

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

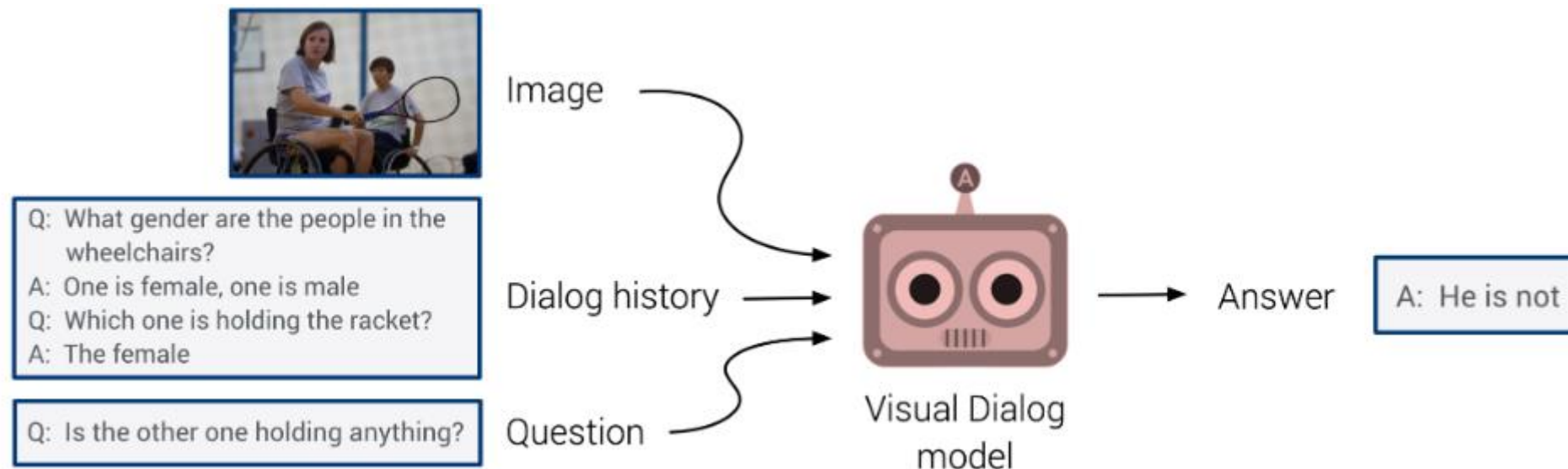
Чернявский Александр (151)

Предметная область

- Задача на стыке Computer Vision, NLP и Artificial Intelligence
- VQA – вопросно-ответная система

Недостаток: здесь нет памяти в системе предыдущих вопросов и логической согласованности ответов.

- Visual Dialog – диалоговая система:



Visual Dialog



VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Captioning

Two people are in a wheelchair and one is holding a racket.

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman



Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right

A: No, it's a woman

Датасет VisDial

Построение датасета:

- Взяты картинки (123к) из Common Objects in Context.
- Получение 10 пар QA на сервисе Amazon Mechanical Turk реальными людьми

Caption: A sink and toilet in a small room.

You have to ASK questions about the image.

Fellow Turker connected.
Now you can send messages.

1.You:
is this a bathroom ?

1.Fellow Turker:
yes, it's a bathroom

2.You:
what color is the room ?

Message **SEND**

Caption: A sink and toilet in a small room.

You have to ANSWER questions about the image.

Fellow Turker connected.
Now you can send messages.

1.Fellow Turker:
is this a bathroom ?

1.You:
yes, it's a bathroom

2.Fellow Turker:
what color is the room ?

2.You:
it looks cream colored

Message **SEND**

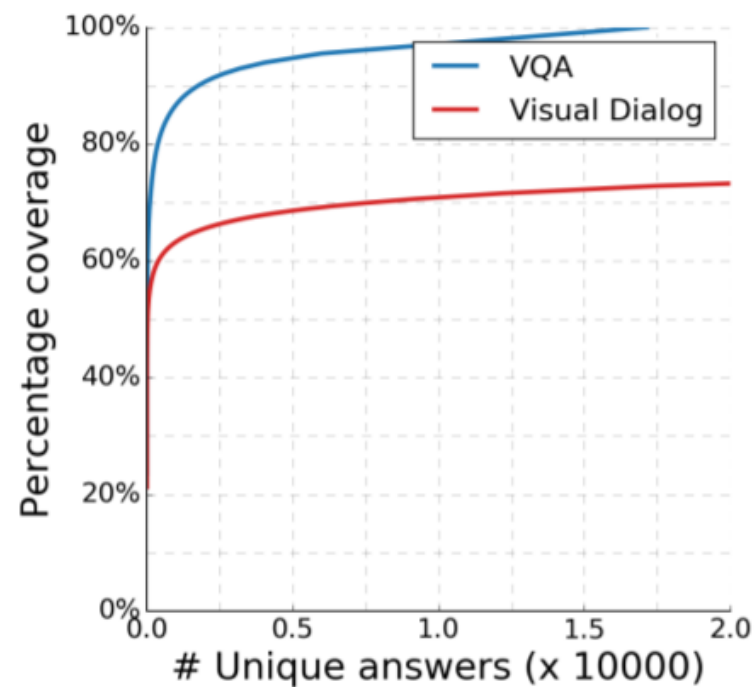
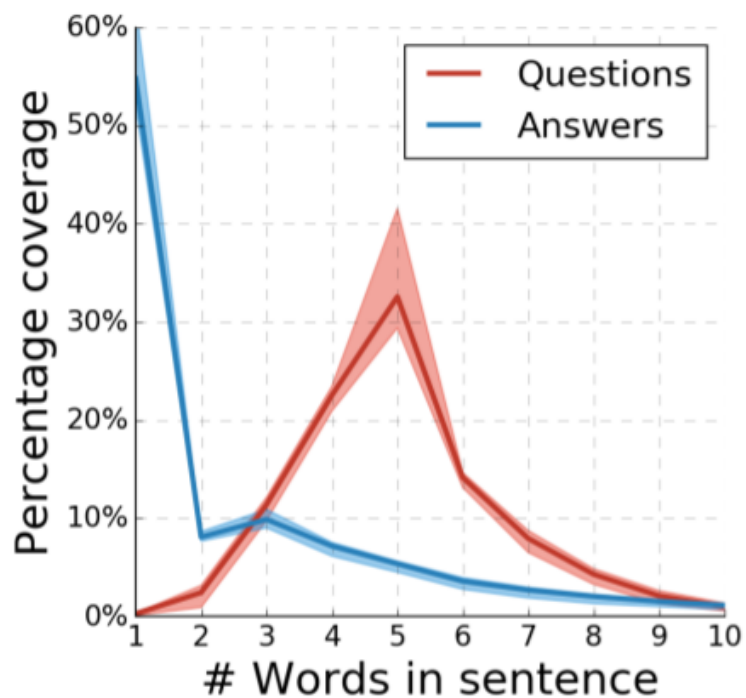


