

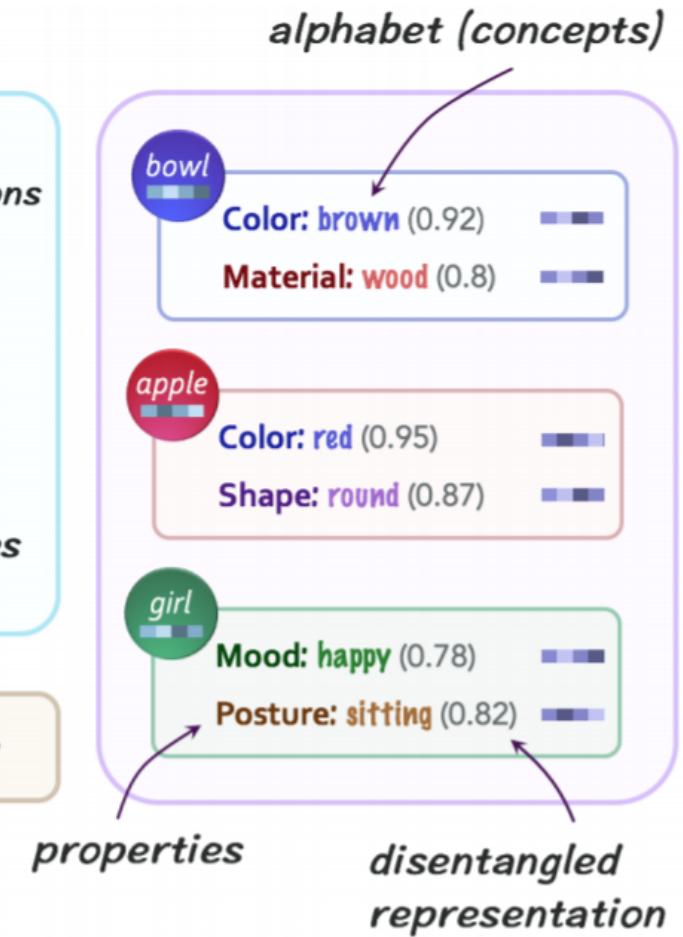
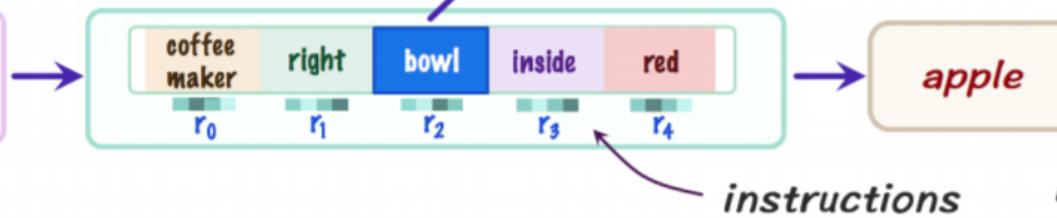
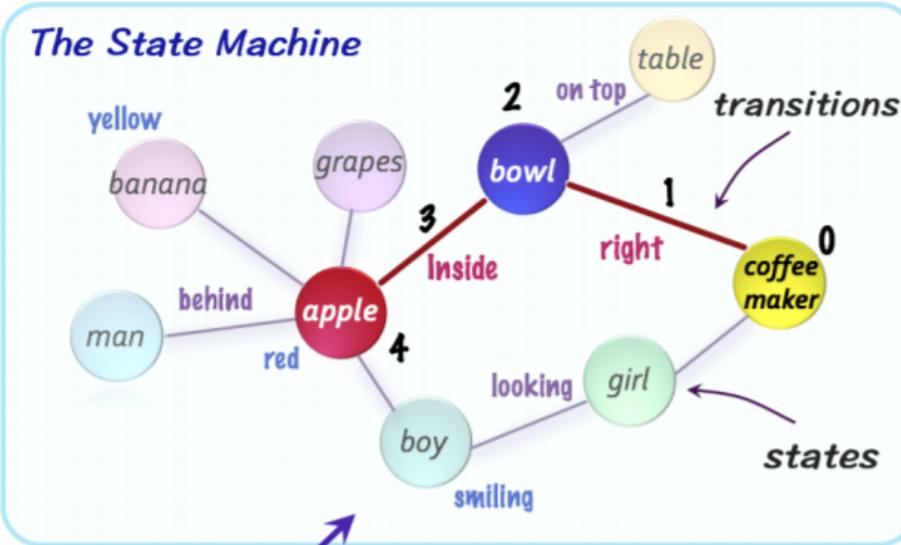
Learning by Abstraction: The Neural State Machine

