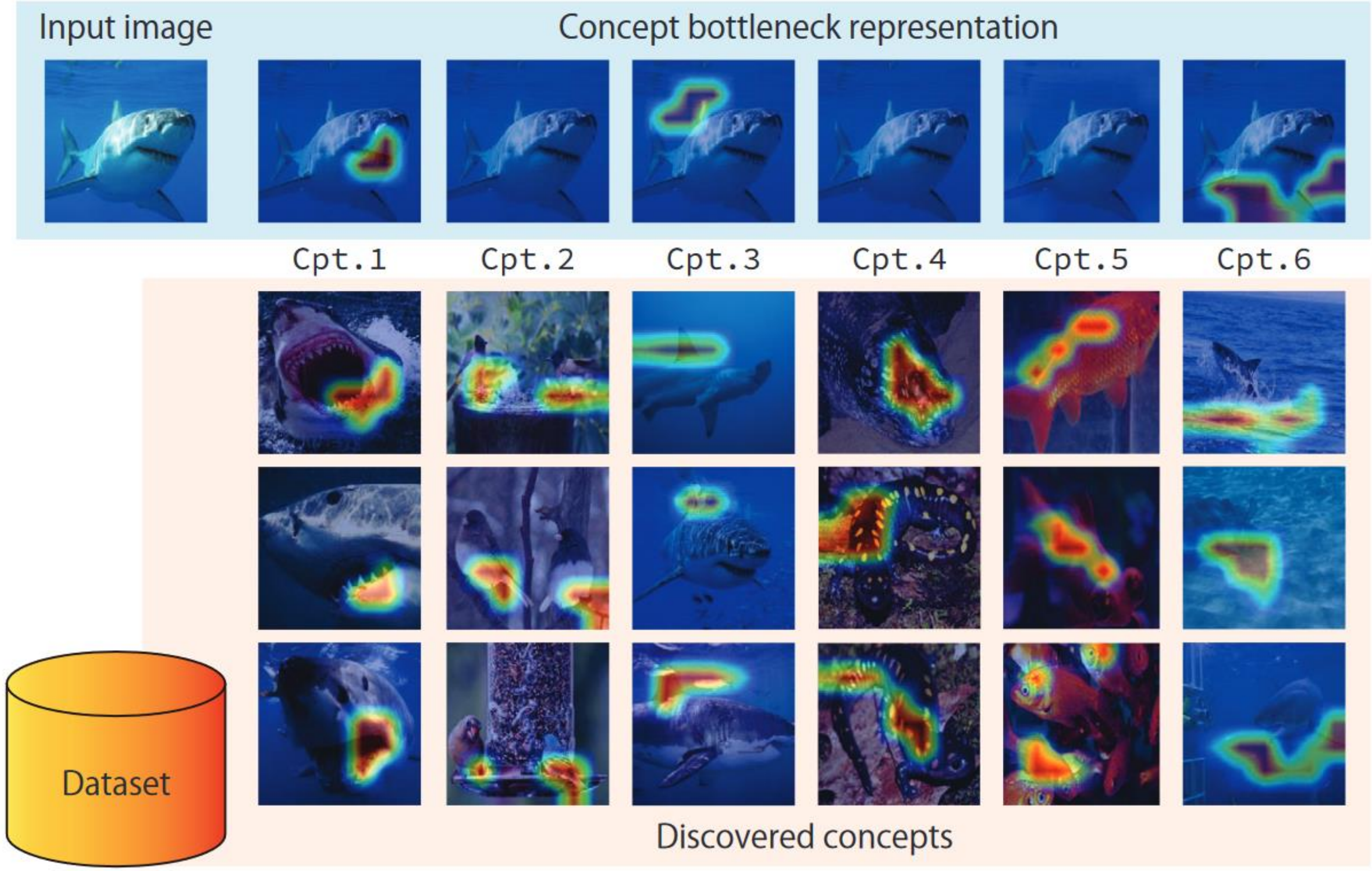
Overview

Previous XAI methods
(per-pixel relevance)

Human Perception
(concept-based explanation)

