1. Summarization/ Captioning

(One turn: Captioning, Story telling, Simplification)

(Multiple turns: Dialog)

Deep Reinforcement Learning-based Image Captioning with Embedding Reward

Zhou Ren[1] Xiaoyu Wang[1] Ning Zhang[1] Xutao Lv[1] Li-Jia Li[2*]
[1]Snap Inc. [2]Google Inc.
{zhou.ren, xiaoyu.wang, ning.zhang, xutao.lv}@snap.com lijiali@cs.stanford.edu

Abstract

Image captioning is a challenging problem owing to the complexity in understanding the image content and diverse ways of describing it in natural language. Recent advances in deep neural networks have substantially improved the performance of this task. Most state-of-the-art approaches follow an encoder-decoder framework, which generates captions using a sequential recurrent prediction model. However, in this paper, we introduce a novel decision-making framework for image captioning. We utilize a "policy network" and a "value network" to collaboratively generate captions. The policy network serves as a local guidance by providing the confidence of predicting the next word according to the current state. Additionally, the value network serves as a global and lookahead guidance by evaluating all possible extensions of the current state. In essence, it adjusts the goal of predicting the correct words towards the goal of generating captions similar to the ground truth captions. We train both networks using an actor-critic reinforcement learning model, with a novel reward defined by visual-semantic embedding. Extensive experiments and analyses on the Microsoft COCO dataset show that the proposed framework outperforms state-of-the-art approaches across different evaluation metrics.

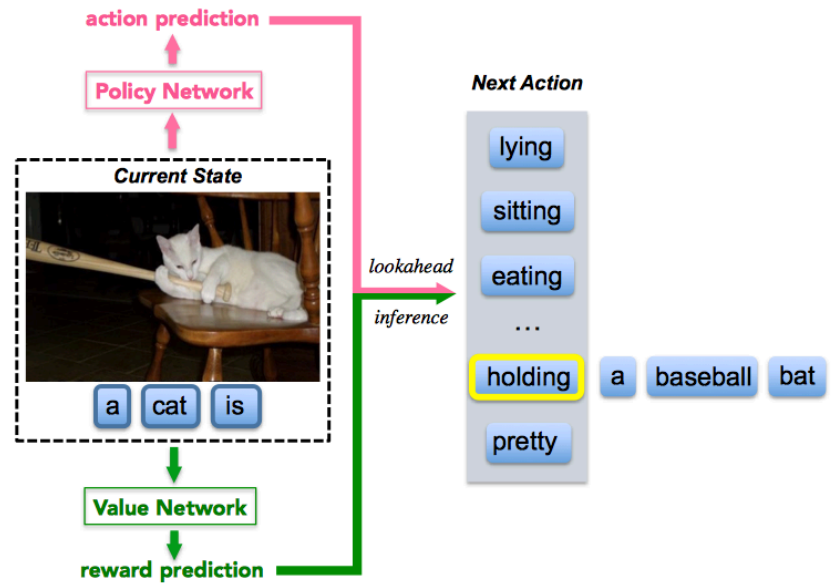


Figure 1. Illustration of the proposed decision-making framework. During our lookahead inference procedure, we utilize a "policy network" and a "value network" to collaboratively predict the word for each time step. The policy network provides an action prediction that locally predicts the next word according to current state. The value network provides a reward prediction that globally evaluates all possible extensions of the current state.

mation to coherent sentences. During training and inference, they try to maximize the probability of the next word based on recurrent hidden state.

Motivation

- Most state-of-the-art approaches follow an encoder-decoder framework, which generates captions using a sequential recurrent prediction model.
- However, in this paper, we introduce a novel decision-making framework for image captioning.
- The policy network, which provides the confidence of predicting the next word according to current state, serves as a local guidance.
- The value network, that evaluates the reward value of all possible extensions of the current state, serves as a global and lookahead guidance.

Policy Network

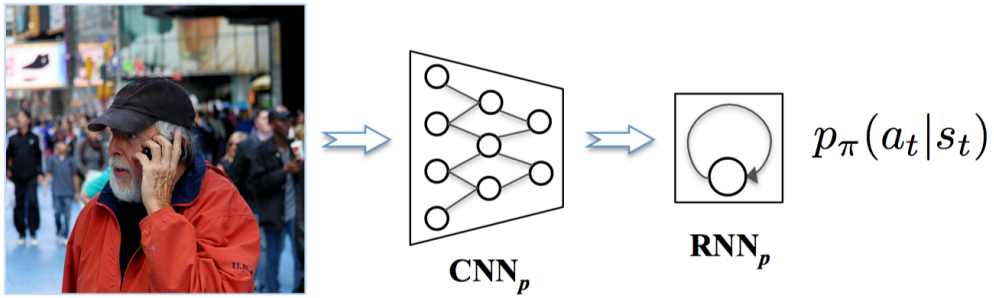


Figure 2. An illustration of our policy network p_π that is comprised of a CNN and a RNN. The CNN$_p$ output is fed as the initial input of RNN$_p$. The policy network computes the probability of executing an action a_t at a certain state s_t, by $p_\pi(a_t|s_t)$.

$$\boldsymbol{x}_0 = \boldsymbol{W}^{x,v}\mathrm{CNN}_p(\mathbf{I}) \tag{1}$$

$$\boldsymbol{h}_t = \mathrm{RNN}_p(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) \tag{2}$$

$$\boldsymbol{x}_t = \phi(w_{t-1}), \quad t > 0 \tag{3}$$

$$p_\pi(a_t|s_t) = \varphi(\boldsymbol{h}_t) \tag{4}$$

Value network

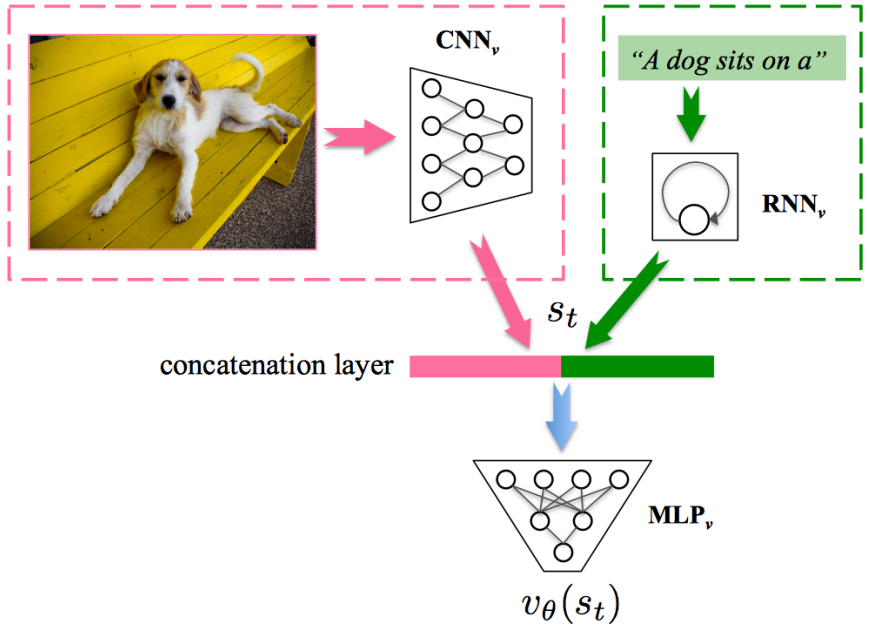


Figure 3. An illustration of our value network v_θ that is comprised of a CNN, a RNN and a MLP. Given a state s_t which contains raw image input $\mathbf{I}$ and a partially generated raw sentence until t, the value network $v_\theta(s_t)$ evaluates its value.

