PaLM-E : An Embodied Multimodal Language Model

유런 5기 고급심화팀 박지연

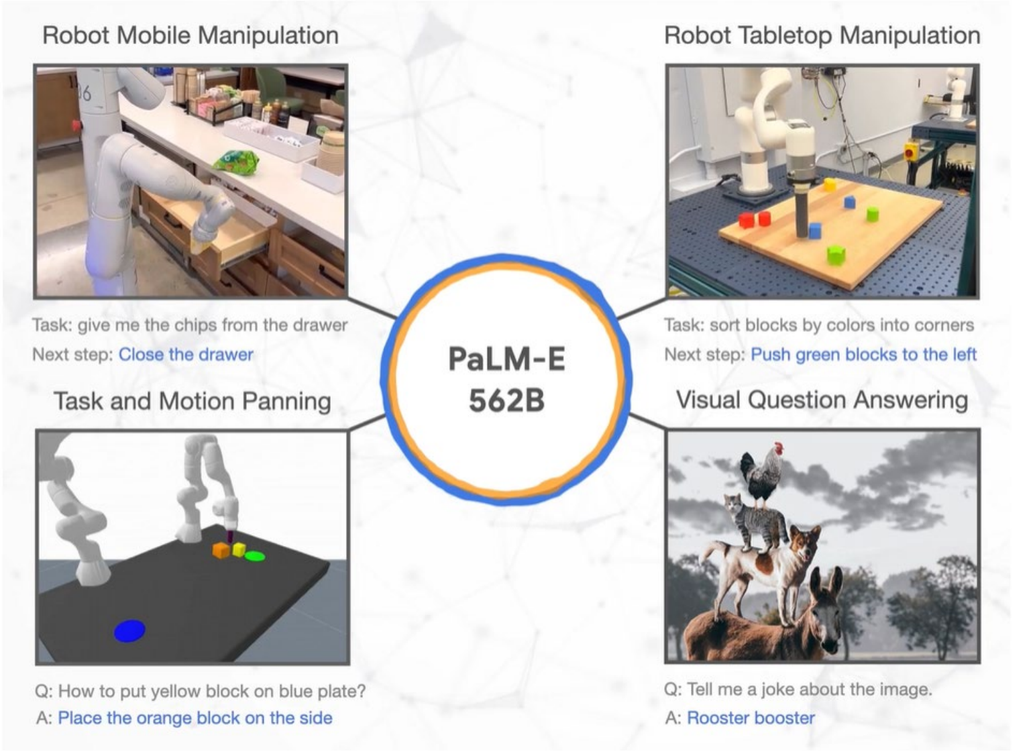
EWHA EURON

#00 Intro

EWHA EURON

#00 Intro


ViT
22B
PaLM
540B


Robot Mobile Manipulation
Task: give me the chips from the drawer
Next step: Close the drawer
Robot Tabletop Manipulation
Task: sort blocks by colors into corners
Next step: Push green blocks to the left
PaLM-E
562B
Task and Motion Panning
Q: How to put yellow block on blue plate?
A: Place the orange block on the side
Visual Question Answering
Q: Tell me a joke about the image.
A: Rooster booster

#00 Intro

1. 임베디드 데이터를 멀티모달 LLM에 혼합해서 학습시켜 범용적 모델, 전이 학습, 다중 구현 의사 결정 에이전트를 교육할 수 있음

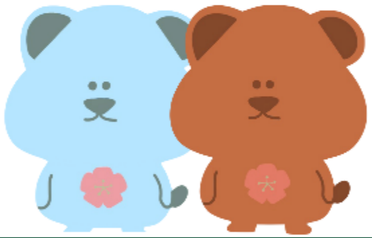
2. 현재 SOTA VLM(vision-language model)은 zero-shot 추론 문제를 잘 다루지 못 함. 하지만 유능한 범용 VLM을 훈련하는 것이 가능함.

3. Neural scene representation, entity-labeling multimodal token 같은 학습 방법에서의 새로운 아키텍쳐를 제안함

4. PaLM-E는 visual 과 language와 같이 다방면에 대해 경쟁적인 모델임

5. 모델의 크기를 늘리는 것이 멀티 모달 파인튜닝에서 catastrophic forgetting 이 더 적어지게 함

#01 Background

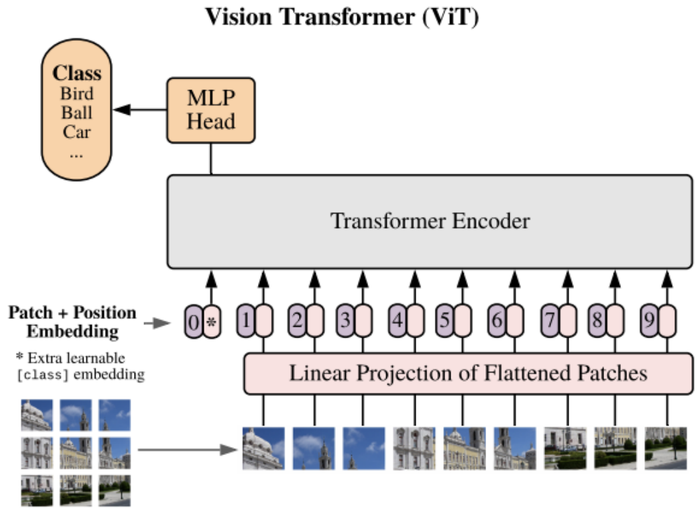


EWHA EURON

#01 Background

PALM(Pathways Language Model)
- 540B 파라미터 (GPT-3는 175B)
- Pathway ML이라는 방법으로 큰 사이즈임에도 좋은 컴퓨팅 효율로 학습을 가능케 함
- Chain-of-thought prompting으로 추론 성능 향상