Caption:
A sink and toilet in a small room.

- Q3: can you see anything else ?
A3: there is a shelf with items on it
- Q4: is anyone in the room ?
A4: nobody is in the room
- Q5: can you see on the outside ?
A5: no, it is only inside
- Q6: what color is the sink ?
A6: the sink is white
- Q7: is the room clean ?
A7: it is very clean
- Q8: is the toilet facing the sink ?
A8: yes the toilet is facing the sink
- Q9: can you see a door ?
A9: yes, I can see the door
- Q10 what color is the door ?
A10 the door is tan colored

Преимущества датасета VisDial

- 1) Отсутствует Visual Priming Bias
- 2) Вопросы и ответы достаточно длинные, а самые популярные ответы покрывают относительно небольшую часть вопросов



Оценка качества

При тестировании на вход подаются: изображение I , 'ground-truth' диалог (включая caption) H , вопрос Q_t и множество из 100 вариантов ответа, которые необходимо отсортировать по релевантности.

Метрики:

1) Rank = Ранг ответа человека

2) recall@k = [ответ существует в top-k]

3) reciprocal rank (RR) = $1 / \text{Rank}$

Усреднение

Mean Rank (ниже лучше)

R@k (выше лучше)

MRR (выше лучше)

Encoder-Decoder модель

Memory Network (MN) Encoder + Discriminative Decoder

Вход: такой же как и для оценки качества.

Энкодер строит эмбединг для тройки (I, H, Q_t) , после чего декодер конвертирует этот вектор в ответ.

MN Encoder



Image I

Do you think
the woman is
with him?