$$v^p(s) = \mathbb{E}[r|s_t = s, \ \ a_{t...T} \sim p] \qquad (5)$$

Reward defined by visual-semantic embedding

Visual-semantic embedding has been successfully applied to image classification [11, 37], retrieval [19, 36, 33], *etc*. Our embedding model is comprised of a CNN, a RNN and a linear mapping layer, denoted as CNN_e, RNN_e and f_e. By learning the mapping of images and sentences into one semantic embedding space, it provides a measure of similarity between images and sentences. Given a sentence S, its embedding feature is represented using the last hidden state of RNN_e, *i.e.*, $\boldsymbol{h}'_T(S)$. Let $\boldsymbol{v}$ denote the feature vector of image $\mathbf{I}$ extracted by CNN_e, and $f_e(\cdot)$ is the mapping function from image features to the embedding space. We train the embedding model using the same image-sentence pairs as in image captioning. We fix the CNN_e weight, and learn the RNN_e weights as well as $f_e(\cdot)$ using a bi-directional ranking loss defined as follows:

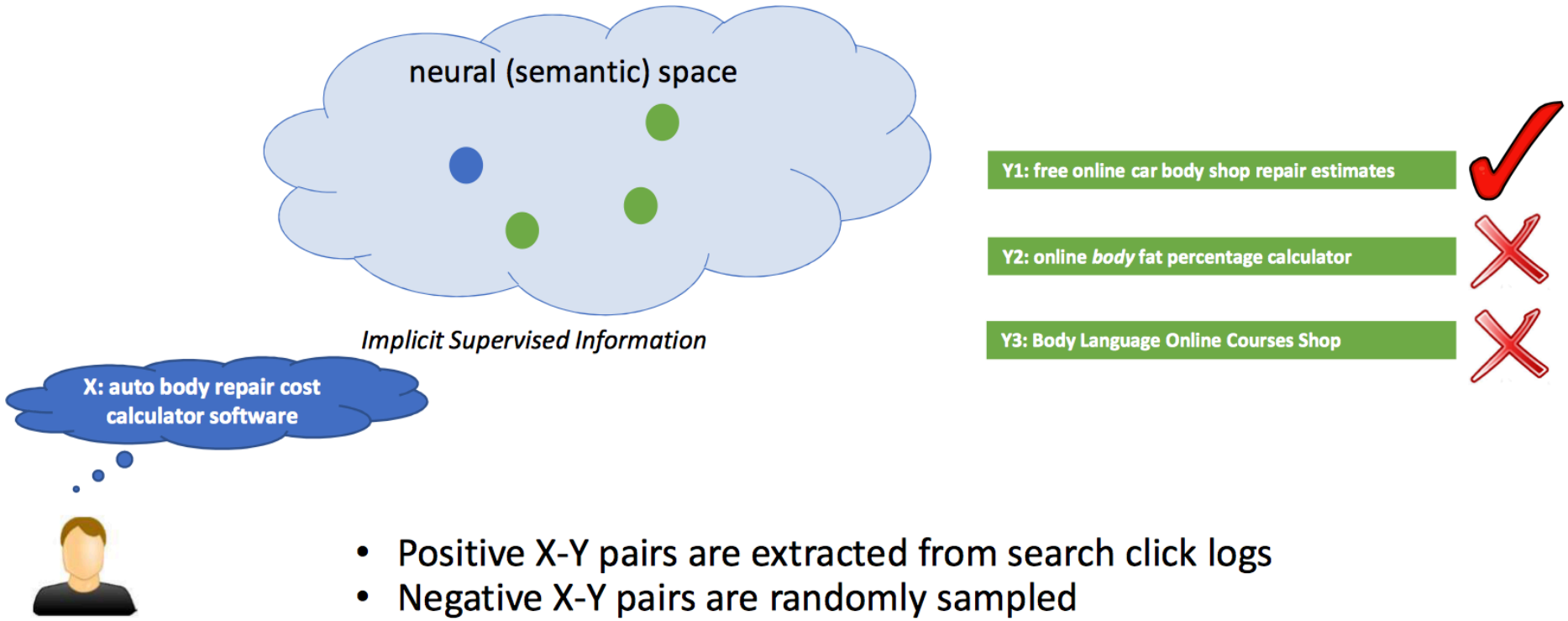
$$L_e = \sum_{\boldsymbol{v}} \sum_{S^-} \max(0,\ \beta - f_e(\boldsymbol{v}) \cdot \boldsymbol{h}'_T(S) + f_e(\boldsymbol{v}) \cdot \boldsymbol{h}'_T(S^-)) + \sum_{S} \sum_{\boldsymbol{v}^-} \max(0,\ \beta - \boldsymbol{h}'_T(S) \cdot f_e(\boldsymbol{v}) + \boldsymbol{h}'_T(S) \cdot f_e(\boldsymbol{v}^-)) \quad (6)$$

where β is the margin cross-validated, every $(\boldsymbol{v}, S)$ are a ground truth image-sentence pair, S^- denotes a negative description for the image corresponding to $\boldsymbol{v}$, and vice-versa with $\boldsymbol{v}^-$.

Given an image with feature $\boldsymbol{v}^*$, we define the reward of a generated sentence $\widehat{S}$ to be the embedding similarity between $\widehat{S}$ and $\boldsymbol{v}^*$:

$$r = \frac{f_e(\boldsymbol{v}^*) \cdot \boldsymbol{h}'_T(\widehat{S})}{\|f_e(\boldsymbol{v}^*)\| \|\boldsymbol{h}'_T(\widehat{S})\|} \quad (7)$$

Learning DSSM from Labeled X-Y Pairs

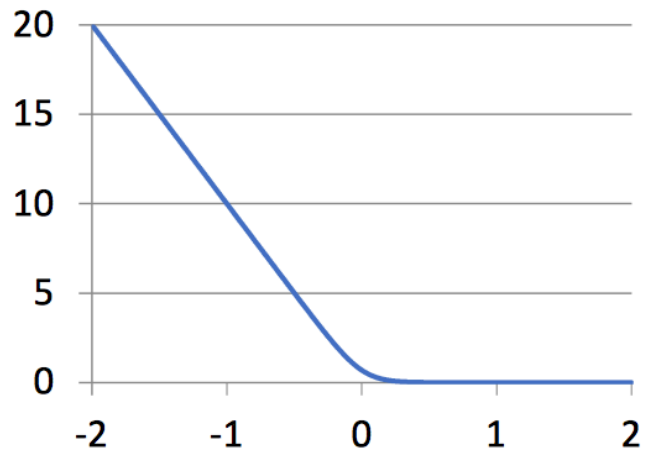


- Positive X-Y pairs are extracted from search click logs
- Negative X-Y pairs are randomly sampled
- Map X and Y into the same semantic space via deep neural net
- Positive Y are closer to X than negative Y in that space

