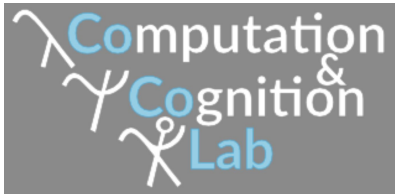
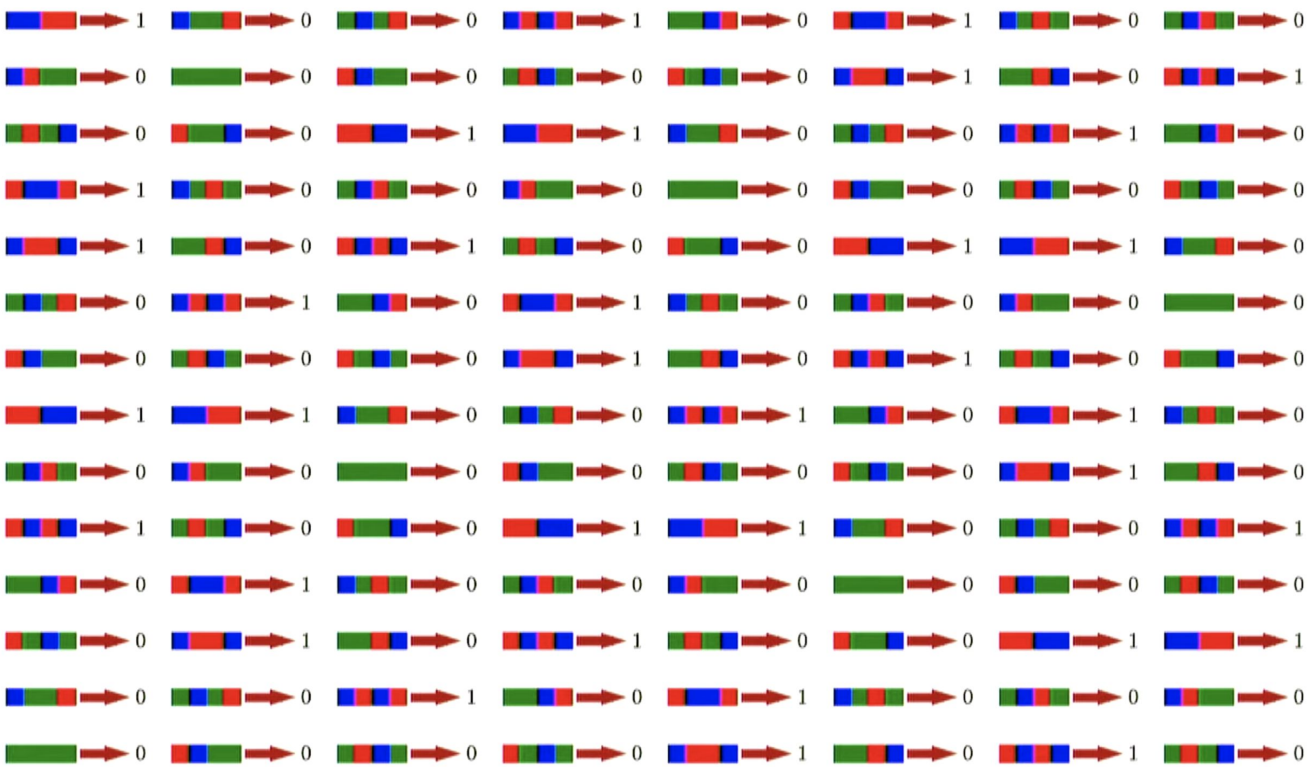
Shaping Visual Representations with Language for Few-Shot Classification



Jesse Mu Percy Liang Noah Goodman

ACL 2020



(courtesy Percy Liang)

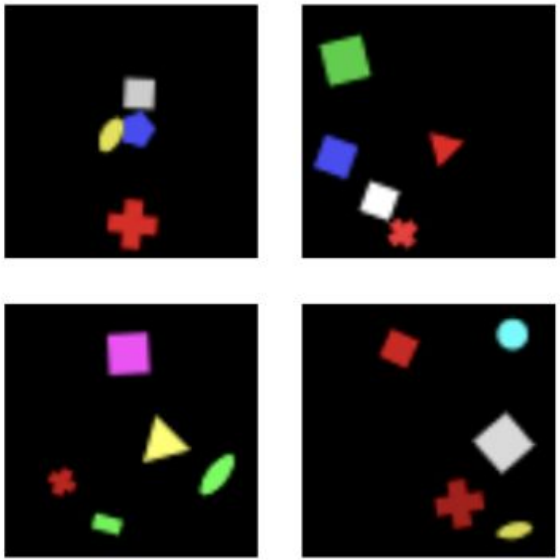
at least 2 red squares

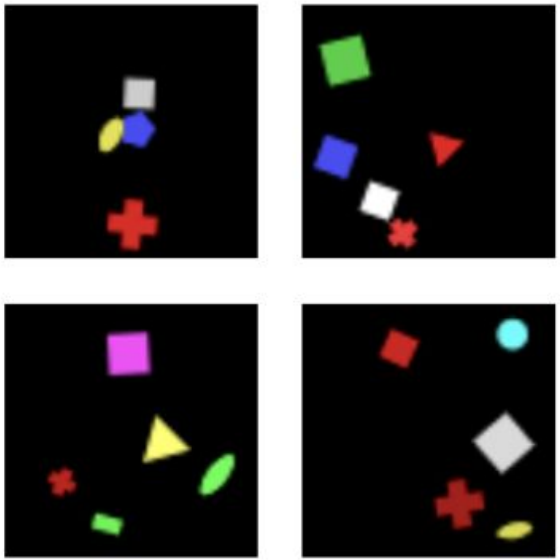
(courtesy Percy Liang)

How can language **guide representation learning**, especially when data is scarce?

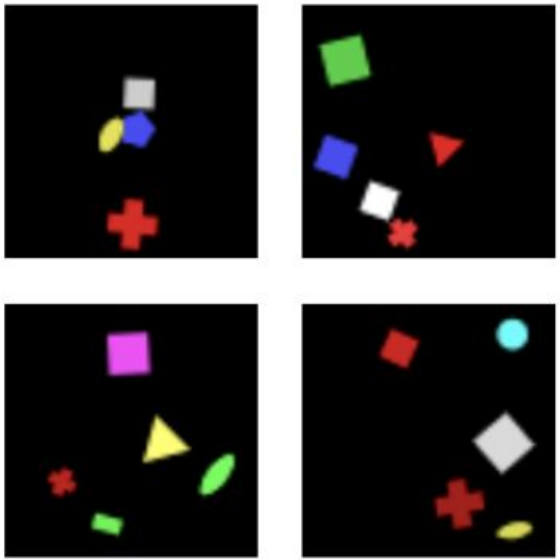
How can language **guide representation learning**, especially when data is scarce?