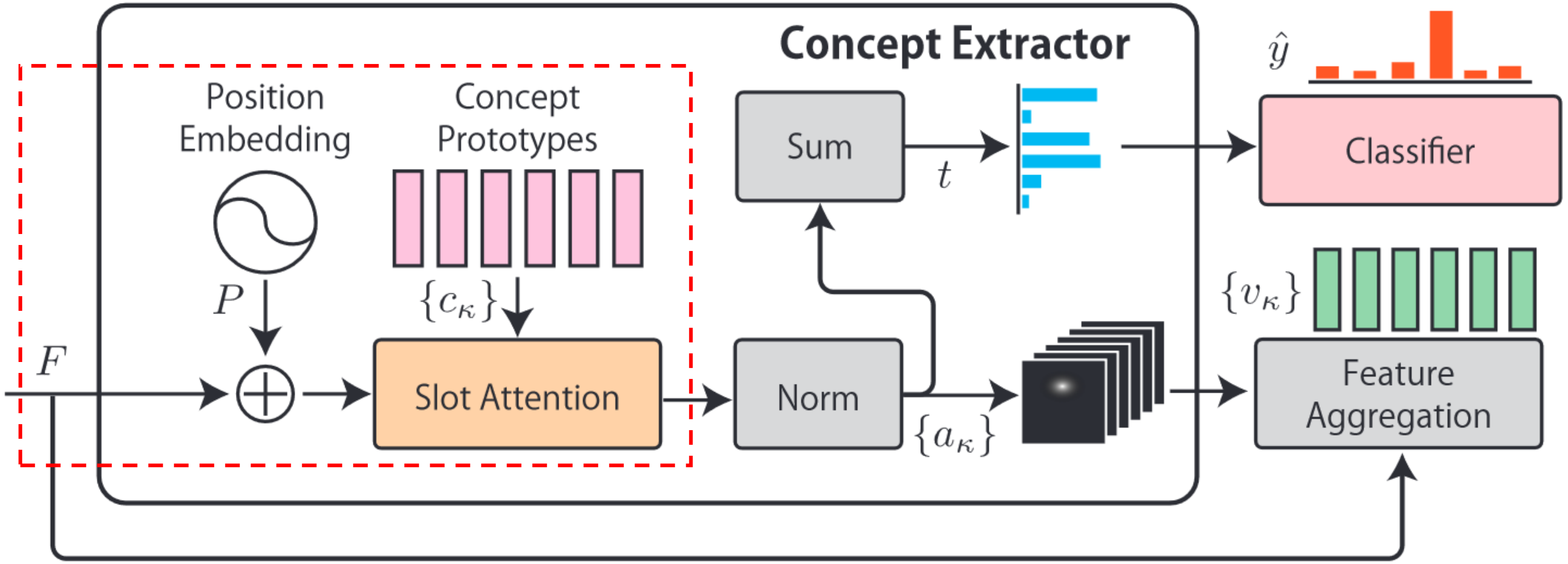


Overview



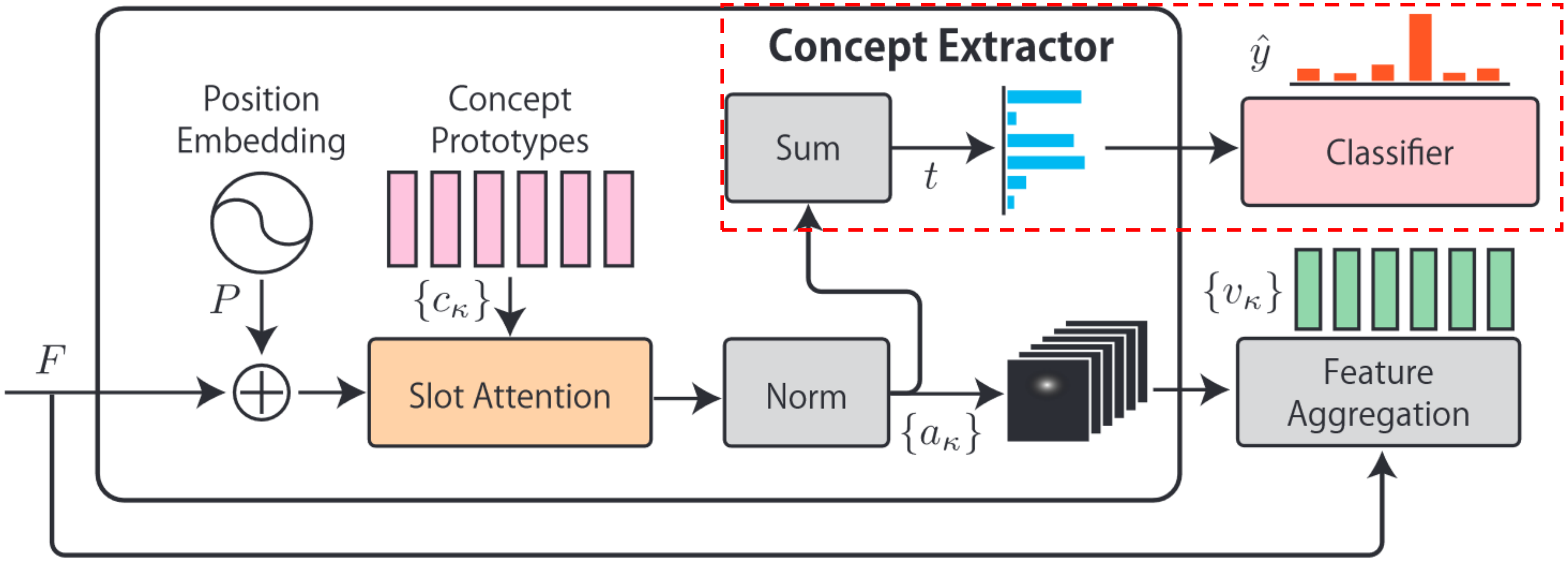
We want to explore the possibility of a deep model learning concepts spontaneously. And thus, designed bottleneck concept learner (BotCL).

