Concept discovery from an image dataset

Toward image representation with an emergent language

Yuta Nakashima

Institute for Datability Science

Osaka University, Japan

June 28, 2023

Question

🤔 Does a model really see an image/video?

In visual question answering?

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

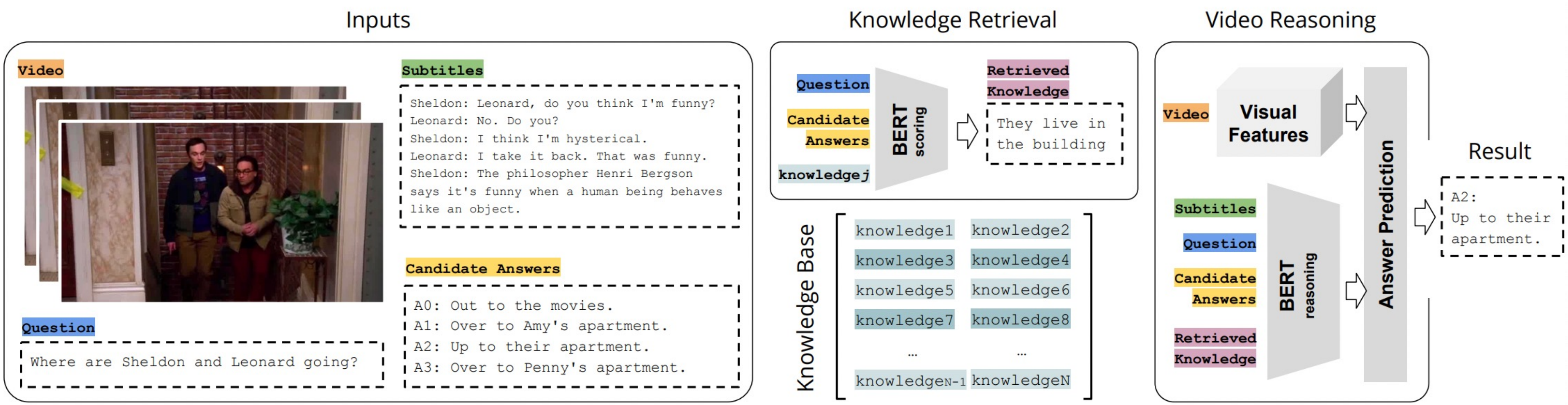
From: [Antol et al., "VQA: Visual question answering," ICCV 2015]

In visual question answering?

Group	Model	Dataset	Overall	Yes/No	Number	Other
Text-only	per Q-type prior [5]	VQA v1	35.13	71.31	31.93	08.86
		VQA-CP v1	08.39	14.70	08.34	02.14
	d-LSTM Q [5]	VQA v1	48.23	79.05	33.70	28.81
		VQA-CP v1	20.16	35.72	11.07	08.34
Text and image	d-LSTM Q + norm I [21]	VQA v1	54.40	79.82	33.87	40.54
		VQA-CP v1	23.51	34.53	11.40	17.42
	NMN [3]	VQA v1	54.83	80.39	33.45	41.07
		VQA-CP v1	29.64	38.85	11.23	27.88
	SAN [36]	VQA v1	55.86	78.54	33.46	44.51
		VQA-CP v1	26.88	35.34	11.34	24.70
	MCB [8]	VQA v1	60.97	81.62	34.56	[illegible]
		VQA-CP v1	34.39	37.96	11.80	[illegible]

From: [Agrawal et al., "Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering," CVPR 2018]

In video question answering?



Penny: What are you doing at work these days?
Sheldon: Oh. I'm working on time-dependent backgrounds in string theory. Specifically, quantum field theory in D-dimensional de Sitter space. in D-dimensional de Sitter space. (...)

What night is it?

- Wednesday
- Monday
- Friday
- **Saturday** ✓

R. Knowledge

Saturday is Sheldon's laundry night.

Leonard: You gotta admit, I'm delightful.
Penny: Why are you making this so difficult?
Leonard: It's not difficult for me, I'm having fun.
Penny: What do you want me to do? (...)

Where is this taking place?

- A local Dance Center
- **A bowling alley** ✓
- Sheldon's bedroom
- A blues dance club

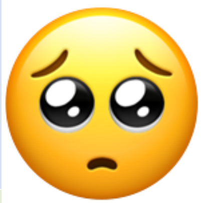
R. Knowledge

It's the closest thing to sports they like to partake in.

[Garcial et al., "KnowIT VQA: Answering Knowledge-Based Questions about Videos," AAAI 2020]

In video question answering?