We study the (underexplored!) setting where language is **available** at train time, but **unavailable** for new tasks at test time

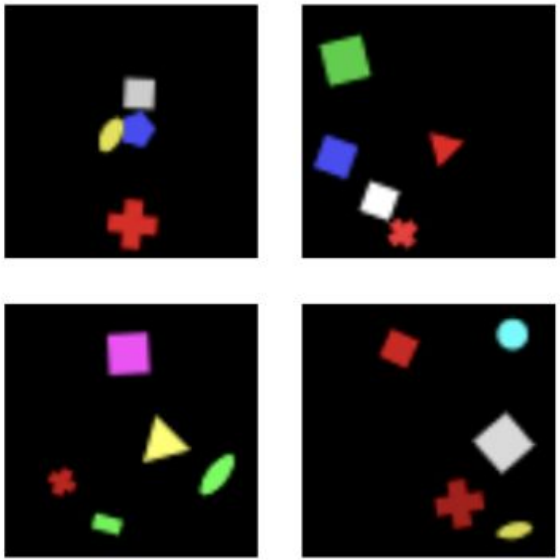




?



true

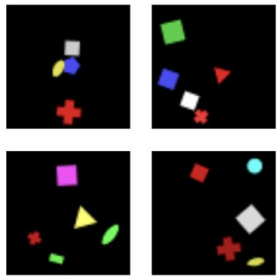


a red cross is below a square

true

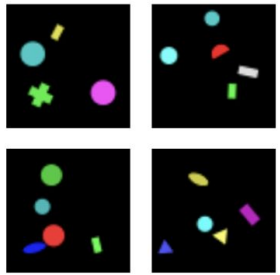
Meta-Train

Task 1



true

Task 2



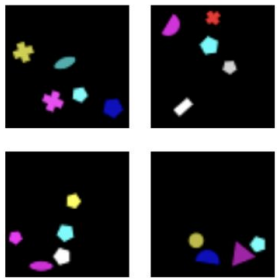
false

a red cross is below a square

a cyan circle is to the left of a rectangle

Language descriptions

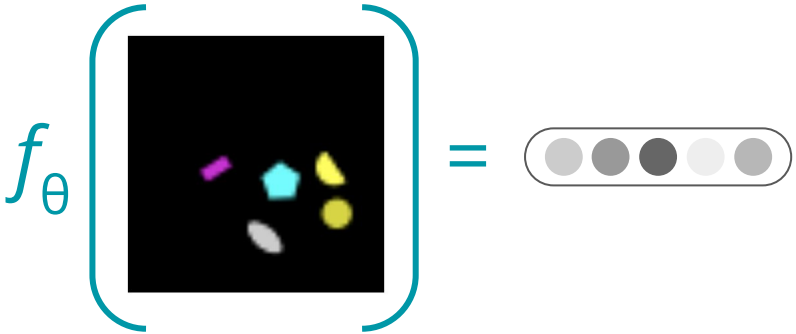
Meta-Test

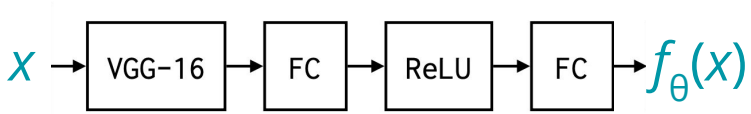


true

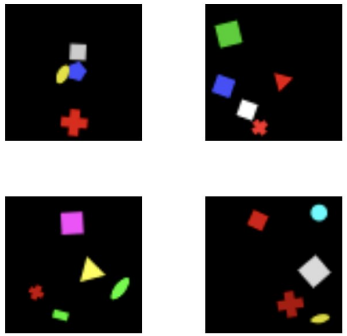
Prototype networks (Snell et al., 2017)

Prototype networks (Snell et al., 2017)



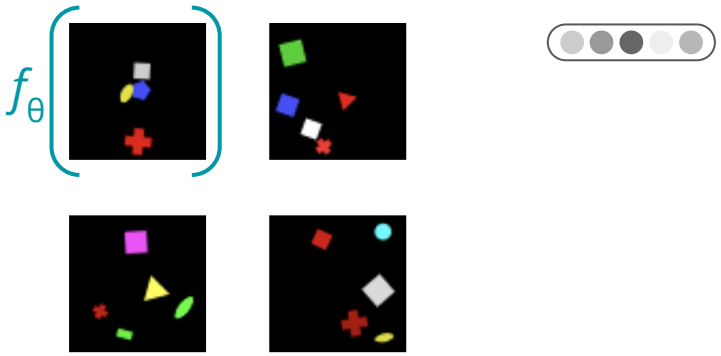


Prototype networks (Snell et al., 2017)



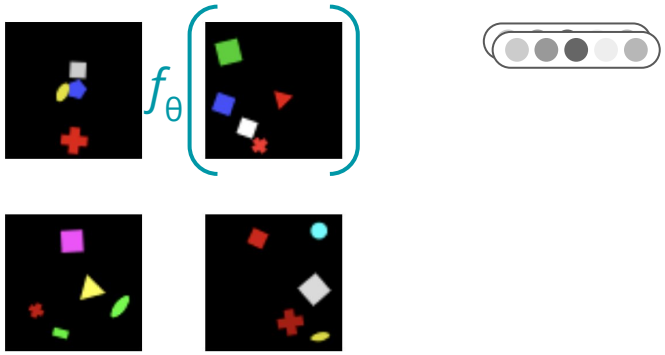
Train

Prototype networks (Snell et al., 2017)

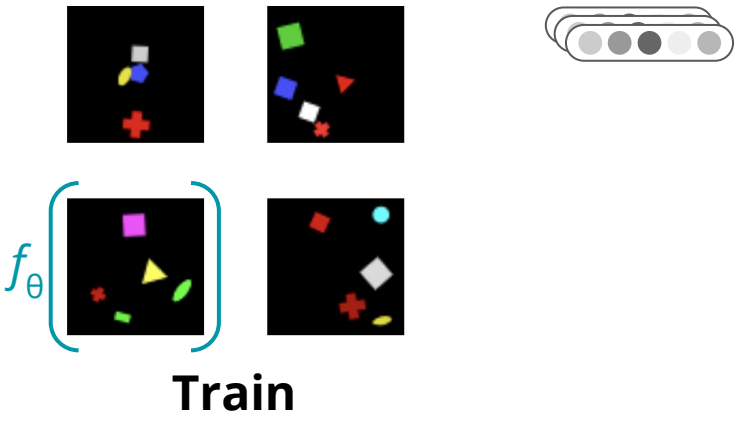


Train

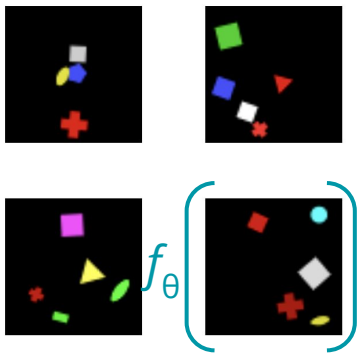
Prototype networks (Snell et al., 2017)

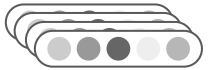


Prototype networks (Snell et al., 2017)