Pipeline



Attention to each concept $a_\kappa = \phi(Q(c_\kappa)^\top K(F'))$

Pipeline

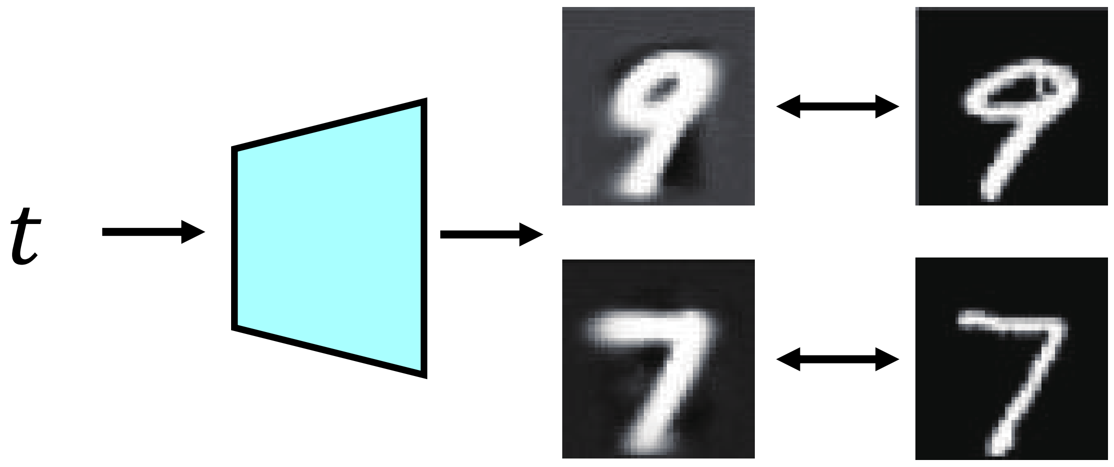


Quantization loss for bottleneck $l_{\text{qua}} = \frac{1}{k|\mathcal{B}|} \sum_{x \in \mathcal{B}} \left\| \text{abs}(\hat{t}) - \mathbf{1}_\kappa \right\|^2$

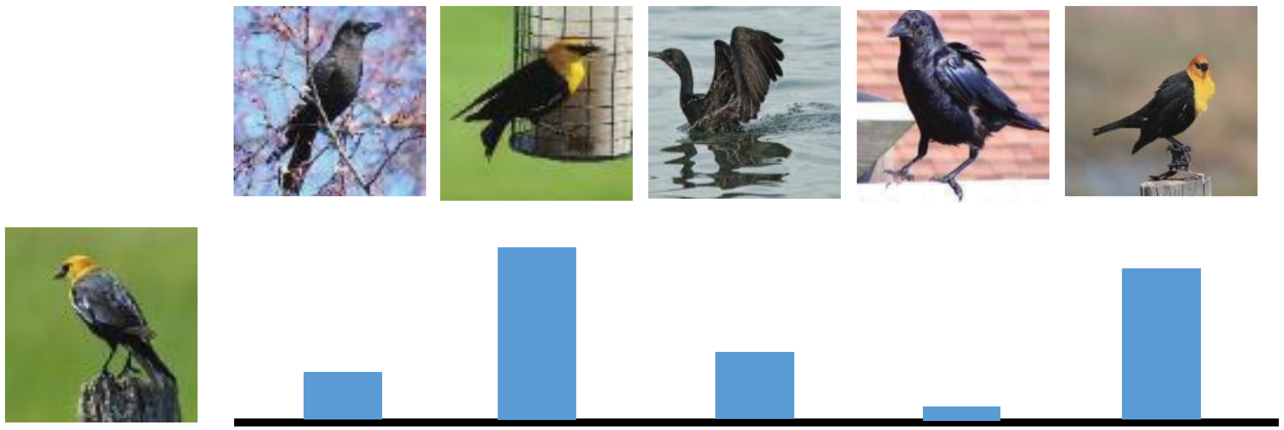
One layer FC for classification $\hat{y} = Wt$