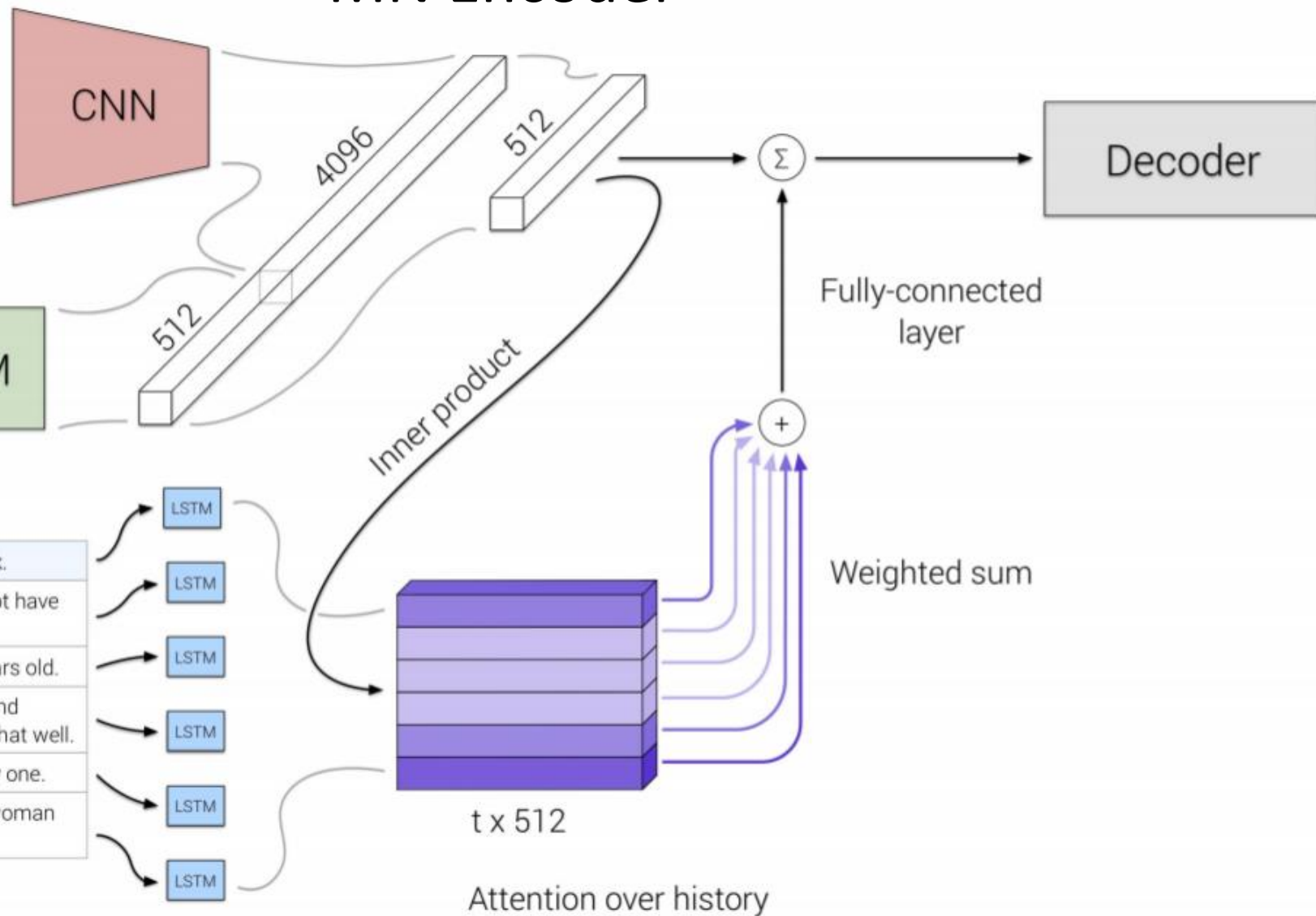
Question Q_t

LSTM

The man is riding his bicycle on the sidewalk.
Is the man wearing a helmet? No he does not have a helmet on.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.

t rounds of history

$\{(\text{Caption}), (Q_1, A_1), \dots, (Q_{t-1}, A_{t-1})\}$




No I don't think
they are together

Answer A_t

Варианты декодера

1) Дискриминативный:

Эмбединги вариантов ответа


$$probs = softmax(\{\langle y_{enc}, LSTMA_i \rangle | i = 1..100\})$$

2) Генеративный: $A_t = LanguageLSTM(y_{enc})$

Проблемы подхода

- Модель не может управлять диалогом и не видит будущих последствий своих высказываний во время обучения.
- Ограниченность оценки для высказываний не из датасета:



В выборке:

Вопрос: «How's the weather?»

Ответ: «Sunny»



Ответы не из выборки:

«Clear», «Looks warm»,

«It's not raining»...

Guessing Game



Guessing Game



Guessing Game

Questioner Q-BOT Answerer A-BOT

Two zebra are walking around their pen at the zoo.

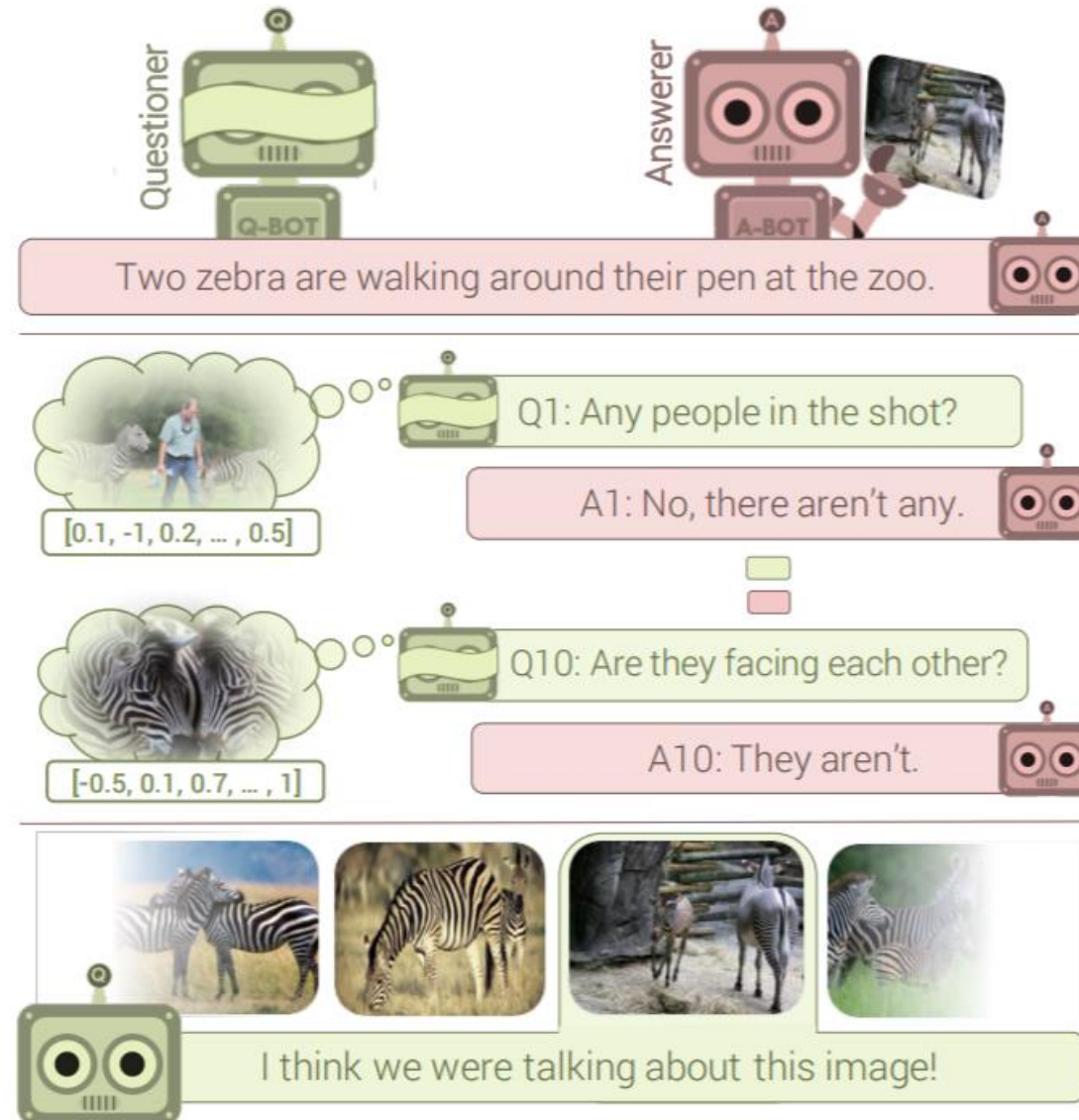
Q1: Any people in the shot?
[0.1, -1, 0.2, ..., 0.5]