Кудрявцева Софья

Нейронная машина состояний



What is the red fruit inside the bowl to the right of the coffee maker?



Словарь

Инициализирован GloVe и состоит из

Матрицы С:

- Объектов (кошка, тарелка) $C_O = C_0$
- Атрибутов объектов (цвета, материалы) $C_A = \bigcup_{i=1}^L C_i$
- Отношений между объектами (держит, справа от) $C_R = C_{L+1}$

Матрицы D:

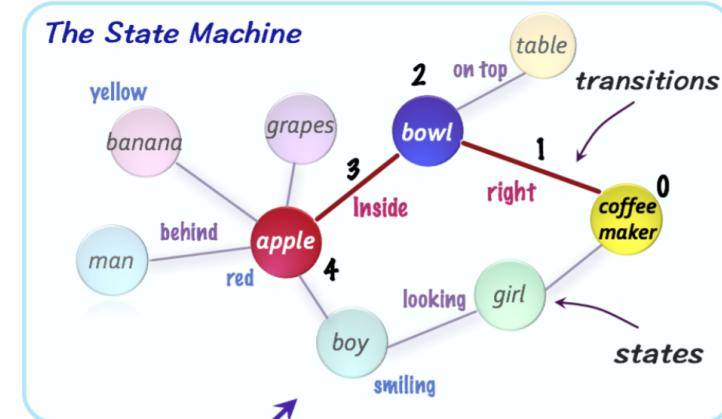
- 1(признак объекта) + L (типов объектов) + 1(признак отношения) = L + 2

Построение графа состояний

Узлы: Определение объектов и атрибутов объектов с помощью Mask R-CNN

$$s^j = \sum_{c_k \in C_j} P_j(k)c_k$$

Ребра:



- 1) Соединяются каждые 2 объекта, которые находятся друг от друга на расстоянии не больше 15 % изображения
- 2) Тип отношений у каждого ребра определяется graph attention network $e' = \sum_{c_k \in C_{L+1}} \dot{P}_{L+1}(k)c_k$

Перевод вопроса в инструкции

- Перевод каждого слова из вопроса в эмбеддинг
- Сравнение полученных эмбеддингов со словарем $P_i = \text{softmax}(w_i^T \mathbf{W} C)$
- Перевод каждого слова в концептуальное представление:

$$v_i = P_i(c') w_i + \sum_{c \in C \setminus \{c'\}} P_i(c) c$$

Перевод вопроса в инструкции

- Нормализованный вопрос проходит через LSTM attention-based encoder-decoder и превращается в набор инструкций