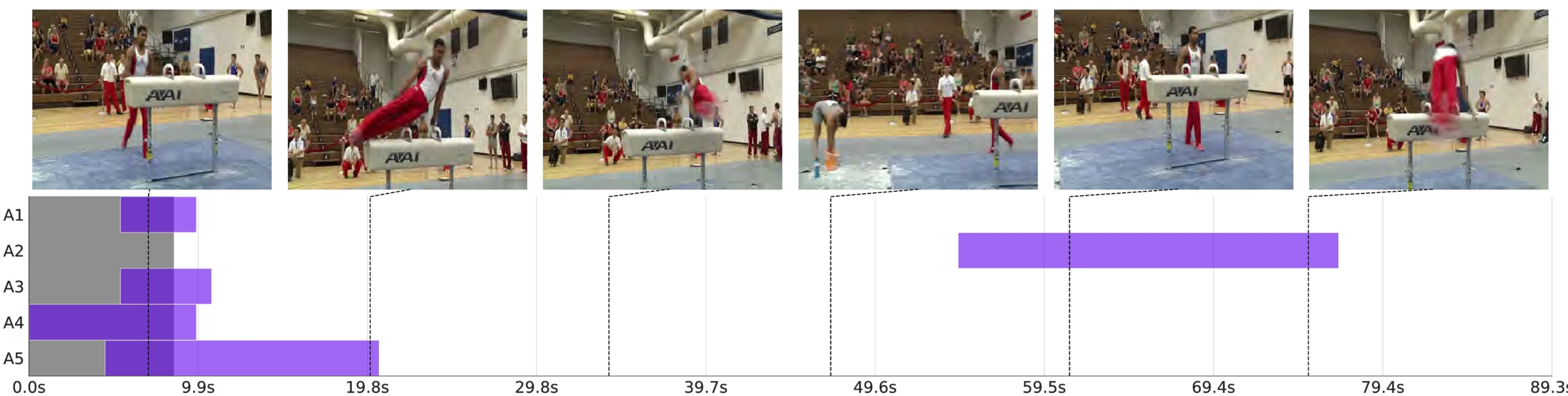
		Model	Vis.	Text.	Temp.	Know.	All
		Random	0.250	0.250	0.250	0.250	0.250
Text-only	QA	word2vec sim	0.108	0.163	0.151	0.180	0.161
		BERT sim	0.174	0.264	0.209	0.190	0.196
		TFIDF	0.434	0.377	0.488	0.485	0.461
		LSTM Emb.	0.444	0.428	0.512	0.515	0.489
		LSTM BERT	0.446	0.464	0.500	0.532	0.504
		◇ ROCK_{QA}	0.542	0.475	0.547	0.535	0.530
		Humans (Rookies, Blind)	0.406	0.407	0.418	0.461	0.440
	Subs, QA	LSTM Emb.	0.432	0.362	0.512	0.496	0.467
		LSTM BERT	0.452	0.446	0.547	0.530	0.504
		TVQA_{SQA}	0.602	0.551	0.512	0.468	0.509
		◇ ROCK_{SQA}	0.651	0.754	0.593	0.534	0.587
		Humans (Rookies, Subs)	0.618	0.837	0.453	0.498	0.562
Text and image	Vis, Subs, QA	TVQA	0.612	0.645	0.547	0.466	0.522
		◇ $\text{ROCK}_{\text{VSQA}}$ Image	0.643	0.739	0.581	0.539	0.587
		◇ $\text{ROCK}_{\text{VSQA}}$ Concepts	0.647	0.743	0.581	0.538	0.587
		◇ $\text{ROCK}_{\text{VSQA}}$ Facial	0.649	0.743	0.581	0.537	0.587
		◇ $\text{ROCK}_{\text{VSQA}}$ Caption	0.666	0.772	0.581	0.514	0.580
		Humans (Rookies, Video)	0.936	0.932	0.624	0.655	0.748



[Garcial et al., "KnowIT VQA: Answering Knowledge-Based Questions about Videos," AAAI 2020]

In video moment retrieval?

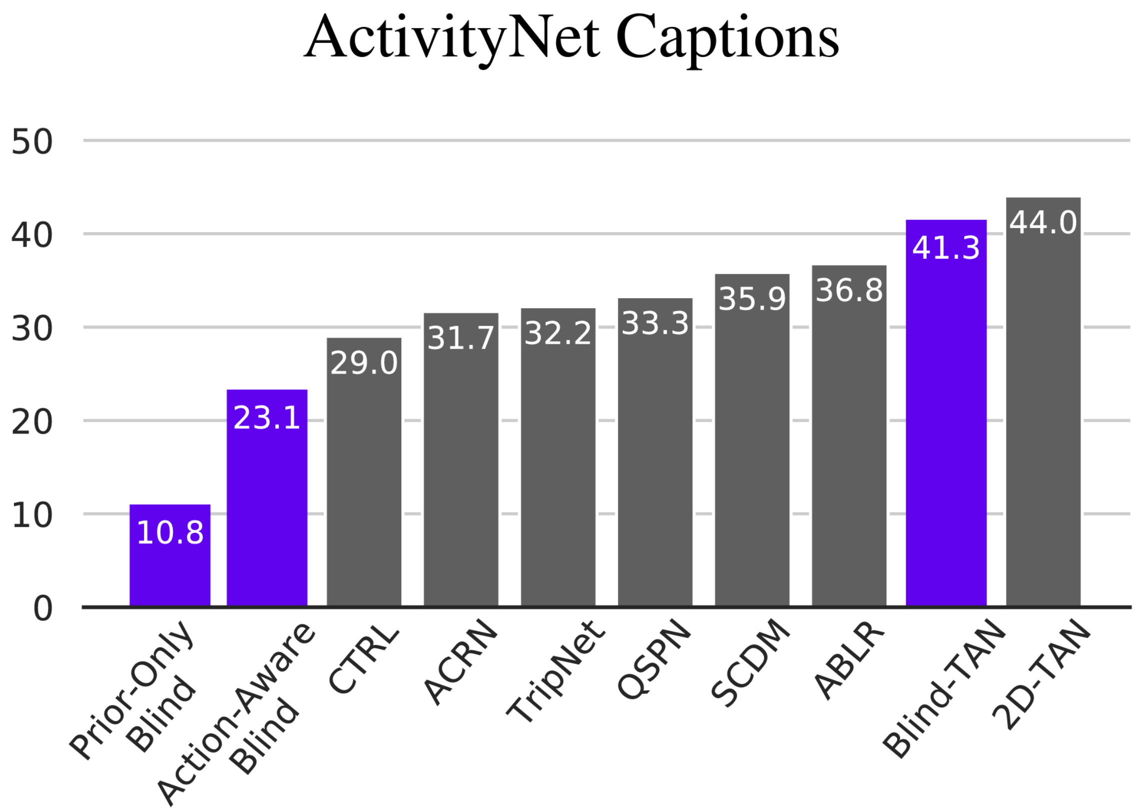
As the walk continues,the cat stops and begins staring at a parked car with large red flames painted on the side.



