at least 2 red squares
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.
a red cross is below a square

true
Meta-Train

Task 1

Task 2

true

false

Meta-Test

Language descriptions

a red cross is below a square

a cyan circle is to the left of a rectangle

true
Prototype networks (Snell et al., 2017)
Prototype networks (Snell et al., 2017)

\[ f_\theta(x) = \text{VGG-16} \rightarrow \text{FC} \rightarrow \text{ReLU} \rightarrow \text{FC} \rightarrow f_\theta(x) \]
Prototype networks (Snell et al., 2017)

Train
Prototype networks (Snell et al., 2017)

Train

$f_\theta$
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)
Prototype networks (Snell et al., 2017)
Prototype networks (Snell et al., 2017)

Train

Test

\[ f_0 \]

\[ \text{avg} \]

\[ c \]

\( \text{(dotp)} \)

True
Prototype networks (Snell et al., 2017)

Train

Test

Minimize

$$\arg\min_{\theta} L_{CLS}(\theta)$$
Language-shaped learning (LSL): **Train**

**Description**
- A red cross is below a square

**Minimize**
$$\arg\min_{\theta} \mathcal{L}_{CLS}(\theta)$$
Language-shaped learning (LSL): \textbf{Train}

\[ f_\theta \]

**Train**

\[ c \]

\[ \text{avg} \]

\[ \text{True} \]

**Test**

\[ g_\phi \]

LSTM-Dec

**Description**

a red cross is below a square

**Minimize**

\[ \arg\min_{\theta} \mathcal{L}_{CLS}(\theta) \]
Language-shaped learning (LSL): **Train**

**Train**

**Test**

Jointly minimize

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

**Description**

A red cross is below a square
Language-shaped learning (LSL): **Test**

Train

\[ f_{\theta} \]

Test

\[ \text{avg} \]

\[ c \]

True
Learning with latent language (L3): Train

Train

Test

avg

true

Description

a red cross is below a square

Andreas et al., 2018
Learning with latent language (L3): Train

Andreas et al., 2018

Description

a red cross is below a square
Learning with latent language (L3): **Train**

- **Train**
  - Use language as a concept
  - avg

- **Test**

**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*

**Use language as a concept**

\[ f_\theta \]

\[ h_\eta \]

**LSTM-Enc**

\[ c \]

Andreas et al., 2018
Learning with latent language (L3): **Train**

Train

Test

Description

a red cross is below a square

Andreas et al., 2018
Learning with latent language (L3): Test

\[ f_\theta \quad \text{avg} \quad g_\phi \quad \text{LSTM-Dec} \]

\[ \text{Train} \quad ? \quad \text{Test} \]

\[ h_\eta \quad \text{LSTM-Enc} \]

\[ c \quad \text{True} \]

Andreas et al., 2018
Learning with latent language (L3): Test

Sample descriptions

a square is above a red cross

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

![Graph showing accuracy over epochs for Train and Val sets for Meta, LSL, and L3]
ShapeWorld: Results

- **Meta**: Val 62, Test 61
- **LSL**: Val 69, Test 67 (+6)
- **L3**: Val 70, Test 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)

Assume limited, class-level language:
sample $D = 20$ captions per class (~2000 captions total)
Birds: results
5-way, 1-shot classification

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (± 95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta</td>
<td>58.0 ± .96</td>
</tr>
<tr>
<td>LSL</td>
<td>61.2 ± .96</td>
</tr>
<tr>
<td>L3</td>
<td>54.0 ± 1.1</td>
</tr>
</tbody>
</table>
Birds: results
5-way, 1-shot classification

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (± 95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta</td>
<td>58.0 ± .96</td>
</tr>
<tr>
<td>LSL</td>
<td>61.2 ± .96 +3.3</td>
</tr>
<tr>
<td>L3</td>
<td>54.0 ± 1.1 -4.0</td>
</tr>
</tbody>
</table>

![Graph showing accuracy over D descriptions/class](image)
What about language helps?

### ShapeWorld

<table>
<thead>
<tr>
<th>Original</th>
<th>Birds</th>
</tr>
</thead>
<tbody>
<tr>
<td>a cyan pentagon is to the right of a magenta shape</td>
<td>The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Only Color</th>
<th>cyan magenta</th>
<th>white black white</th>
</tr>
</thead>
</table>

| 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. |

### ShapeWorld Accuracy

![Accuracy Chart for ShapeWorld]

### Birds Accuracy

![Accuracy Chart for Birds]
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.