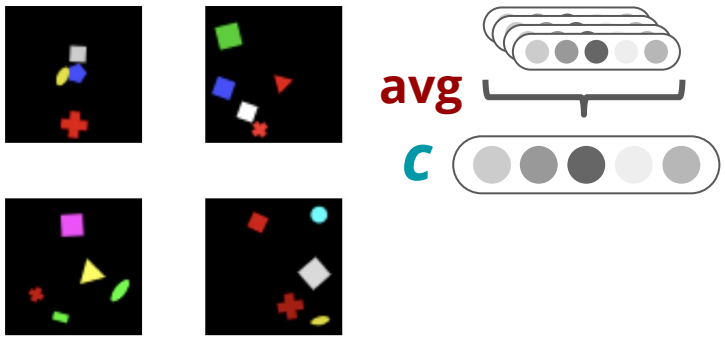


Prototype networks (Snell et al., 2017)

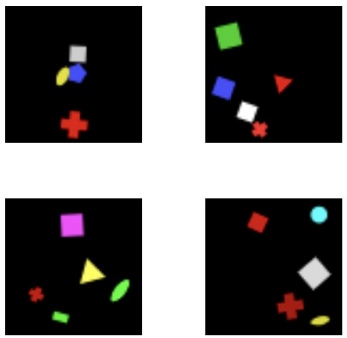


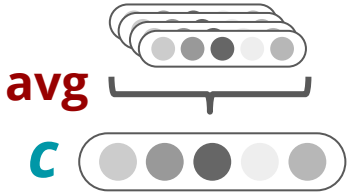


Prototype networks (Snell et al., 2017)



Prototype networks (Snell et al., 2017)

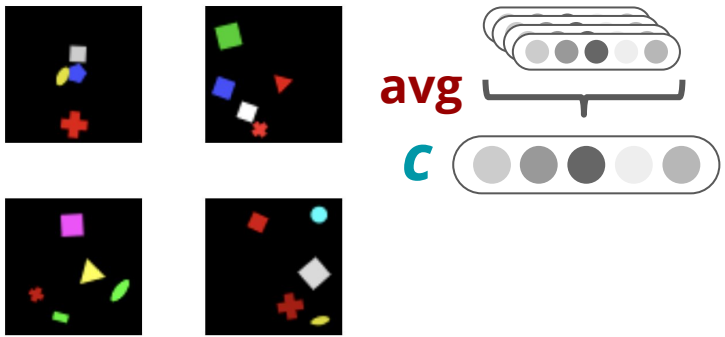




Train

Test

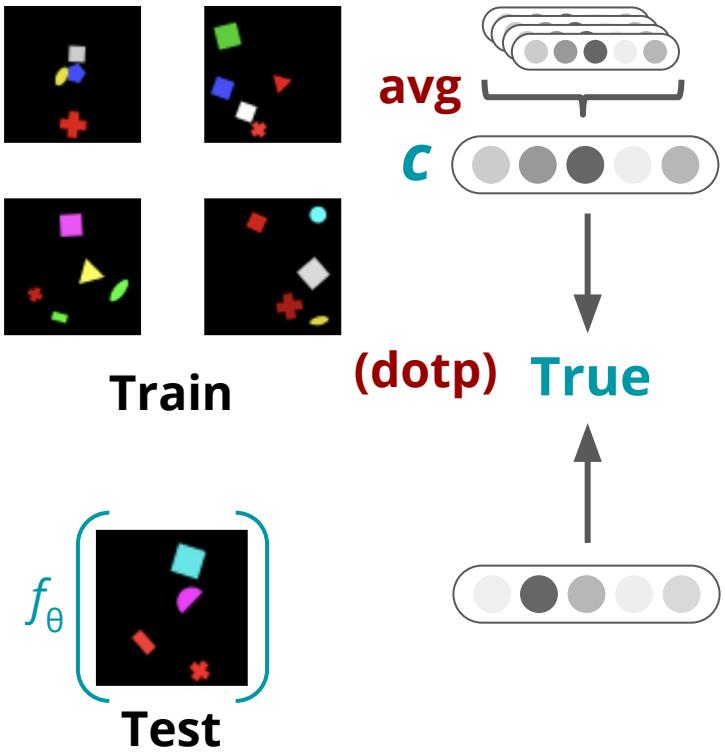
Prototype networks (Snell et al., 2017)



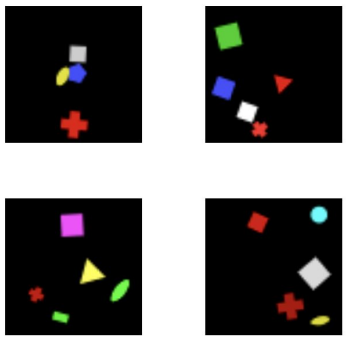
Train

Test

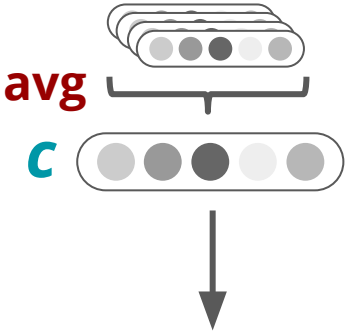
Prototype networks (Snell et al., 2017)



Prototype networks (Snell et al., 2017)