A male gymnast walks up to a beam.

[Otani et al., “Uncovering hidden challenges in query-based video moment retrieval,” BMVC 2020]

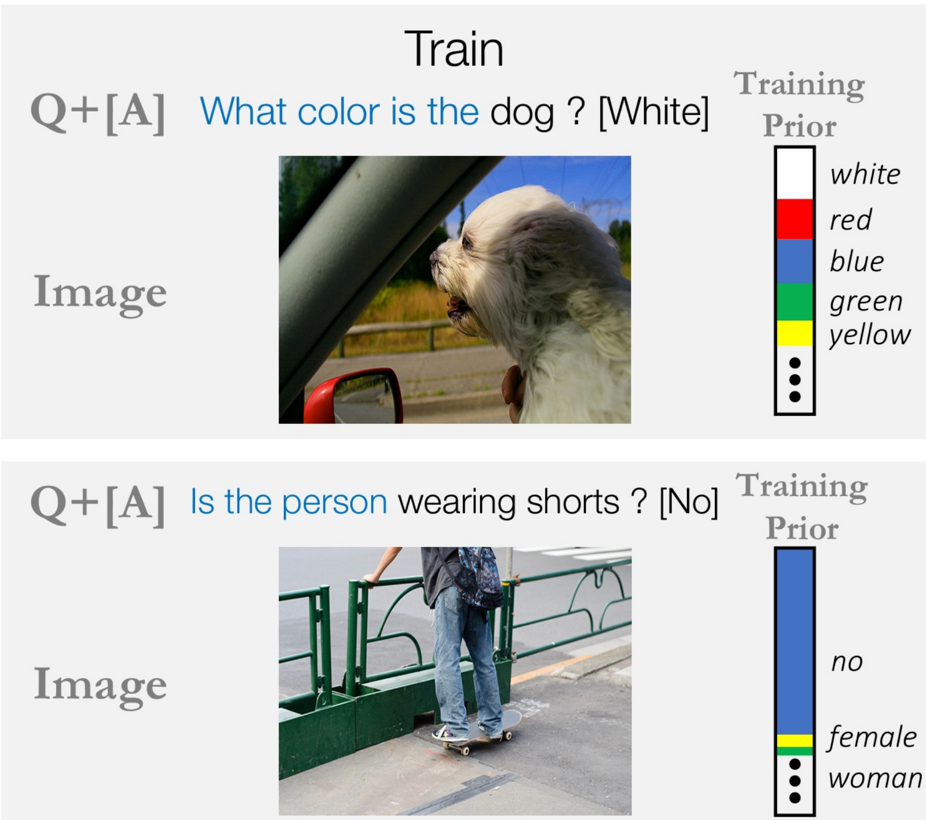
In video moment retrieval?



[Otani et al., "Uncovering hidden challenges in query-based video moment retrieval," BMVC 2020]

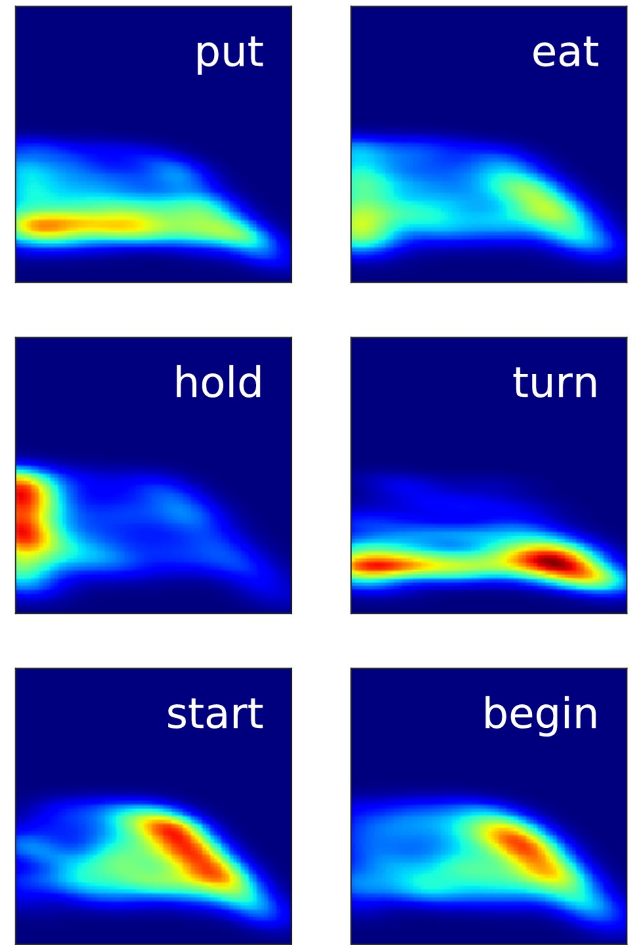
Spurious correlation matters

Visual question answering



From: [Agrawal et al., “Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering,” CVPR 2018]

Video moment retrieval



[Otani et al., “Uncovering hidden challenges in query-based video moment retrieval,” BMVC 2020]