Self-supervision

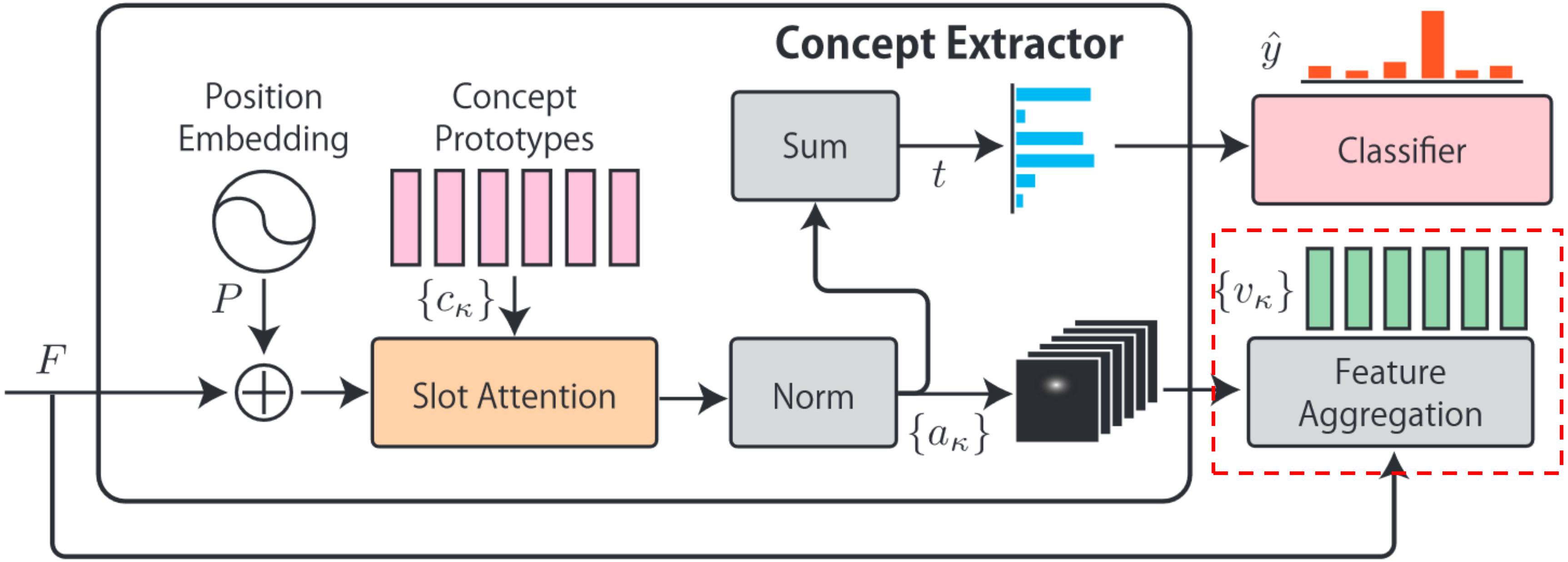
Reconstruction Loss



Contrastive Loss

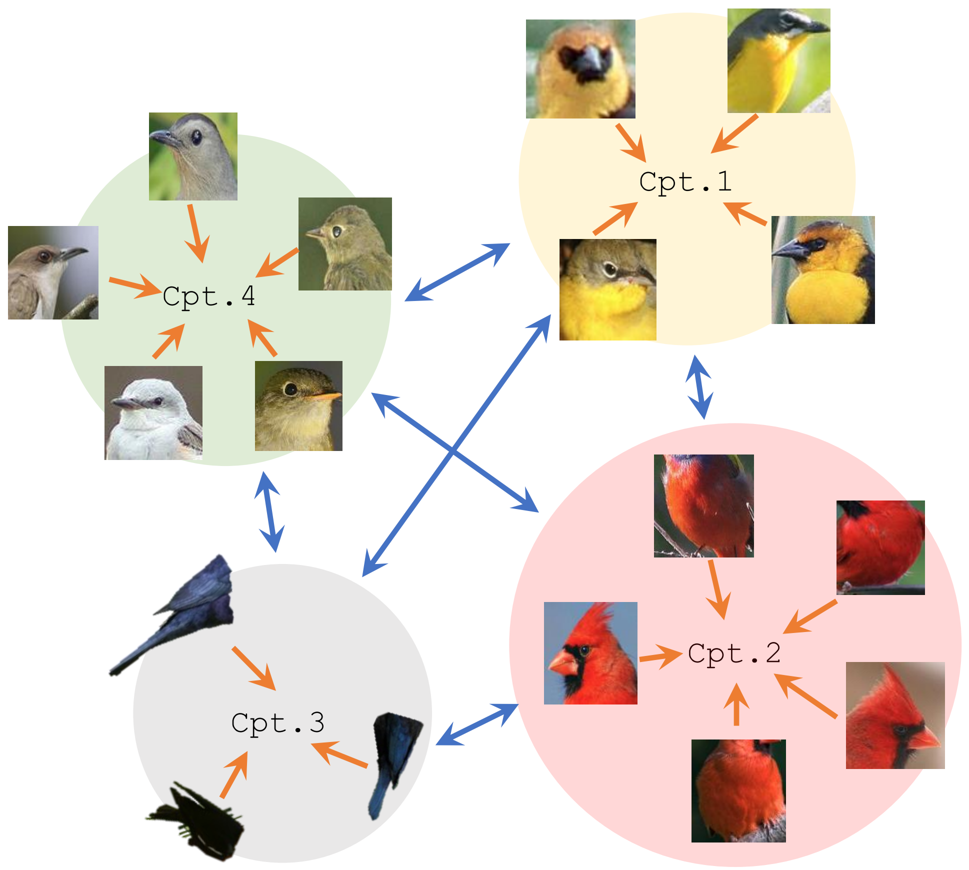


Pipeline



Feature aggregation for regularizers $v_\kappa = F a_\kappa$

Regularizers



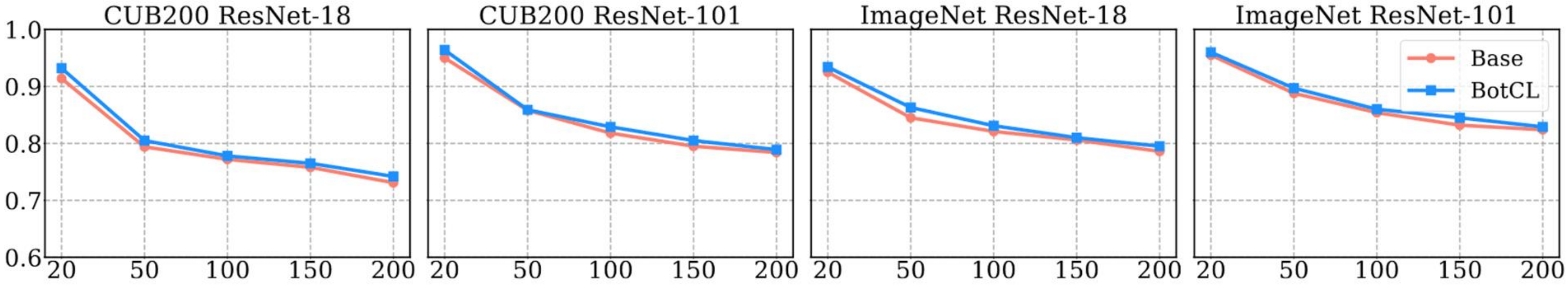
Individual Consistency

Designed to force a concept to learn similar features.

Mutual Distinctiveness

Let concepts cover different visual elements.

Accuracy vs. The Number of Classes