ViT(Vision Transformer)
- 이미지 분류에 Transformer를 적용시킨 모델
- CNN에 비해 inductive bias가 부족하여 일반화 성능이 떨어지지만, 많은 데이터양으로 극복함
- 많은 데이터로 pretrain 후 작은 데이터로 전이 학습을 할 경우 좋은 성능을 보임

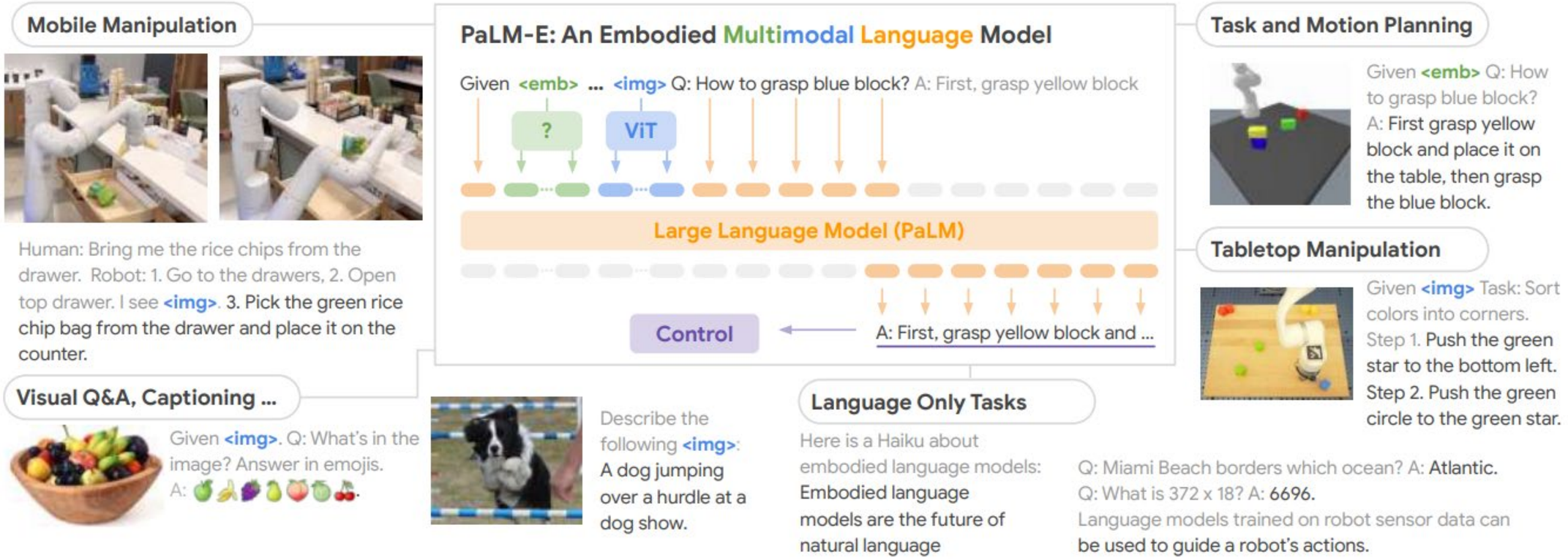


https://jiho-ml.com/weekly-nlp-54/
https://daebaq27.tistory.com/108
https://velog.io/@tobigs-nlp/PaLM-Scaling-Language-Modeling-with-Pathways-1

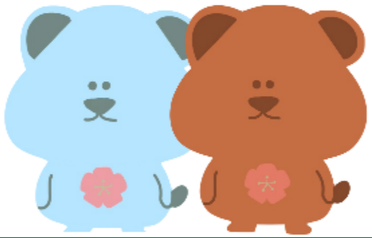
#01 Background

PALM-E

- Embedding된 PaLM로, 디코더 온리 LLM
- 입력 : multimodal sentence, 출력 : 모델이 생성한 텍스트



#02 Related Work



EWHA EURON

#02 Related Work

General vision-language modeling
- 기존의 VLM들은 이미지와 텍스트를 동시에 이해
- 하지만 PaLM-E는 “multimodal sentence”의 형태로 이해 : 더 유연한 사고 가능
- VQAv2 벤치마크에서 모델 Frozen을 45% 넘게 앞서감

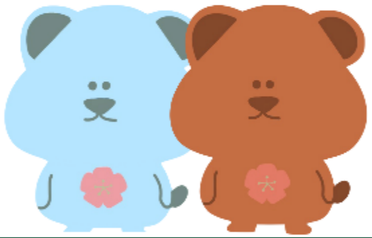
Actions-output models
- 시각 및 언어 입력을 합쳐 직접적 행동 예측을 목표로 연구가 진행됨
- PaLM-E는 고차원 지시를 텍스트로 생성함
- 이는 다양한 도메인을 교차하여 태스크 수행이 가능하게 함