A1: No, there aren't any.

Q10: Are they facing each other?
[-0.5, 0.1, 0.7, ..., 1]

A10: They aren't.

I think we were talking about this image!



The diagram illustrates a guessing game between two robots, Q-BOT (green) and A-BOT (red). The game starts with a description: "Two zebra are walking around their pen at the zoo." Q-BOT asks a series of questions about the image, and A-BOT provides answers. The questions and answers are as follows:

- Q1: Any people in the shot? (Answer: No, there aren't any.)
- Q10: Are they facing each other? (Answer: They aren't.)

At the bottom, a set of four images is shown, and A-BOT identifies the correct one: "I think we were talking about this image!"

	Q-BOT	A-BOT
Среда	Изображения для выбора	Исходное изображение I
Действие	1) Вопрос q_t - предложение из токенов словаря V 2) Предсказание эмбединга исходной картинки \hat{y}_t .	Ответ a_t - предложение из токенов словаря V
Состояние	$s_t^Q = [c, q_1, a_1, \dots, q_{t-1}, a_{t-1}]$	$s_t^A = [I, c, q_1, a_1, \dots, q_{t-1}, a_{t-1}, q_t]$
Стратегия	$\pi_t^Q(q_t s_t^Q, \theta_Q)$ $f(\cdot)$ — Feature Regression Network: $\hat{y}_t = f(s_t^Q, q_t, a_t; \theta_f) = f(s_{t+1}^Q; \theta_f)$	$\pi_t^A(a_t s_t^A, \theta_A)$

Целью обучения стратегий является обучение параметров θ_Q , θ_A и θ_f .

RL награда

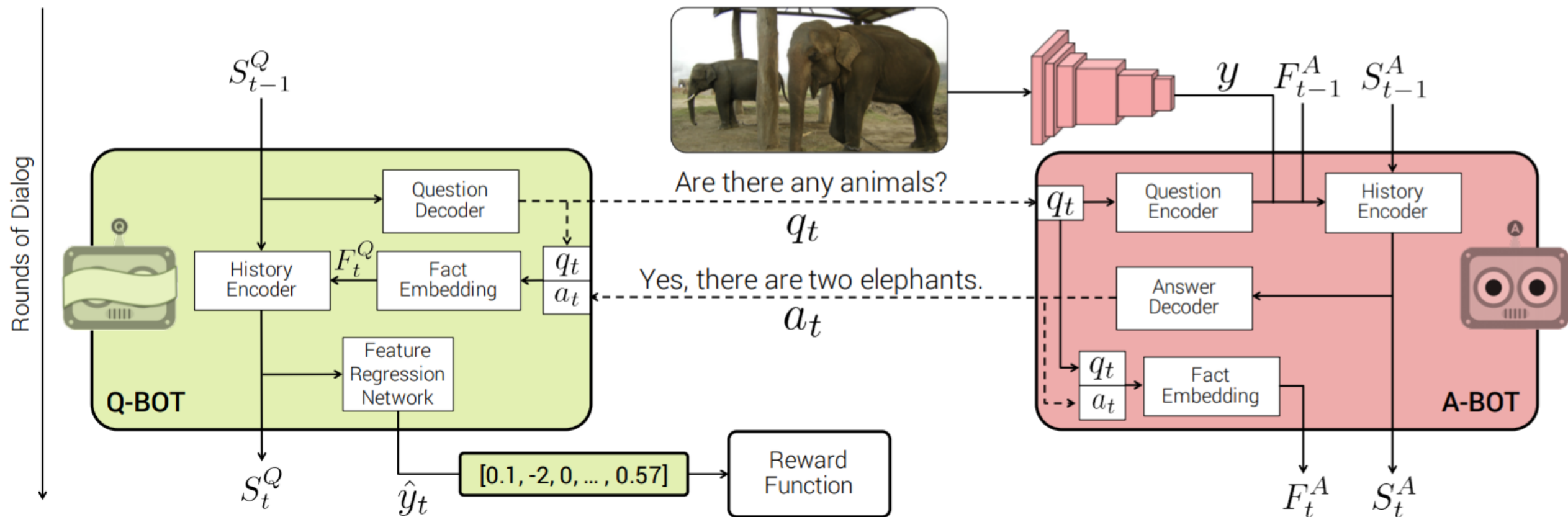
Поскольку игра кооперативная, то оба бота получают одинаковую награду.