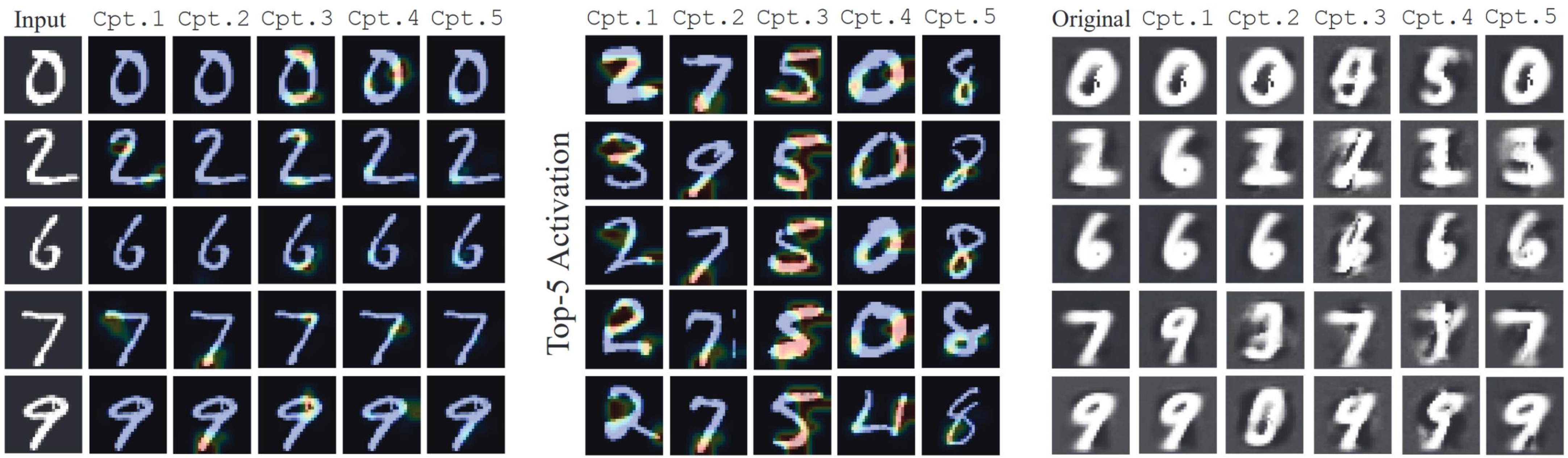


Classification Performance

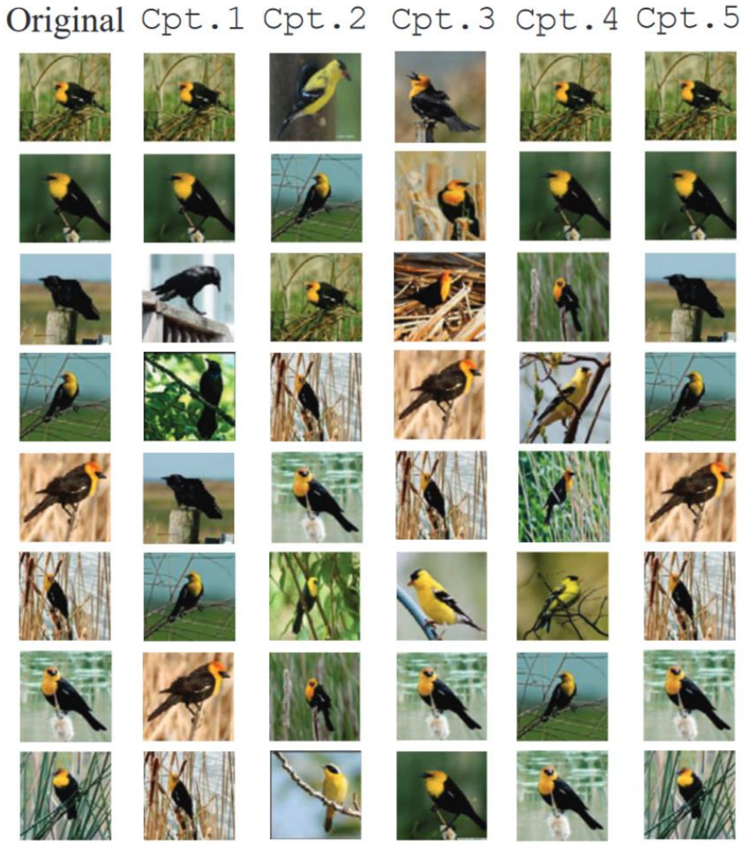
	CUB200	ImageNet	MNIST	Synthetic
Baseline	0.731	0.786	0.988	0.999
k-means* [46]	0.063	0.427	0.781	0.747
PCA* [46]	0.044	0.139	0.653	0.645
SENN [1]	0.642	0.673	**0.985**	0.984
ProtoPNet [8]	0.725	0.752	0.981	0.992
$\text{BotCL}_{\text{Rec}}$	0.693	0.720	0.983	0.785
$\text{BotCL}_{\text{Cont}}$	**0.740**	**0.795**	0.980	**0.998**

