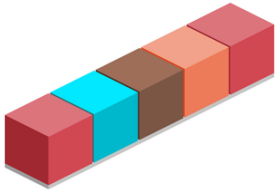
Pre-Learning Environment Representations for Data-Efficient Neural Instruction Following

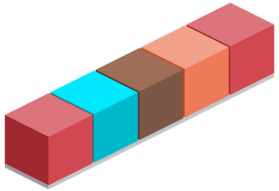
David Gaddy and Dan Klein

End-to-End Instruction Following

End-to-End Instruction Following

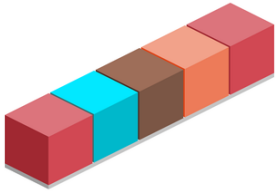


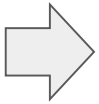
End-to-End Instruction Following



“Remove the block on the right”

End-to-End Instruction Following





“Remove the block on the right”

How do we do?

How do we do?

Baseline Neural

17.9%

How do we do?

Baseline Neural

17.9%

Logical Form Gap

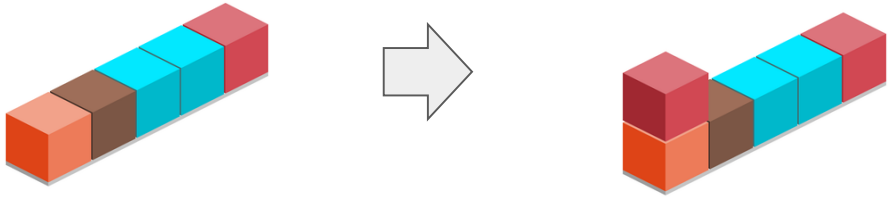
Logical Forms (Wang et al. 2016)

33.8%

Inductive Bias

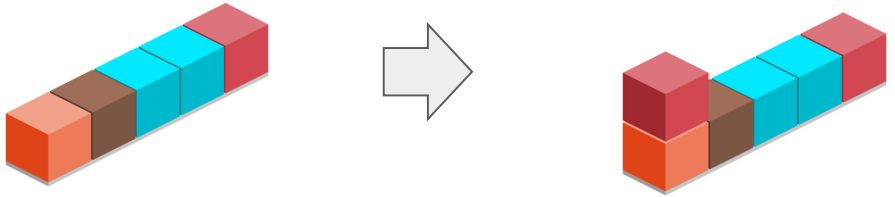
Inductive Bias

“Układaj czerwone bloki na niebieskich blokach”



Inductive Bias

“Układaj czerwone bloki na niebieskich blokach”

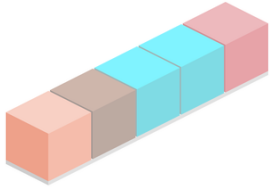


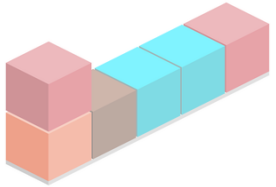
`stack(red, with(orange))`

`stack(red, leftmost)`

Inductive Bias

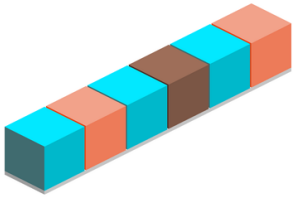
“Układaj czerwone bloki na niebieskich blokach”

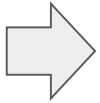




`stack(red, with(orange))`

`stack(red, leftmost)`

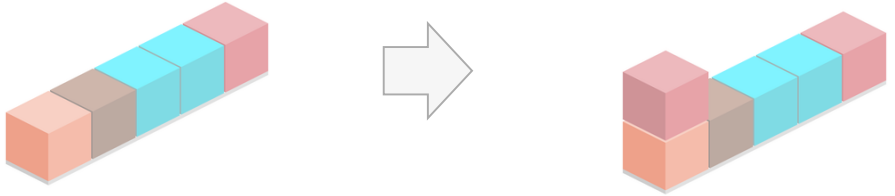




?

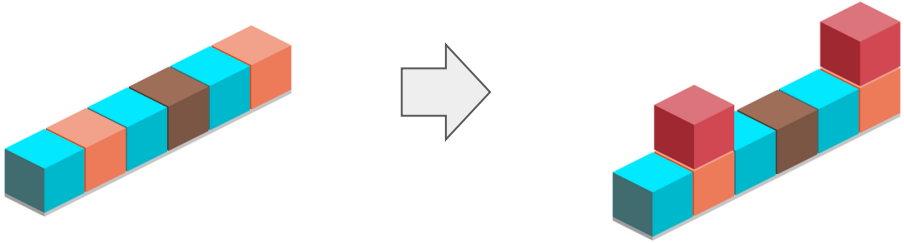
Inductive Bias

“Układaj czerwone bloki na niebieskich blokach”



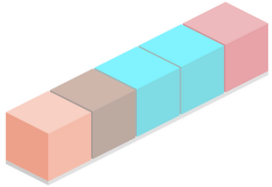
`stack(red, with(orange))`

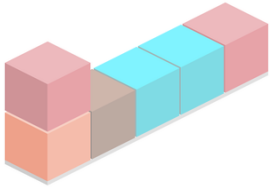
`stack(red, leftmost)`

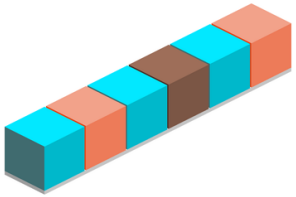


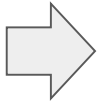
Inductive Bias

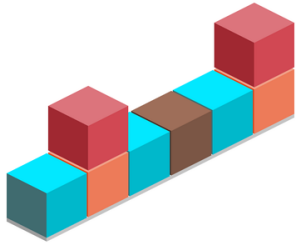
“Układaj czerwone bloki na niebieskich blokach”



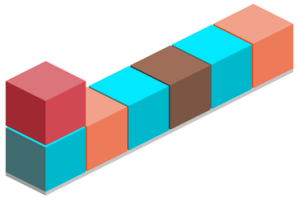






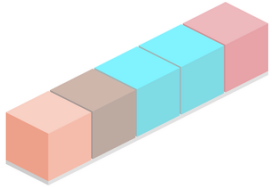


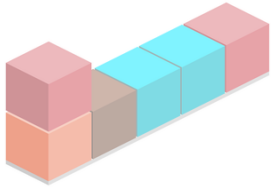
?

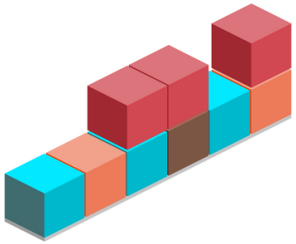


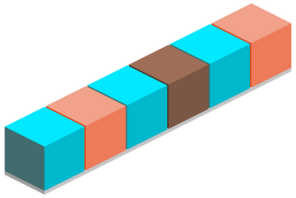
Inductive Bias

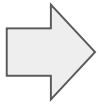
“Układaj czerwone bloki na niebieskich blokach”

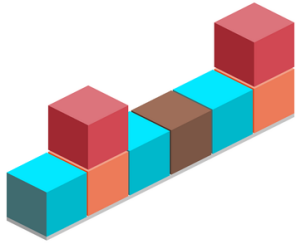




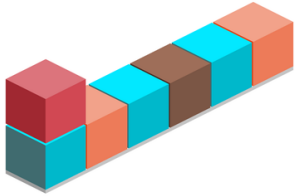






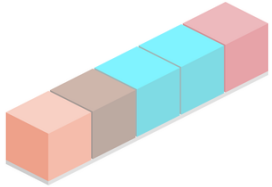


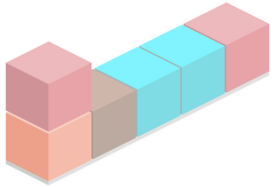
?

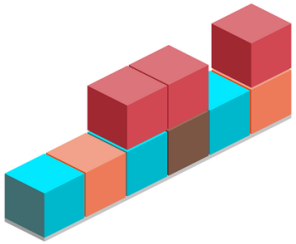


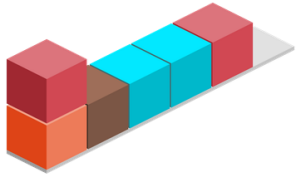
Inductive Bias

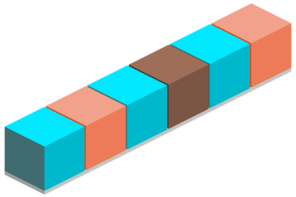
“Układaj czerwone bloki na niebieskich blokach”

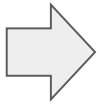


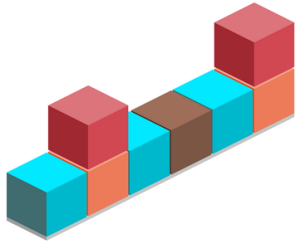




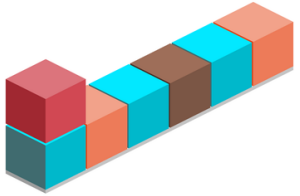






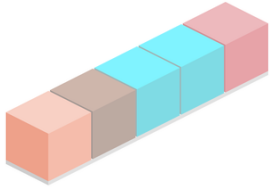


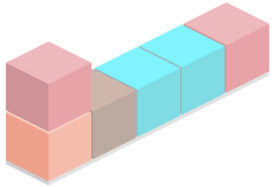
?

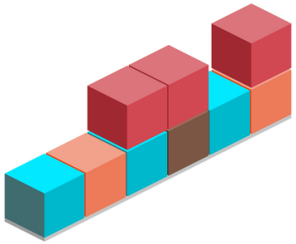


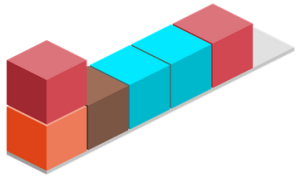
Inductive Bias

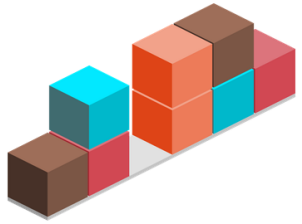
“Układaj czerwone bloki na niebieskich blokach”

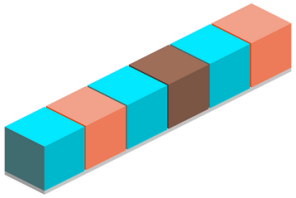


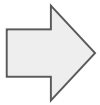


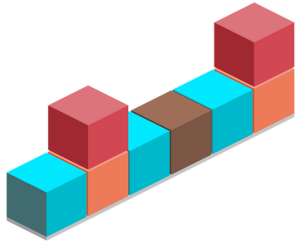




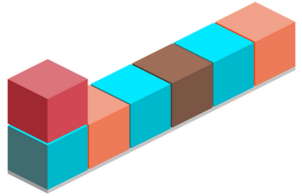






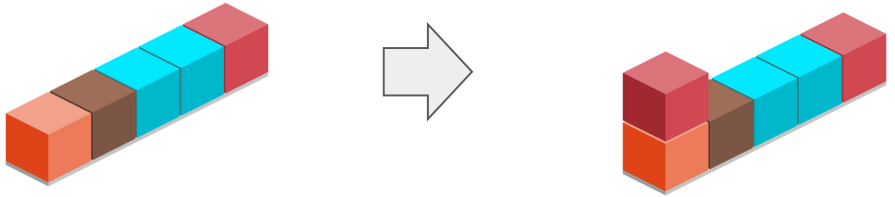


?



