SmashCuts: Random Compositional Data Augmentation for Instruction Following

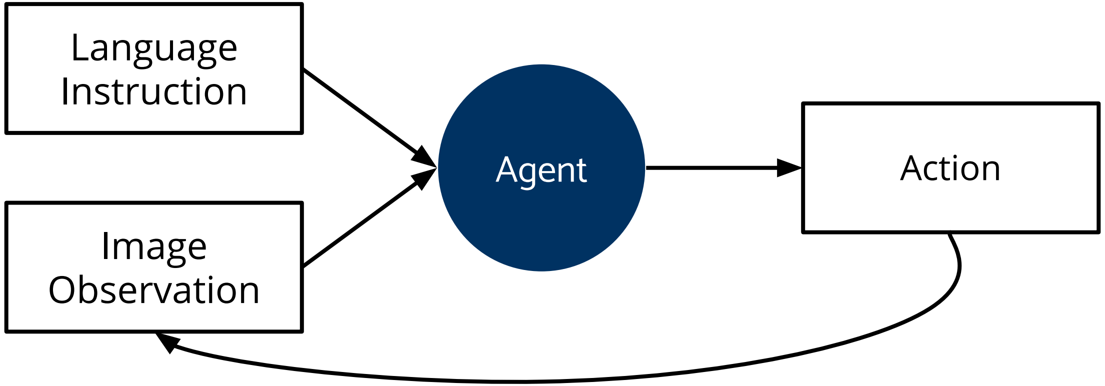
Anjali Kantharuban, Rudy Corona, Coline Devin, Dan Klein, and Trevor Darrell

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

[Anderson et al. 2017]

Visual Language Navigation



- Language instructions provide high-level information on what goals are
- Image observations provide low-level environment information how goals can be achieved

[Chevalier-Boisvert et al. 2019]

Sample Complexity in Imitation Learning

- Imitation learning has a high sample complexity
- Even after training, no guarantee that it will generalize to novel tasks that are composed of previously seen components

  “Pick up the blue ball” + “Go to the door” → “Pick up the blue ball and go the door”

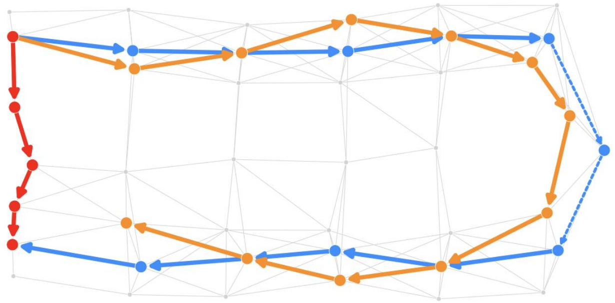
- Increasingly difficult to generate/annotate expert demonstrations as tasks get more complex
- Humans are much better at this generalization [Lake et al. 2019]

[Peng et al. 2018]

[Jain et al. 2019]

