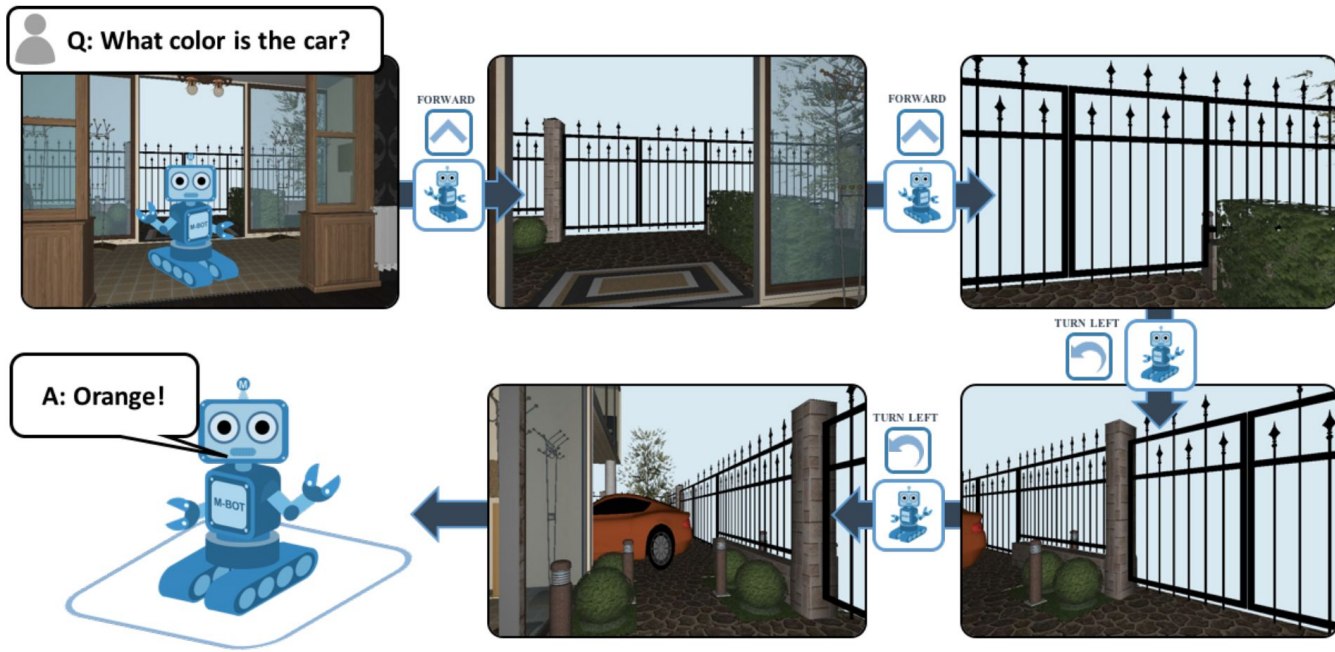


EMBODIED QUESTION ANSWERING

Авторы статьи: A Das et Al.

Georgia Institute of Technology, Facebook AI Research, 2017

Автор доклада: Данилова Юлия, БПМИ141



Задача: ориентация в 3D-среде с целью ответа на вопрос, поставленный на естественном языке.

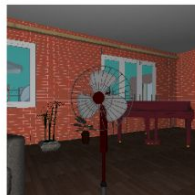
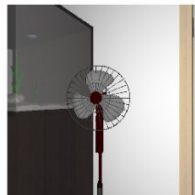
- Наблюдение от первого лица (в каждый момент доступно одно RGB изображение)
- Вопросы вида "Какого цвета машина?" или "В какой комнате находится телефон?"

СХОЖИЕ НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ

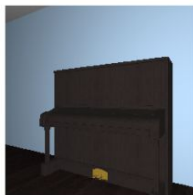
Направление	Отличие от EQA
VQA: Vision + Language	Только ответ на вопрос по картинке - нет навигации.
Visual Navigation: Vision + Action	Нет вопроса на естественном языке
Situated Language Learning: Language + Action	У агента есть доступ к плану помещения, лейблам объектов и др. информации
Embodiment: Language + Vision + Action	Есть работа (еще не опубликована) с очень схожей постановкой. Меньше комнат, агента просят совершить последовательность команд.