$$V^{P \times d} = \{v_i\}_{i=1}^P$$
$$r_i = \text{softmax}(h_i V^T) V$$

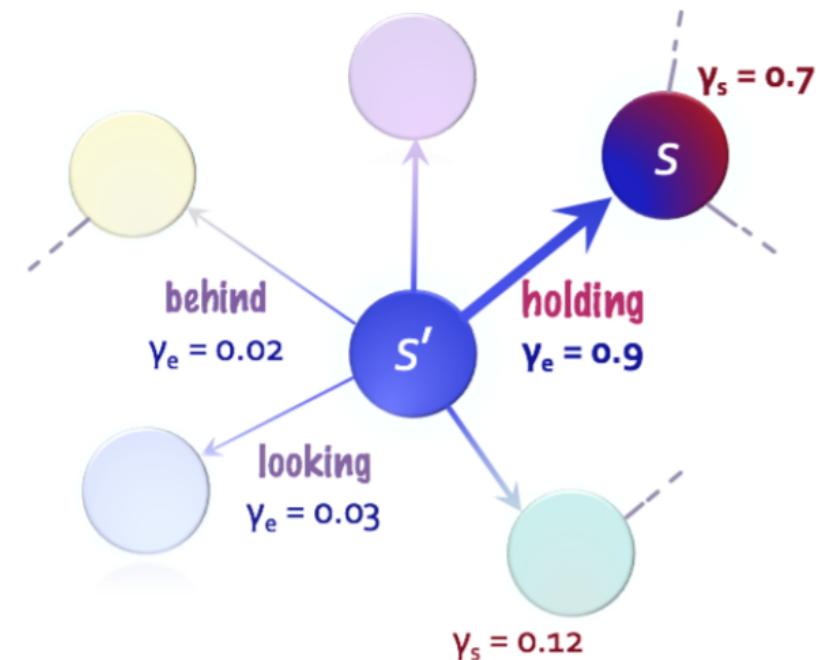
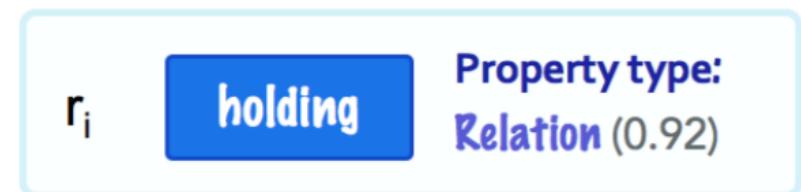
Проход инструкций по графу

- Функция перехода состояния $\delta_{S,E} : p_i \times r_i \rightarrow p_{i+1}$

каждом шаге i модуль принимает распределение r_i по состояниям в качестве входных данных и вычисляет обновленное распределение r_{i+1} , руководствуясь инструкцией r_i .

$$1) R_i = \text{softmax}(r_i^T \circ D)$$

Считаем вероятность принадлежности инструкции r_i к определенному свойству. D – эмбеддинги типов свойств из словаря.



Проход инструкций по графу

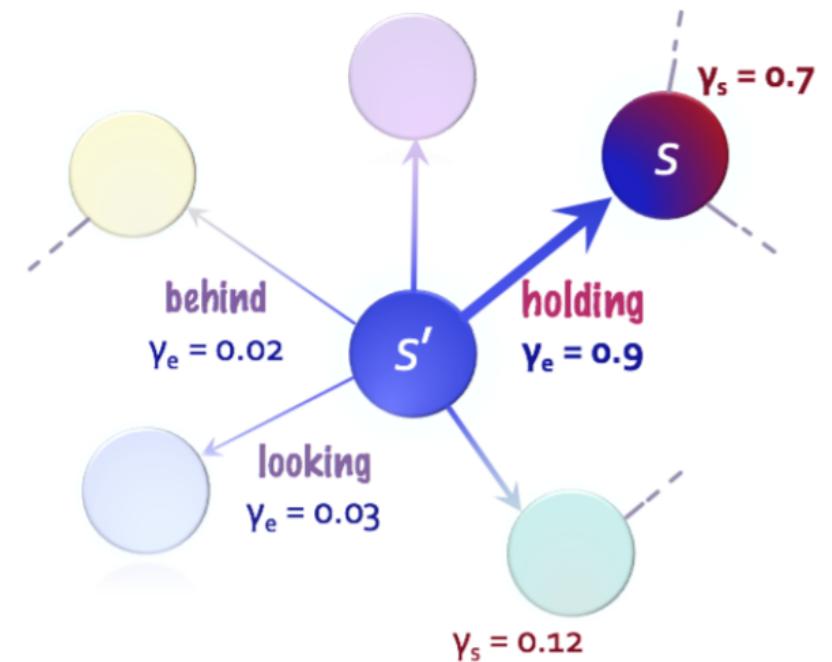
- Функция перехода состояния $\delta_{S,E} : p_i \times r_i \rightarrow p_{i+1}$