[1] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. *NeurIPS, 2018.*
[8] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. *NeurIPS, 2019.*
[46] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *NeurIPS, 2020.*

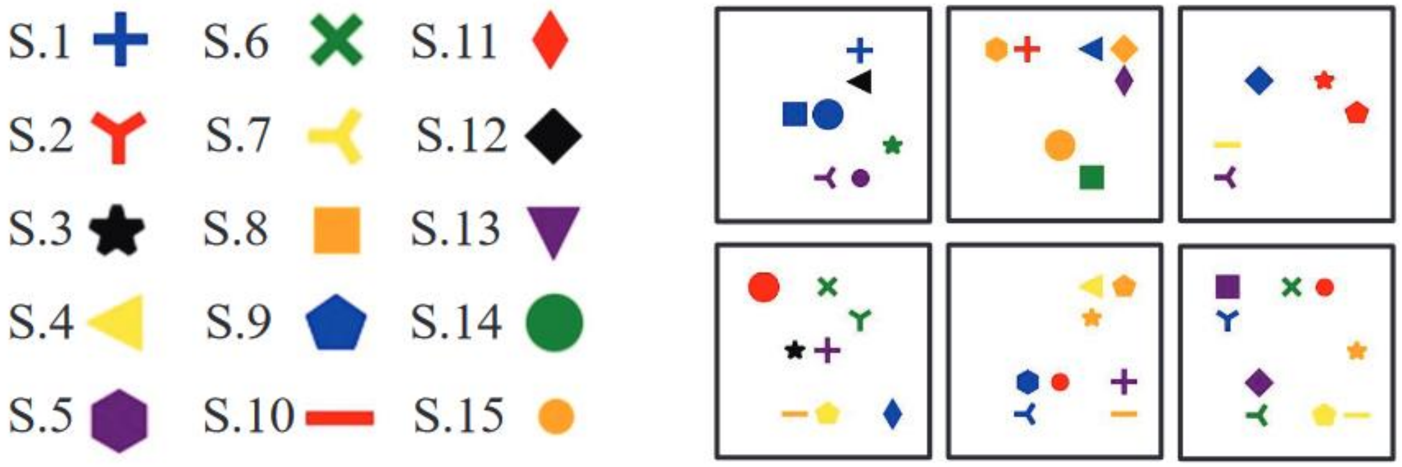
Concepts Learned in MNIST



Concepts Learned in CUB200