ИНТЕРАКТИВНЫЕ КВАРТИРЫ HOUSE3D

pedestal fan



piano



fish tank



Доступна информация о **географии** помещения (напр., расстояние для объекта) и **семантике** (классы объектов)

Для задачи:

- Агент не может проходить сквозь стены и объекты
- Отобраны квартиры среднего размера и "сложности"
- Схожие объекты объединены в одну категорию (напр., чайник и кофейник)

СОСТАВЛЕНИЕ ВОПРОСОВ

В EQAv1 вопросы могут принадлежать только к определенным типам:

$$\text{EQAv1} \left\{ \begin{array}{ll} \text{location:} & \text{'What room is the } \langle \text{OBJ} \rangle \text{ located in?'} \\ \text{color:} & \text{'What color is the } \langle \text{OBJ} \rangle \text{'?'} \\ \text{color_room:} & \text{'What color is the } \langle \text{OBJ} \rangle \text{ in the } \\ & \text{\langle ROOM \rangle?'} \\ \text{preposition:} & \text{'What is } \langle \text{on/above/below/next-to} \rangle \text{ the } \\ & \text{\langle OBJ \rangle in the } \langle \text{ROOM} \rangle \text{'?'} \end{array} \right.$$

И включать только ограниченный набор комнат / объектов:

gym	dining room
patio	living room
office	bathroom
lobby	bedroom
garage	elevator
kitchen	balcony

rug	piano	dryer	computer	fireplace	whiteboard	bookshelf	wardrobe	cabinet
pan	toilet	plates	ottoman	fish tank	dishwasher	microwave	water dispenser	
bed	table	mirror	tv stand	stereo set	chessboard	playstation	vacuum cleaner	
cup	xbox	heater	bathtub	shoe rack	range oven	refrigerator	coffee machine	
sink	sofa	kettle	dresser	knife rack	towel rack	loudspeaker	utensil holder	
desk	vase	shower	washer	fruit bowl	television	dressing table	cutting board	
ironing board	food processor							

ДАТАСЕТ (ИТОГИ)

После балансировки (отфильтровывание сложных вопросов / редких объектов) датасет включает:

5000 вопросов

по более чем **750 средам**

относящихся к **45 уникальным объектам**

в **7 типах комнат**

МОДУЛЬНАЯ АРХИТЕКТУРА

Модуль навигации (navigation)		Модуль зрения (vision)
	Embodied Question Answering	
Модуль ответа на вопросы (answering)		Языковой модуль (language)

МОДУЛЬ ЗРЕНИЯ

CNN из 4 блоков вида {5*5 Conv, ReLU, BatchNorm, 2*2 MaxPool}

Вход: 224*224 RGB изображение

Обучение: encoder-decoder

Decoder возвращает тензор размера [191+1+3, img.shape]

191 канал - семантическая сегментация

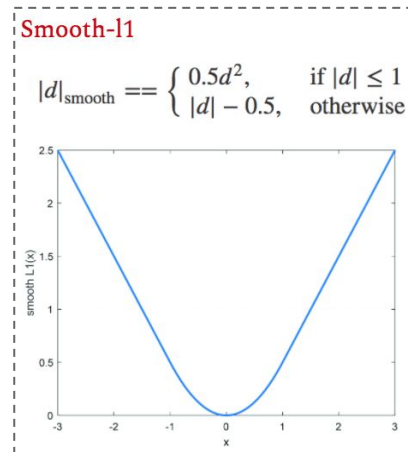
1 канал - расстояния для объектов

3 канала - восстановленное исходное изображение






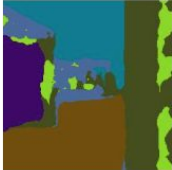
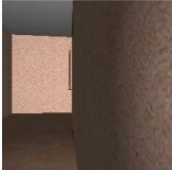









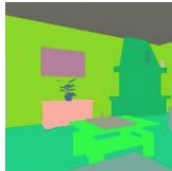
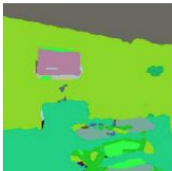