$$2) \quad \gamma_i(s) = \sigma \left(\sum_{j=0}^L R_i(j)(r_i \circ \mathbf{W}_j s^j) \right)$$

$$\gamma_i(e) = \sigma(r_i \circ \mathbf{W}_{L+1} e')$$

$\Upsilon(s)$ – оценка релевантности для узлов,

$\Upsilon(e)$ – оценка релевантности для ребер,



Проход инструкций по графу

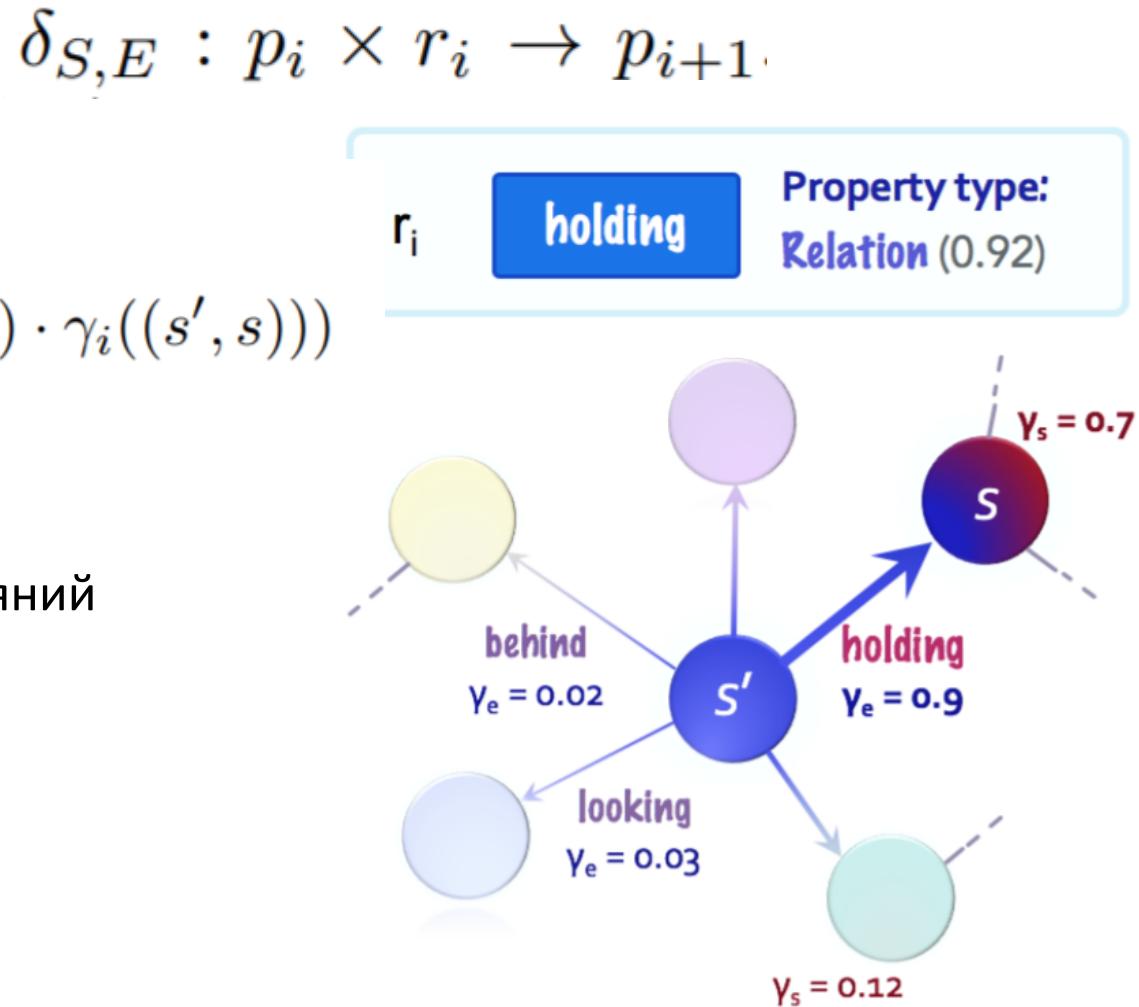
- Функция перехода состояния $\delta_{S,E} : p_i \times r_i \rightarrow p_{i+1}$