Train

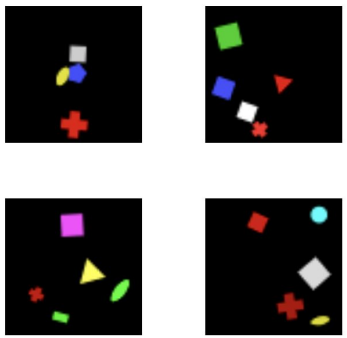


Test

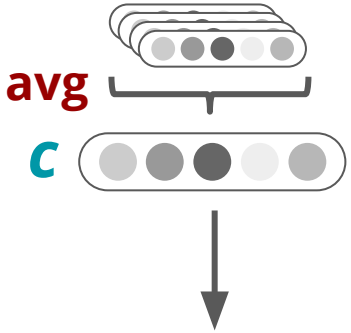
Minimize

$$\arg\min_{\theta} \mathcal{L}_{\text{CLS}}(\theta)$$

Language-shaped learning (LSL): **Train**



Train



Test

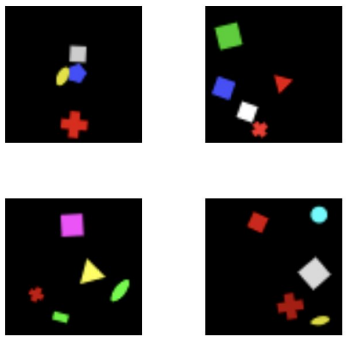
Description

a red cross is below a square

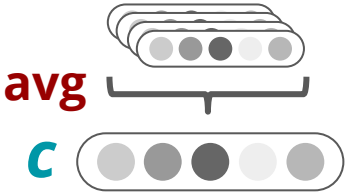
Minimize

$$\arg\min_{\theta} \mathcal{L}_{\text{CLS}}(\theta)$$

Language-shaped learning (LSL): **Train**



Train



g_ϕ

LSTM-Dec

Description

a red cross is below a square

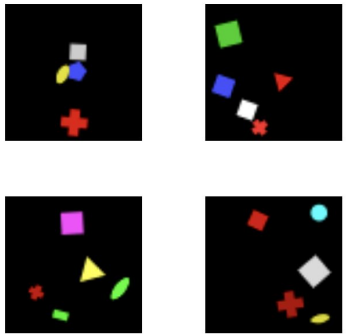
True

Test

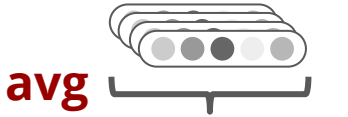
Minimize

$$\arg\min_{\theta} \mathcal{L}_{\text{CLS}}(\theta)$$

Language-shaped learning (LSL): **Train**



Train



True

Test

Auxiliary training

g_ϕ

LSTM-Dec

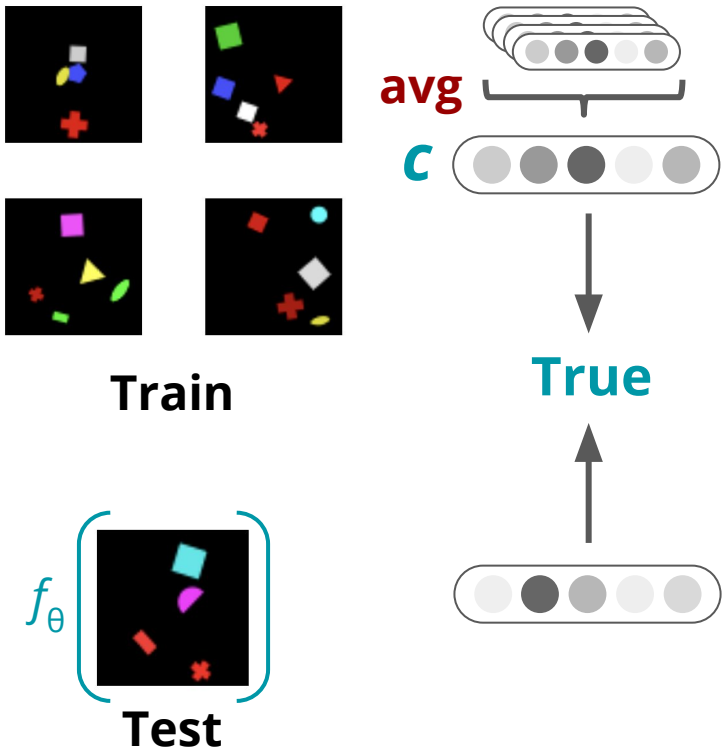
Description

a red cross is below a square

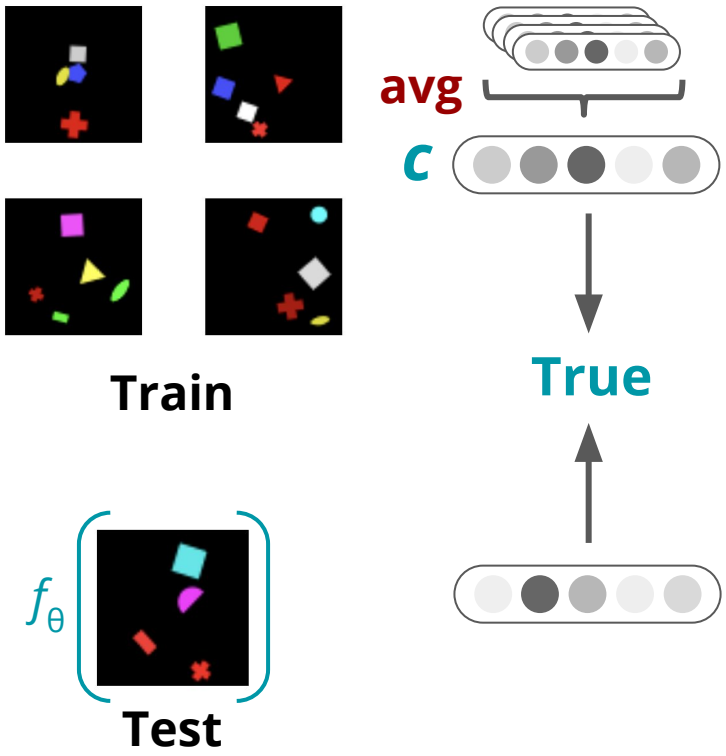
Jointly minimize

$$\arg\min_{\theta,\phi} \left[\mathcal{L}_{\text{CLS}}(\theta) + \beta_{\text{NL}}\mathcal{L}_{\text{NL}}(\theta,\phi)\right]$$