Learning DSSM from Labeled X-Y Pairs

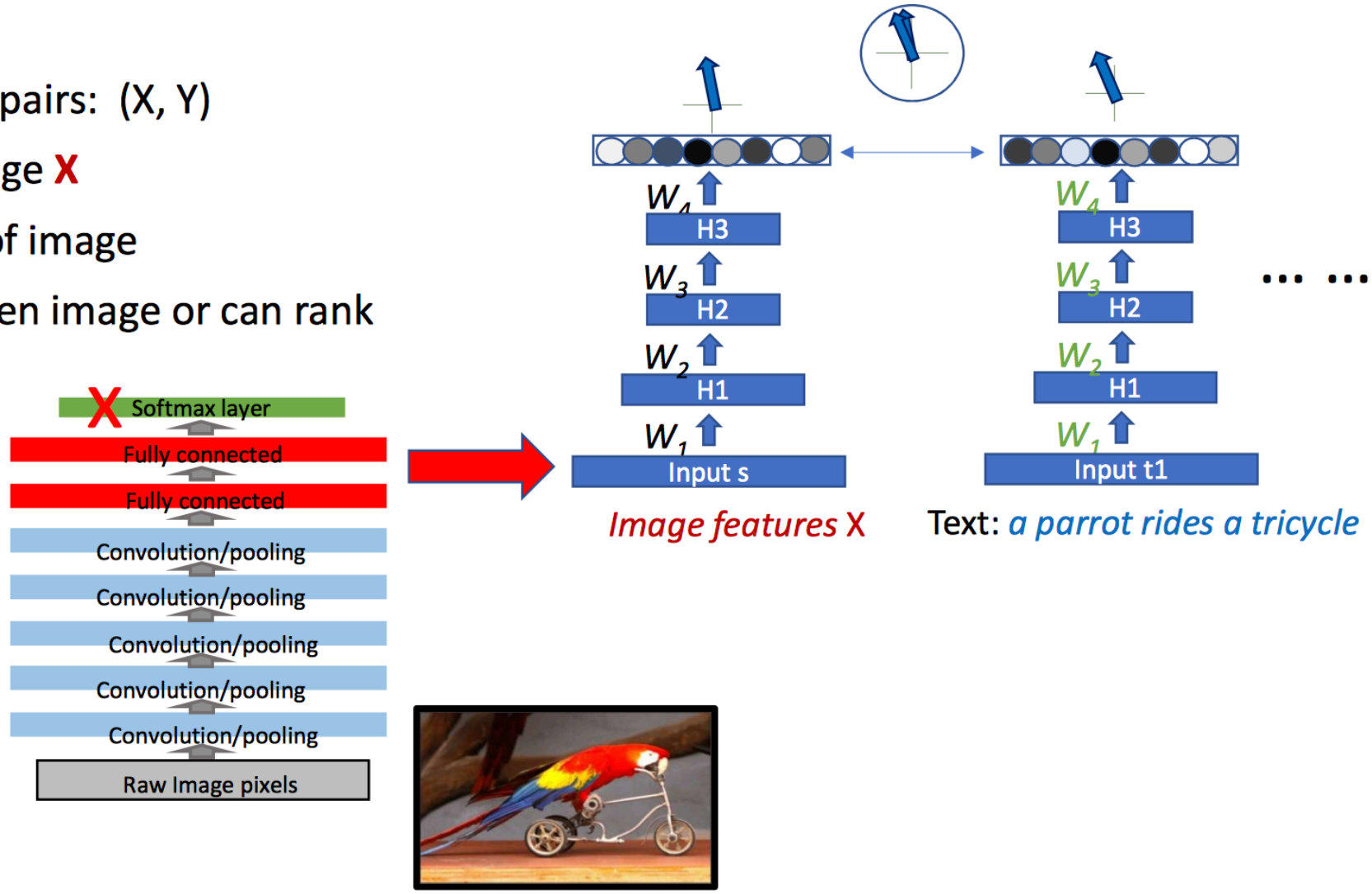
- Consider a query X and two docs Y^+ and Y^-
  - Assume Y^+ is more relevant than Y^- with respect to X
- $\text{sim}_{\boldsymbol{\theta}}(X, Y)$ is the cosine similarity of X and Y in semantic space, mapped by DSSM parameterized by $\boldsymbol{\theta}$

- $\Delta = \text{sim}_{\boldsymbol{\theta}}(X, Y^+) - \text{sim}_{\boldsymbol{\theta}}(X, Y^-)$
  - We want to maximize Δ
- $Loss(\Delta; \boldsymbol{\theta}) = \log(1 + \exp(-\gamma\Delta))$
- Optimize $\boldsymbol{\theta}$ using mini-batch SGD on GPU



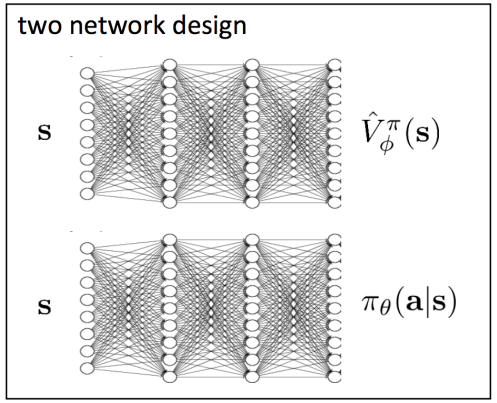
Go beyond text:
DSSM for multi-modal representation learning

- Recall DSSM for text input pairs: (X, Y)
- Now: replace text X by image **X**
- Using DNN/CNN features of image
- Can rank/generate text given image or can rank images given text.

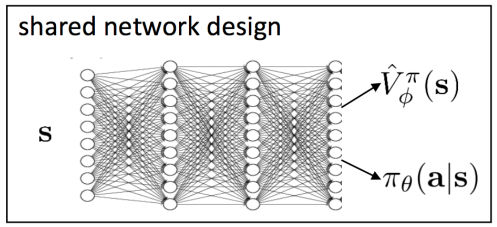


[Fang+ 15]

Value (critic) network architecture



+ simple & stable
- no shared features between actor & critic



4.4. Value network architecture analysis

In this paper we propose a novel framework that involves value network, whose architecture is worth looking into. As in Figure 3, we use CNN_v and RNN_v to extract visual and semantic information from the raw image and sentence inputs. Since the hidden state in policy network at each time step is a representation of each state as well, a natural question is "can we directly utilize the policy hidden state?". To answer this question, we construct two variants of our value network: the first one, named as (**hid-VN**), is comprised of a MLP_v on top of the policy hidden state of RNN_p; the second variant, (**hid-Im-VN**), is comprised of a MLP_v on top of the concatenation of the policy hidden state of RNN_p and the visual input $\boldsymbol{x}_0$ of policy RNN_p. The results are shown in Table 2. As we see, both variants that utilize policy hidden state do not work well, comparing to our **Full-model**. The problem of the policy hidden state is that it compresses and also loses lots of information. Thus, it is reasonable and better to train independent CNN, RNN for value network it-

A DEEP REINFORCED MODEL FOR ABSTRACTIVE SUMMARIZATION

Romain Paulus, Caiming Xiong*& Richard Socher
Salesforce Research
575 High Street
Palo Alto, CA 94301, USA
{rpaulus,cxiong,rsocher}@salesforce.com

ABSTRACT

Attentional, RNN-based encoder-decoder models for abstractive summarization have achieved good performance on short input and output sequences. For longer documents and summaries however these models often include repetitive and incoherent phrases. We introduce a neural network model with a novel intra-attention that attends over the input and continuously generated output separately, and a new training method that combines standard supervised word prediction and reinforcement learning (RL). Models trained only with supervised learning often exhibit "exposure bias" – they assume ground truth is provided at each step during training. However, when standard word prediction is combined with the global sequence prediction training of RL the resulting summaries become more readable. We evaluate this model on the CNN/Daily Mail and New York Times datasets. Our model obtains a 41.16 ROUGE-1 score on the CNN/Daily Mail dataset, an improvement over previous state-of-the-art models. Human evaluation also shows that our model produces higher quality summaries.

Motivation

- First, extractive summarization sys- tems form summaries by copying parts of the input (Dorr et al., 2003; Nallapati et al., 2017).
- Second, abstractive summarization systems generate new phrases, possibly rephrasing or using words that were not in the original text (Chopra et al., 2016; Nallapati et al., 2016).
- Attentional, RNN-based encoder-decoder models for abstractive summarization have achieved good performance on short input and output sequences.
- For longer documents and summaries however these models often include repetitive and incoherent phrases.