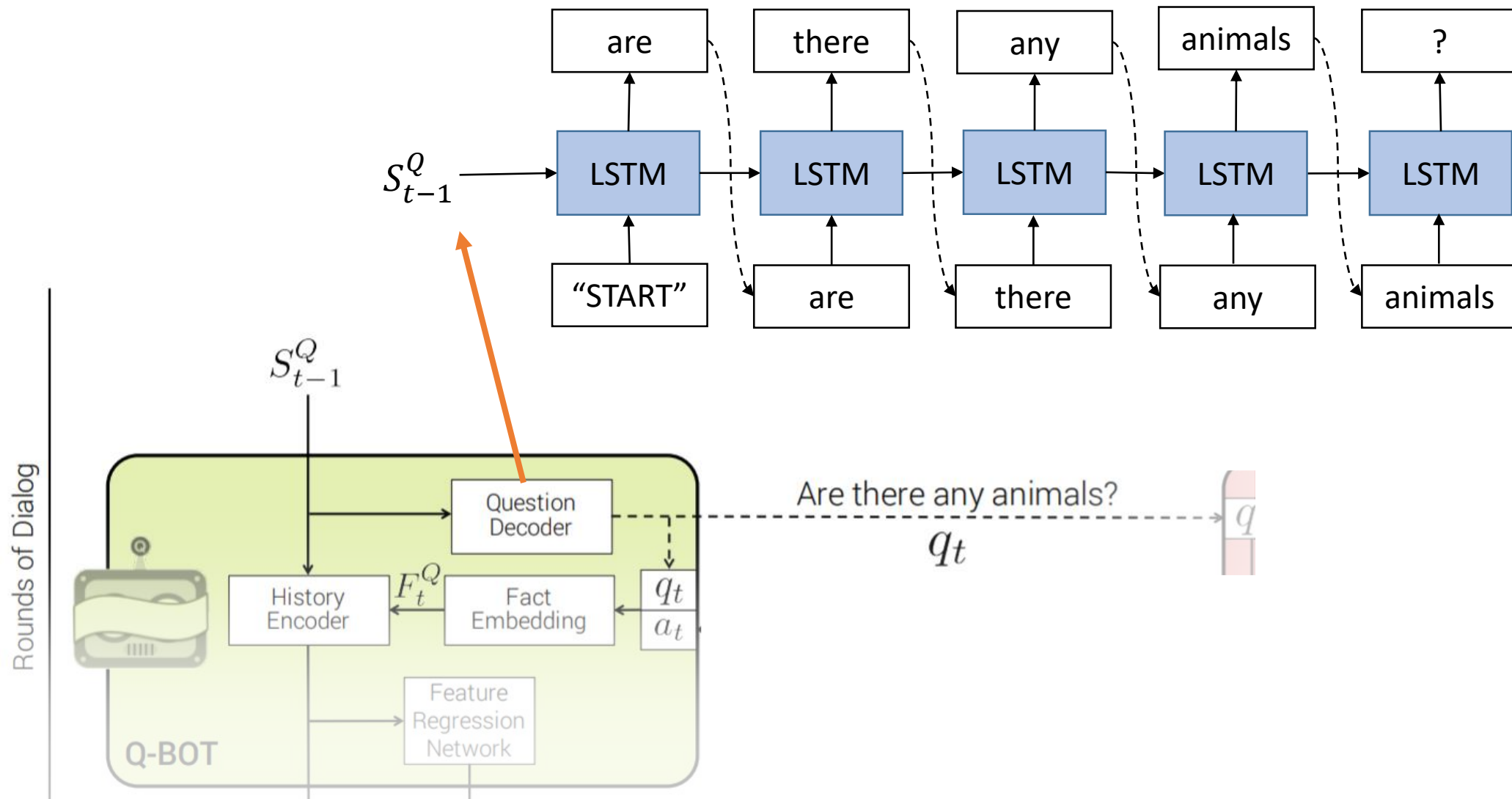
$$r_t \left(s_t^Q, (q_t, a_t, y_t) \right) = l(\hat{y}_{t-1}, y^{gt}) - l(\hat{y}_t, y^{gt})$$
$$l(x, y) = \text{dist}(x, y) = \| x - y \|_2^2$$