A similar happens in visual-input-only tasks

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

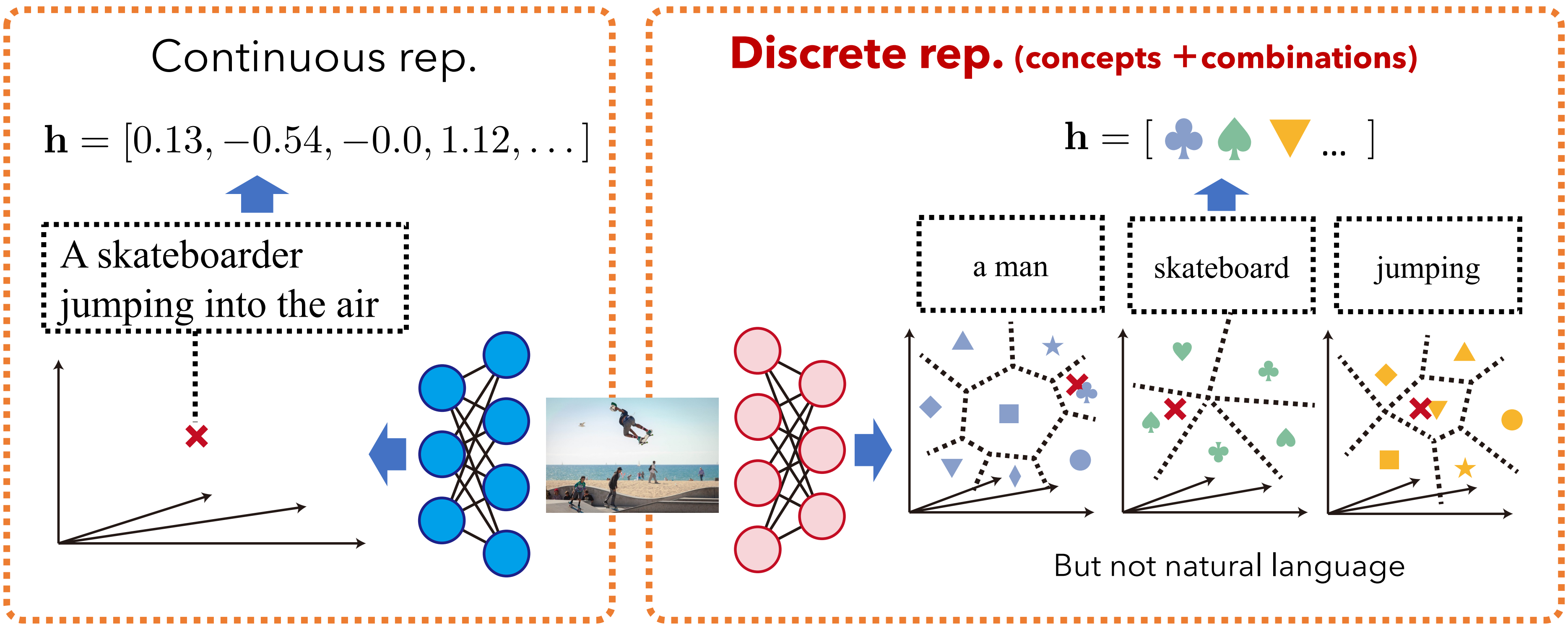
From: [Beery et al., "Recognition in Terra Incognita" ECCV 2018]

Mitigating the spurious correlation problem

- Considering certain attributes; for example:
  - [Burns et al., “Women also snowboard: Overcoming bias in captioning models,” ECCV 2018]
  - [Agarwal et al., “Does data repair lead to fair models? Curating contextually fair data to reduce model bias,” WACV 2022]
- Background matters; for example:
  - [Sagawa et al., “Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization,” ICLR 2020]
  - [Taghanaki et al., “Robust representation learning via perceptual similarity metrics,” NeurIPS 2022]
- Identifying confounders; for example:
  - [Liu et al., “Show, deconfound and tell: Image captioning with causal inference,” CVPR 2022]

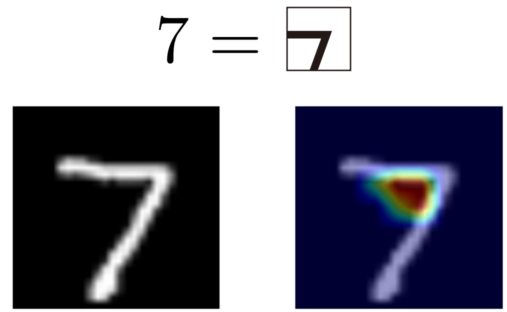
Yet another (perhaps crazy) idea?

- Representing an image/video with a set or a sequence of discrete labels, but not with a continuous vector

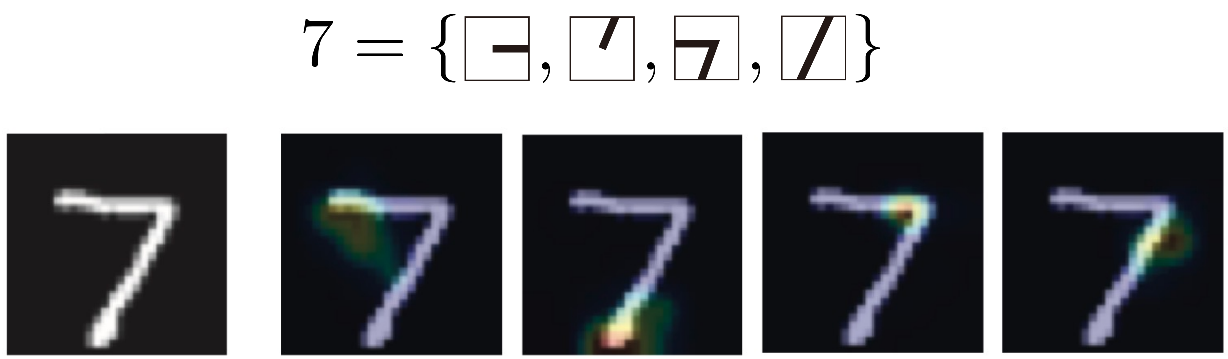


Discovering concepts in an image dataset

- SCOUTER [Li et al., ICCV 2021]
  - Better explainability with a *SINGLE* distinctive visual feature per class