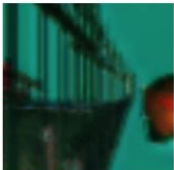




Cross-entropy loss для сегментации, Smooth-l1 для остальных.

overall_loss = seg_loss + 10 * depth_loss + 10 * reconstruction_loss

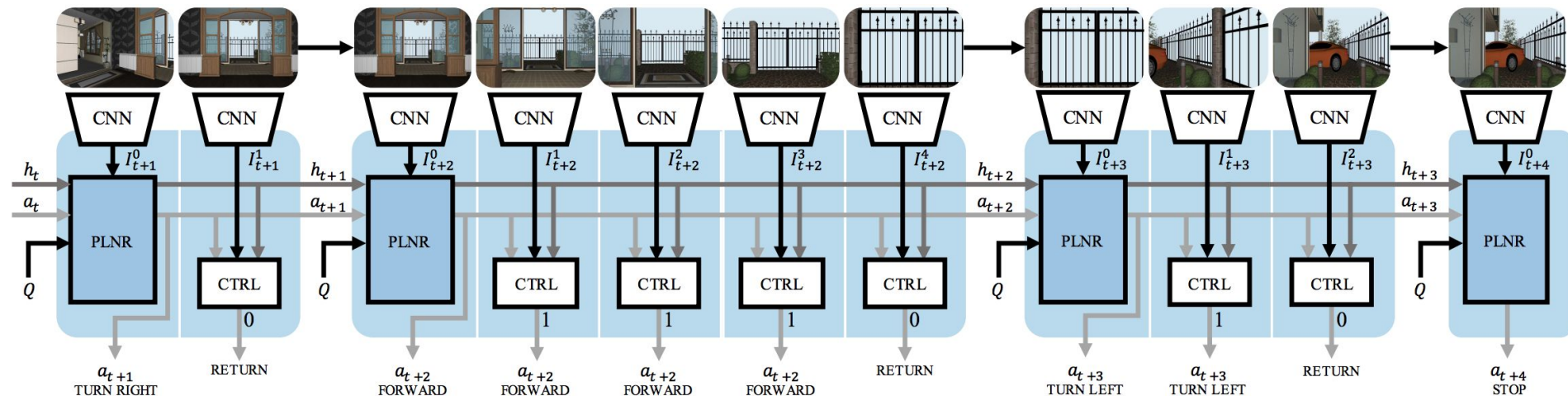
5 эпох на датасете из 100 тысяч RGB картинок



МОДУЛЬ ЗРЕНИЯ: ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ ПРЕДОБУЧЕНИЯ

GT RGB	Pred. RGB	GT Depth	Pred. Depth	GT Seg.	Pred. Seg.
					
					
					
					

МОДУЛЬ НАВИГАЦИИ



Используют **Adaptive Computation Time (ACT) RNN** [Graves 2016]

Плanner (PLNR) определяет действие (прямо, влево).

-- **LSTM** $a_t, h_t \leftarrow \text{PLNR}(h_{t-1}, I_t^0, Q, a_{t-1})$

Контроллер (CTRL) выполняет выбранное действие переменное число раз.

-- **Multilayer perceptron w. 1 hidden layer** $\{0, 1\} \ni c_t^n \leftarrow \text{CTRL}(h_t, a_t, I_t^n)$