LLMs in embodied task planning
- 많은 연구들이 planning 보다는 goal 이해에 초점을 둠
- Grounding을 위한 보조 모델 없이 하나의 모델로 직접 적용 가능한 planning 가능

*Grounding : 사람 사이에 효과적인 소통을 위해 필수적인 공통 된 이해와 기반을 다지는 과정

https://jiho-ml.com/weekly-nlp-34

#03 PaLM-E



EWHA EURON

#03 PaLM-E

Decoder-only LLM

$$p(w_{1:L}) = \prod_{l=1}^{L} p_{\mathrm{LM}}(w_l | w_{1:l-1}),$$

"I am a girl." 이라는 문장이 생성될 확률은
P(I) * P(am | I) * P(a | am) * P(girl | a)

Prefix-decoder-only LLM

- prefix 혹은 프롬프트로 추가 정보 제공

$$p(w_{n+1:L} | w_{1:n}) = \prod_{l=n+1}^{L} p_{\mathrm{LM}}(w_l | w_{1:l-1}).$$

'Girl은 이대생' 이라는 추가 정보가 제공 되었을 때,
"I am a girl"이라는 문장이 생성될 확률을
'girl 이 이대생' 이라는 맥락의 조건부 확률 안에서 확률을 계산

https://wikidocs.net/166790

#03 PaLM-E

Multi-modal sentences

- 문장 내에 어디에 위치해도 상관 없음

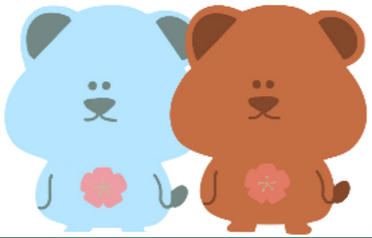
$$x_i = \begin{cases} \gamma(w_i) & \text{if } i \text{ a is text token, or} \\ \phi_j(O_j)_i & \text{if } i \text{ corresponds to observation } O_j. \end{cases}$$

Embodying the output

1. 단순 텍스트로 해결되는 태스크 : 바로 적용
2. 임베디드 planning이나 조종 태스크 : low-level 커맨드 추가 생성

- 로봇이 실행가능한 명령어를 추리는 필터가 없이도, PaLM-E는 알아서 로봇이 실행가능한 명령어가 무엇인지 판단하고 새로운 정보가 들어오면 다시 실행 계획을 짬

#04 Modalities

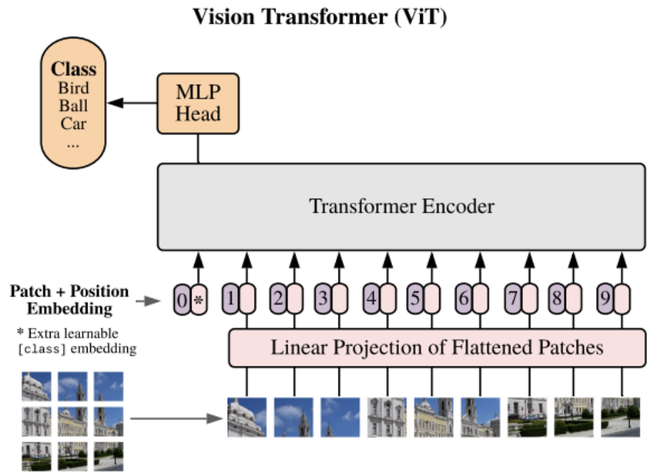


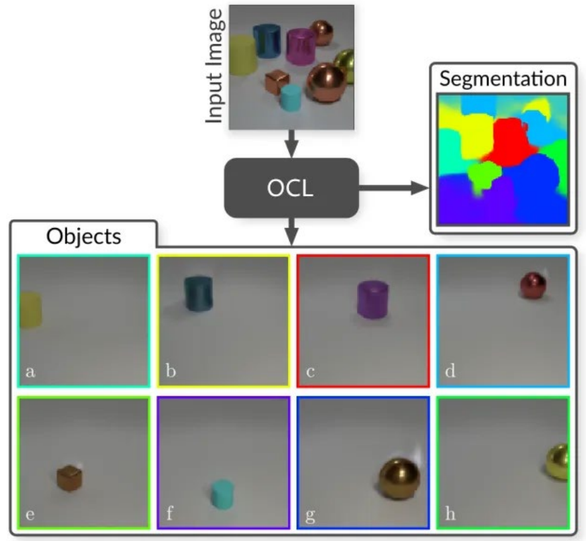
EWHA EURON

#04 Modalities

1. Vision Transformer(ViT)

- 이미지를 토큰화 하여 임베딩 하며, Token Learner 구조 함께 사용





Object Centric Representation

1,2번 인코더 모두 위와 같은 방식으로 Object 들을 인식하여 구별 할 수 있도록 함. 단, 1번은 Ground truth가 필요한 반면, 2번은 필요 없음.