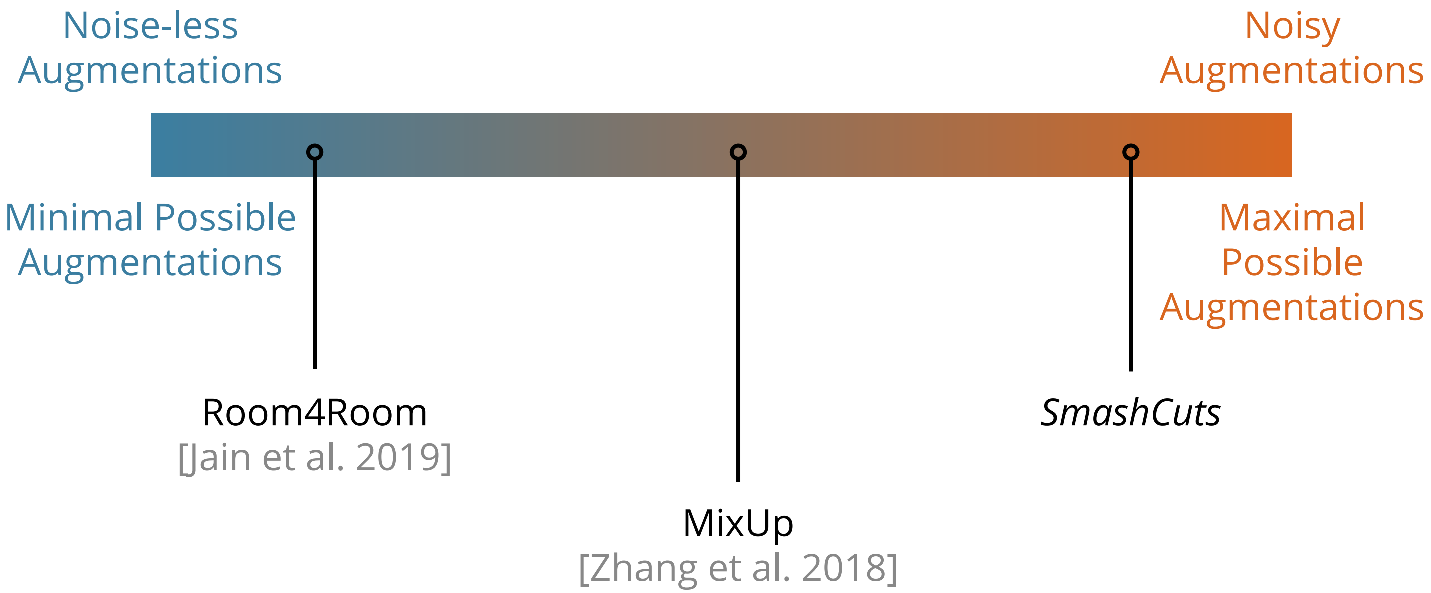
Stitching Together Sequences

- Taking short (atomic) sequences and stitching them together would give expert demonstrations for longer sequences
- When it comes to language conditioned instruction following, we can combine atomic sequences using conjunctions (“and”, “then”, etc.)
- Room4Room improves navigation using this method
- Current methods require the stitched together sequences to “line up” so that the transition is smooth



What’s worse: less data or noisy data?

Exploring the Size-Noise Tradeoff



Noise-less Augmentations

Noisy Augmentations

Minimal Possible Augmentations

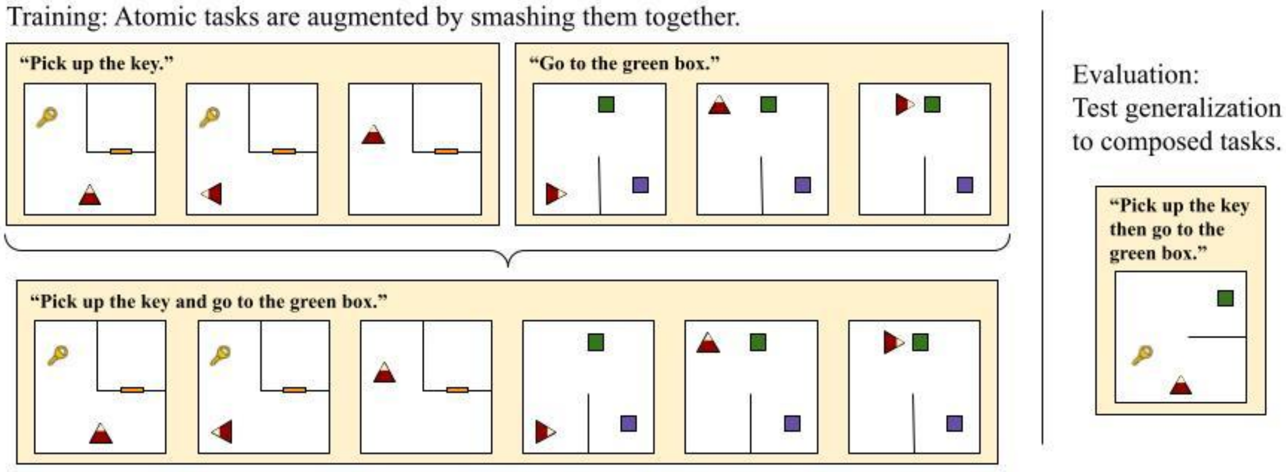
Maximal Possible Augmentations

Room4Room
[Jain et al. 2019]

MixUp
[Zhang et al. 2018]

SmashCuts

SmashCuts for Instruction Following

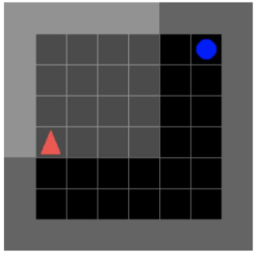

Training: Atomic tasks are augmented by smashing them together.
"Pick up the key."
"Go to the green box."
"Pick up the key and go to the green box."
Evaluation:
Test generalization
to composed tasks.
"Pick up the key then go to the green box."

We augment instruction following demonstrations by concatenating multiple samples together. While these new trajectories are often physically infeasible (because the setting or object configurations abruptly change partway through), we find this augmentation strategy significantly improves model generalization to more complex tasks without extra environment information.