Языковой модуль

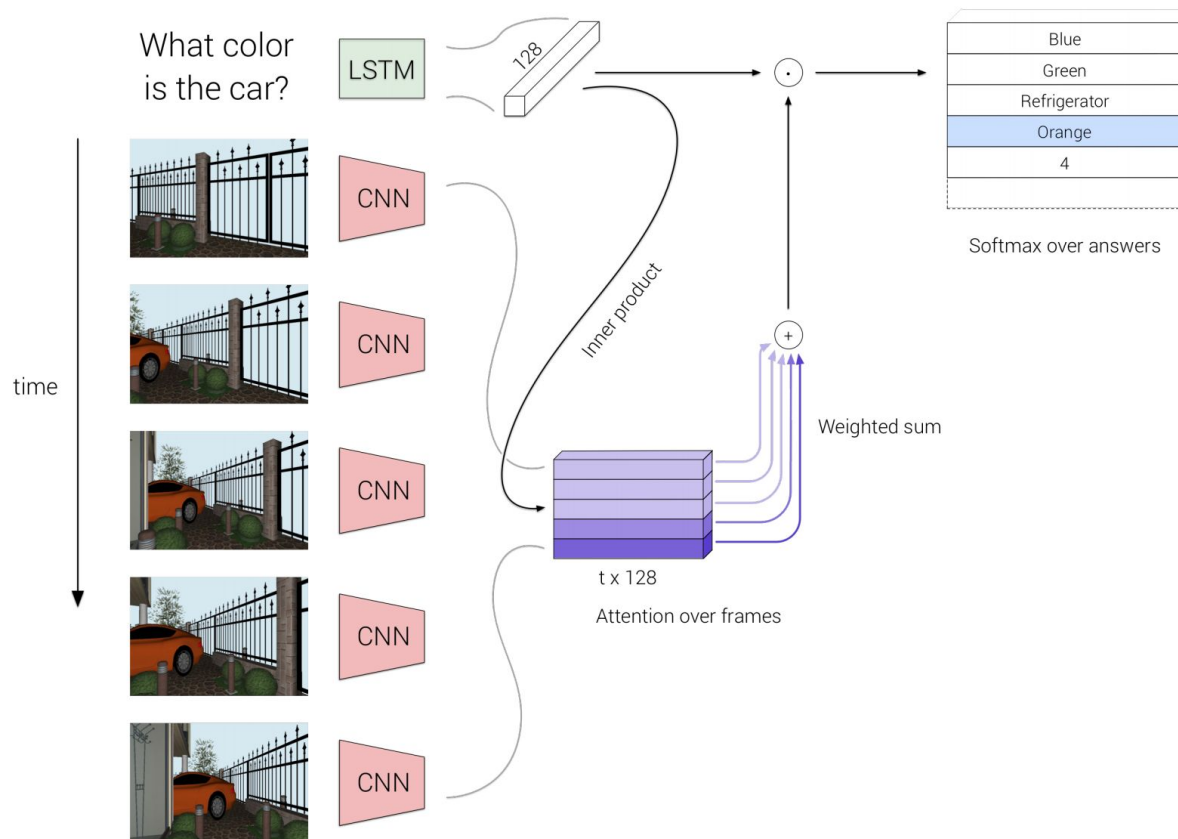
2-слойный LSTM со 128 скрытыми состояниями.

! Отдельно обучаются энкодеры вопроса для модуля навигации и модуля ответов.

"Какого *цвета стул* на *кухне*?"

"Цвет" и "стул" важнее для модуля ответа на вопрос, "кухня" - для навигации.

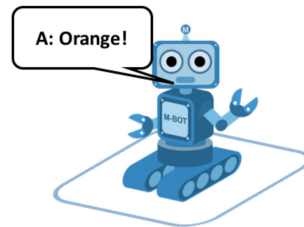
МОДУЛЬ ОТВЕТА НА ВОПРОС



ОБУЧЕНИЕ: ПРЕДОБУЧЕНИЕ ОТДЕЛЬНЫХ МОДУЛЕЙ

Предобучение модуля **навигации** с помощью **imitation learning**:

- С помощью House3D можно построить кратчайший путь до объекта
- Агенту доступна информация:
 - history encoding
 - question encoding
 - текущий кадр
- Учится имитировать эту "идеальную" траекторию
- Cross-entropy loss
- Начинают близко от цели (10 шагов), после каждой успешной эпохи ± 10 шагов
- 15 эпох с батчем от 5 до 20 вопросов



Модуль **ответа на вопрос** обучается на 5 последних кадрах идеальной траектории.

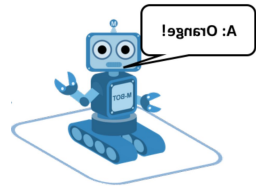
ОБУЧЕНИЕ: FINE-TUNING МОДУЛЕЙ

В результате предыдущего шага модели **не очень взаимодействуют между собой**. Модуль ответа на вопрос замораживают и фاین-тюнят модуль навигации.