2. Object Scene Representation Transformer(OSRT)

- 3D 정보를 받음

https://www.thedatahunt.com/trend-insight/vision-transformer
https://osrt-paper.github.io/

#04 Modalities

3. State estimation Vectors

- 물체의 위치, 크기, 색상 등의 정보

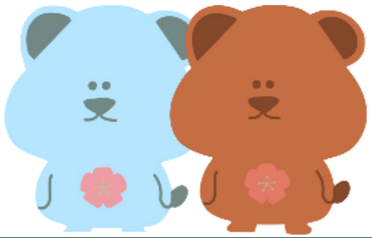
4. Entity Referrals

- 각 object를 구별하기 위한 레이블을 프롬프트로 함께 입력

예시) <prefix> = Obj 1 is <obj$_1$> <prefix> = Obj 1 is <obj$_1$>

https://www.thedatahunt.com/trend-insight/vision-transformer
https://osrt-paper.github.io/

#05 Training



EWHA EURON

#05 Training

학습 데이터 형태

$$\left\{\left(I^i_{1:u_i}, w^i_{1:L_i}, n_i\right)\right\}_{i=1}^{N}$$

Given <emb> ... <img> Q: How to grasp blue block?

파라미터 크기

- 8B PaLM + 4b ViT = PaLM-E-12B
- 62B PaLM + 22b ViT = PaLM-E-84B
- 540B PaLM + 22b ViT = PaLM-E-562B