Framework

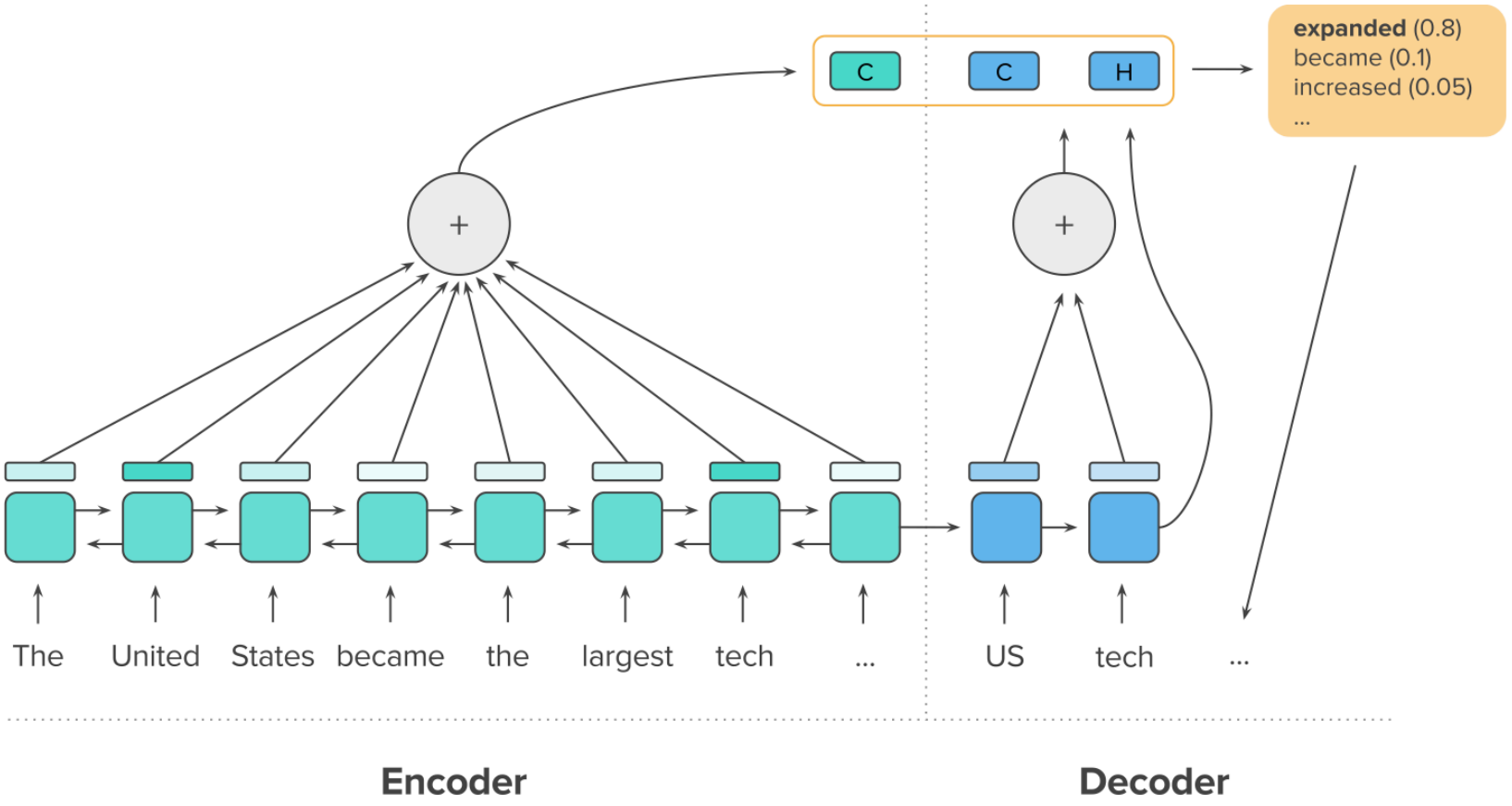


Figure 1: Illustration of the encoder and decoder attention functions combined. The two context vectors (marked "C") are computed from attending over the encoder hidden states and decoder hidden states. Using these two contexts and the current decoder hidden state ("H"), a new word is generated and added to the output sequence.

NEURALINTRA - ATTENTION MODEL

- INTRA - TEMPORAL ATTENTION ON INPUT SEQUENCE

$$e_{ti} = f(h_t^d, h_i^e), \qquad f(h_t^d, h_i^e) = h_t^{dT} W_{\text{attn}}^e h_i^e. \qquad e'_{ti} = \begin{cases} exp(e_{ti}) & \text{if } t = 1 \\ \frac{exp(e_{ti})}{\sum_{j=1}^{t-1} \exp(e_{ji})} & \text{otherwise.} \end{cases}$$

$$\alpha_{ti}^e = \frac{e'_{ti}}{\sum_{j=1}^{n} e'_{tj}} \qquad (4) \qquad c_t^e = \sum_{i=1}^{n} \alpha_{ti}^e h_i^e. \qquad (5)$$

- INTRA - DECODER ATTENTION

$$e_{tt'}^d = h_t^{dT} W_{\text{attn}}^d h_{t'}^d \qquad (6) \qquad \alpha_{tt'}^d = \frac{exp(e_{tt'}^d)}{\sum_{j=1}^{t-1} exp(e_{tj}^d)} \qquad (7) \qquad c_t^d = \sum_{j=1}^{t-1} \alpha_{tj}^d h_j^d \qquad (8)$$

TOKEN GENERATION AND POINTER

- We define u_t as a binary value, equal to 1 if the pointer mechanism is used to output y_t , and 0 otherwise.

$$p(u_t = 1) = \sigma(W_u[h_t^d \| c_t^e \| c_t^d] + b_u),$$

$$p(y_t | u_t = 0) = \text{softmax}(W_{\text{out}}[h_t^d \| c_t^e \| c_t^d] + b_{\text{out}})$$

$$p(y_t = x_i | u_t = 1) = \alpha_{ti}^e$$

$$p(y_t) = p(u_t = 1)p(y_t | u_t = 1) + p(u_t = 0)p(y_t | u_t = 0).$$

HYBRID LEARNING OBJECTIVE

$$L_{ml} = -\sum_{t=1}^{n'} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*, x)$$

For this training algorithm, we produce two separate output sequences at each training iteration: y^s, which is obtained by sampling from the $p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x)$ probability distribution at each decoding time step, and $\hat{y}$, the baseline output, obtained by maximizing the output probability distribution at each time step, essentially performing a greedy search. We define $r(y)$ as the reward function for an output sequence y, comparing it with the ground truth sequence y^* with the evaluation metric of our choice.

$$L_{rl} = (r(\hat{y}) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x)$$

$$L_{mixed} = \gamma L_{rl} + (1 - \gamma) L_{ml},$$

2. Text QA (Machine Reading Comprehension)

TODO

Open-Domain Question Answering (QA)

Will I qualify for OSAP if I'm new in Canada?

Selected Passages from Bing

"Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free."

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/

"To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD)."

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/

"You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students."

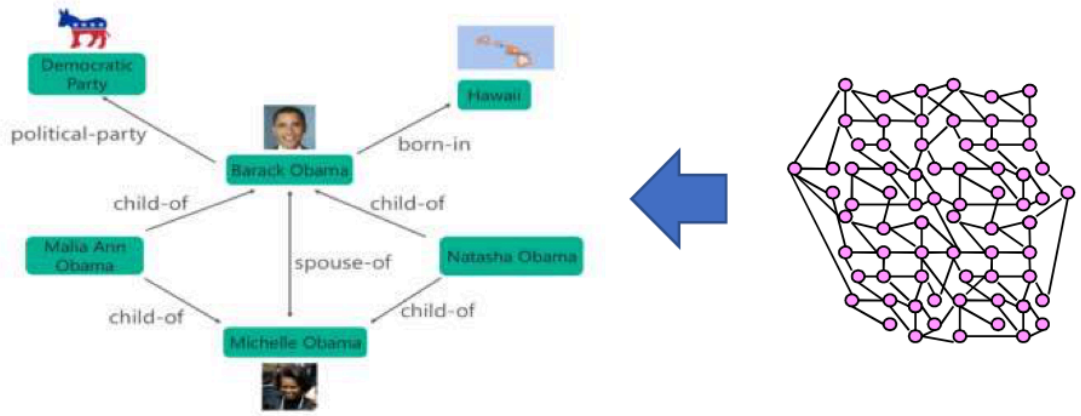
Source: http://www.campusaccess.com/financial-aid/osap.html

Answer
No. You won't qualify.