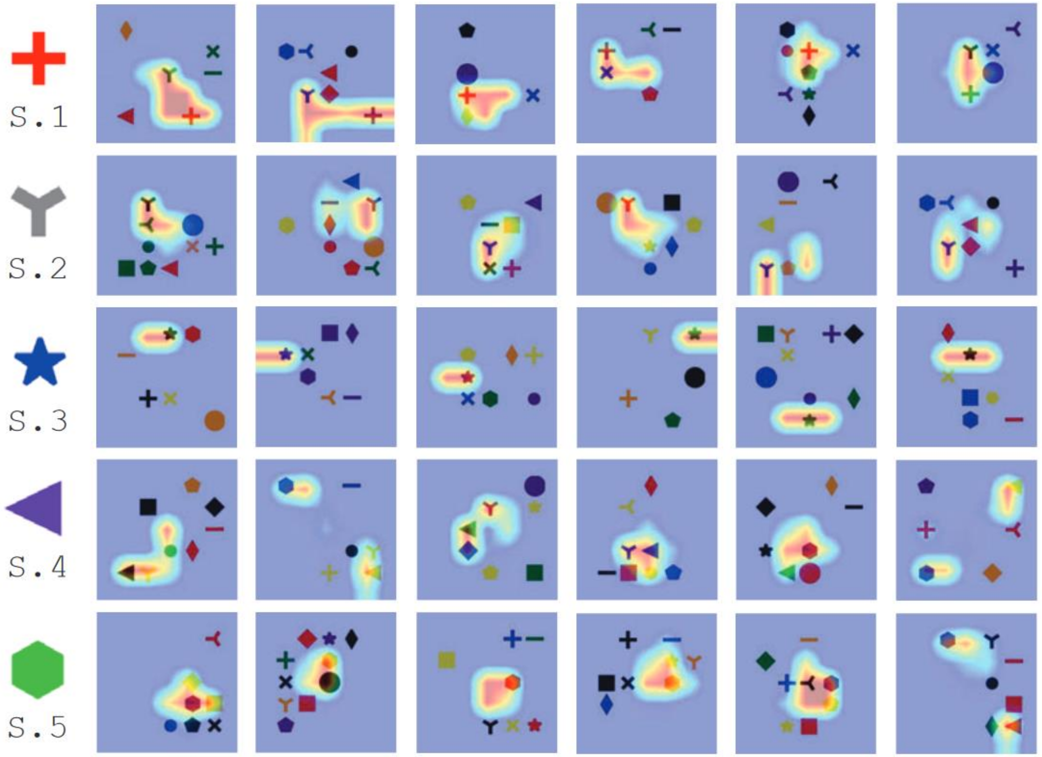


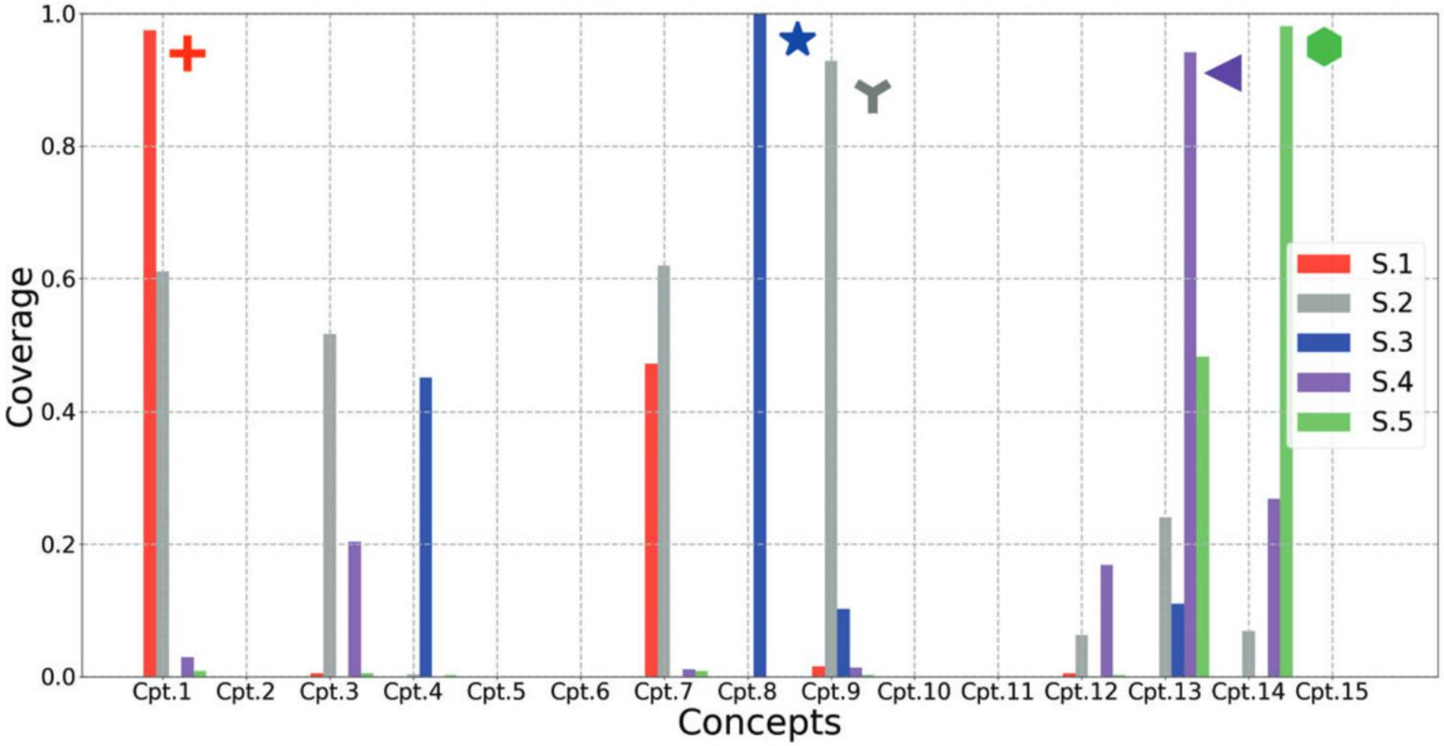
Quantitative Evaluation of Synthetic Dataset



The task is a multi-label classification that involves 15 shapes. Combinations of the 5 shapes (shown in Figure a, S.1 to S.5) form 15 classes, and the other 10 shapes are noises