Inductive Bias

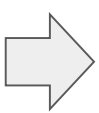
“Układaj czerwone bloki na niebieskich blokach”

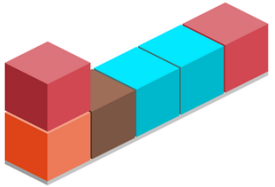


```
stack(red, leftmost)
```

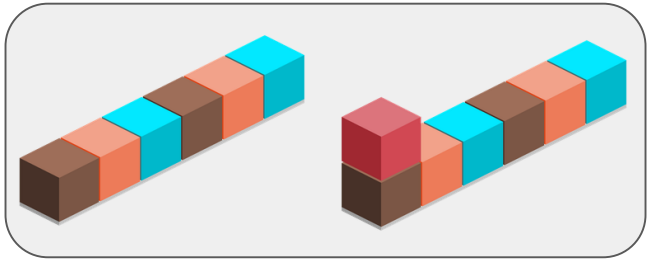
Inductive Bias

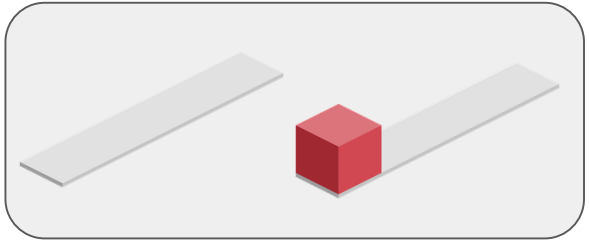
“Układaj czerwone bloki na niebieskich blokach”

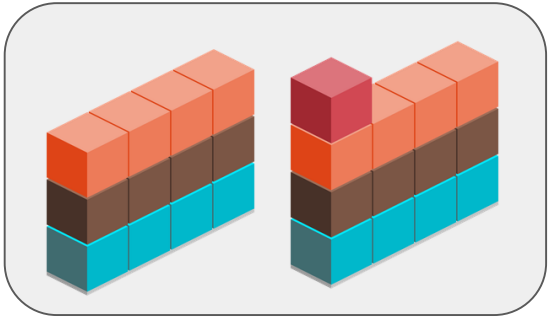




`stack(red, leftmost)`

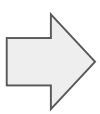


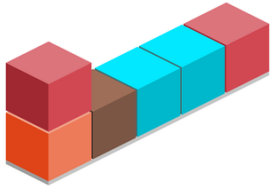


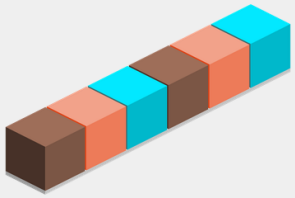


Inductive Bias

“Układaj czerwone bloki na niebieskich blokach”

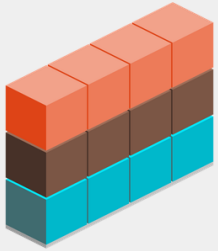






?

?



?

Logical Form Gap

Phase 1:
Environment Learning

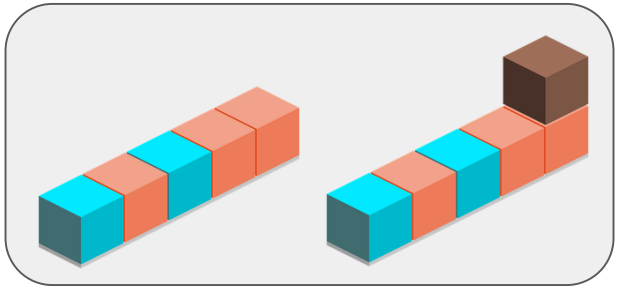
Learn abstractions
No Language data needed