Language-shaped learning (LSL): **Test**



Learning with latent language (L3): **Train**

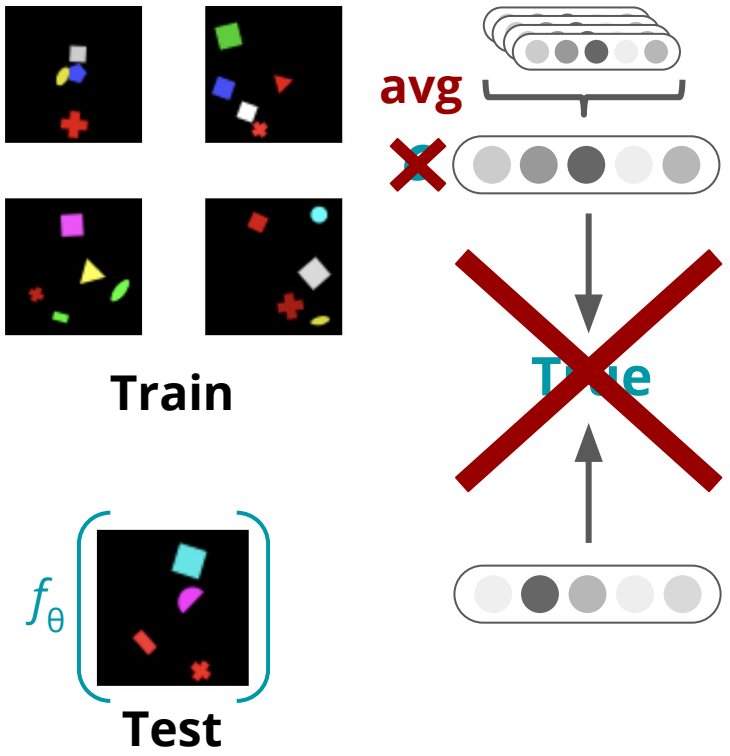


Description

a red cross is below a square

Andreas et al., 2018

Learning with latent language (L3): **Train**

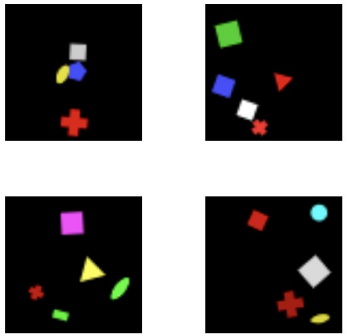


Description

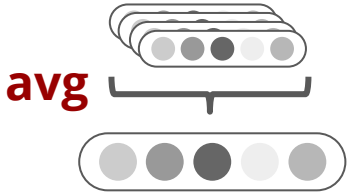
a red cross is below a square

Andreas et al., 2018

Learning with latent language (L3): **Train**



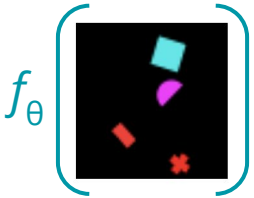
Train



Use language as a concept

Description

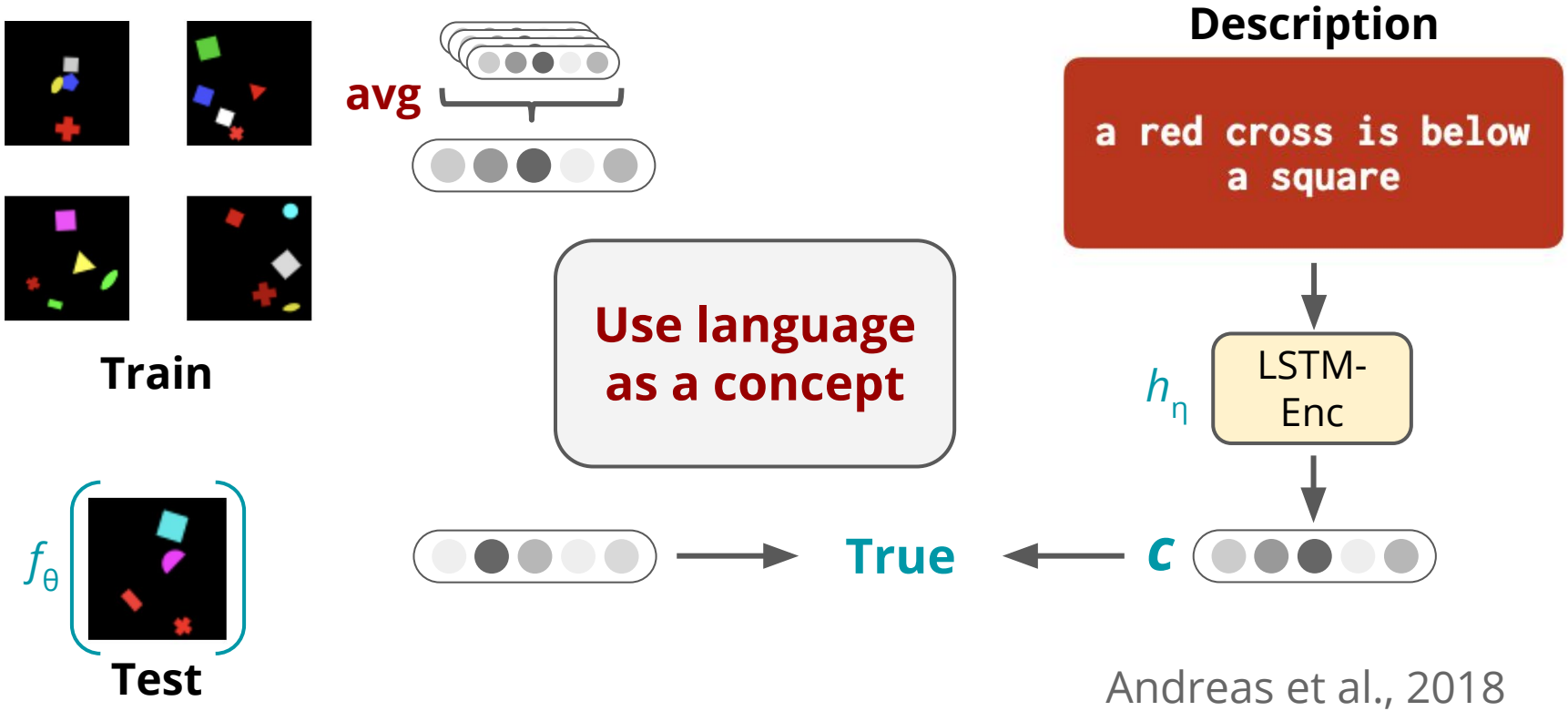
a red cross is below a square



Test

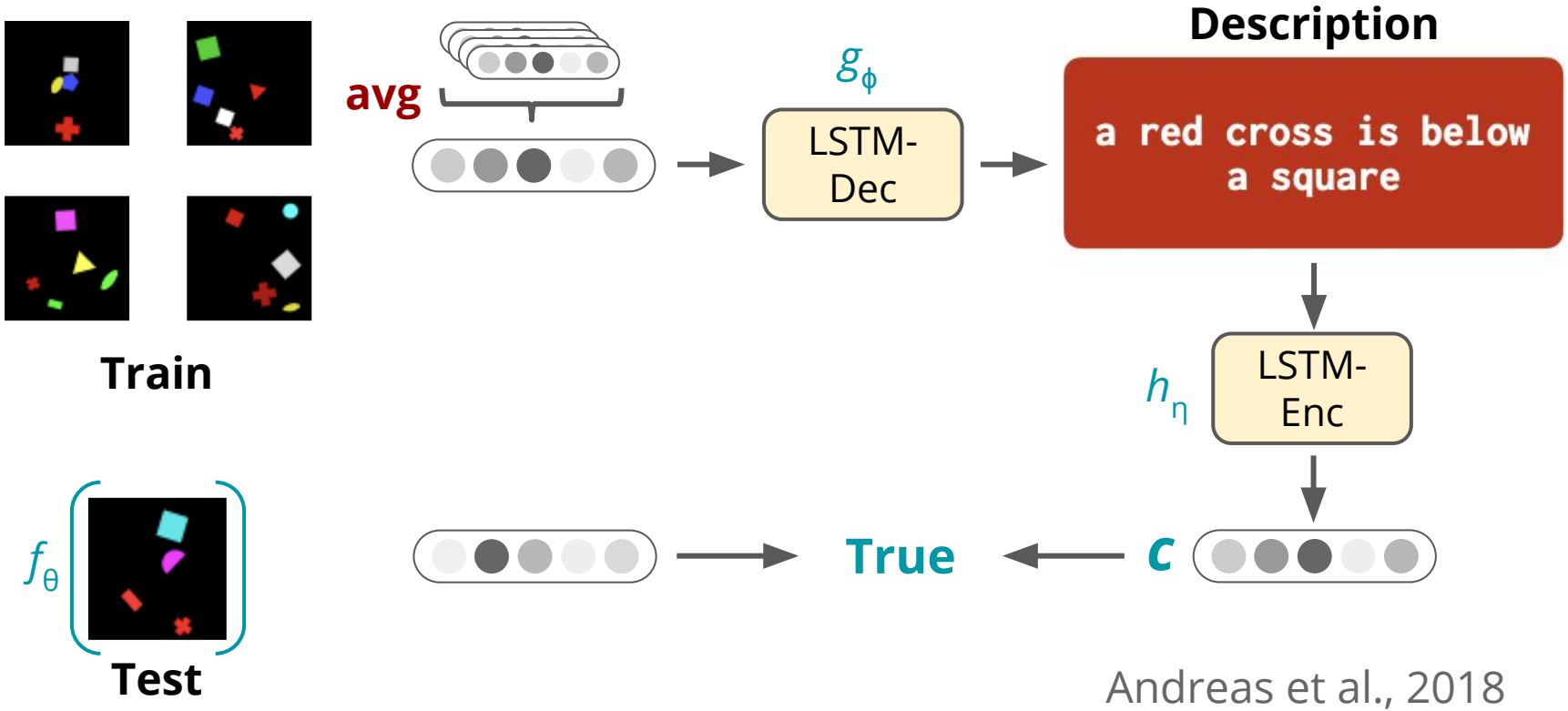