Два типа наград (rewards):

- 1) Question answering accuracy, равная 5 в случае правильного ответа и 0 - неправильного
- 2) "навигационная" награда за приближение к цели, равная 0.05

Обучают с помощью **REINFORCE policy gradients**.



СРАВНЕНИЕ С ДРУГИМИ МОДЕЛЯМИ

		Navigation																		QA		
		d _T			d _Δ			d _{min}			%r _T			%r _↵			%stop			MR		
		T ₋₁₀	T ₋₃₀	T ₋₅₀	T ₋₁₀	T ₋₃₀	T ₋₅₀	T ₋₁₀	T ₋₃₀	T ₋₅₀	T ₋₁₀	T ₋₃₀	T ₋₅₀	T ₋₁₀	T ₋₃₀	T ₋₅₀	T ₋₁₀	T ₋₃₀	T ₋₅₀	T ₋₁₀	T ₋₃₀	T ₋₅₀
Baselines	Reactive	2.09	2.72	3.14	-1.44	-1.09	-0.31	0.29	1.01	1.82	50%	49%	47%	52%	53%	48%	-	-	-	3.18	3.56	3.31
	LSTM	1.75	2.37	2.90	-1.10	-0.74	-0.07	0.34	1.06	2.05	55%	53%	44%	59%	57%	50%	80%	75%	80%	3.35	3.07	3.55
	Reactive+Q	1.58	2.27	2.89	-0.94	-0.63	-0.06	0.31	1.09	1.96	52%	51%	45%	55%	57%	54%	-	-	-	3.17	3.54	3.37
	LSTM+Q	1.13	2.23	2.89	-0.48	-0.59	-0.06	0.28	0.97	1.91	63%	53%	45%	64%	59%	54%	80%	71%	68%	3.11	3.39	3.31
Us	ACT+Q	0.46	1.50	2.74	0.16	0.15	0.12	0.42	1.42	2.63	58%	54%	45%	60%	56%	46%	100%	100%	100%	3.09	3.13	3.25
	ACT+Q-RL	1.67	2.19	2.86	-1.05	-0.52	0.01	0.24	0.93	1.94	57%	56%	45%	65%	62%	52%	32%	32%	24%	3.13	2.99	3.22
Oracle	HumanNav*	0.81	0.81	0.81	0.44	1.62	2.85	0.33	0.33	0.33	86%	86%	86%	87%	89%	89%	-	-	-	-	-	-
	ShortestPath+VQA	-	-	-	0.85	2.78	4.86	-	-	-	-	-	-	-	-	-	-	-	-	3.21	3.21	3.21

MR - mean rank (MR) правильного ответа, усреднение по вопросам и средам из тестовой выборки

d_T - расстояние до цели

d_Δ/delta - изменение расстояния от изначальной до конечной точки

d_{min} - наименьшее расстояние до цели в эпизоде (d_{min})

%r_T - процент эпизодов, в котором агент завершил игру

%r_{arrow} - процент эпизодов, в котором вошел в нужную комнату (r_{arrow})

%stop - процент эпизодов, в котором ответил вопрос до достижения конца эпизода (maximum episode length)

ВЫВОДЫ

- Предоставили **новую постановку задачи** в сфере искусственного интеллекта: Embodied Question Answering (EmbodiedQA)
- Предоставили **вариант решения этой задачи** с помощью **модульной архитектуры**: модуль навигации на основе Adaptive Computation Time (ACT), модуль зрения (CNN), языковые модули и ответа на вопрос.
- Заставили все это работать: **предобучали модули** с использованием информации о географии помещения от House3D + (опционально) **последующий fine-tuning** модели целиком.
- **Сравнили с бейзлайнами** (насколько смогли их найти - все же, вроде как, первые в своем роде) - вариант без fine-tuning подбирался ближе других к цели, с fine-tuning давал хорошие результаты в ответе на вопрос.

ИСТОЧНИК

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2017). Embodied question answering. *arXiv preprint arXiv:1711.11543*

<https://arxiv.org/pdf/1711.11543.pdf>