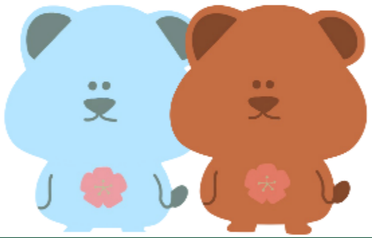
Model Freezing

- 인코더 + 프로젝터 + LLM 구조에서 LLM은 고정 시킨 채 나머지 부분만 학습시키는 방법

Task 교차 공동학습

- 다양한 task 관련 이미지 및 언어 데이터를 모두 섞은 데이터를 학습시켜 그 효과를 실험해 봄

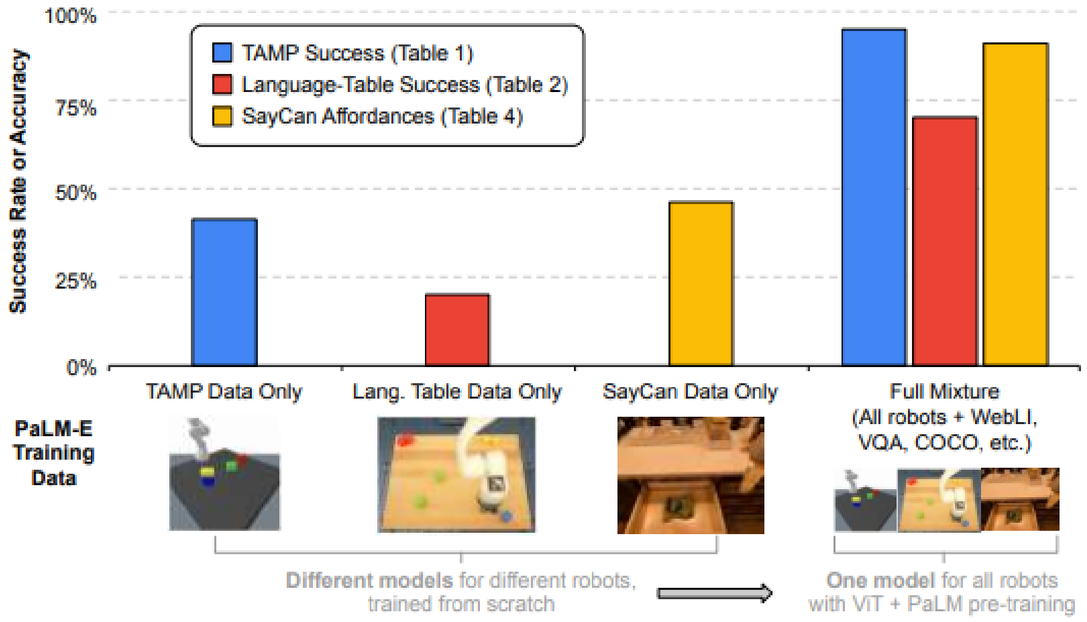
#06 Results

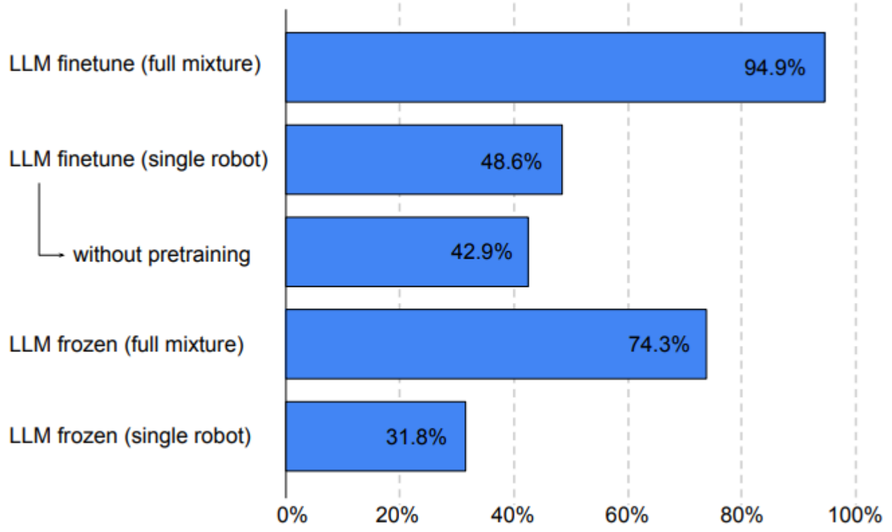


EWHA EURON

#06 Results

Full Mixture





#06 Results

TAMP(Task and Motion Planning)

VQA(Visual Question Answering)

1. **Detection** Given <img>. Q : 이 물체의 색은?
2. **Object-table relation** : Given <img>. Q : 빨간색 물체는 책상의 왼쪽, 오른쪽, 중간 중 어디에 있는가?
3. **Object-object relation** : Given <img>. Q : 노란색 물체가 파란색 물체 아래에 있나?
4. **Plan feasibility** : Given <img> Q : 파란색 물체를 먼저 잡고, 노란색 물체 위에 올려둔 뒤, 노란색 물체를 잡는 것이 가능한가?

Planning

1. **Grasping** : Given <img>. Q : 초록색 물체를 잡으려면 어떻게 해야 해?
2. **Stacking** : Given <img>. Q : 하얀색 물체를 빨간색 물체 위에 놓으려면 어떻게 해야 해?

	Object-centric	LLM pre-train	Embodied VQA				Planning	
			q_1	q_2	q_3	q_4	p_1	p_2
SayCan (oracle afford.) (Ahn et al., 2022)		✓	-	-	-	-	38.7	33.3
PaLI (zero-shot) (Chen et al., 2022)		✓	-	0.0	0.0	-	-	-
PaLM-E (ours) w/ input enc:								
State	✓(GT)	✗	99.4	89.8	90.3	88.3	45.0	46.1
State	✓(GT)	✓	**100.0**	96.3	95.1	93.1	55.9	49.7
ViT + TL	✓(GT)	✓	34.7	54.6	74.6	91.6	24.0	14.7
ViT-4B single robot	✗	✓	-	45.9	78.4	92.2	30.6	32.9
ViT-4B full mixture	✗	✓	-	70.7	93.4	92.1	74.1	74.6
OSRT (no VQA)	✓	✓	-	-	-	-	71.9	75.1
OSRT	✓	✓	99.7	**98.2**	**100.0**	**93.7**	**82.5**	**76.2**

EWHA EURON

#06 Results

TAMP(Task and Motion Planning)

	ϕ	LLM pre-trained	q_1	q_2	q_3	q_4	p_1	p_2
3 - 5 objects	SayCan (w/ oracle affordances)	✓	-	-	-	-	38.7	33.3
	state	✗	100.0	99.3	98.5	99.8	97.2	95.5
	state	✓(unfrozen)	100.0	98.8	100.0	97.6	97.7	95.3
	state	✓	100.0	98.4	99.7	98.5	97.6	96.0
	state (w/o entity referrals)	✓	100.0	98.8	97.5	98.1	94.6	90.3
	ViT + TL (obj. centric)	✓	99.6	98.7	98.4	96.8	9.2	94.5
	ViT + TL (global)	✓	-	60.7	90.8	94.3	70.7	69.2
	ViT-4B (global)	✓	-	98.2	99.4	99.0	96.0	93.4
	ViT-4B generalist	✓	-	97.1	100.0	98.9	97.5	95.2
	OSRT	✓	99.6	99.1	100.0	98.8	98.1	95.7
6 objects	state	✗	20.4	39.2	71.4	85.2	56.5	34.3
	state	✓	100.0	98.5	94.0	89.3	95.3	81.4
	state (w/o entity referrals)	✓	77.7	83.7	93.6	91.0	81.2	57.1
8 objects	state	✗	18.4	27.1	38.1	87.5	24.6	6.7
	state	✓	100.0	98.3	95.3	89.8	91.3	89.3
	state (w/o entity referrals)	✓	60.0	67.1	94.1	81.2	49.3	49.3
6 objects + OOD tasks	state (8B LLM)	✗	-	0	0	72.0	0	0
	state (8B LLM)	✓	-	49.3	89.8	68.5	28.2	15.7
	state (62B LLM)	✓	-	48.7	92.5	88.1	40.0	30.0

Table 7: Success rates on TAMP environment for different input representations. 3-5 objects in the scene correspond to the training distribution. OOD tasks means out-of-distribution tasks where the objects are referenced by color, although in the trainig data they have been referenced by their special tokens $\texttt{obj}_j$ in the object-centric case. The SayCan baseline (Ahn et al., 2022) utilizes oracle, one-step affordance functions.

#06 Results

Language-Table

1. Q: There is a block that is closest to {i.e., top right corner}. Push that block to the other block of the same color.
2. Q: How to sort the blocks by colors into corners?
3. Q: How to push all the blocks that are on the {left/right} side together, without bringing over any of the blocks that are on the {right/left} side?