Andreas et al., 2018

Learning with latent language (L3): **Train**



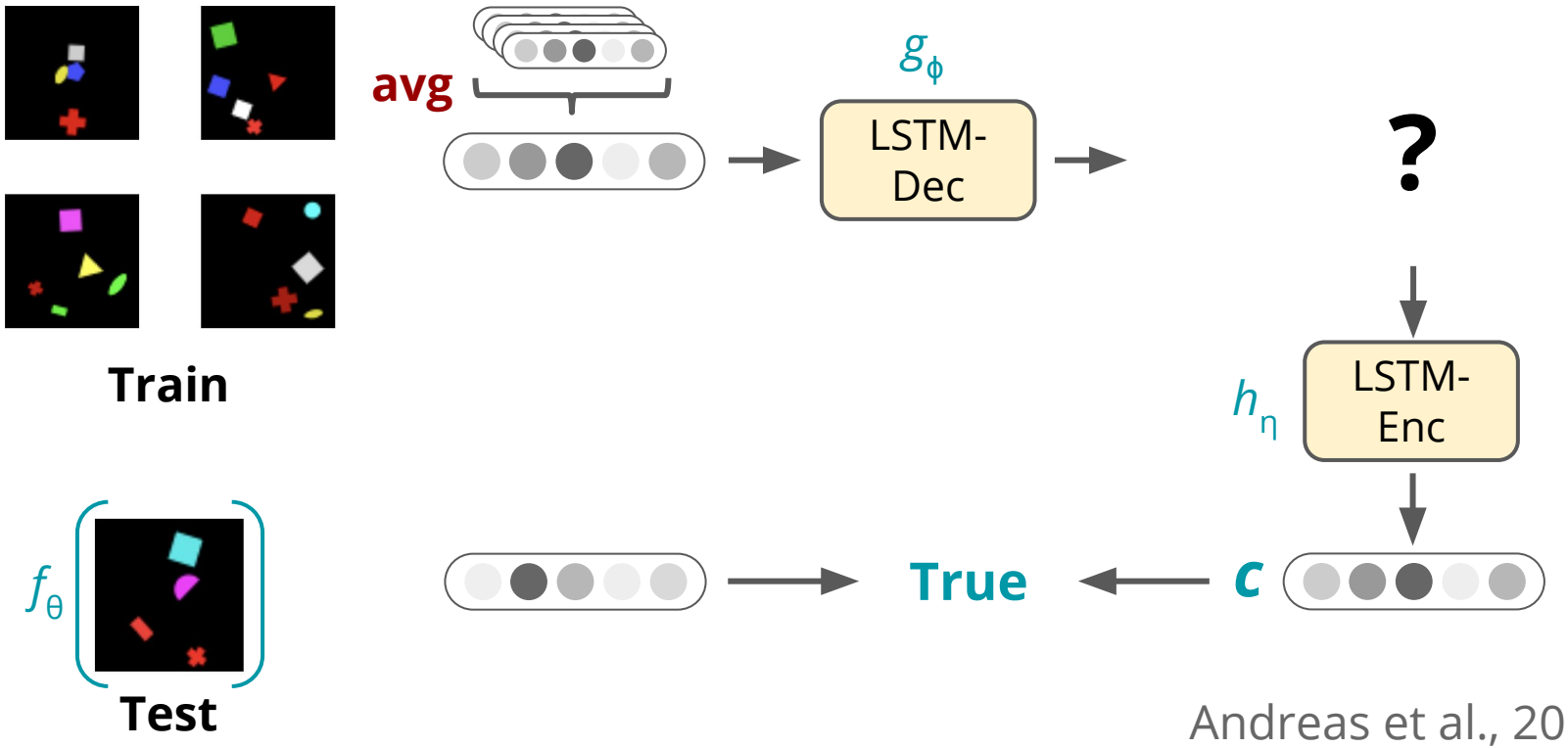
Andreas et al., 2018

Learning with latent language (L3): **Train**



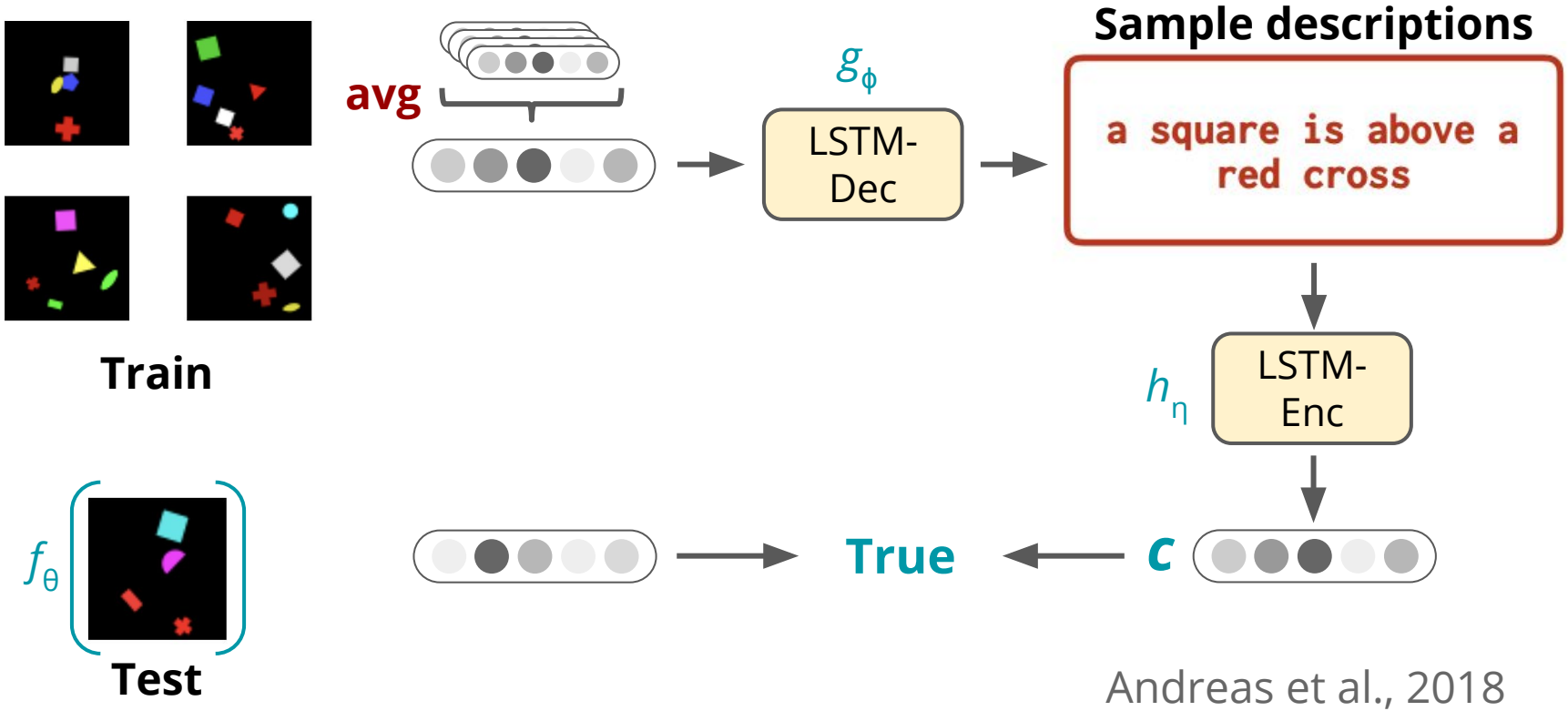
Andreas et al., 2018

Learning with latent language (L3): **Test**



Andreas et al., 2018

Learning with latent language (L3): **Test**



Andreas et al., 2018

Two Questions

Two Questions

1. Does a model trained **with language** (LSL) do better than a model trained **without** (Meta)?

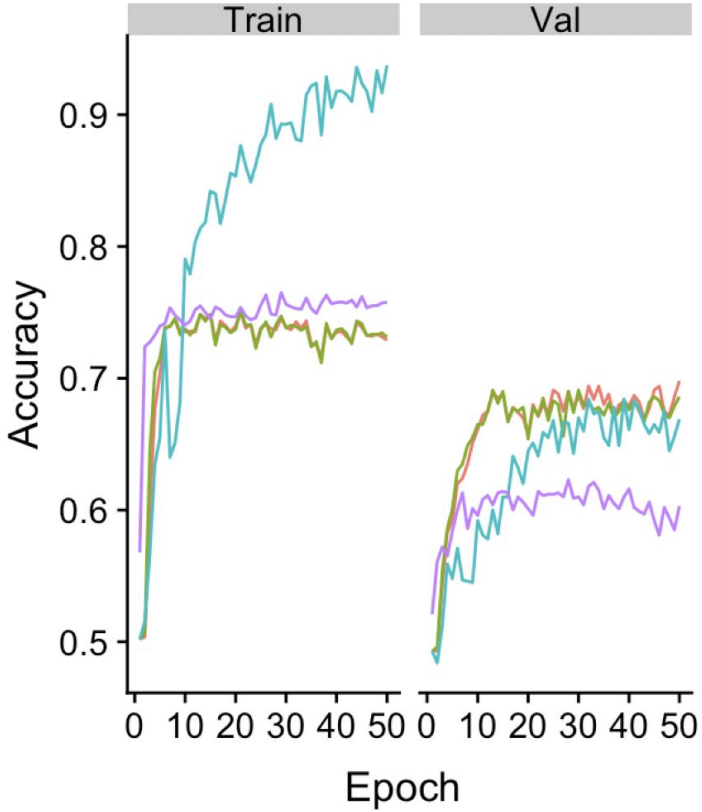
Two Questions

1. Does a model trained **with language** (LSL) do better than a model trained **without** (Meta)?