Text-QA

What is Obama's citizenship?

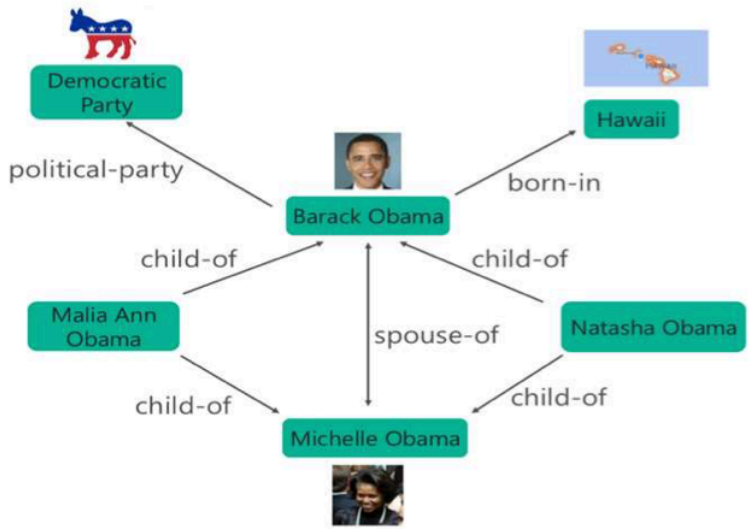
Selected subgraph from Microsoft's Satori



Answer
USA

Knowledge Base (KB)-QA

Question Answering (QA) on Knowledge Base



Large-scale knowledge graphs

- Properties of billions of entities
- Plus relations among them

An QA Example:

Question: `what is Obama's citizenship?`

- Query parsing:
  `(Obama, Citizenship,?)`
- Identify and infer over relevant subgraphs:
  `(Obama, BornIn, Hawaii)`
  `(Hawaii, PartOf, USA)`
- correlating semantically relevant relations:
  `BornIn ~ Citizenship`

Answer: `USA`

Symbolic approaches to QA

- Understand the question via **semantic parsing**
  - Input: what is Obama's citizenship?
  - Output (LF): (Obama, Citizenship,?)
- Collect relevant information via fuzzy **keyword matching**
  - (Obama, BornIn, Hawaii)
  - (Hawaii, PartOf, USA)
  - Needs to know that BornIn and Citizenship are semantically related
- Generate the answer via **reasoning**
  - (Obama, Citizenship, **USA**)
- **Challenges**
  - **Paraphrasing in NL**
  - **Search complexity of a big KG**

[Richardson+ 98; Berant+ 13; Yao+ 15; Bao+ 14; Yih+ 15; etc.]

Neural MRC Models on SQuAD

What types of European groups were able to avoid the plague?

From Italy, the disease spread northwest across Europe, striking France, Spain, Portugal and England by June 1348, then turned and spread east through Germany and Scandinavia from 1348 to 1350. It was introduced in Norway in 1349 when a ship landed at Askøy, then spread to Bjørgvin (modern Bergen) and Iceland. Finally it spread to northwestern Russia in 1351. The plague was somewhat less common in parts of Europe that had smaller trade relations with their neighbours, including the Kingdom of Poland, the majority of the Basque Country, isolated parts of Belgium and the Netherlands, and isolated alpine villages throughout the continent.

A limited form of comprehension:

- No need for extra knowledge outside the paragraph
- No need for clarifying questions
- The answer must exist in the paragraph
- The answer must be a text span, not synthesized

- Encoding: map each text span to a semantic vector
- Reasoning: rank and re-rank semantic vectors
- Decoding: map the top-ranked vector to text

3. Text Dialog

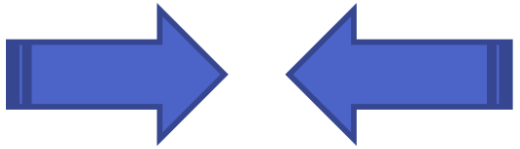
Two Branches of Bots

Task-Oriented Bot

- Personal assistant, helps users achieve a certain task
- Combination of rules and statistical components
  - POMDP for spoken dialog systems (Williams and Young, 2007)
  - End-to-end trainable task-oriented dialogue system (Wen et al., 2016)
  - End-to-end reinforcement learning dialogue system (Li et al., 2017; Zhao and Eskenazi, 2016)

Chit-Chat Bot

- No specific goal, focus on natural responses
- Using variants of seq2seq model
  - A neural conversation model (Vinyals and Le, 2015)
  - Reinforcement learning for dialogue generation (Li et al., 2016)
  - Conversational contextual cues for response ranking (Al-Rfou et al., 2016)