Phase 2:
Language Learning

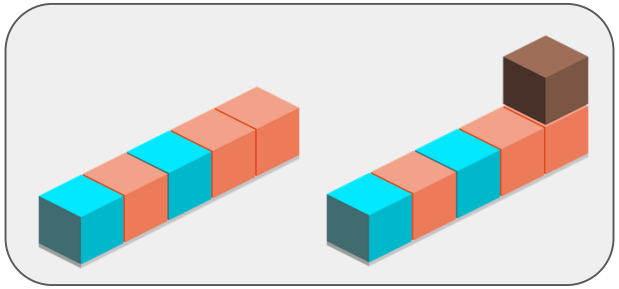
Map to abstractions
Needs less data

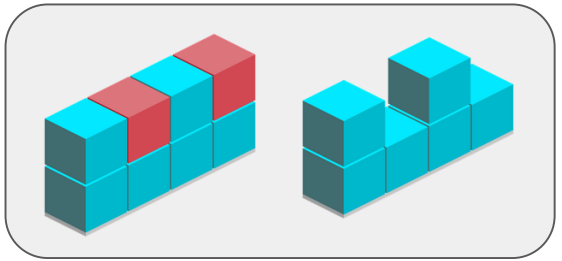
Watching the Environment

Watching the Environment

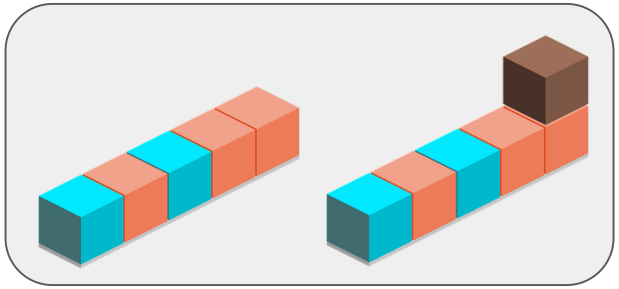


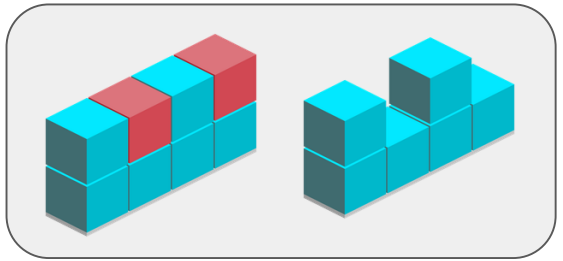
Watching the Environment

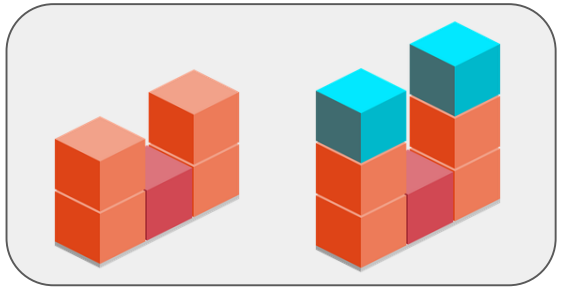




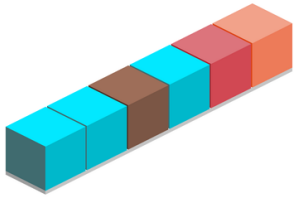
Watching the Environment

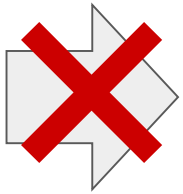


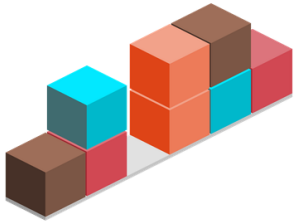




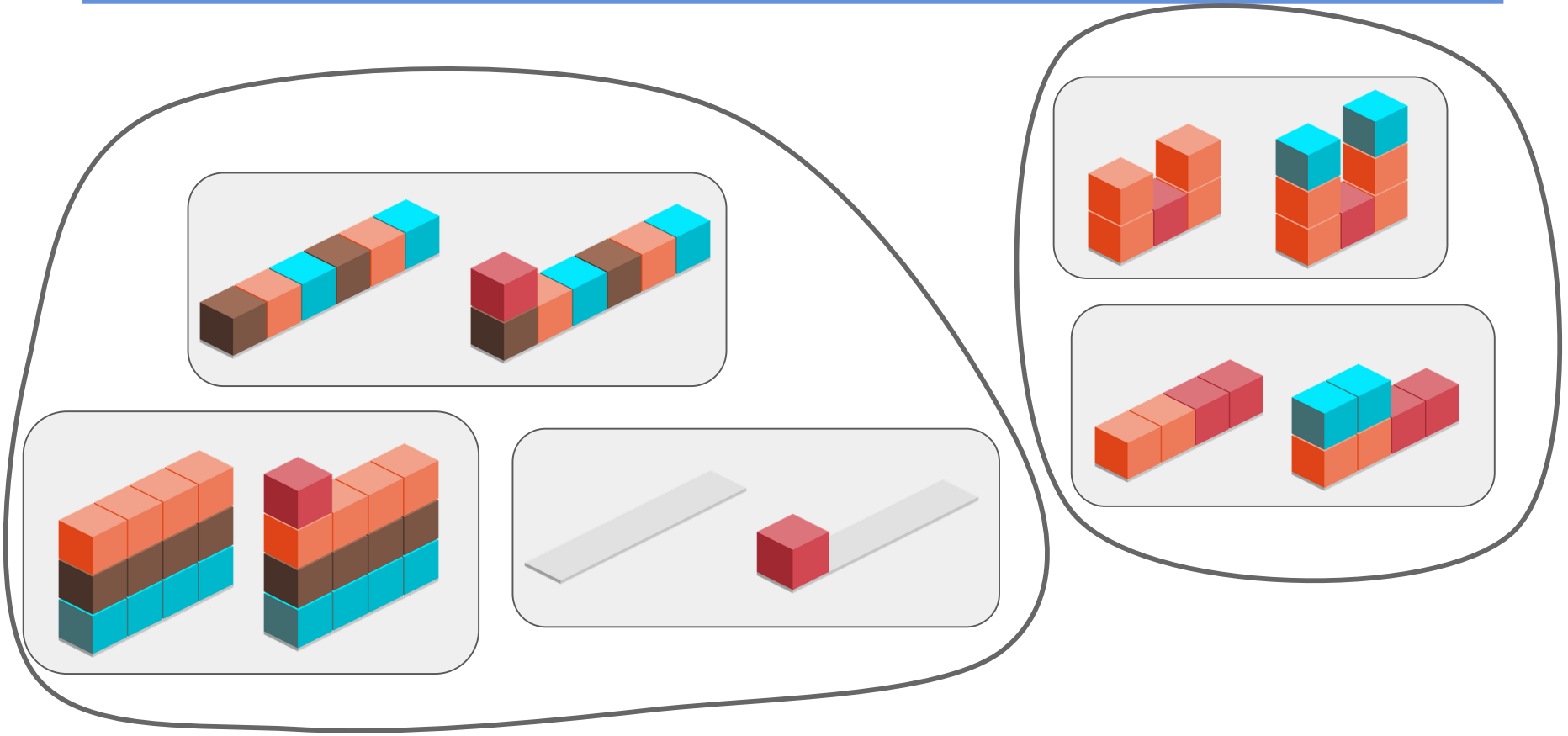
Watching the Environment







Watching the Environment

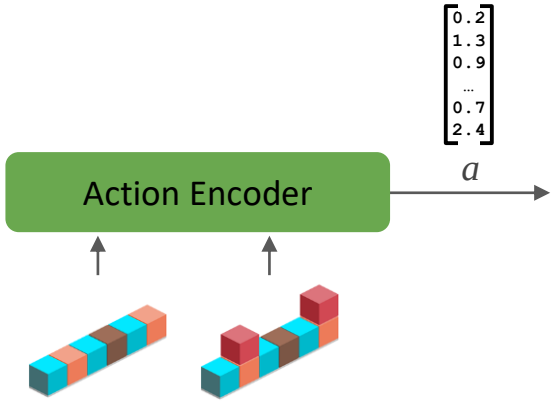


Environment Learning

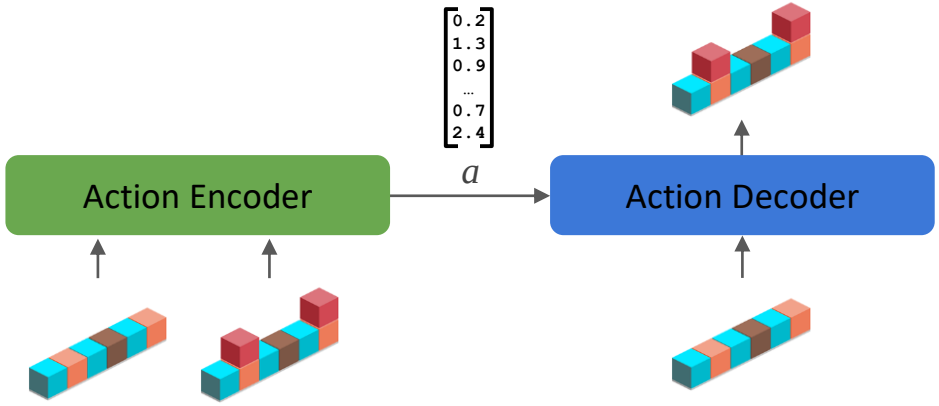
Language Learning

Environment Learning

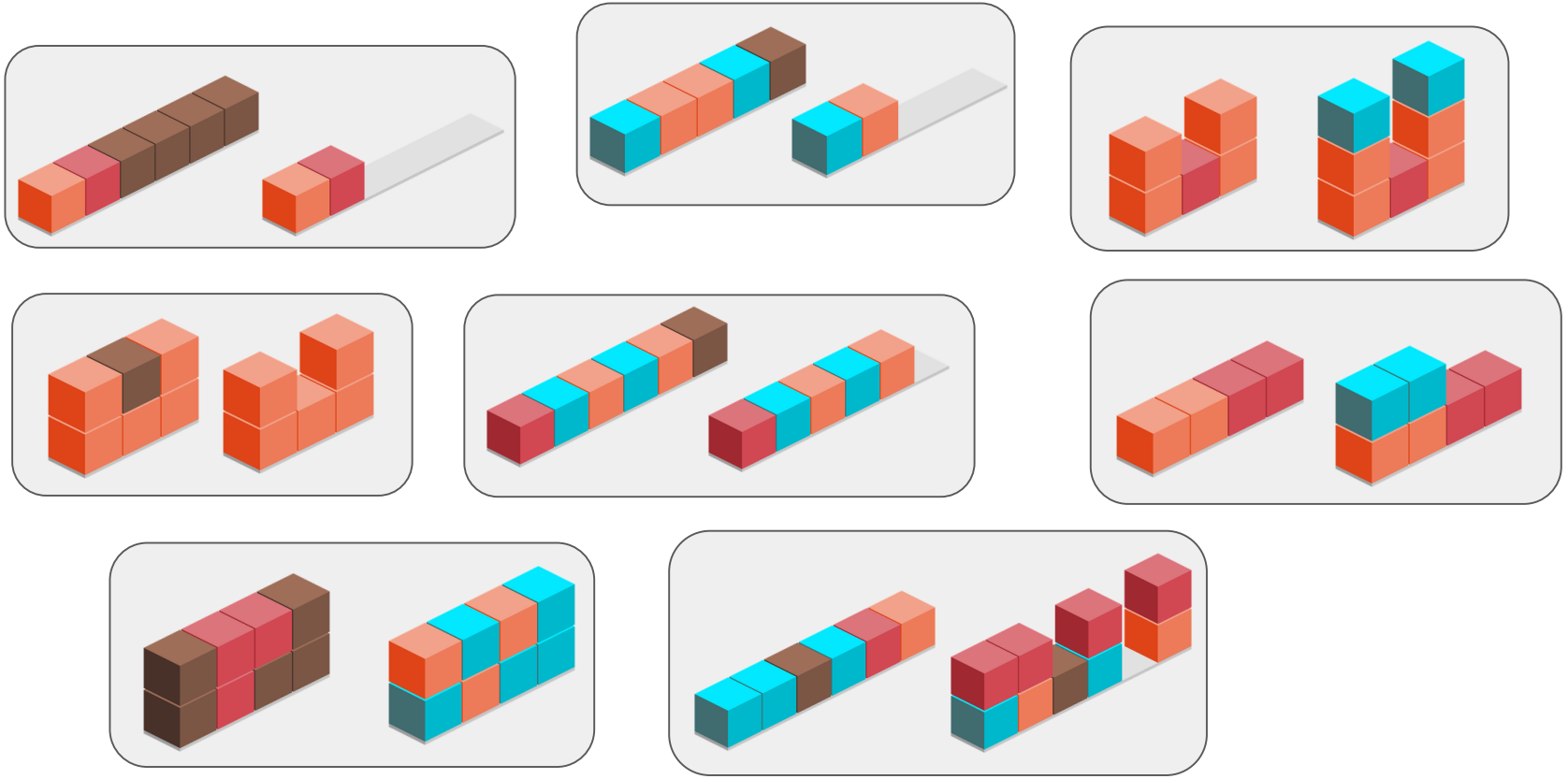
Environment Learning



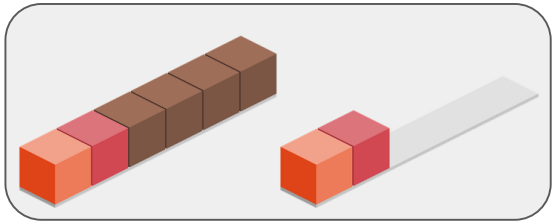