Zero-shot Baselines						Task 1			Task 2			Task 3		
SayCan (oracle afford.) (Ahn et al., 2022)						0.0			-			-		
PaLI (Chen et al., 2022)						0.0			-			-		
PaLM-E-	trained on	from scratch	LLM+ViT pretrain	LLM frozen	Task finetune	*# Demos 10*	*20*	*40*	*10*	*20*	*40*	*10*	*20*	*80*
12B	Single robot	✓	✗	n/a	✓	20.0	30.0	50.0	2.5	6.3	2.5	11.3	16.9	28.3
12B	Full mixture	✗	✓	✓	✗	-	-	20.0	-	-	36.3	-	-	29.4
12B	Full mixture	✗	✓	✗	✗	-	-	80.0	-	-	57.5	-	-	50.0
12B	Full mixture	✗	✓	✗	✓	**70.0**	**80.0**	80.0	**31.3**	**58.8**	**58.8**	**57.5**	**54.4**	56.3
84B	Full mixture	✗	✓	✗	✗	-	-	**90.0**	-	-	53.8	-	-	**64.4**

Table 2: Results on planning tasks in the simulated environment from Lynch et al. (2022).

#06 Results

Language-Table

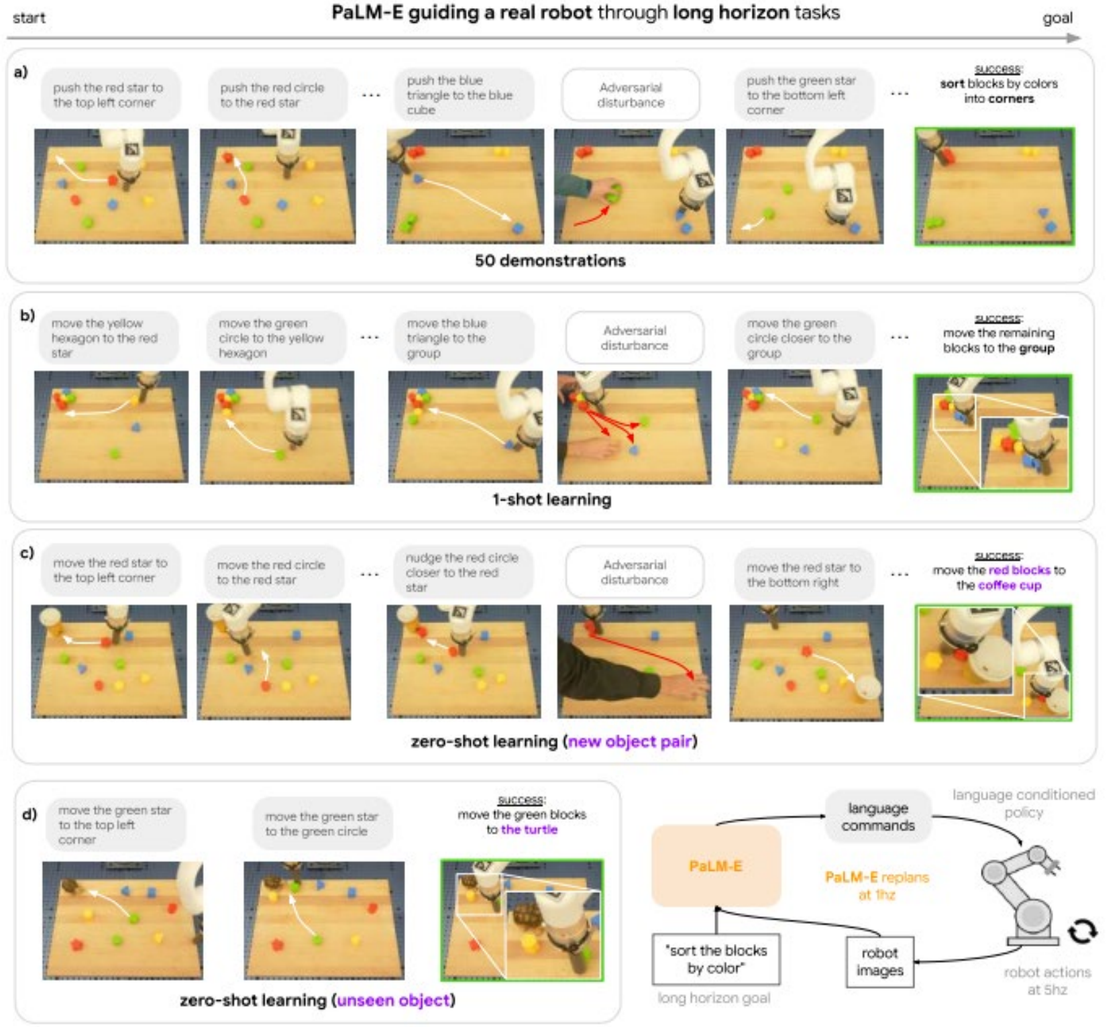


Figure 7: PaLM-E interactively guides a real robot through long-horizon manipulation tasks on Language-Table, while remaining robust to adversarial disturbances. We find evidence that PaLM-E is capable of one-shot and zero shot generalization.