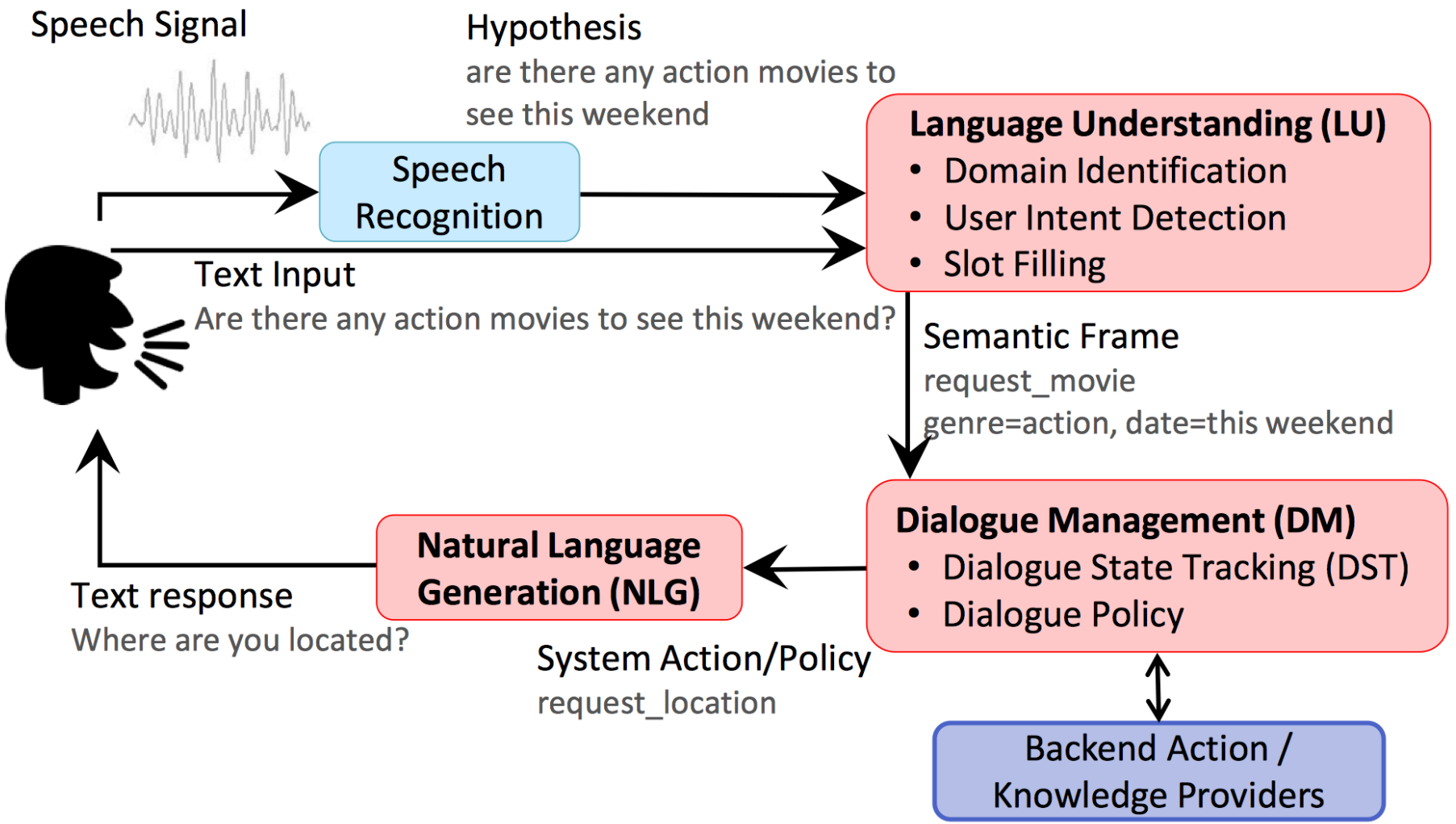


Modality

- Text QA -> Dialog
- VQA -> Visual Dialog
- Video QA -> Video Dialog

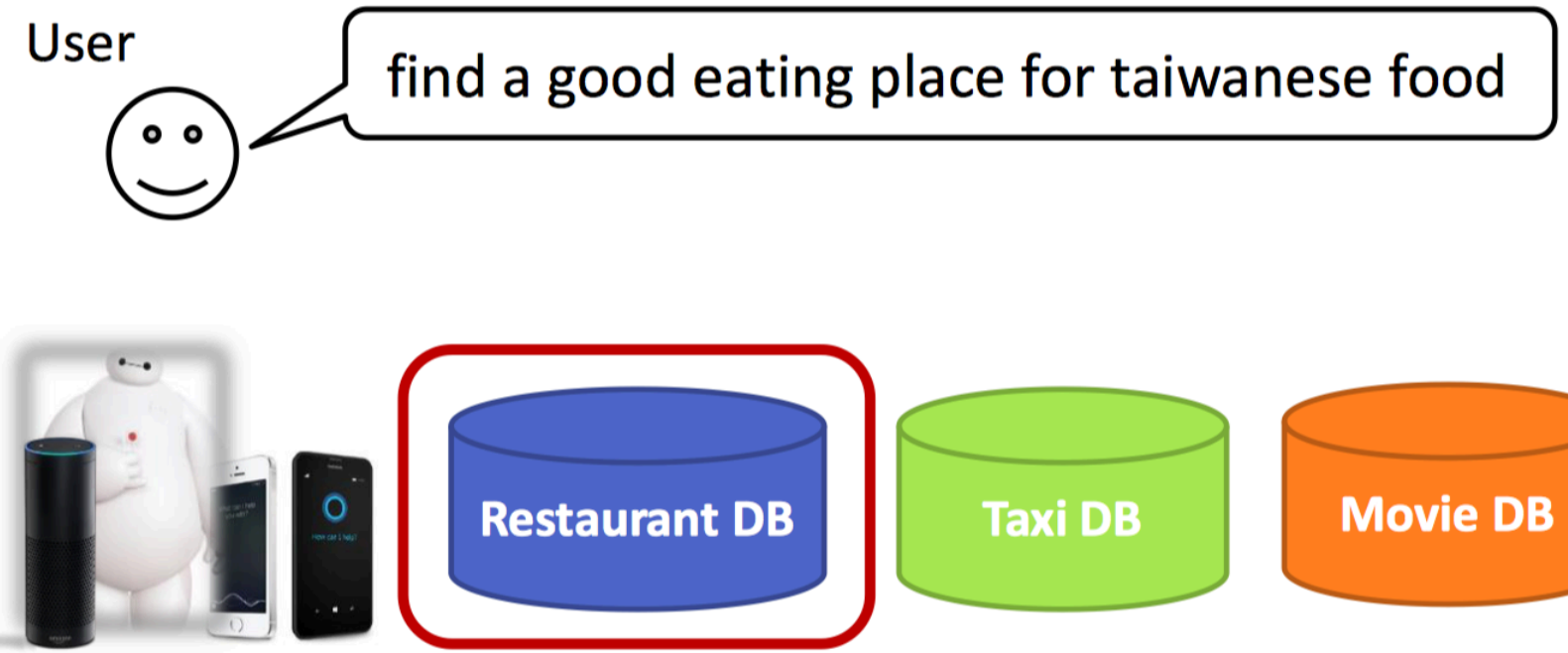
Task-Oriented Dialogue System (Young, 2000)

http://rsta.royalsocietypublishing.org/content/358/1769/1389.short



1. Domain Identification

Requires Predefined Domain Ontology



2. Intent Detection

Requires Predefined Schema

User

find a good eating place for taiwanese food

Intelligent Agent

Restaurant DB

FIND_RESTAURANT
FIND_PRICE
FIND_TYPE
:

Classification!

3. Slot Filling

Requires Predefined Schema

O O B-rating O O O B-type O

User

find a good eating place for taiwanese food

Intelligent Agent

Restaurant DB

Restaurant	Rating	Type
Rest 1	good	Taiwanese
Rest 2	bad	Thai
:	:	:

FIND_RESTAURANT
rating="good"
type="taiwanese"