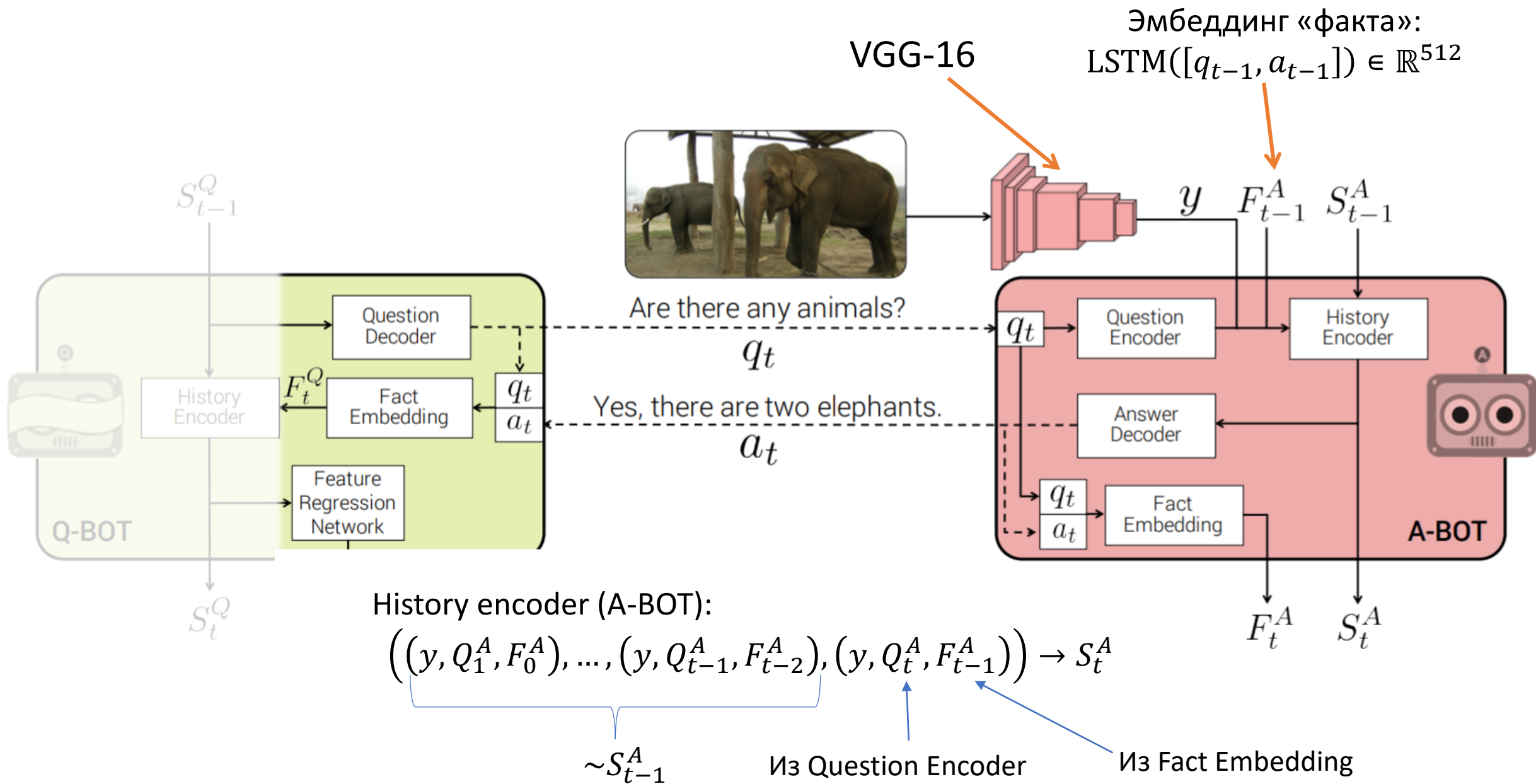
Суммарная награда:

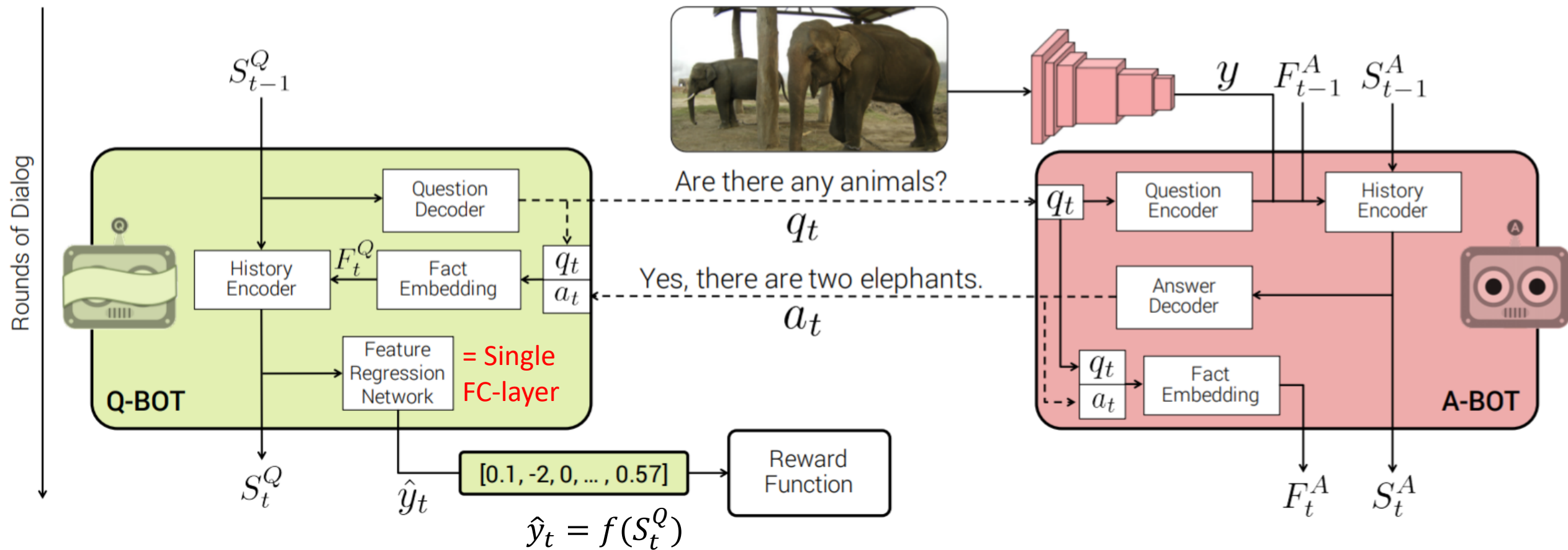
$$\sum_{t=1}^T r_t \left(s_t^Q, (q_t, a_t, y_t) \right) = l(\hat{y}_0, y^{gt}) - l(\hat{y}_T, y^{gt})$$