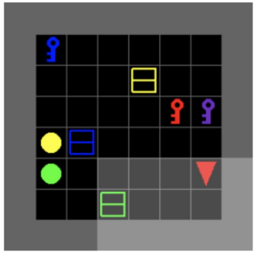
[Chevalier-Boisvert et al. 2019]

BabyAI

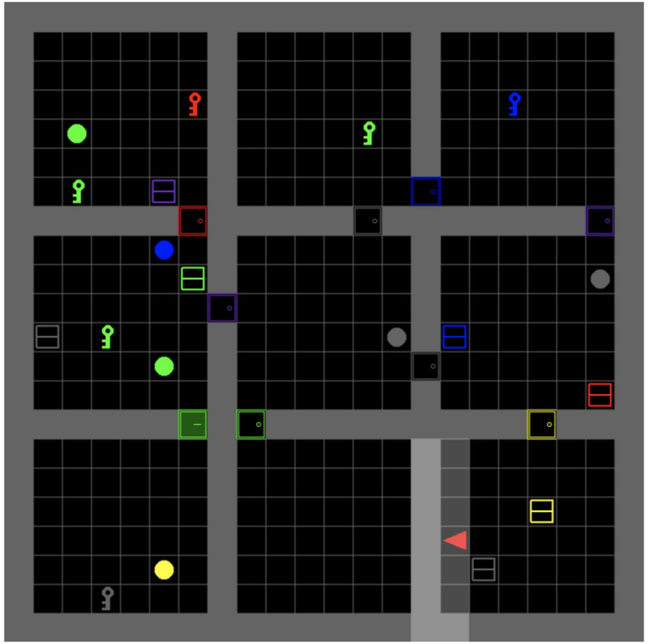
- Platform for testing sample efficiency for RL/IL in a language conditioned setting
- Partially observed (7x7 view distance)
- “Images” preprocessed feature map
- Language is synthetic



(a) GoToObj: "go to the blue ball"



(b) PutNextLocal: "put the blue key next to the green ball"



(c) BossLevel: "pick up the grey box behind you, then go to the grey key and open a door". Note that the green door near the bottom left needs to be unlocked with a green key, but this is not explicitly stated in the instruction.

BabyAI

- Currently training on non compositional tasks to learn as many competencies as possible
- Compositional tasks are tasks in which two subgoals are present, combined with “and,” “then,” or “after you”
- Evaluating on compositional tasks to evaluate generalization in agent