Semantic Frame

SELECT restaurant {
rest.rating="good"
rest.type="taiwanese"
}

Sequence Labeling

State Tracking

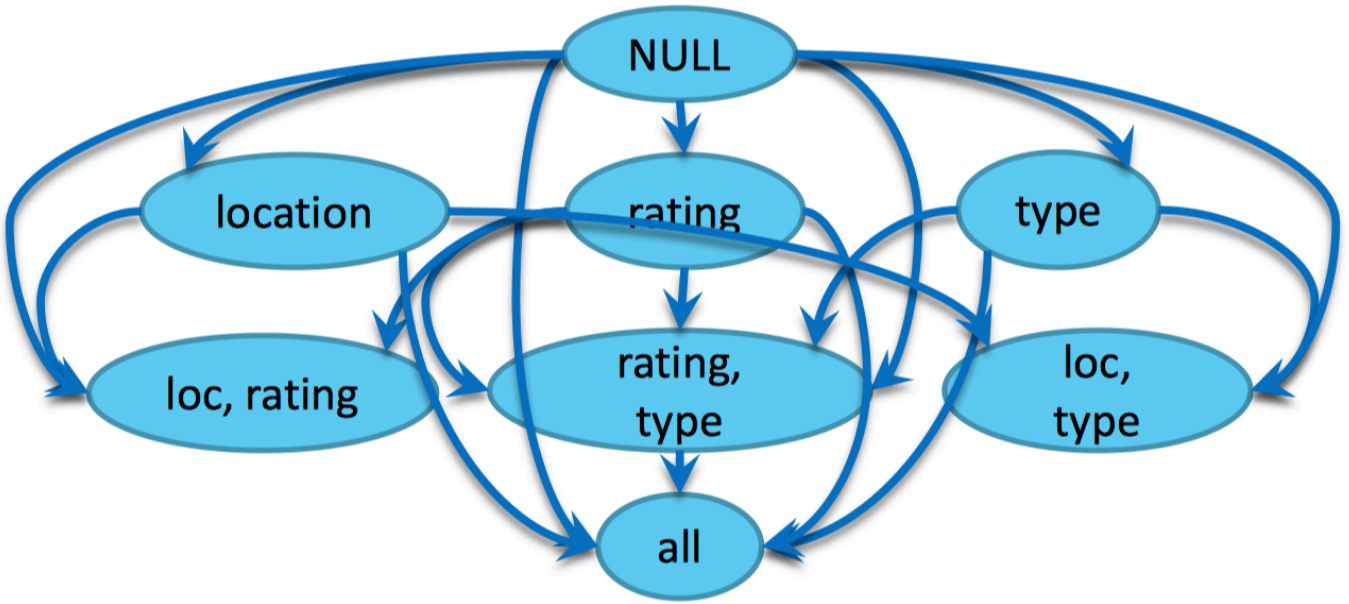
Requires Hand-Crafted States

User

find a good eating place for taiwanese food

i want it near to my office

Intelligent Agent



State Tracking

Handling Errors and Confidence

User

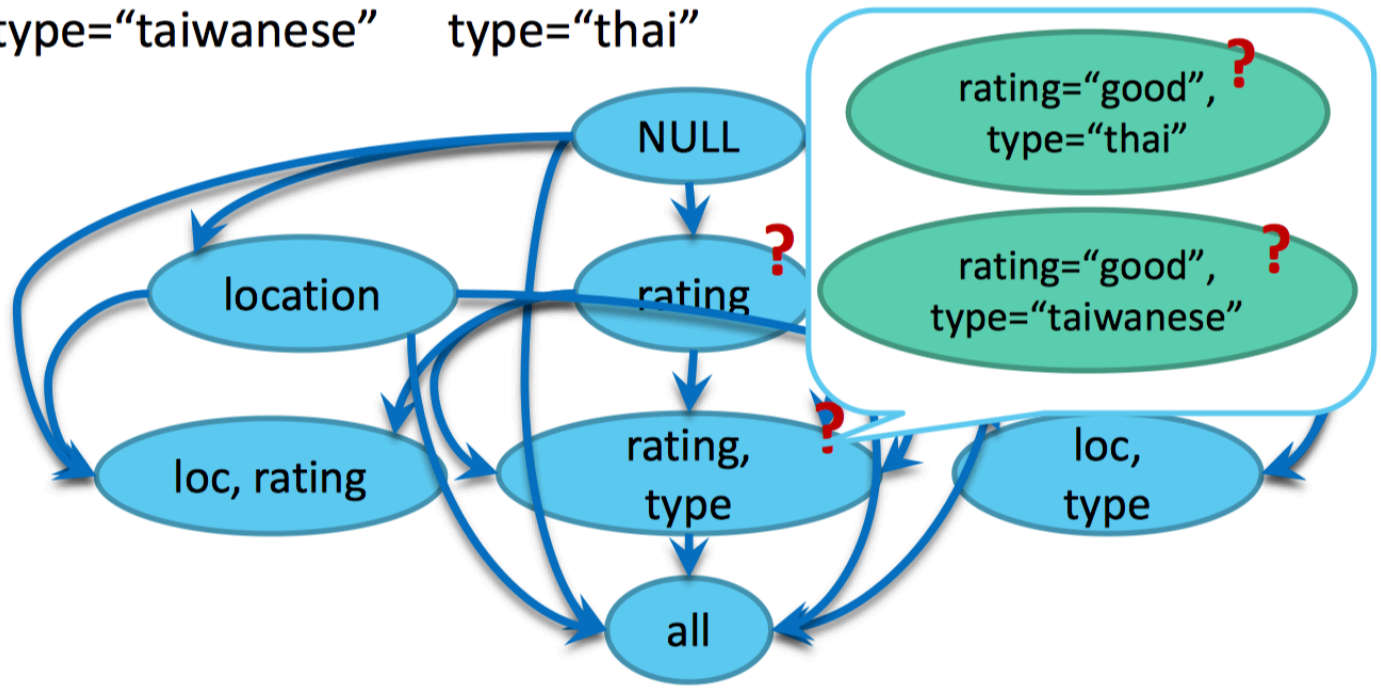
find a good eating place for taixxxx food

FIND_RESTAURANT
rating="good"
type="taiwanese"

FIND_RESTAURANT
rating="good"
type="thai"

FIND_RESTAURANT
rating="good"

Intelligent Agent



Dialogue Policy for Agent Action

- Inform(location="Taipei 101")
  - "The nearest one is at Taipei 101"
- Request(location)
  - "Where is your home?"
- Confirm(type="taiwanese")
  - "Did you want Taiwanese food?"

Variational Inference

Log-likelihood and Evidence Lower Bound (ELOB)

- It is universally true that:

$$\ln\left(p(X)\right) = \ln\left(p(X,Z)\right) - \ln\left(p(Z|X)\right)$$

- It's also true (a bit silly) that:

$$\ln\left(p(X)\right) = \left[\ln(p(X,Z)) - \ln(q(Z))\right] - \left[\ln(p(Z|X)) - \ln(q(Z))\right]$$

- The above is so that we can insert an arbitrary pdf $q(Z)$ into, now we get:

$$\ln\left(p(X)\right) = \ln\left(\frac{p(X,Z)}{q(Z)}\right) - \ln\left(\frac{p(Z|X)}{q(Z)}\right)$$

- Taking the expectation on both sides, given $q(Z)$:

$$\begin{aligned}
\ln\left(p(X)\right) &= \int q(Z)\ln\left(\frac{p(X,Z)}{q(Z)}\right)\mathrm{d}Z - \int q(Z)\ln\left(\frac{p(Z|X)}{q(Z)}\right)\mathrm{d}Z \\
&= \underbrace{\int q(Z)\ln(p(X,Z))\mathrm{d}Z - \int q(Z)\ln(q(Z))\mathrm{d}Z}_{\mathcal{L}(q)} + \underbrace{\left(-\int q(Z)\ln\left(\frac{p(Z|X)}{q(Z)}\right)\mathrm{d}Z\right)}_{\mathbb{KL}(q\|p)} \\
&= \mathcal{L}(q) + \mathbb{KL}(q\|p)
\end{aligned}$$

Variational Inference

Alternative Evidence Lower Bound (ELOB)

We often see the following alternative derivation:

$$
\begin{aligned}
\ln\left(p(X)\right) &= \log \int_Z p(X,Z)\mathrm{d}z \\
&= \log \int_Z p(X,Z)\frac{q(Z)}{q(Z)}\mathrm{d}z \\
&= \log\left(\mathbb{E}_q\left[\frac{p(X,Z)}{q(Z)}\right]\right) \\
&\geq \mathbb{E}_q\left[\log\left(\frac{p(X,Z)}{q(Z)}\right)\right] \text{ using Jensen's inequality} \\
&= \mathbb{E}_q\left[\log(p(X,Z))\right] - \mathbb{E}_q\left[\log(q(Z))\right] \\
&\triangleq \mathcal{L}(q)
\end{aligned}
$$

It can be proven easily that the "missing" part, i.e., $\ln\left(p(X)\right) - \mathcal{L}(q) = \mathbb{KL}(q\|p)$.

Variational Inference

Maximize Evidence Lower Bound (ELOB)

$$\ln\left(p(X)\right) = \mathcal{L}(q) + \mathbb{KL}(q||p)$$

- We can give a name to both terms:

Evidence Lower Bound (ELOB): $$\mathcal{L}(q) = \int q(Z)\ln(p(X,Z))\mathrm{d}Z - \int q(Z)\ln(q(Z))\mathrm{d}Z$$

KL divergence: $$\mathbb{KL}(q||p) = \int q(Z)\ln\left(\frac{p(Z|X)}{q(Z)}\right)dZ$$

- Notice $p(X)$ is fixed with respect to the choice of $q(Z)$. We wanted to choose a $q(Z)$ function that minimize KL divergence, so that $q(Z)$ becomes closer and closer to $p(Z|X)$. Of course, let's see what happens when $q(Z) = p(Z|X)$:

$$\mathbb{KL}(q||p) = -\int p(Z|X)\ln\left(\frac{p(Z|X)}{p(Z|X)}\right)\mathrm{d}Z = 0$$

- We know that $p(X) = \mathcal{L}(q) + \mathbb{KL}(q||p)$. Minimizing $\mathbb{KL}(q||p)$ is the same as maximizing the Evidence Lower Bound $\mathcal{L}(q)$.