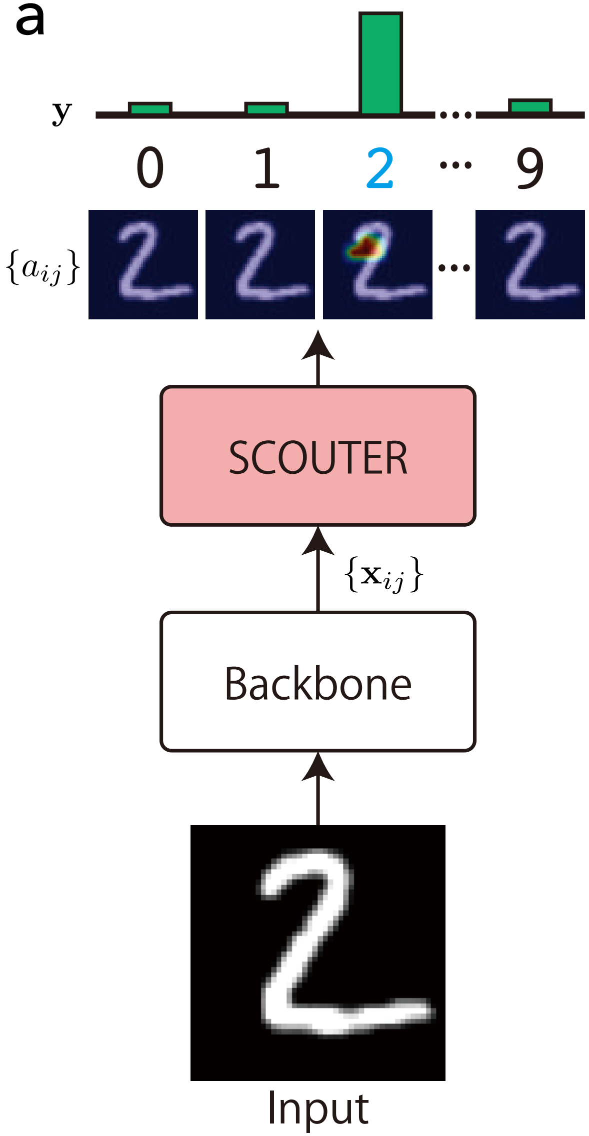


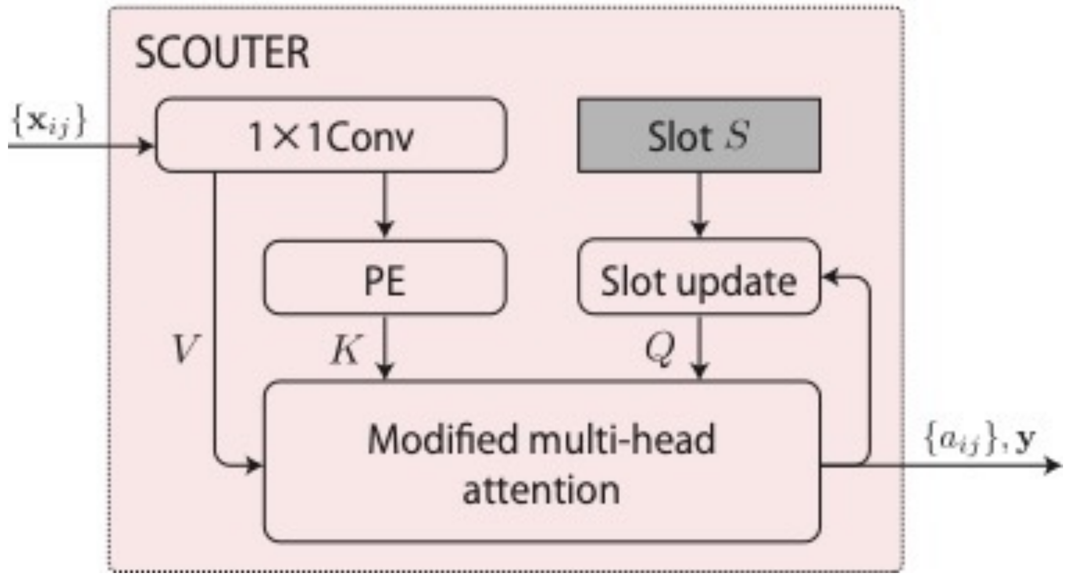
- BotCL [Wang et al., CVPR 2023]
  - Discovering concepts that describe image for a given classification task