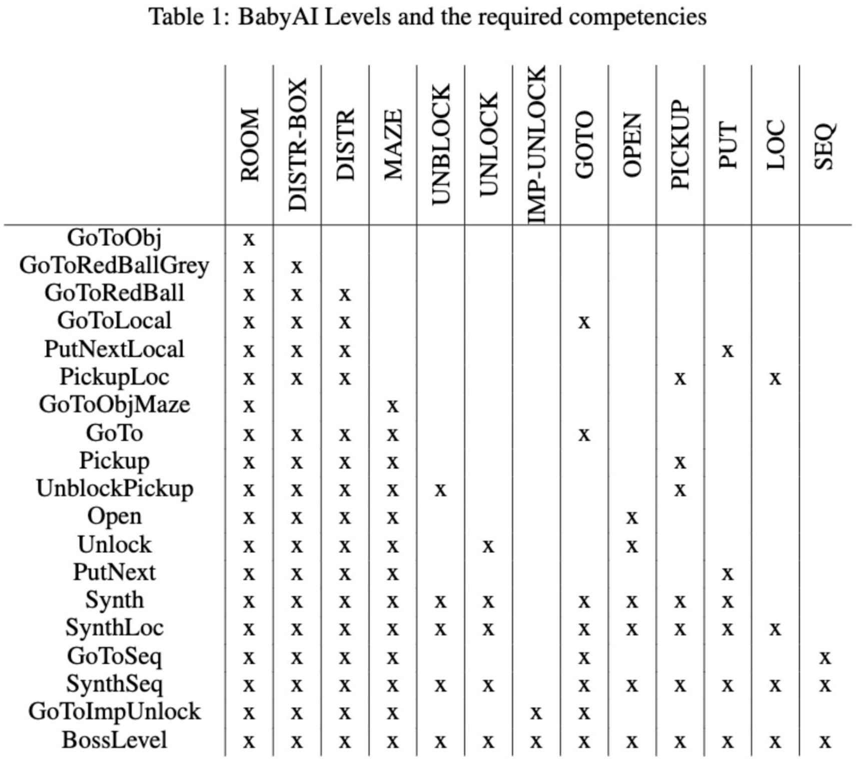


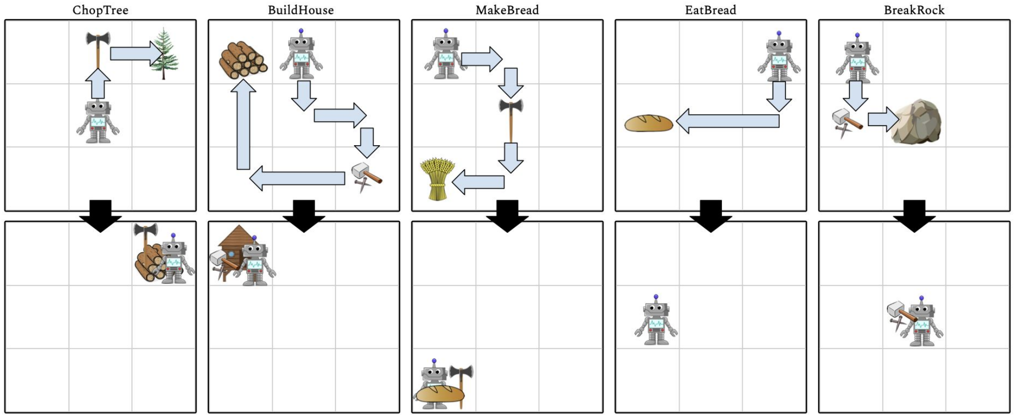
Table 1: BabyAI Levels and the required competencies

	ROOM	DISTR-BOX	DISTR	MAZE	UNBLOCK	UNLOCK	IMP-UNLOCK	GOTO	OPEN	PICKUP	PUT	LOC	SEQ
GoToObj	x												
GoToRedBallGrey	x	x											
GoToRedBall	x	x	x										
GoToLocal	x	x	x					x					
PutNextLocal	x	x	x								x		
PickupLoc	x	x	x							x		x	
GoToObjMaze	x			x									
GoTo	x	x	x	x				x					
Pickup	x	x	x	x						x			
UnblockPickup	x	x	x	x	x					x			
Open	x	x	x	x					x				
Unlock	x	x	x	x		x			x				
PutNext	x	x	x	x							x		
Synth	x	x	x	x	x	x		x	x	x	x		
SynthLoc	x	x	x	x	x	x		x	x	x	x	x	
GoToSeq	x	x	x	x				x					x
SynthSeq	x	x	x	x	x	x		x	x	x	x	x	x
GoToImpUnlock	x	x	x	x			x	x					
BossLevel	x	x	x	x	x	x	x	x	x	x	x	x	x

[Devin et al. 2020]

Crafting

- Platform for testing performance on compositional instruction following tasks
- Fully observed
- “Images” preprocessed feature map
- Language is synthetic



Results

	Atomic	**SynthSeq**	**BossLevel**
Baseline (No Augmentation)	99 ± 0.1	52 ± 1	50 ± 0.5
SmashCuts	99 ± 0.6	**64 ± 3**	**62 ± 2**

Average success rate ± the standard deviation over three different seeds. Atomic refers to only one-goal instructions (except for GoToSeq, which allows the baseline to learn conjunctions). Statistically significant differences (p<0.05) are bolded.

- Augmented dataset performs comparably on atomic tasks (good, since we train to convergence)
- Augmented dataset significantly improves performance on the two purely compositional tasks

Results

	Atomic	**SynthLoc**	**NLSynthLoc**
Baseline (No Augmentation)	99 ± 0.1	90 ± 0.5	64 ± 2
SmashCuts	99 ± 0.6	89 ± 1	**71 ± 2**

Average success rate ± the standard deviation over three different seeds. Atomic refers to only one-goal instructions (except for GoToSeq, which allows the baseline to learn conjunctions). Statistically significant differences (p<0.05) are bolded.

- SynthLoc is a non-compositional (atomic) task purposely withheld during training
- Augmented dataset performs comparably on SynthLoc
- Augmented dataset significantly improves performance when SynthLoc instructions are from a natural language dataset
- Natural language instructions include more compositionality than generated language instructions

Results

	1-2 Goals	**3-4 Goals**	**8 Goals**	**16 Goals**
Baseline (No Augmentation)	98 ± 0.1	80 ± 7	52 ± 10	18 ± 6
SmashCuts	98 ± 0.5	93 ± 1	**83 ± 2**	**47 ± 2**

Average success rate ± the standard deviation over three different seeds in Crafting environment. We consider 1-2 subgoals to be atomic, which results in 3-4 subgoals when augmented. Statistically significant differences (p<0.05) are bolded.

- Augmented dataset performs comparably on atomic dataset
- Benefits propagate past the # of subgoals seen in training with the data augmentation