#06 Results

Mobile Manipulation

Affordance Prediction

- 현 상황에서 수행해야 하는 하위 규칙이 가능한지 판정
- Given <img>. Q : Is it possible to <skill> here?

Failure Detection

- 반복문 종료를 위한 수행 결과 확인
- Given <img>. Q : Was <skill> successful

Baselines				Failure det.	Affordance
PaLI (Zero-shot) (Chen et al., 2022)				0.73	0.62
CLIP-FT (Xiao et al., 2022)				0.65	-
CLIP-FT-hindsight (Xiao et al., 2022)				0.89	-
QT-OPT (Kalashnikov et al., 2018)				-	0.63
PaLM-E-12B trained on	from scratch	LLM+ViT pretrain	LLM frozen		
Single robot	✓	✗	n/a	0.54	0.46
Single robot	✗	✓	✓	**0.91**	0.78
Full mixture	✗	✓	✓	**0.91**	0.87
Full mixture	✗	✓	✗	0.77	**0.91**

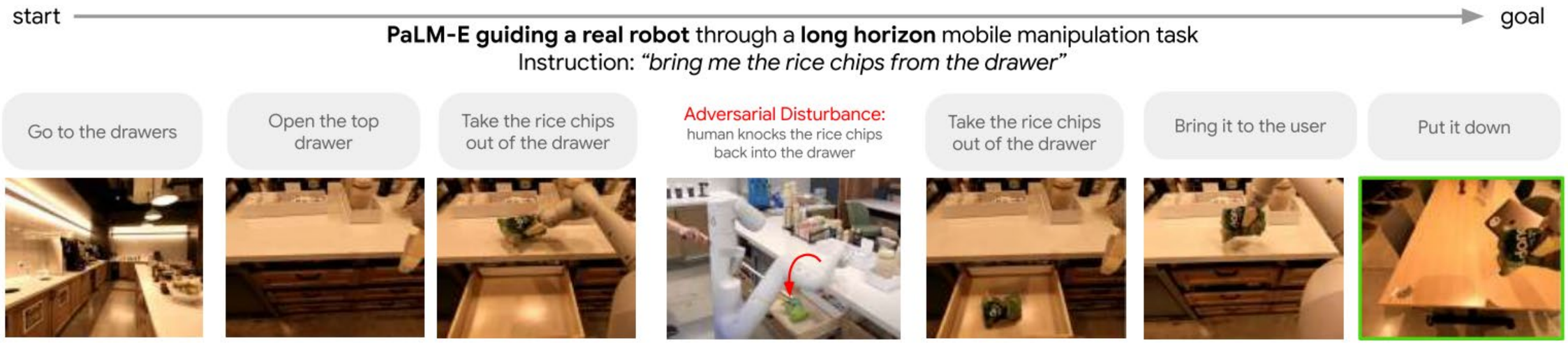
Table 4: Mobile manipulation environment: failure detection and affordance prediction (F1 score).

#06 Results

Mobile Manipulation

Long-horizon planning

- Human : <instruction> Robot : <step history>. I see <img>.



#06 Results

General Visual-Language Tasks

- 로봇 임베딩 모델임에도 일반적인 시각-언어 모델들의 성능 비교에서 크게 뒤지지 않음

Model	VQAv2 test-dev	VQAv2 test-std	OK-VQA val	COCO Karpathy test
Generalist (one model)				
PaLM-E-12B	76.2	-	55.5	135.0
PaLM-E-562B	80.0	-	**66.1**	138.7
Task-specific finetuned models				
Flamingo (Alayrac et al., 2022)	82.0	82.1	57.8†	138.1
PaLI (Chen et al., 2022)	84.3	84.3	64.5	149.1
PaLM-E-12B	77.7	77.9	60.1	136.0
PaLM-E-66B	-	-	62.9	-
PaLM-E-84B	80.5	-	63.3	138.0
Generalist (one model), with frozen LLM				
(Tsimpoukelli et al., 2021)	48.4	-	-	-
PaLM-E-12B frozen	70.3	-	51.5	128.0

Table 5: Results on general visual-language tasks. For the generalist models, they are the same checkpoint across the different evaluations, while task-specific finetuned models use different-finetuned models for the different tasks. COCO uses Karpathy splits. † is 32-shot on OK-VQA (not finetuned).