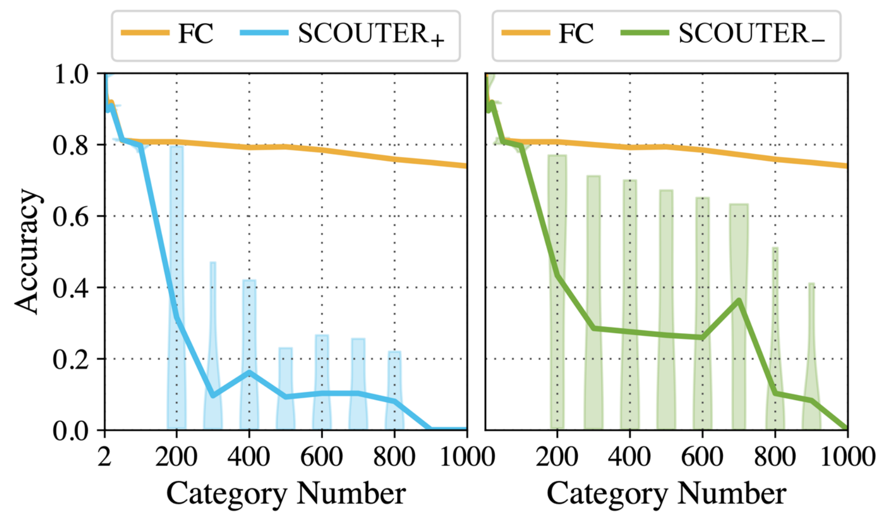


SCOUTER 🧐 [Li et al. ICCV 2021]