Архитектура









History Encoder (Q-BOT):

$$(F_1^Q, F_2^Q, \dots, F_t^Q) \rightarrow S_t^Q$$

Обучение

Глобальная цель:

$$\max_{\theta_Q, \theta_A, \theta_f} J(\theta_Q, \theta_A, \theta_f)$$
$$J(\theta_Q, \theta_A, \theta_f) = \mathbb{E}_{\pi_Q, \pi_A} \left[\sum_{t=1}^T r_t \left(s_t^Q, (q_t, a_t, y_t) \right) \right]$$

Для конкретного RL эпизода параметры обновляются согласно

$$J(\theta_Q, \theta_A, \theta_f) = \mathbb{E}_{\pi_Q, \pi_A} \left[r_t \left(s_t^Q, (q_t, a_t, y_t) \right) \right]$$

REINFORCE

Для параметров Q-BOT:

$$\begin{aligned}\nabla_{\theta_Q} J &= \nabla_{\theta_Q} \left[\mathbb{E}_{\pi_Q, \pi_A} r_t(\cdot) \right] = \nabla_{\theta_Q} \left[\sum_{q_t, a_t} \pi_Q(q_t | s_t^Q) \pi_A(q_t | s_t^A) r_t(\cdot) \right] = \\ &= \sum_{q_t, a_t} \pi_Q(q_t | s_t^Q) \nabla_{\theta_Q} \log \left(\pi_Q(q_t | s_t^Q) \right) \pi_A(q_t | s_t^A) r_t(\cdot) = \\ &= \mathbb{E}_{\pi_Q, \pi_A} r_t(\cdot) \nabla_{\theta_Q} \log \pi_Q(q_t | s_t^Q)\end{aligned}$$

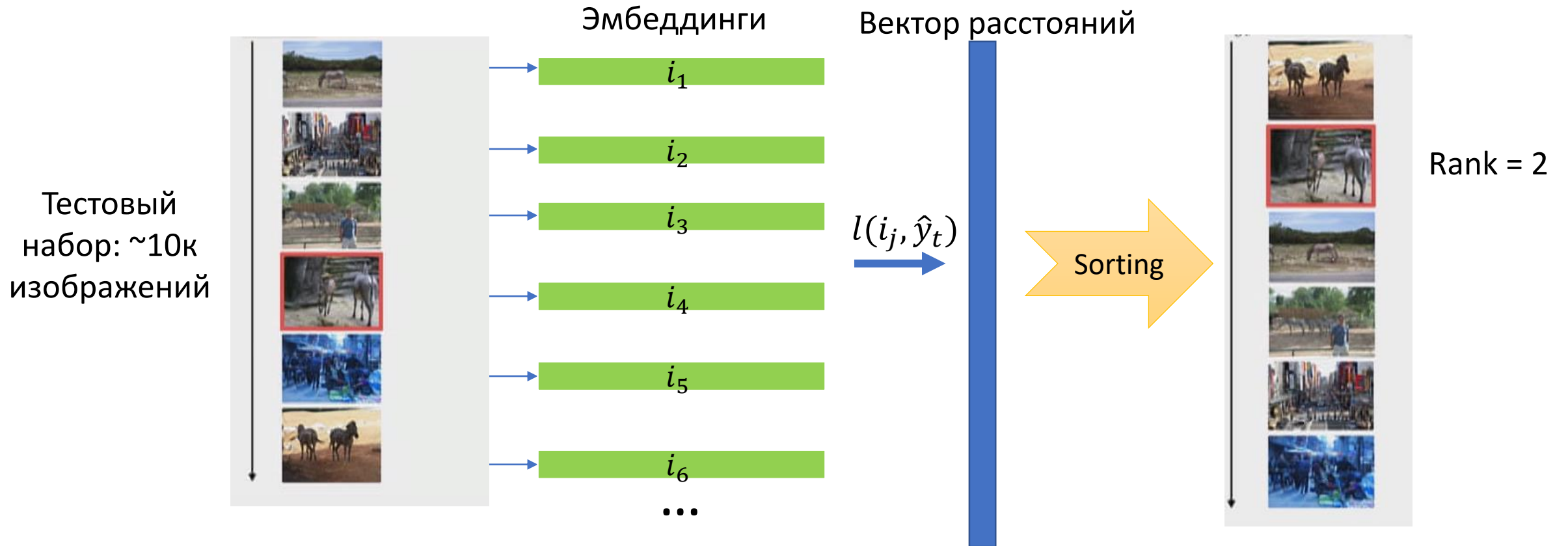
Аналогично для A-BOT:

$$\nabla_{\theta_A} J = \mathbb{E}_{\pi_Q, \pi_A} r_t(\cdot) \nabla_{\theta_A} \log \pi_A(a_t | s_t^A)$$