#06 Results

General Language Tasks

- 모델의 사이즈를 늘릴수록 PaLM-E로 옮겨가는 과정에서의 정보 손실이 덜함

C. Natural Language Generation and Understanding Results

1-shot evals	PaLM-8B	PaLM-E-12B *(unfrozen)*	PaLM-62B	PaLM-E-84B *(unfrozen)*	PaLM-540B	PaLM-E-562B *(unfrozen)*	Category
TriviaQA (wiki) (EM)	48.5	10.1	72.7	31.8	81.4	74.6	NLG
Natural Questions (EM)	10.6	1.6	23.1	7.6	29.3	27.2	NLG
WebQuestions (EM)	12.6	3.4	19.8	7.9	22.6	21.8	NLG
Lambada	57.8	1.4	75.5	26.1	81.8	83.3	NLG
HellaSwag	68.2	48.4	79.7	75.3	83.6	83.5	NLU
StoryCloze	78.7	68.7	83.8	83.9	86.1	86.3	NLU
Winograd	82.4	71.8	85.3	86.4	87.5	89.0	NLU
Winogrande	68.3	55.3	76.8	72.5	83.7	83.0	NLU
RACE-M	57.7	43.2	64.1	57.4	69.3	70.3	NLU
RACE-H	41.6	33.2	48.7	42.3	52.1	52.8	NLU
PIQA	76.1	68.1	80.9	78.2	83.9	84.9	NLU
ARC-e	71.3	53.4	78.9	71.4	85.0	86.3	NLU
ARC-c	42.3	30.9	51.8	46.7	60.1	62.6	NLU
OpenBookQA	47.4	41.4	51.2	51.6	53.6	55.8	NLU
BoolQ	64.7	61.6	83.1	81.6	88.7	89.4	NLU
Copa	82.0	77.0	93.0	91.0	91.0	93.0	NLU
RTE	57.8	54.9	71.5	59.6	78.7	75.1	NLU
Wic	50.6	50.0	48.6	50.2	63.2	64.1	NLU
WSC	81.4	68.4	84.9	75.8	86.3	85.6	NLU
ReCoRD	87.8	71.2	91.0	78.5	92.8	92.5	NLU
CB	41.1	37.5	55.4	73.2	83.9	80.3	NLU
Avg NLU	64.7	55.0	72.3	69.2	78.2	78.5	
Avg NLG	32.4	4.1	47.8	18.4	53.8	51.7	
NLU delta (%, relative)		-15.0%		-4.3%		+0.4%	
NLG delta (%, relative)		-87.3%		-61.6%		-3.8%	

Table 8: Full language evaluation task results on both NLU and NLG tasks, for both the original PaLM models and for associated PaLM-E (unfrozen) models. The PaLM-E models with a frozen LLM have the same performance as their corresponding underlying PaLM models.

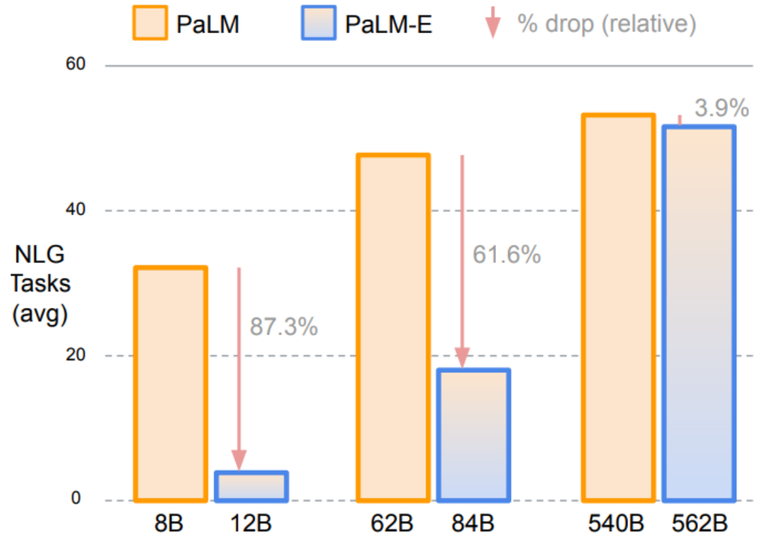
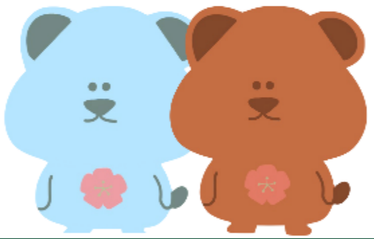


Figure 6: Results on general language tasks (NLG = natural language generation): increasing scale leads to less catastrophic forgetting between a corresponding PaLM-E model and its inherited PaLM model. See full suite of tasks and results in Tab. 8.

#07 Discussion



EWHA EURON

#07 Discussion

1. “full mixture” 데이터셋으로 transfer의 효과를 경험함

2. 데이터 양이 많지 않았는데도 좋은 성능을 보임

3. LLM을 freezing 하거나 모델의 크기를 늘리는 것으로 언어 성능을 보존할 수 있음

#07 Discussion

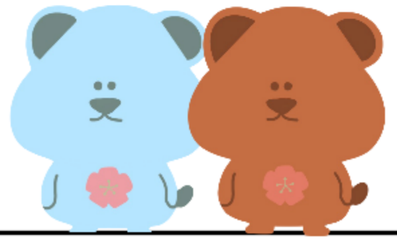
언어 모델의 확장성은 어디까지일까?

- 기존에는 번역, 챗봇 등만 생각했지, 로봇에게 명령을 내릴 수 있는 주체라고 생각하지 못함

서로 다른 태스크 데이터를 다 섞어 학습하는 것이
따로따로 학습하는 것보다 더 성능이 좋은 이유는 무엇일까?

- 전반적인 내용에 대한 이해도가 특정 태스크를 위한 이해도까지 늘리는 것일까

THANK YOU



EWHA EURON