Environment Learning

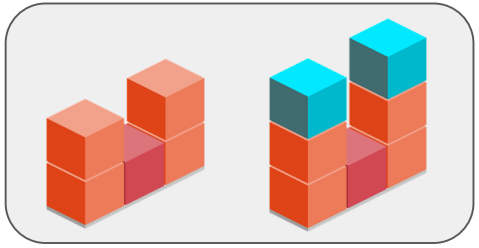


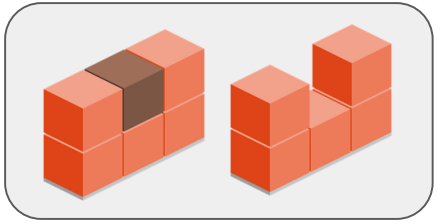
Environment Learning

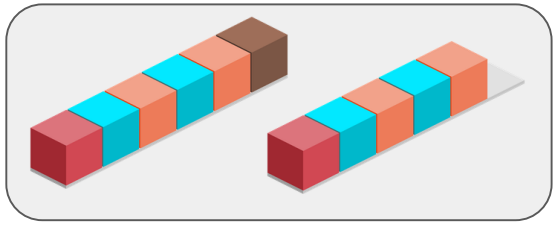


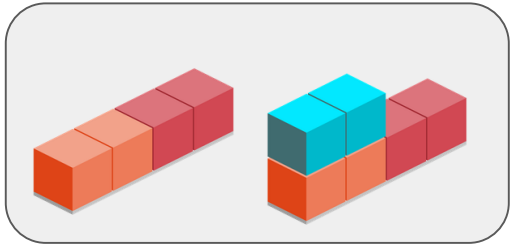
Environment Learning



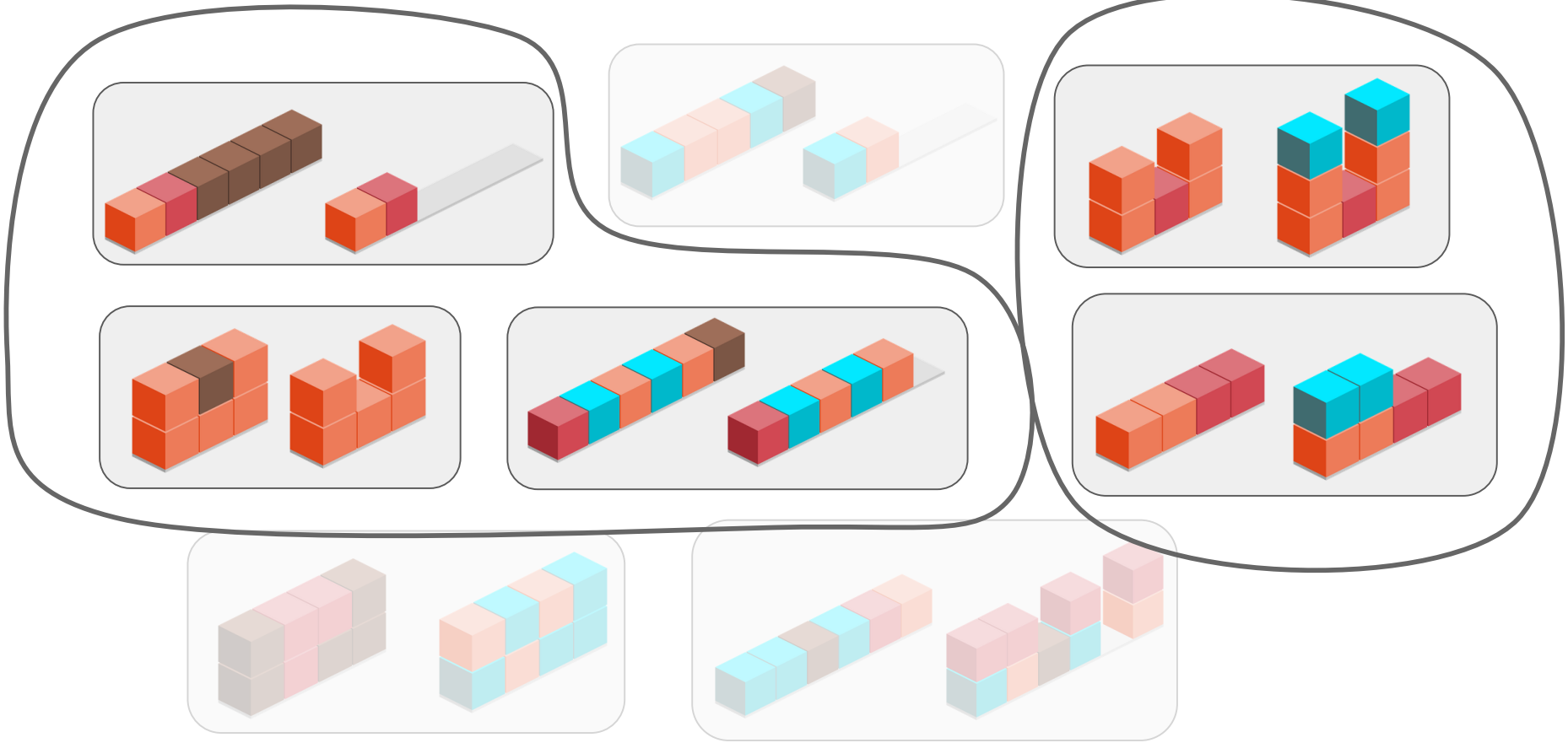








Environment Learning



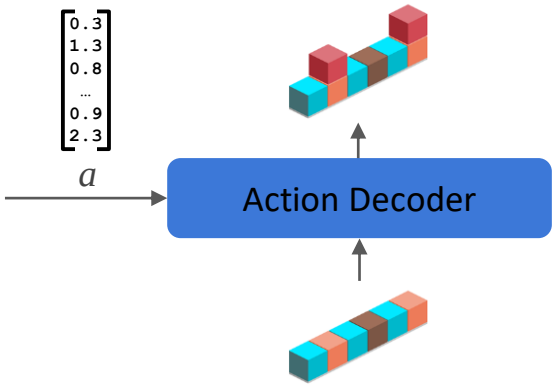
Environment Learning

Language Learning

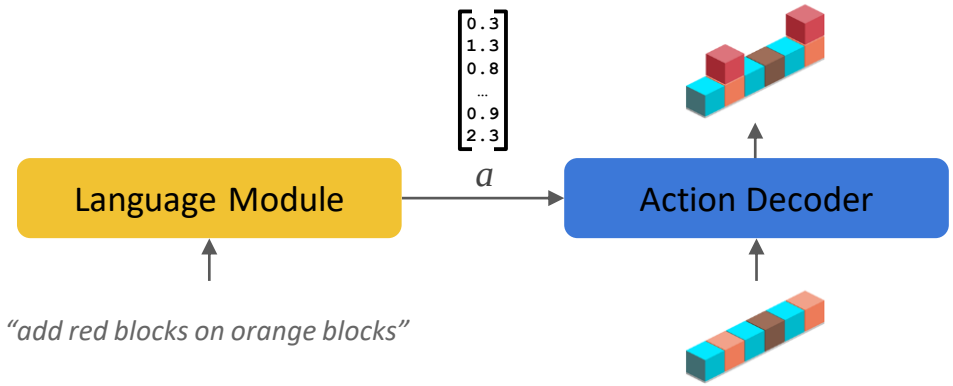
Environment Learning

Language Learning

Language Learning



Language Learning



Environment Learning

Language Learning

Results

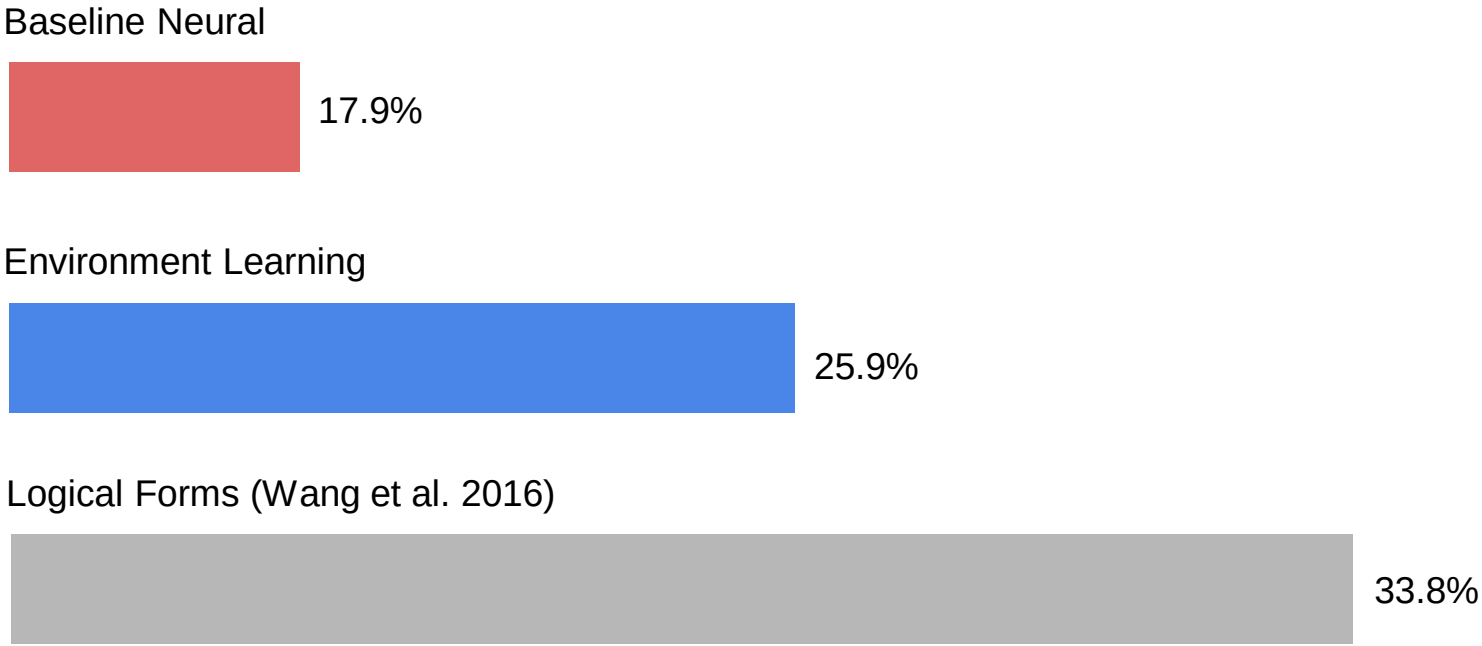
Baseline Neural

17.9%

Logical Forms (Wang et al. 2016)

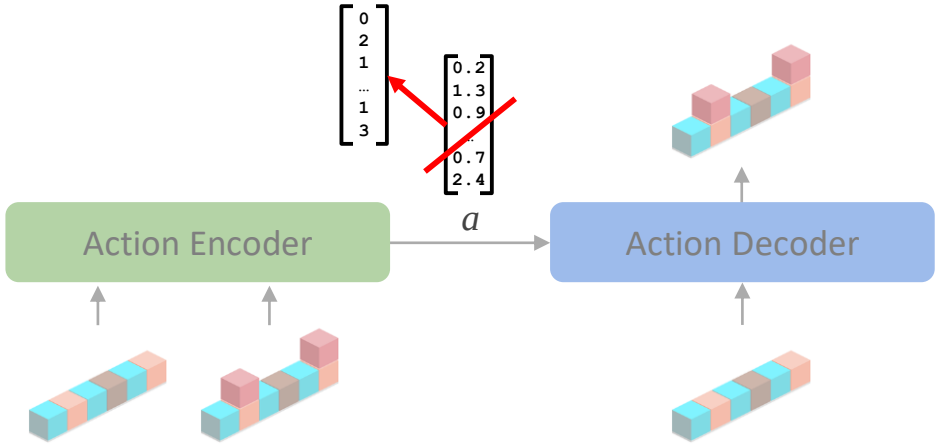
33.8%

Results

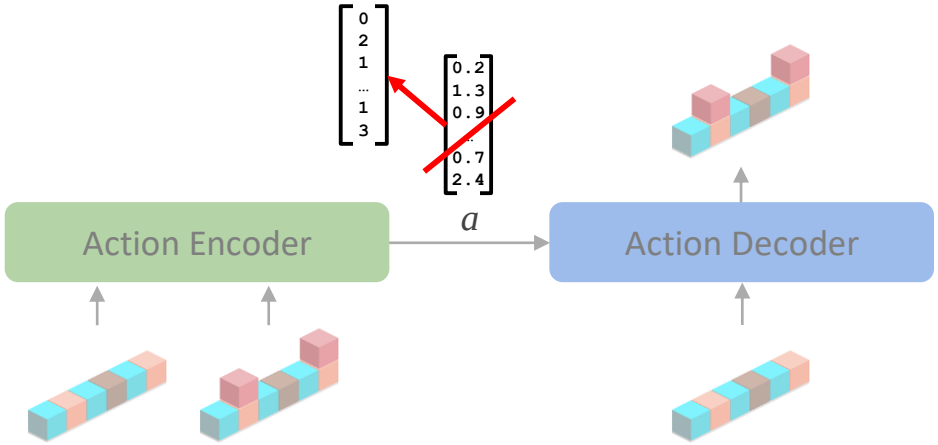

Baseline Neural
17.9%
Environment Learning
25.9%
Logical Forms (Wang et al. 2016)
33.8%