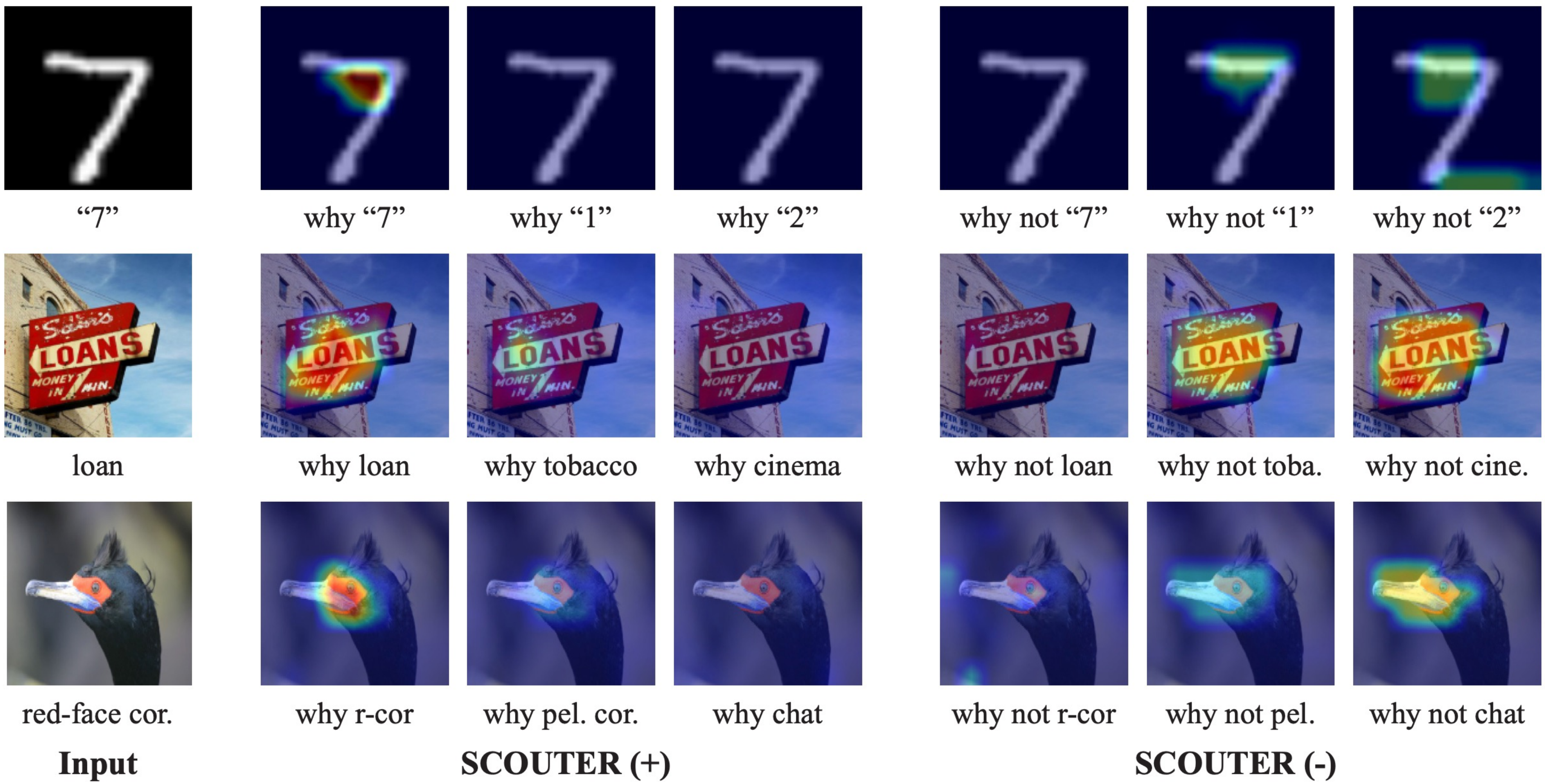
- a



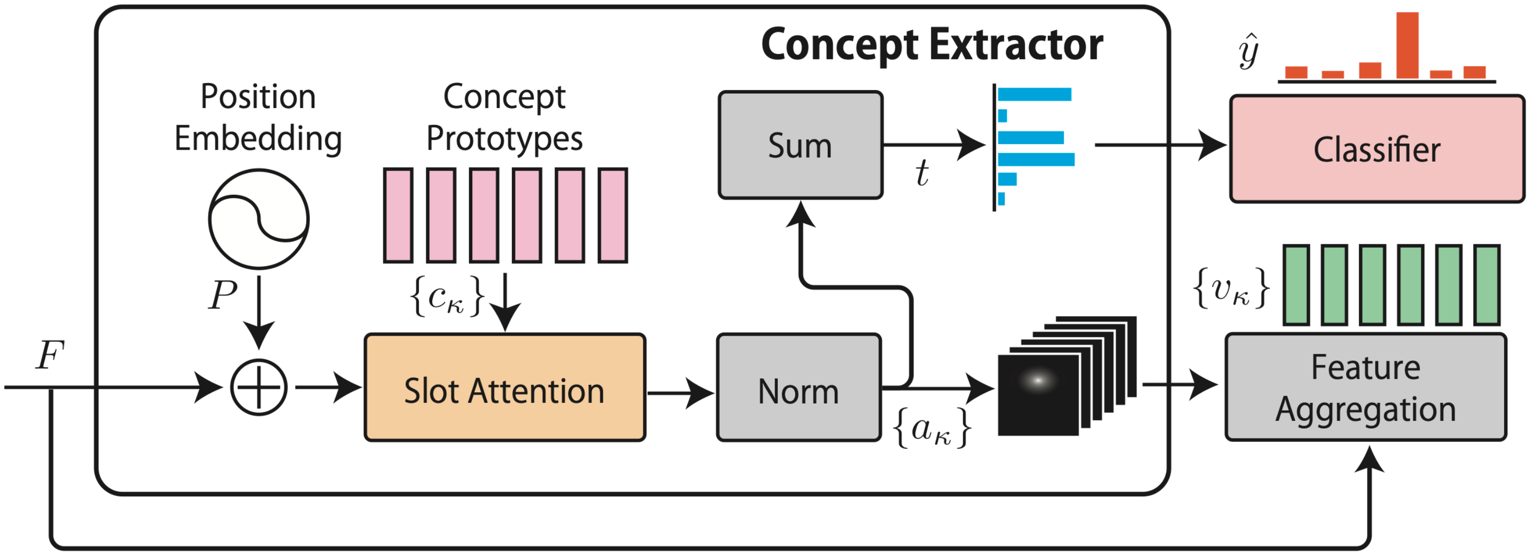


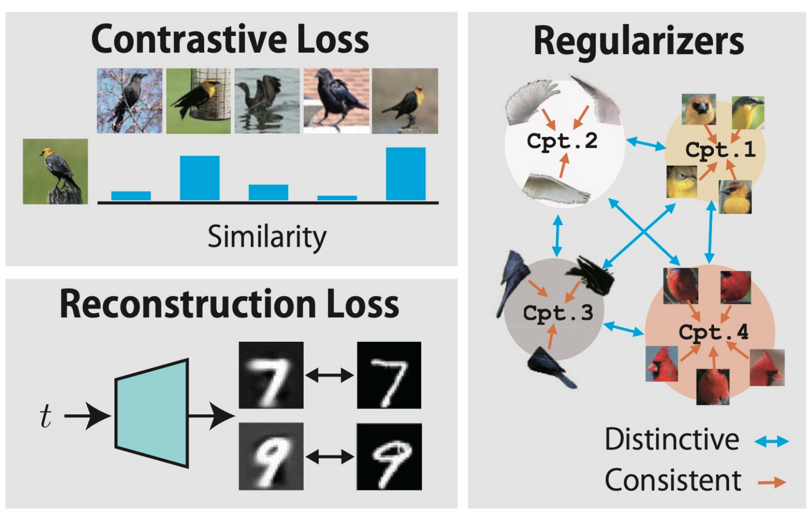


SCOUTER 🧐 [Li et al. ICCV 2021]



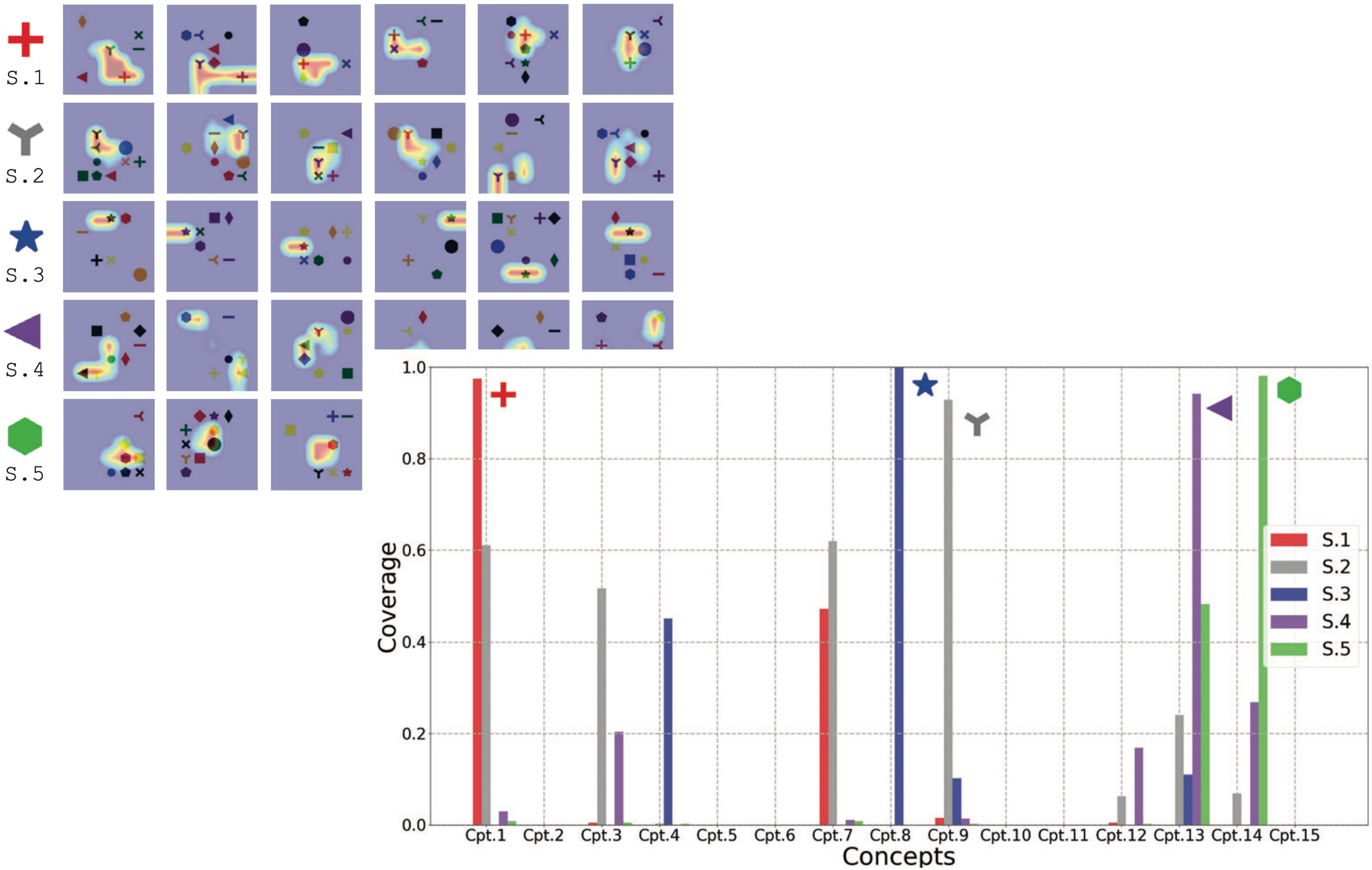
BotCL [Wang et al., CVPR 2023]