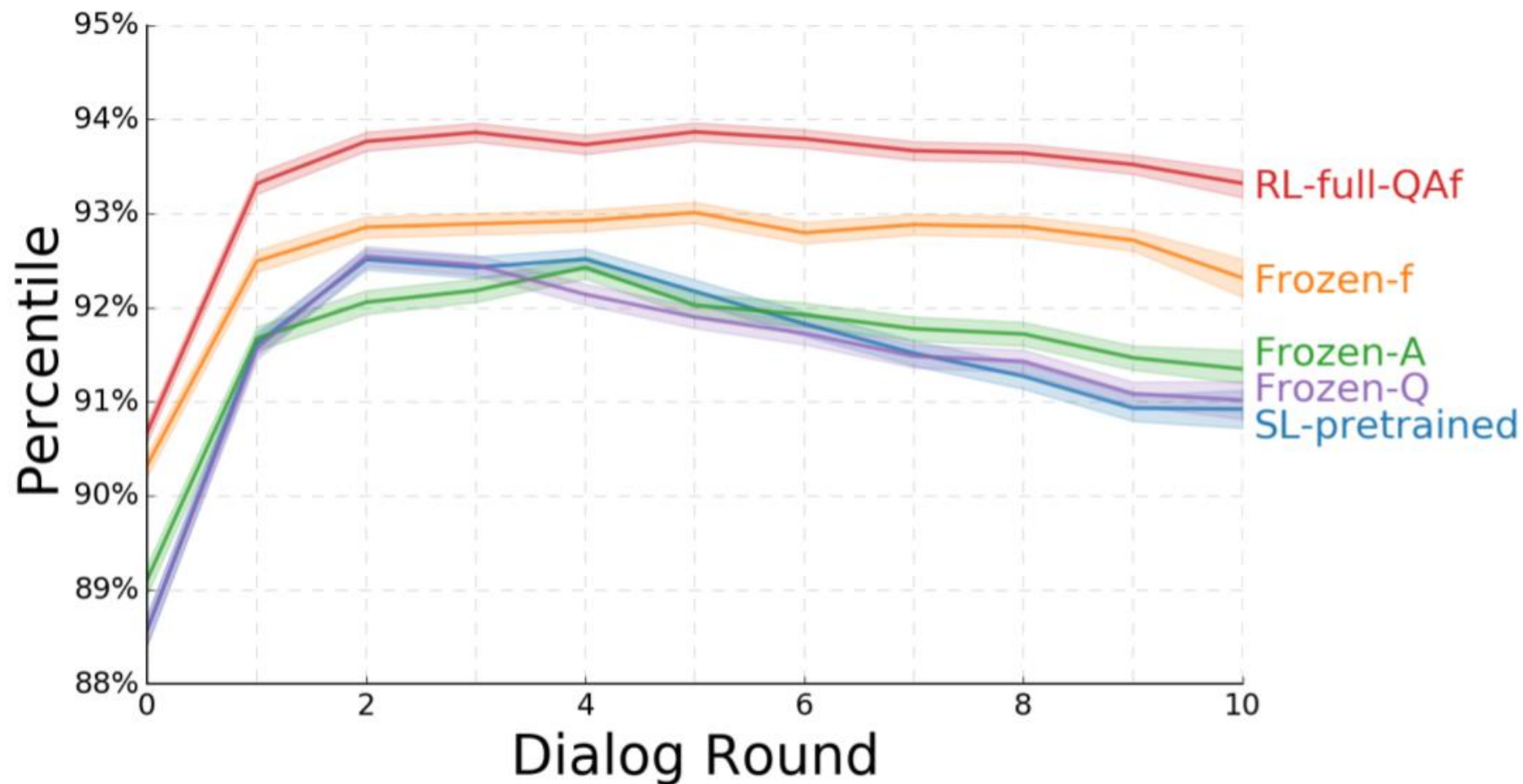
Особенности обучения

- Supervised pretraining для построения вопросов/ответов на естественном языке, а также функции f и VGG-16.
- Curriculum learning – сглаживание RL.

Guessing Game Evaluation



Результаты для Guessing Game





An airplane sitting on an asphalt landing strip.

SL

Q: What color is plane?

A: White

Q: Is it large or small?

A: Large

Q: Any people?

A: No

Q: Any writing on plane?

A: Yes

Q: Any writing on plane?

A: Yes

Q: Any writing on plane?

A: Yes

Q: What color is plane?

A: White

RL

Q: What color is airplane?

A: White with red and blue stripes

Q: Can you tell what airline it is?

A: No, it's too far away

Q: What time of day is it?

A: It is daytime

Q: Can you see any people?

A: No

Q: Can you see any buildings?

A: Yes, in background

Q: Is it day or night time?

A: It is daytime

Q: It is daytime

A: Yes

Результаты для Visual Dialog

Model	MRR	R@5	R@10	Mean Rank
SL-pretrain	0.436	53.41	60.09	21.83
Frozen-Q	0.428	53.12	60.19	21.52
Frozen-f	0.432	53.28	60.11	21.54
RL-full-QAf	0.428	53.08	60.22	21.54
Frozen-Q-multi	0.437	53.67	60.48	21.13

СПИСОК ИСТОЧНИКОВ

- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh: «Visual Dialog», CVPR 2017 <https://arxiv.org/pdf/1611.08669.pdf>
- Abhishek Das, Satwik Kottur: «Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning», ICCV 2017, <https://arxiv.org/pdf/1703.06585.pdf>
- <https://visualdialog.org> – официальный сайт проекта, есть работающая демо-версия бота