2. Is there any benefit to using language as a **discrete bottleneck** (L3), rather than just an **auxiliary training objective** (LSL)?

ShapeWorld: Results

ShapeWorld: Results

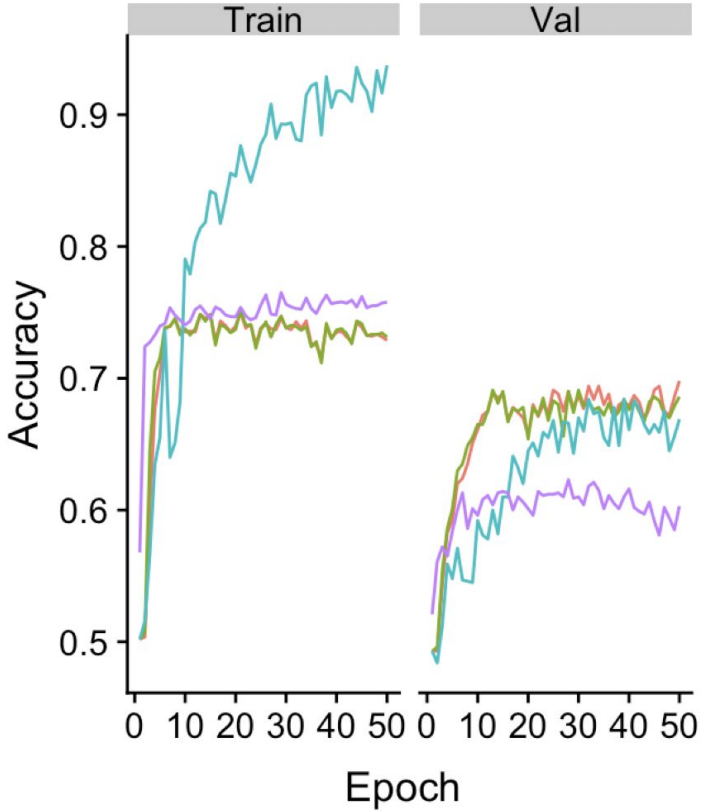


Meta

LSL

L3

ShapeWorld: Results

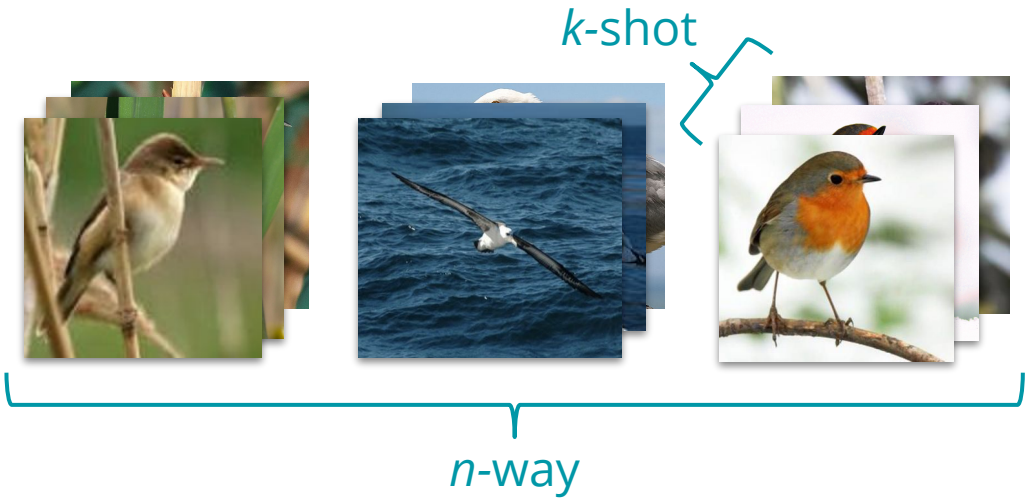


	Val	Test	
Meta	62	61	
LSL	69	**67**	**+6**
L3	**70**	**67**	**+6**

Scaling up to real vision + language

Scaling up to real vision + language
Caltech-UCSD Birds

n-way, *k*-shot classification



Train

Test

Natural language annotations (Reed et al., 2016)

The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.