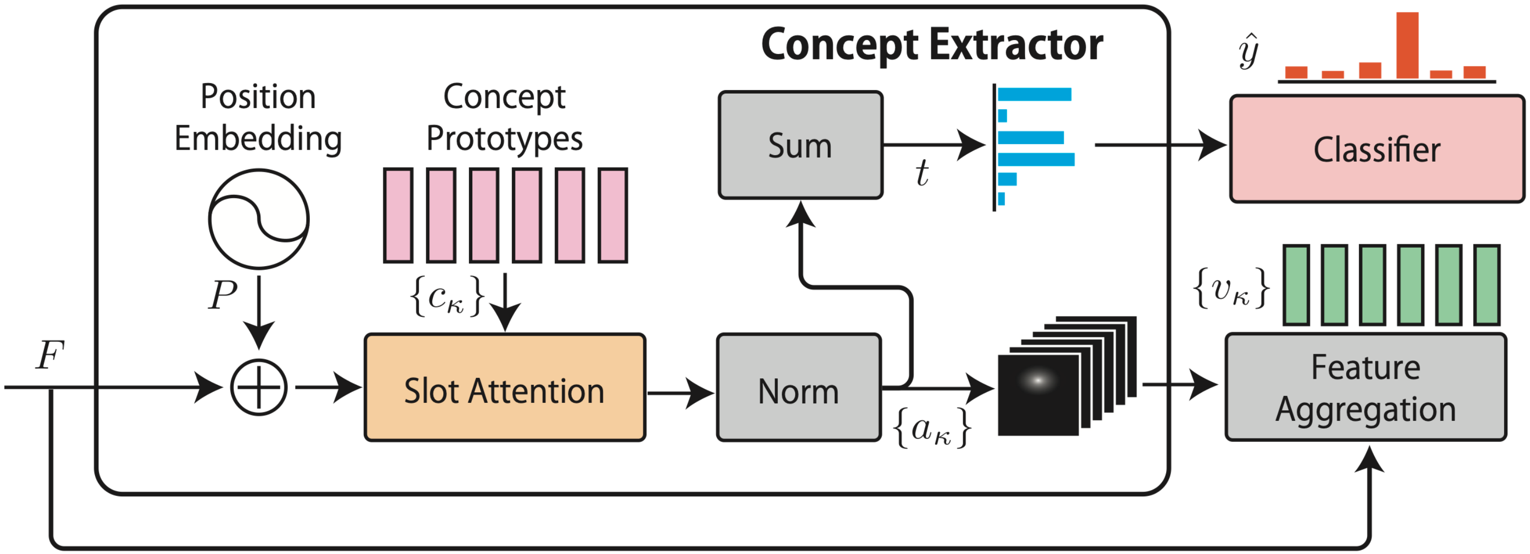


BotCL: Discovered concepts

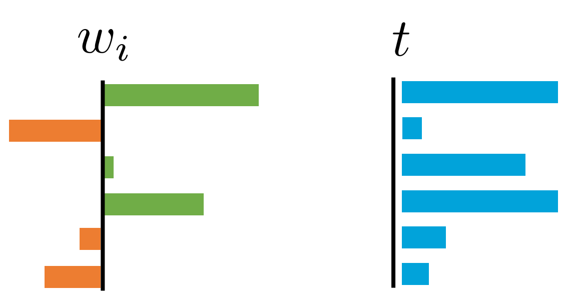
BotCL: Evaluation of discovered concepts



BotCL: Interpretation of classifier



- Our classifier is a single FC layer $y_i = w_i^\top t + b_i$
  - w_i learns the correlations between each class and concept



Next steps

- Efficient representation with a smaller set of concepts
  - More structured
  - Perhaps with grammar

- Without target task
  - Unsupervised (self-supervised) training for concept discovery

- Exploring some ways to identify spurious correlations
  - For vision and language tasks
  - For vision tasks