Problem: The semantics we want to learn may be discrete, not continuous

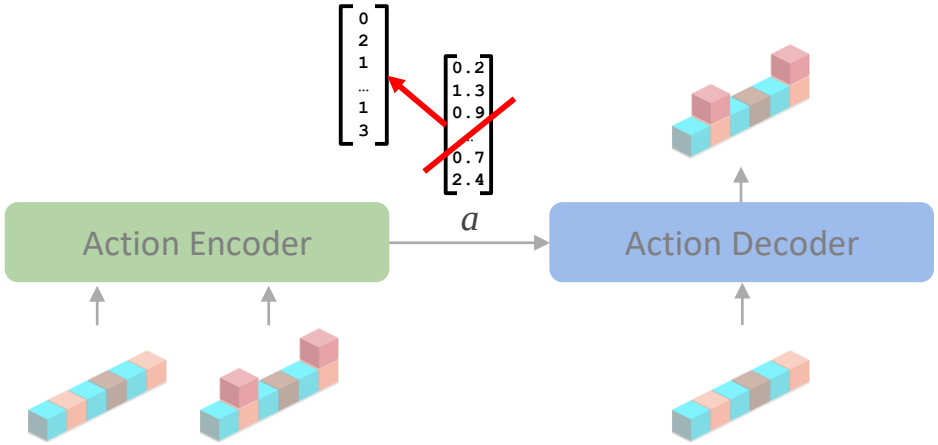
Discrete Representations



Discrete Representations



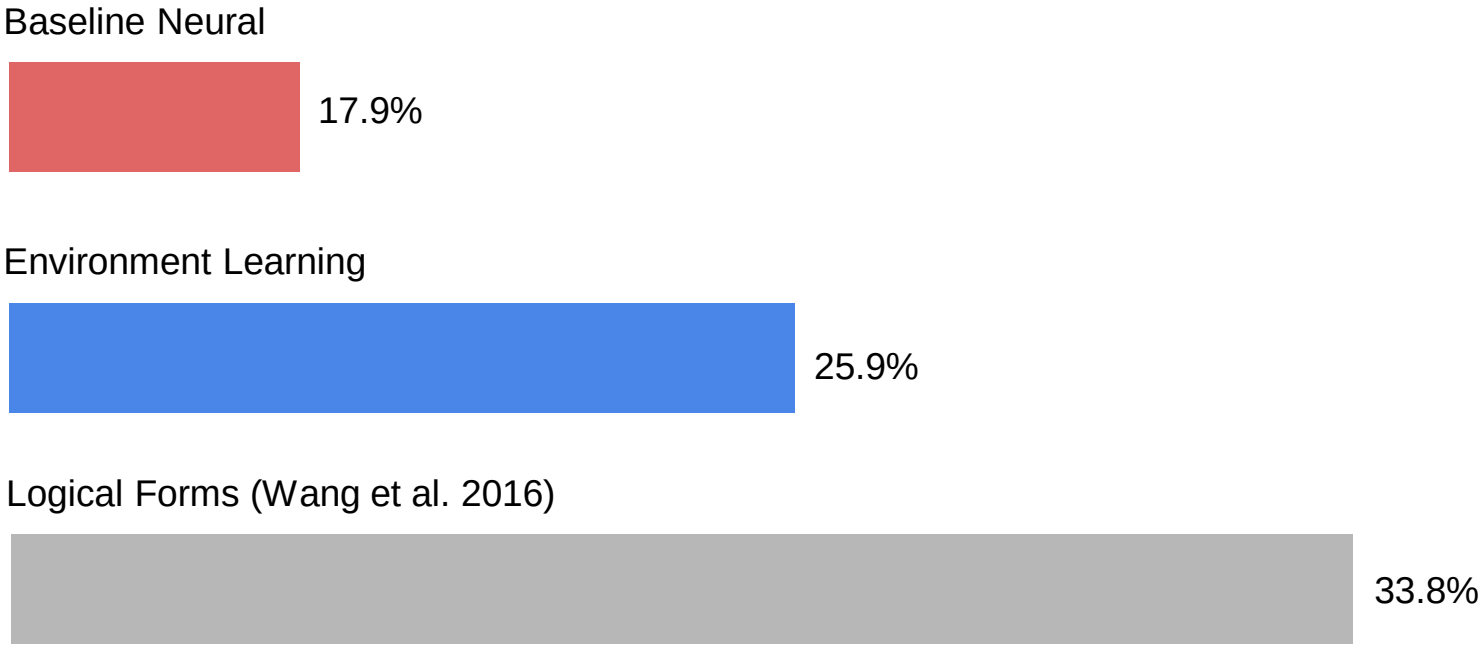
Discrete Representations



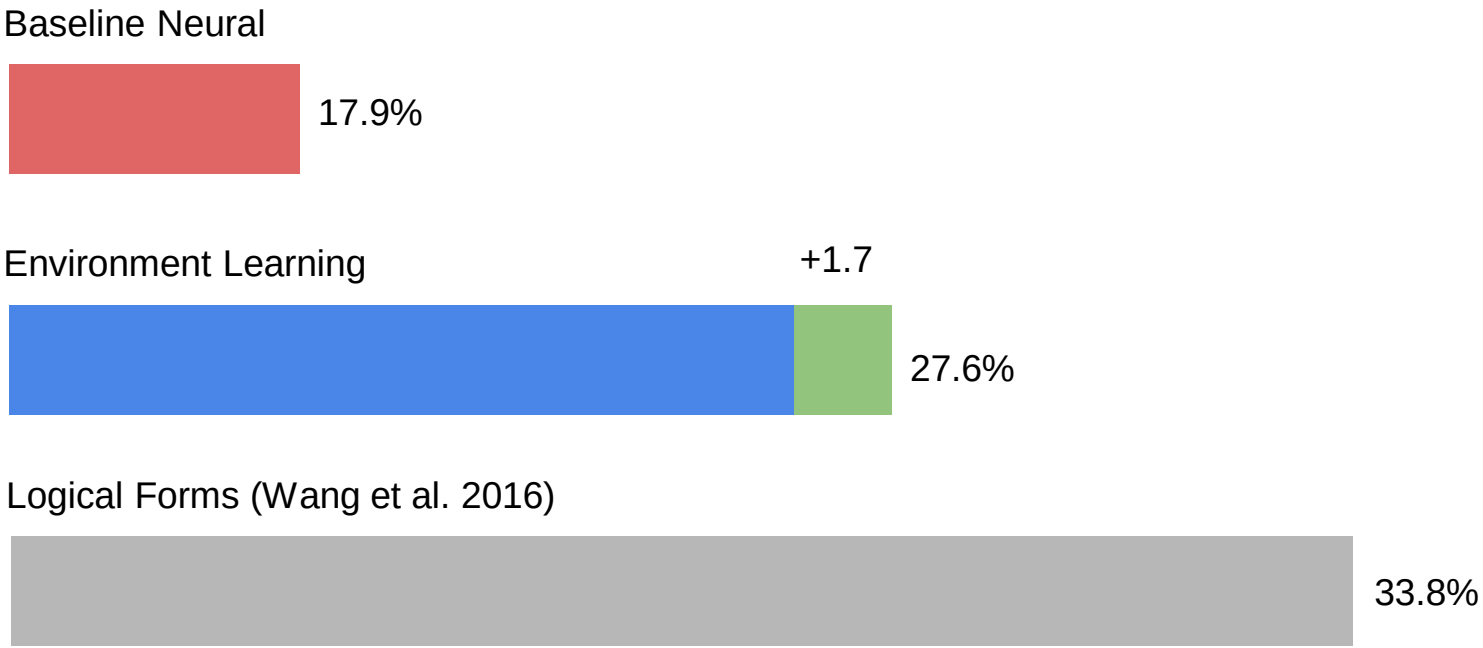
Gumbel Softmax

$$G(x_i) = \frac{\exp(x_i + \epsilon_i)}{\sum_{j=0}^{k} \exp(x_j + \epsilon_j)}$$

Results

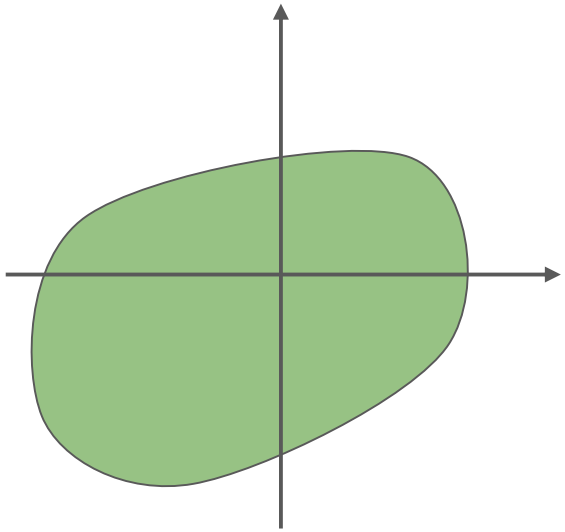

Baseline Neural
17.9%
Environment Learning
25.9%
Logical Forms (Wang et al. 2016)
33.8%

Results


Baseline Neural
17.9%
Environment Learning
+1.7
27.6%
Logical Forms (Wang et al. 2016)
33.8%

Problem: What happens if the language encoder uses a different part of the space than the autoencoder?