Quantitative Evaluation of Synthetic Dataset





$$\mathrm{Coverage}_{s\kappa} = \mathbb{E}[h_{s\kappa}]$$

Quantitative Results

		Comp.	Purity	Dist.	Acc.
$k = 5$	ACE	0.662	0.274	0.084	—
	k-means	0.630	0.724	0.215	0.652
	PCA	0.458	0.170	0.298	0.571
	BotCL	0.618	0.453	0.281	0.835
$k = 15$	ACE	0.614	0.221	0.151	—
	k-means	0.816	**0.978**	0.272	0.747
	PCA	0.432	0.162	0.286	0.645
	BotCL	**0.925**	0.744	**0.452**	**0.998**

We apply k-means or PCA to feature map F of all images in the dataset after flattening the spatial dimensions. (Supplementary for more details)

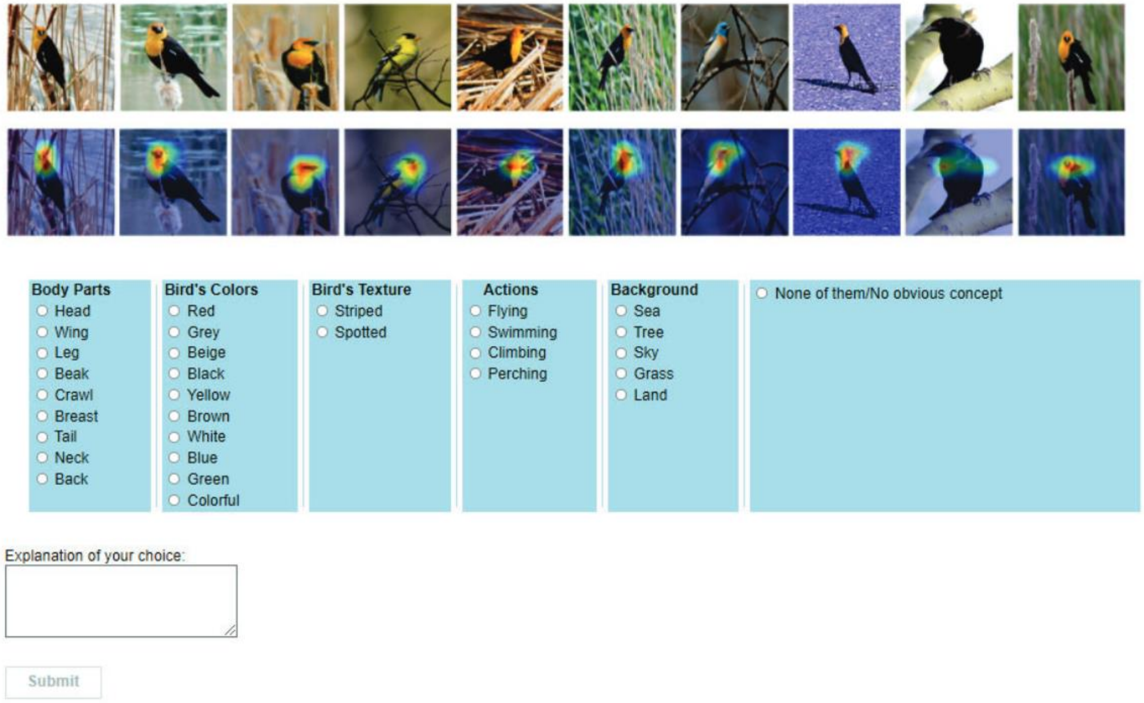
- Completeness: measures how well a concept covers its associated shape in the dataset.
- Purity: shows the ability to discover concepts that only cover a single shape.
- Distinctiveness: quantifies the difference among concepts based on the coverage.

User Study

Defined Concepts

Dataset	Group	Vocabulary
MNIST	Position (3)	upper, middle, lower
	Shape (8)	the end of a slanted vertical line, the end of a vertical line, a (part of) curve, a (part of) right-open curve, a circle, a white-black-white pattern, a horizontal line, the edge around a curve/line
CUB200	Body Part (9)	head, wing, leg, beak, crawl, breast, tail, neck, back
	Color (10)	red, grey, beige, black, yellow, brown, white, blue, green, colorful
	Texture (2)	striped, spotted
	Action (4)	flying, swimming, climbing, perching
	Background (5)	sea, tree, sky, grass, land

One Sample for User



User Study

Dataset	Concepts	CDR ↑ Mean	CDR ↑ Std	CC ↑ Mean	CC ↑ Std	MIC ↓ Mean	MIC ↓ Std
MNIST	Annotated	1.000	0.000	0.838	0.150	0.071	0.047
	BotCL	0.825	0.288	0.581	0.274	0.199	0.072
	Random	0.122	0.070	0.163	0.074	0.438	0.039
CUB200	Annotated	0.949	0.115	0.595	0.113	0.512	0.034
	BotCL	0.874	0.156	0.530	0.116	0.549	0.036
	Random	0.212	0.081	0.198	0.039	0.574	0.031

- Concept discovery rate (CDR): The ratio of the responses that are not "None of them" to all responses.
- Concept consistency (CC): The ratio of exact matches out of all pairs of participants' responses.
- Mutual information between concepts (MIC): The similarity of the response distribution, computed over all possible pairs of concepts.

Paper and Code

Paper: https://arxiv.org/abs/2304.10131

Code: https://github.com/wbw520/BotCL