$$3) p_{i+1}^s = \text{softmax}_{s \in S}(\mathbf{W}_s \cdot \gamma_i(s))$$

$$p_{i+1}^r = \text{softmax}_{s \in S}(\mathbf{W}_r \cdot \sum_{(s',s) \in E} p_i(s') \cdot \gamma_i((s', s)))$$

$$p_{i+1} = r'_i \cdot p_{i+1}^r + (1 - r'_i) \cdot p_{i+1}^s$$

Смещаем внимание модели с текущих состояний
- s' на следующие - s .



Проход инструкций по графу

- Для получения ответа используется полносвязный двухслойный классификатор Softmax. На вход ему подается конкатенация вектора вопроса q , и дополнительный вектор m , агрегирующий информацию из состояний машины

$$m = \sum_{s \in S} p_N(s) \left(\sum_{j=0}^L R_N(j) \cdot s^j \right)$$

Обобщающая способность

Table 2: GQA ensemble

Model	Accuracy
Kakao*	73.33
270	70.23
NSM	67.25
LXRT	62.71
GRN	61.22
MSM	61.09
DREAM	60.93
SK T-Brain*	60.87
PKU	60.79
Musan	59.93

Table 3: VQA-CPv2

Model	Accuracy
SAN [86]	24.96
HAN [59]	28.65
GVQA [3]	31.30
RAMEN [73]	39.21
BAN [46]	39.31
MuRel [15]	39.54
ReGAT [52]	40.42
NSM	45.80

Table 4: GQA generalization

Model	Content	Structure
Global Prior	8.51	14.64
Local Prior	12.14	18.21
Vision	17.51	18.68
Language	21.14	32.88
Lang+Vis	24.95	36.51
BottomUp [5]	29.72	41.83
MAC [40]	31.12	47.27
NSM	40.24	55.72

Structure Generalization		Content Generalization	
training	testing	training	testing
What is the <obj> covered by ?	What is covering the <obj>?	Only questions that do not refer to any type of food or animal (do not include any word from these categories)	Only questions that refer to foods or animals (include a word from one of these categories)
Is there a <obj> in the image ?	Do you see any <obj>s in the photo ?		
What is the <obj> made of ?	What material makes up the <obj>?		
What's the name of the <obj> that is <attr>?	What is the <attr> <obj> called ?		



- 1) What is the **giraffe** looking at?
person ✓
- 2) Is the **fence** in front of the **giraffe** made of metal? **no** ✓
- 3) Is the **woman's shirt** blue or yellow? **blue** ✓
- 4) On which side of the image is the **person**? **right** ✓
- 5) Is there a **child** behind the **giraffe**? **no** ✗



- 1) What is the **fruit** to the right of the **salad**? **strawberries** ✓
- 2) Is the **fork** to the right of the **salad**? **no** ✓
- 3) Is the **plate** white and square?
no ✓
- 4) Is the **cup** behind the round **plate**?
yes ✓
- 5) What is the **plate** made of?
paper ✗



- 1) Are there either **scarves** or **hats** that are not pink? **no** ✓
- 2) Do the **bear's dress** and the **person's shirt** have the same color? **yes** ✓
- 3) Is the **bear** sitting or standing?
sitting ✓
- 4) What is the green **object** that the **bear** is sitting on? **book** ✓
- 5) Is the **bear** wearing white **shoes**?
yes ✗



- 1) Are there either a **chair** or a **clock** in the image? **no** ✓
- 2) Are there any **flowers** behind the **bed** on the left of the **room**? **yes** ✓
- 3) What color is the **appliance** on the right? **black** ✓
- 4) Is the **carpet** brown or blue?
brown ✓
- 5) Is the **TV** turned on? **yes** ✗