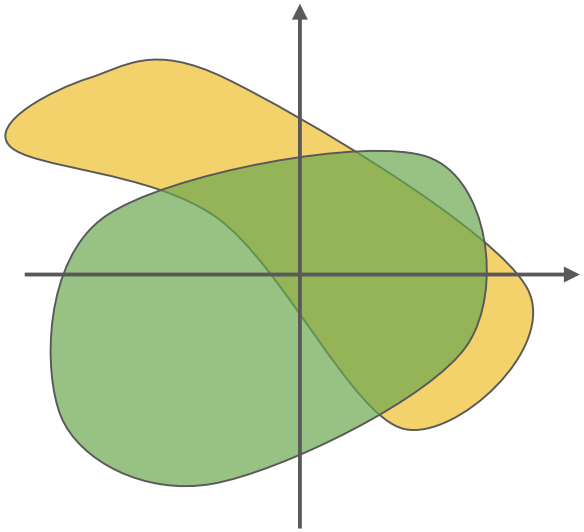
Encoder Matching



Pre-training encoder representations

Encoder Matching

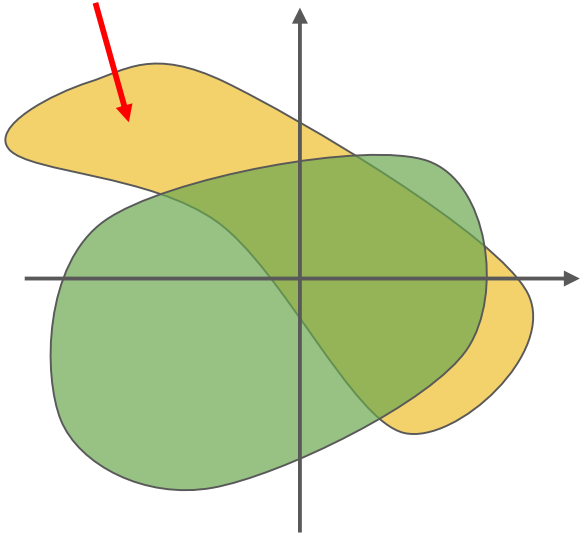
Language encoder representations



Pre-training encoder representations

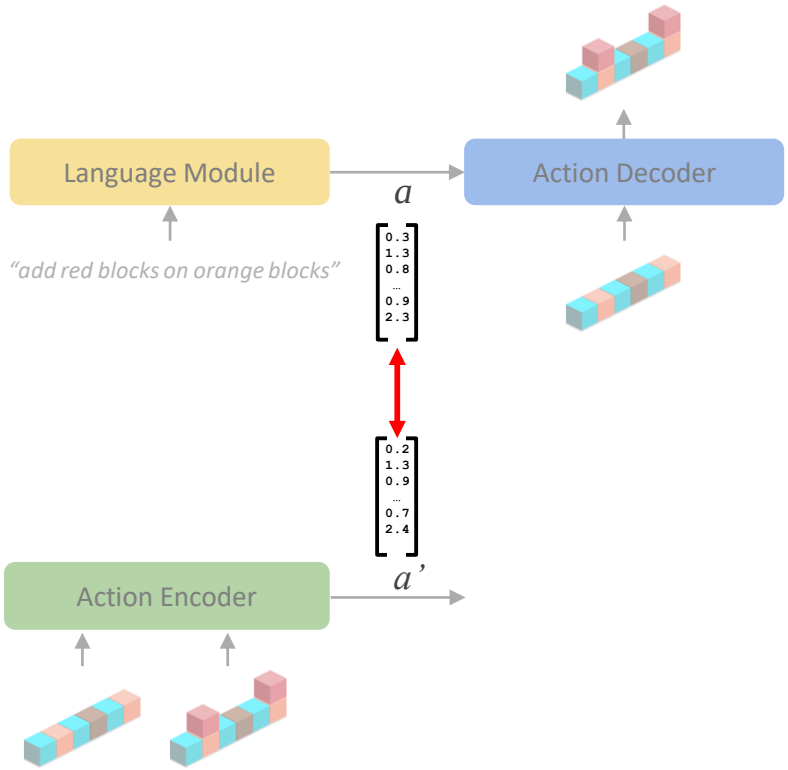
Encoder Matching

Language encoder representations

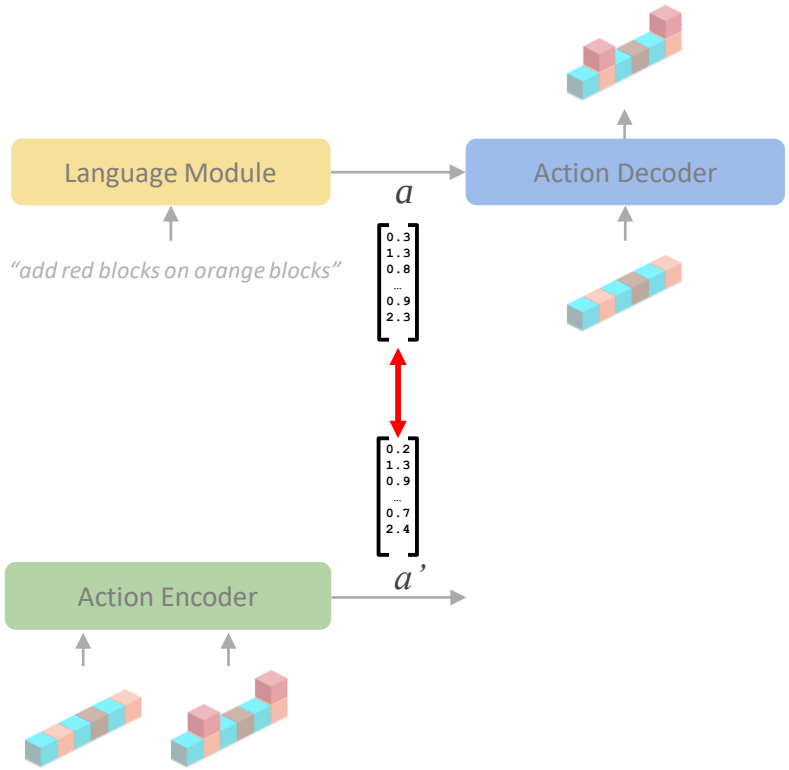


Pre-training encoder representations

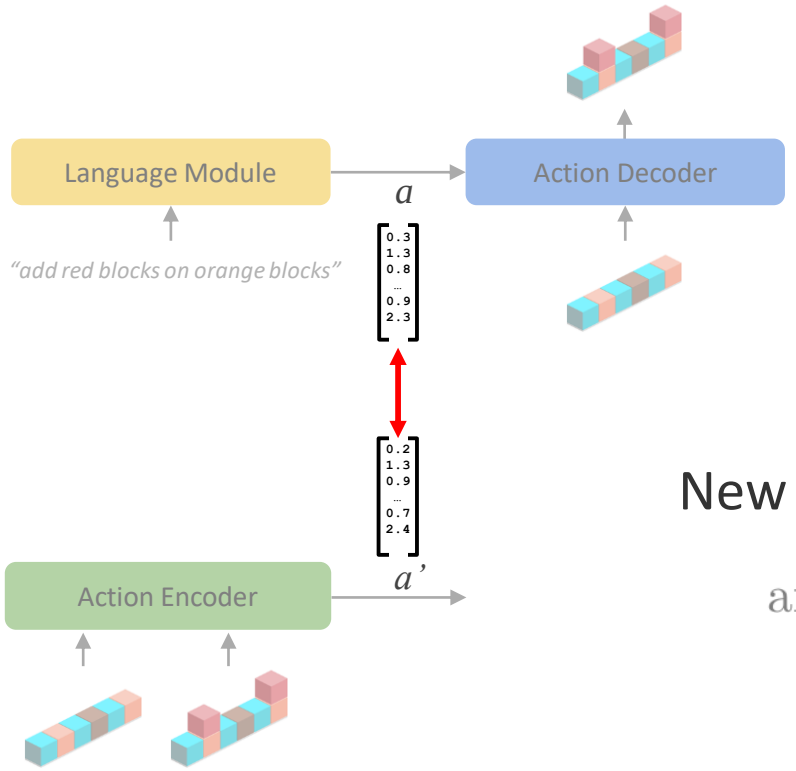
Encoder Matching



Encoder Matching



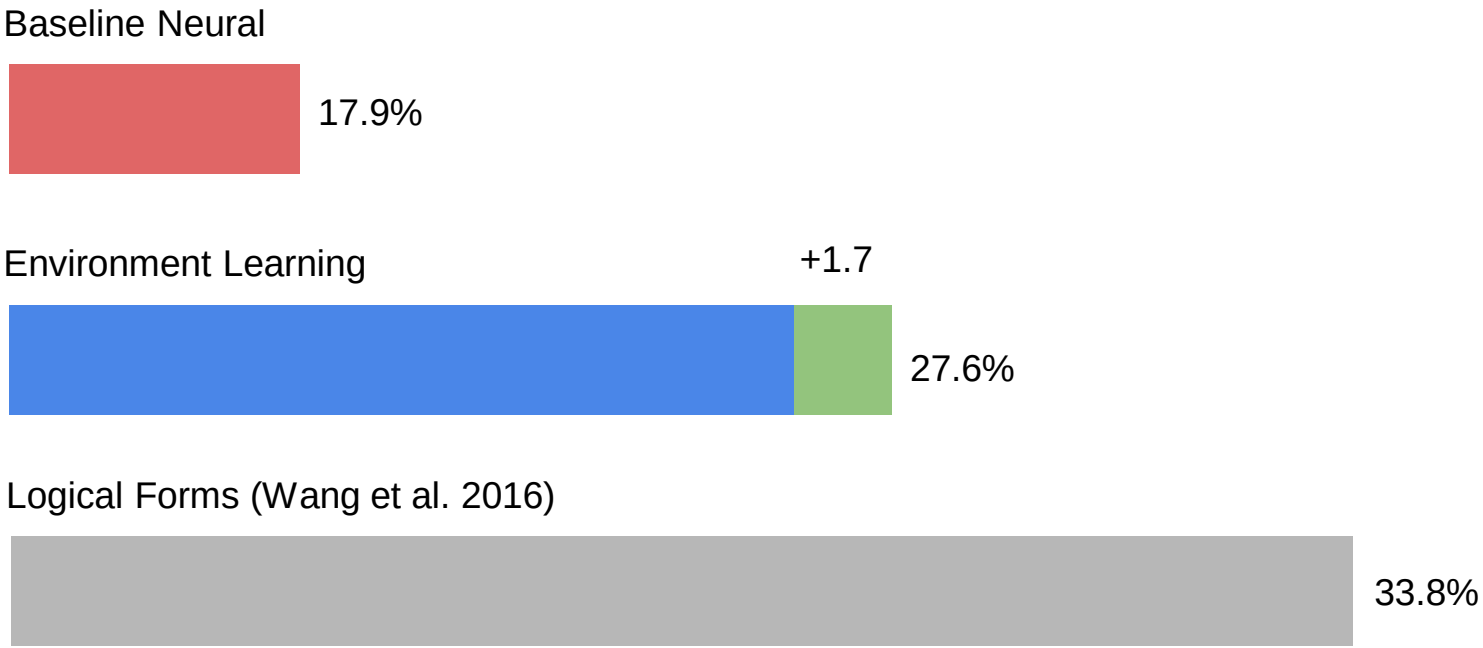
Encoder Matching



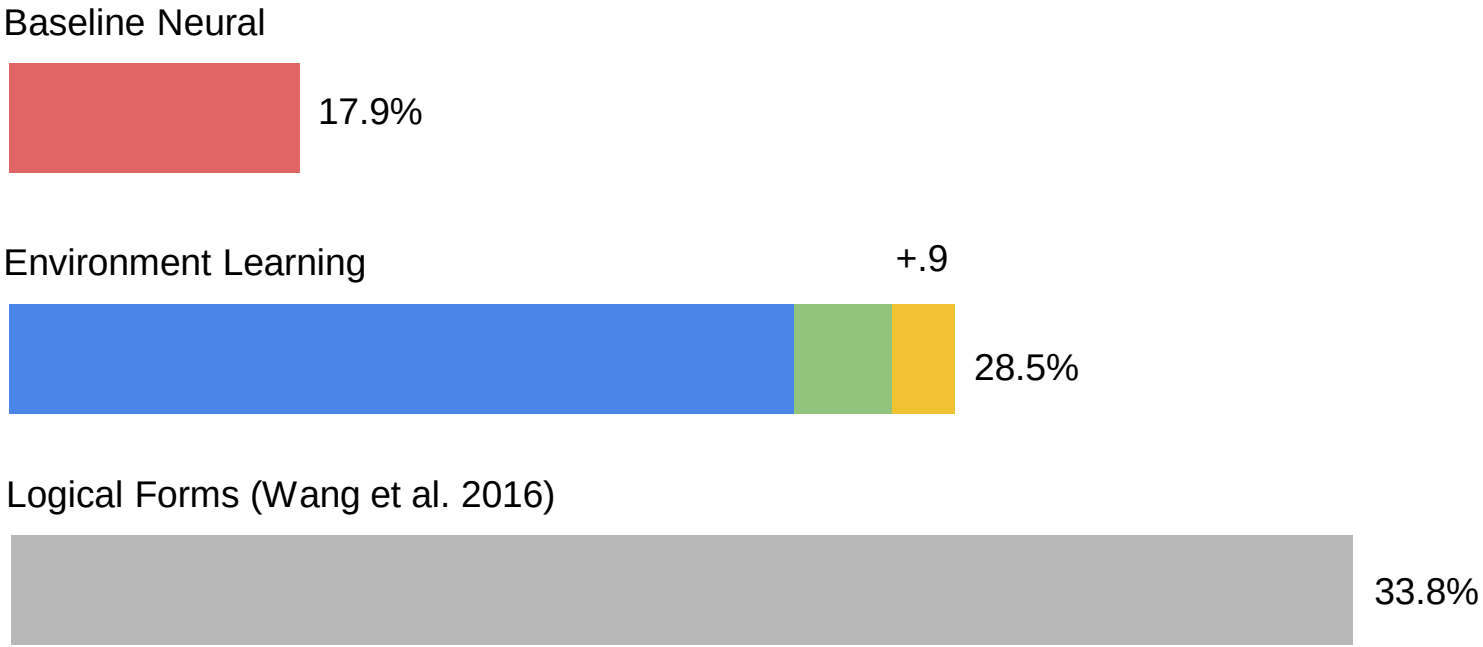
New objective during language learning:

$$\arg\max_{\theta_L} \left[\log P_D(s'|s, L(c)) + \lambda \log P_L(a_E|c)\right]$$

Results


Baseline Neural
17.9%
Environment Learning
+1.7
27.6%
Logical Forms (Wang et al. 2016)
33.8%

Results


Baseline Neural
17.9%
Environment Learning
+.9
28.5%
Logical Forms (Wang et al. 2016)
33.8%

String Manipulation Task

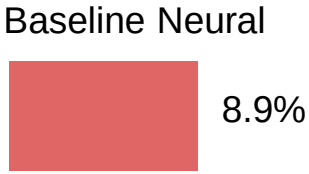
c	replace consonants with p x
s	`fines`
s'	`pxipxepx`

c	add a letter k before every b
s	`rabbles`
s'	`rakbkbles`

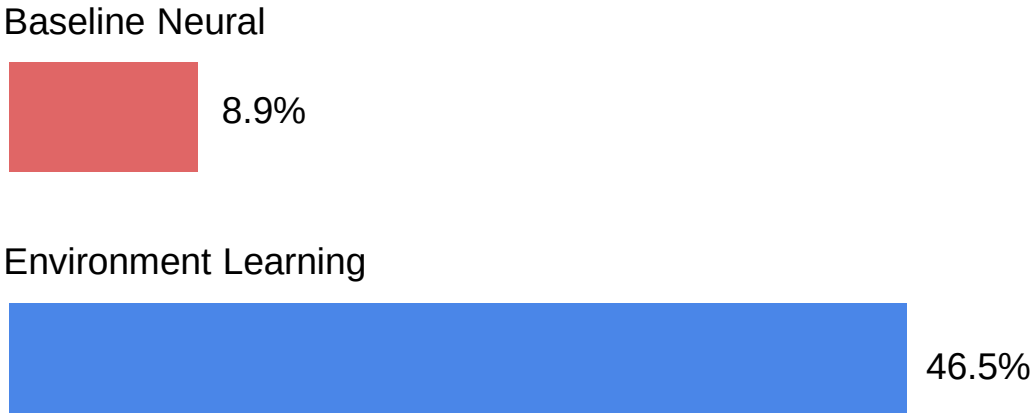
c	replace vowel consonant pairing with v g
s	`thatched`
s'	`thvgchvg`