This bird has distinctive-looking brown and white stripes all over its body, and its brown tail sticks up.

Natural language annotations (Reed et al., 2016)

The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.

This bird has distinctive-looking brown and white stripes all over its body, and its brown tail sticks up.

Assume limited, *class-level* language:
sample *D* = 20 captions per class (~2000 captions total)

Birds: results

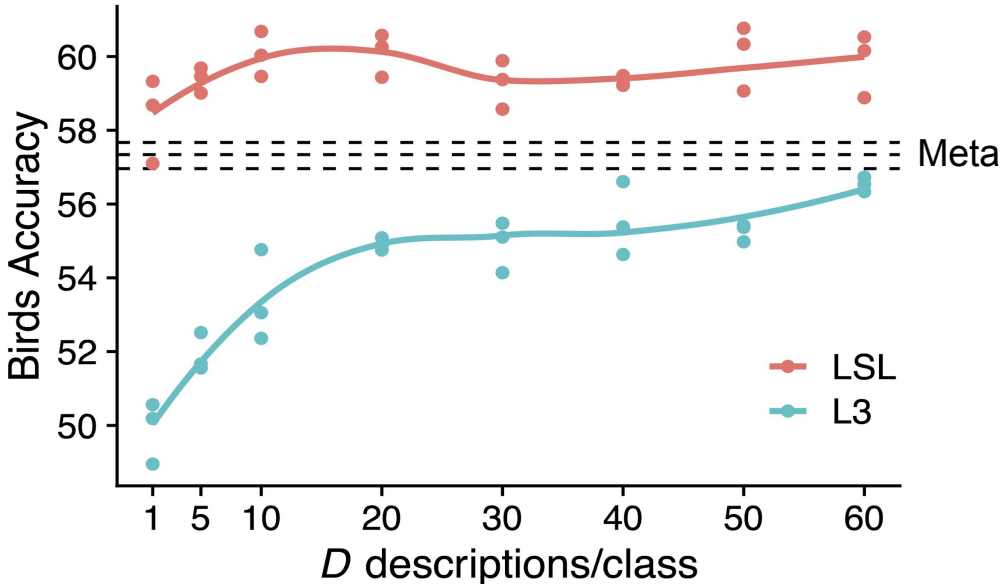
5-way, 1-shot classification

	Accuracy (± 95% CI)	
Meta	**58.0** ± .96	
LSL	**61.2** ± .96	**+3.3**
L3	**54.0** ± 1.1	**-4.0**

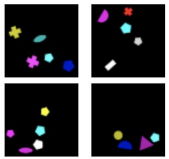
Birds: results

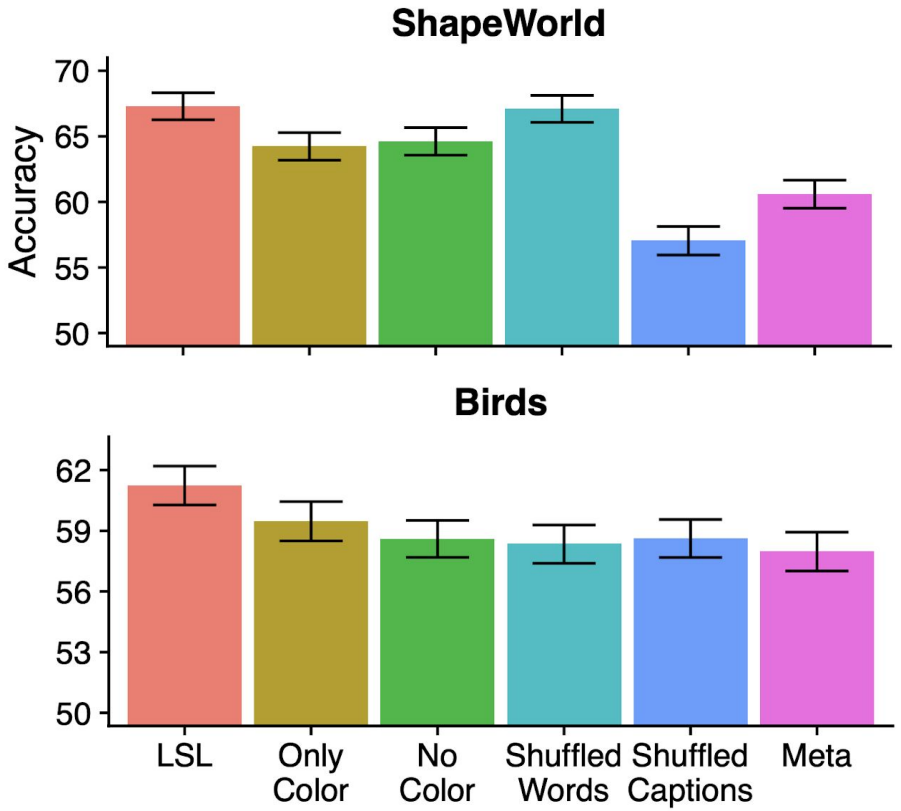
5-way, 1-shot classification

	Accuracy (± 95% CI)	
Meta	**58.0** ± .96	
LSL	**61.2** ± .96	**+3.3**
L3	**54.0** ± 1.1	**-4.0**



What about language helps?

	ShapeWorld	Birds
		
Original	a cyan pentagon is to the right of a magenta shape	The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.
Only Color	cyan magenta	white black white
No Color	a pentagon is to the right of a shape	The bird has a underbelly feathers in the wings, a large wingspan, and a beak.
Shuffled Words	shape right the is a pentagon a of cyan to magenta	The , a and a . , beak bird in wingspan feathers large the black white underbelly has , white a wings
Shuffled Captions	a green square is below a triangle	This magnificent fellow is almost all black with a red crest, and white cheek patch.



Two Questions

1. Does a model trained **with language** (LSL) do better than a model trained **without** (Meta)?

2. Is there any benefit to using language as a **discrete bottleneck** (L3), rather than just an **auxiliary training objective** (LSL)?

Two Questions

1. Does a model trained **with language** (LSL) do better than a model trained **without** (Meta)?
   > Yes! Language is a promising source of supervision for vision models.
2. Is there any benefit to using language as a **discrete bottleneck** (L3), rather than just an **auxiliary training objective** (LSL)?

Two Questions

1. Does a model trained **with language** (LSL) do better than a model trained **without** (Meta)?
   > Yes! Language is a promising source of supervision for vision models.
2. Is there any benefit to using language as a **discrete bottleneck** (L3), rather than just an **auxiliary training objective** (LSL)?
   > No, at least for the tasks explored here.

Questions for discussion

1. This paper looked at using language as (1) a regularizer, or (2) a bottleneck for class-level representations. How / where else could we use language to support the training process?

2. What do we expect to be the comparative strengths of LSL / L3 / other language-based training procedures?

Thanks!

Thanks to Pang Wei Koh, Sebastian Schuster, and Dan Iter for feedback, and Mike Wu and Jacob Andreas for code and discussions.

We gratefully acknowledge support from Toyota Research Institute, the Office of Naval Research, and an NSF Graduate Fellowship.