c	add b for the third letter
s	`thanks`
s'	`thbanks`

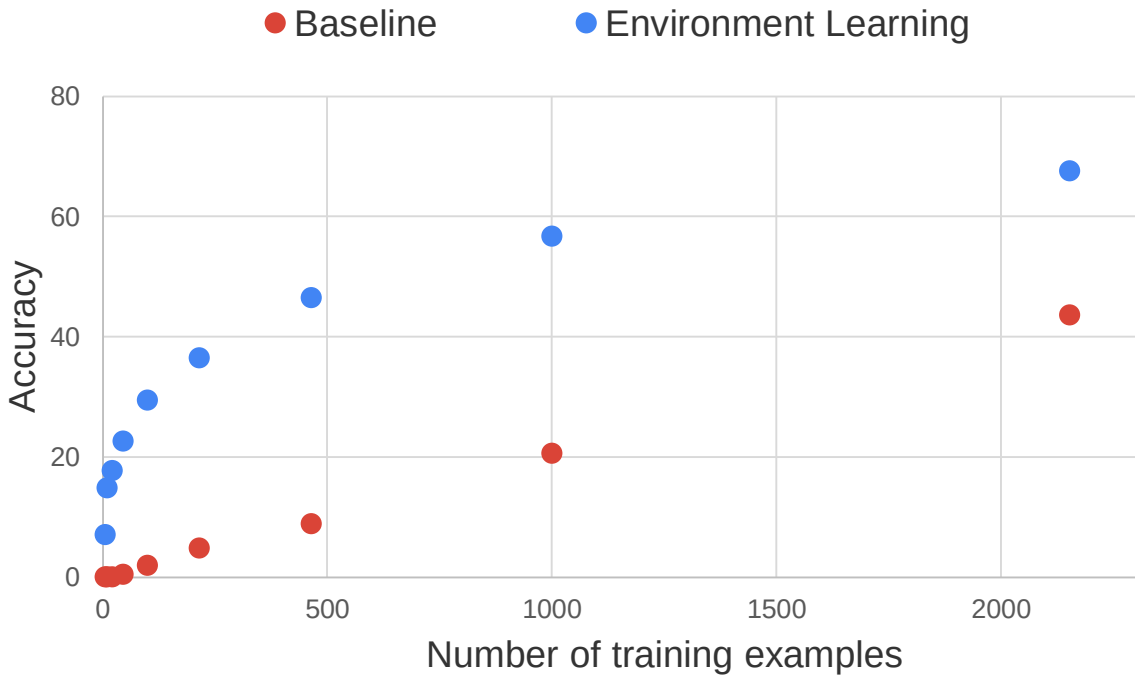
Results - String Manipulation



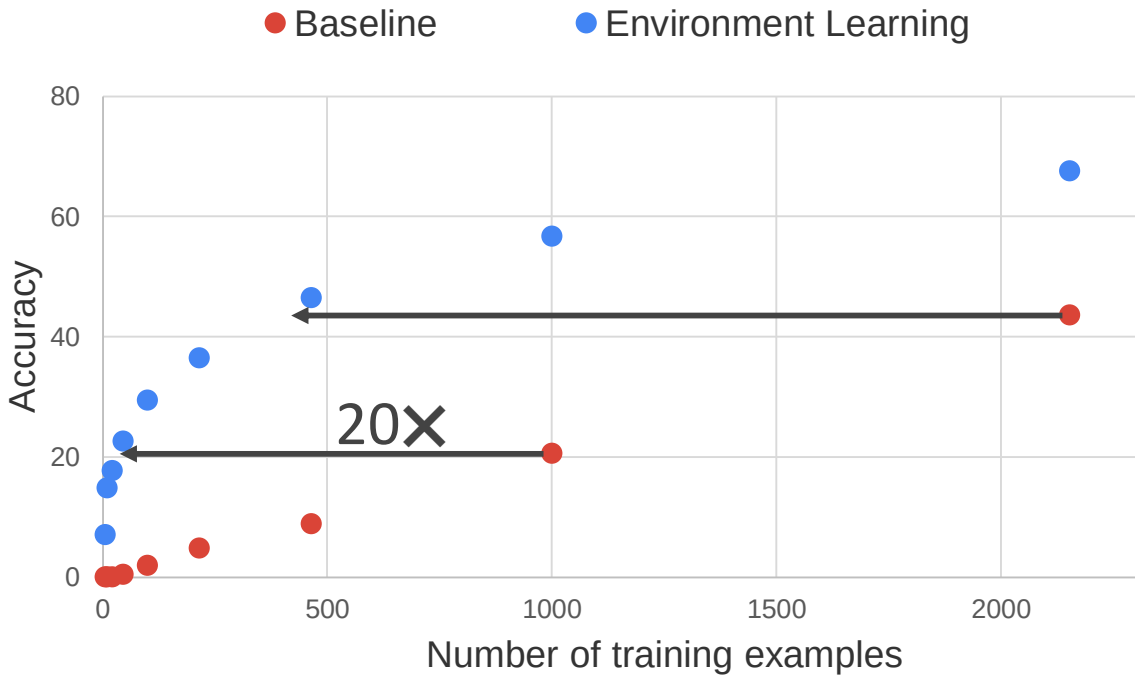
Results - String Manipulation



Results - String Manipulation

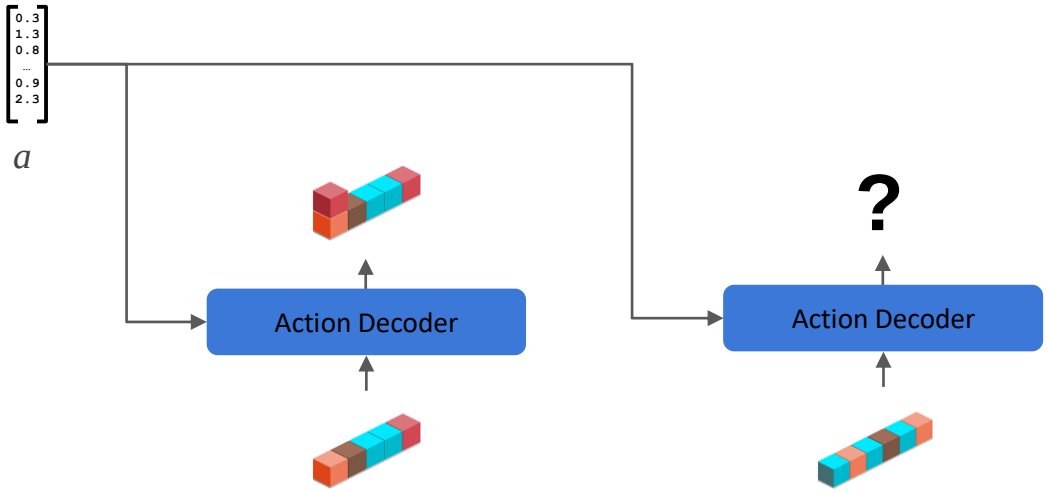


Results - String Manipulation

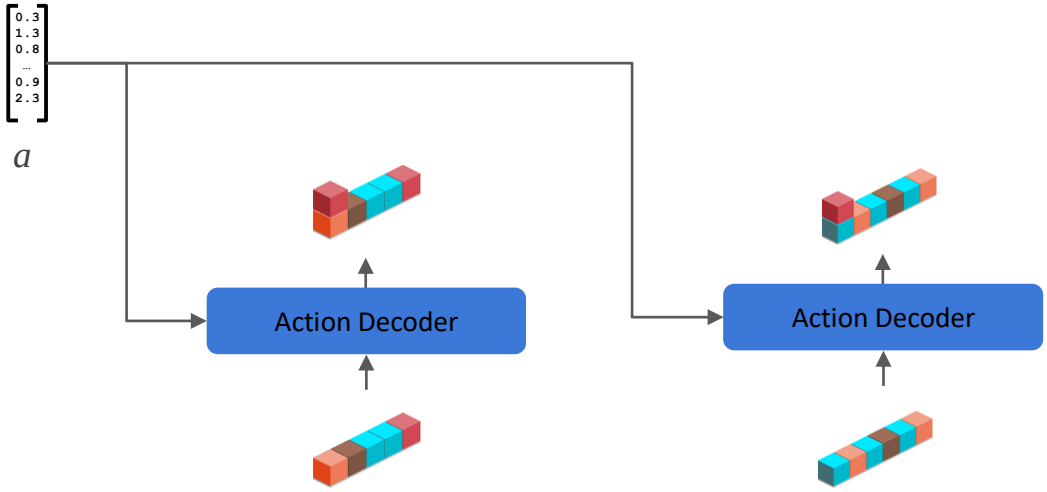


Question: Do our representations behave like logical forms?

Analysis

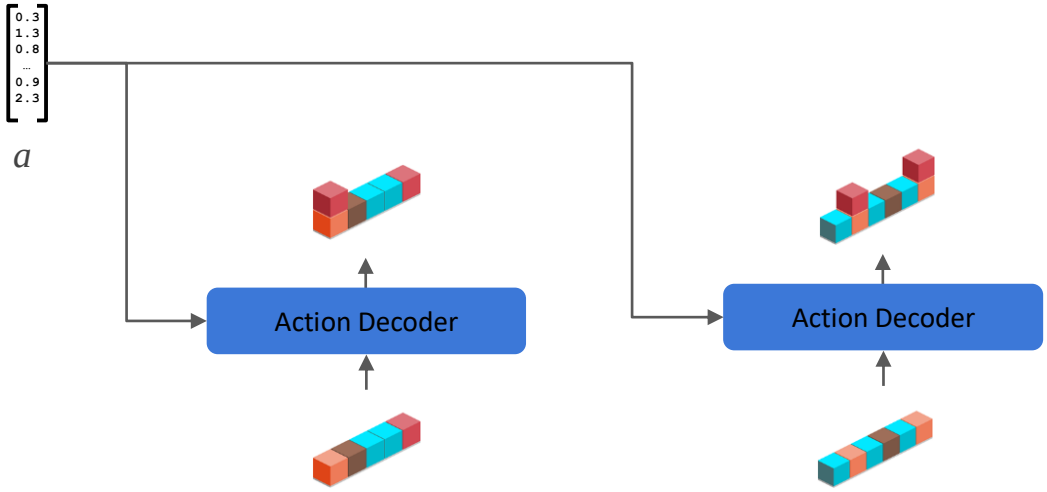


Analysis



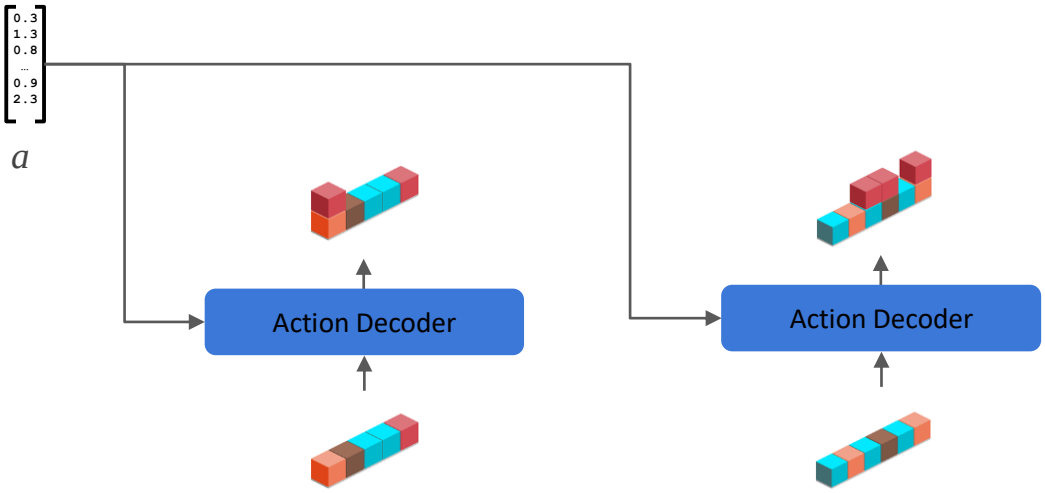
`stack(red, leftmost)`

Analysis



`stack(red, top_color(orange))`

Analysis



Analysis

✓ 84% consistent with logical form

✗ 16%

Conclusions

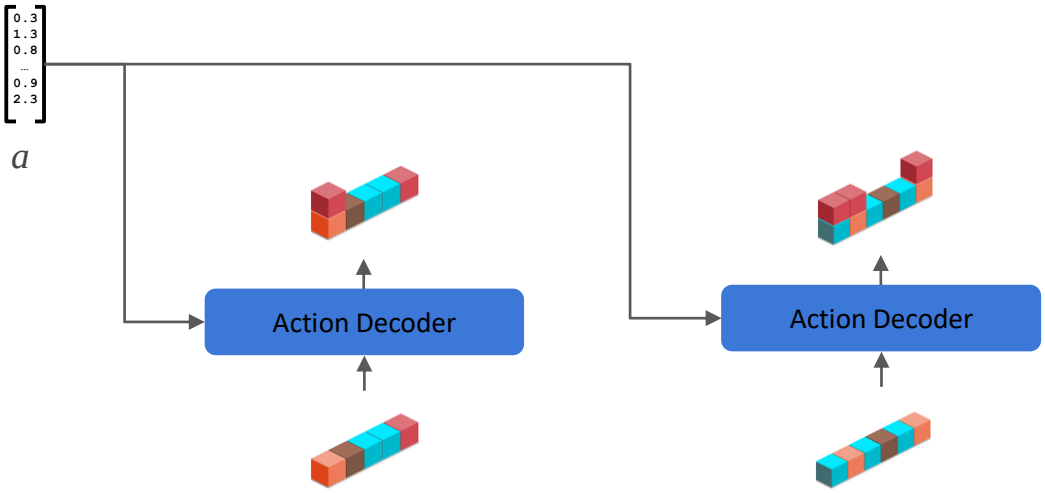
Neural models struggle from lack of inductive bias

It is possible to learn good representations with unsupervised observation

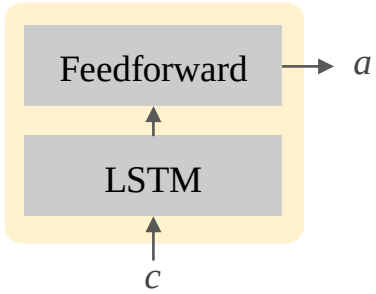
Mapping to pre-learned representations makes instruction following more data efficient

Thanks!

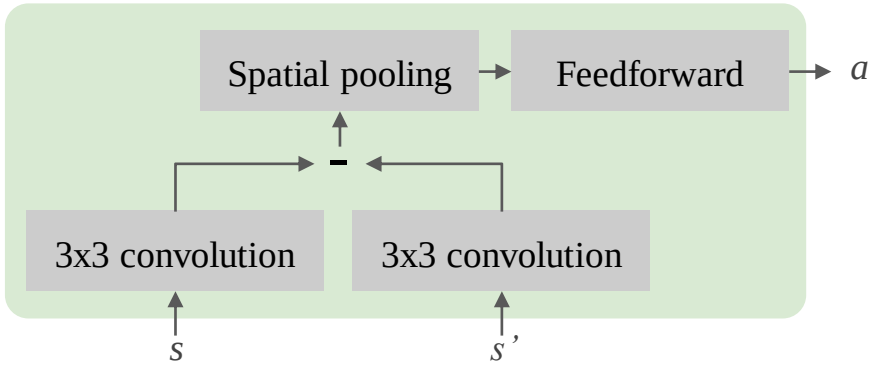
Analysis

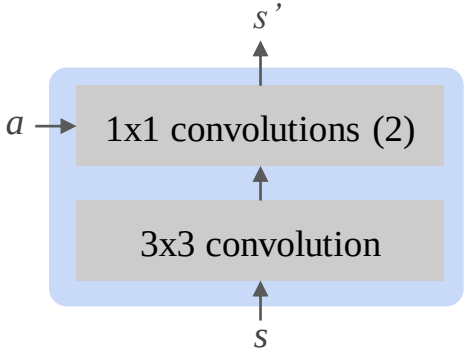


Language Module

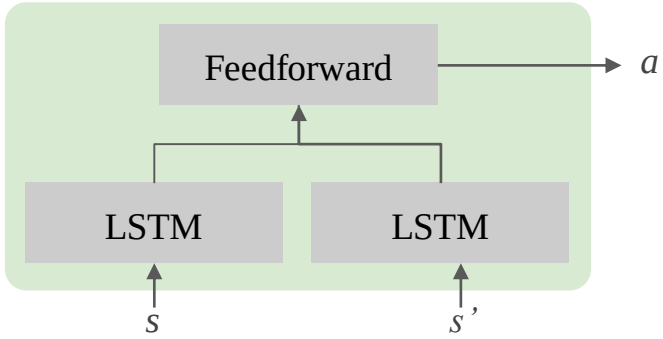


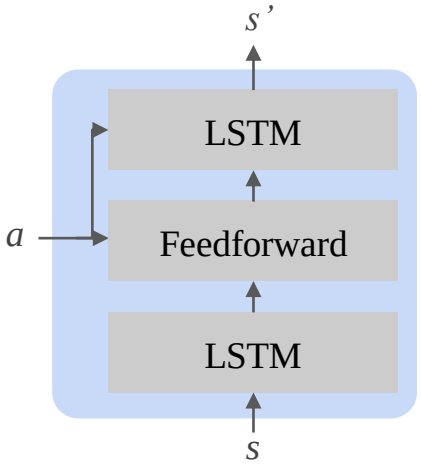
Block Stacking Modules





String Manipulation Modules





Transition Examples