A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues

RNNLM & HRED

$$P_\theta(w_1, \ldots, w_M) = \prod_{m=2}^{M} P_\theta(w_m \mid w_1, \ldots, w_{m-1}) P_\theta(w_1).$$

$$h_m = f_\theta(h_{m-1}, w_m)$$

$$P_\theta(w_{m+1} \mid w_1, \ldots, w_m) = P_\theta(w_{m+1} \mid h_m).$$

$$P_\theta(\mathbf{w}_1, \ldots, \mathbf{w}_N) = \prod_{n=1}^{N} P_\theta(\mathbf{w}_n | \mathbf{w}_{<n}),$$

$$= \prod_{n=1}^{N} \prod_{m=1}^{M_n} P_\theta(w_{n,m} | w_{n,<m}, \mathbf{w}_{<n}), \qquad (2)$$

$$h^{\text{enc}}_{n,0} = \mathbf{0}, \quad h^{\text{enc}}_{n,m} = f^{\text{enc}}_\theta(h^{\text{enc}}_{n,m-1}, w_{n,m}) \; \forall m = 1, \ldots, M_n,$$

issues: The Restricted Generation Proess

- *shadow*: each word is sampled conditioned only on previous words.
- This process is problematic from a probabilistic perspective, because the model is forced to generate all high-level structure locally on a step-by-step basis. For example, for generating dialogue responses such a model has to decide the conversation topic in the middle of the generation process – when it is generating the first topic-related word – and, afterwards, for each future word the model will have to decide whether to change or to remain on the same topic. This makes it difficult for the model to generate long-term structure.
- The shallow generation process is also problematic from a computational learning perspective: the state h_m in the RNNLM—or correspondingly the state of the decoder RNN in HRED—has to summarize all the past information up to time step m in order to (a) generate a probable next token (short-term objective) and (b) occupy a position in embedding space which sustains an output trajectory, for generating probable future tokens (long-term objective).

Method (VHRED)

- two steps:
  - 1. sampling latent variable
  - 2. generate output sequence

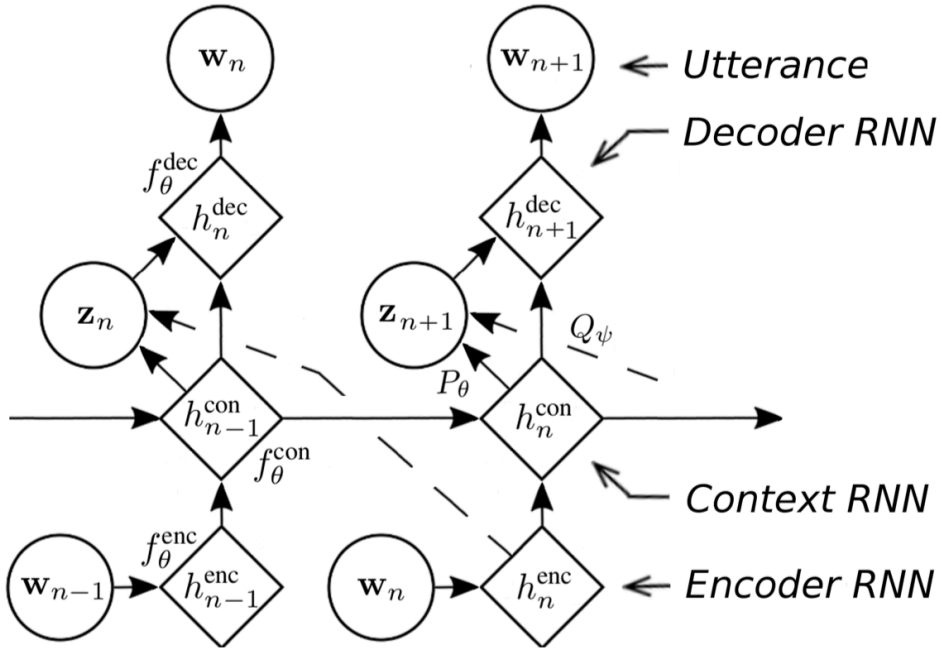


Figure 1: VHRED computational graph. Diamond boxes represent deterministic variables and rounded boxes represent stochastic variables. Full lines represent the generative model and dashed lines represent the approximate posterior model.

VHRED

$$P_\theta(\mathbf{z}_n \mid \mathbf{w}_{<n}) = \mathcal{N}(\boldsymbol{\mu}_{\text{prior}}(\mathbf{w}_{<n}), \Sigma_{\text{prior}}(\mathbf{w}_{<n})),$$

$$P_\theta(\mathbf{w}_n \mid \mathbf{z}_n, \mathbf{w}_{<n}) = \prod_{m=1}^{M_n} P_\theta(w_{n,m} \mid \mathbf{z}_n, \mathbf{w}_{<n}, w_{n,<m}),$$

$$\begin{aligned}
&\log P_\theta(\mathbf{w}_1, \ldots, \mathbf{w}_N) \\
&\geq \sum_{n=1}^{N} -\mathrm{KL}\left[Q_\psi(\mathbf{z}_n \mid \mathbf{w}_1, \ldots, \mathbf{w}_n) || P_\theta(\mathbf{z}_n \mid \mathbf{w}_{<n})\right] \\
&+ \mathbb{E}_{Q_\psi(\mathbf{z}_n|\mathbf{w}_1,\ldots,\mathbf{w}_n)}\left[\log P_\theta(\mathbf{w}_n \mid \mathbf{z}_n, \mathbf{w}_{<n})\right], \qquad (4)
\end{aligned}$$

$$\begin{aligned}
&Q_\psi(\mathbf{z}_n \mid \mathbf{w}_1, \ldots, \mathbf{w}_N) \\
&= \mathcal{N}(\boldsymbol{\mu}_{\text{posterior}}(\mathbf{w}_1, \ldots, \mathbf{w}_n), \Sigma_{\text{posterior}}(\mathbf{w}_1, \ldots, \mathbf{w}_n)) \\
&\approx P_\psi(\mathbf{z}_n \mid \mathbf{w}_1, \ldots, \mathbf{w}_N), \qquad (5)
\end{aligned}$$

$$\mathbf{h}_{t-1}^{\text{enc}} = f_{\boldsymbol{\theta}}^{\text{enc}}(\mathbf{x}_{t-1}) \qquad (3)$$

$$\mathbf{h}_t^{\text{cxt}} = f_{\boldsymbol{\theta}}^{\text{cxt}}(\mathbf{h}_{t-1}^{\text{cxt}}, \mathbf{h}_{t-1}^{\text{enc}}) \qquad (4)$$

$$p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t\mathbf{I}) \qquad (5)$$

$$\text{where } \boldsymbol{\mu}_t = \text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}}) \qquad (6)$$

$$\boldsymbol{\sigma}_t = \text{Softplus}(\text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}})) \qquad (7)$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{<t}) = f_{\boldsymbol{\theta}}^{\text{dec}}(\mathbf{x}|\mathbf{h}_t^{\text{cxt}}, \mathbf{z}_t^{\text{utt}}) \qquad (8)$$

$$\begin{aligned}
&\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) \qquad (2) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})] \\
&= -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})]
\end{aligned}$$

最后一步: 同时减去p(z)

VAE and VHRED

VAE:

$$
\begin{aligned}
&-\log p_\theta(x) \leq -\log p_\theta(x) + \mathbf{KL}(q_\phi(z|x)||p_\theta(z|x)) \\
&= -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathbf{KL}(q_\phi(z|x||p(z))
\end{aligned} \quad (1)
$$

CVAE:

$$
-\mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|c,z)] + \mathbf{KL}(q_\phi(z|x,c)\|p_\theta(z|c)) \quad (2)
$$

Specially, to some extent, when both the context c and output x are sequential data, CVAE can also be treated as a seq2seq model

The variational hierarchical recurrent encoder-decoder (VHRED) (Serban et al. 2017b) is a CVAE with hierarchical RNN encoders, where the first-layer RNN encodes token- level variations and the second-layer RNN captures sentence- level topic shifts. I

Issues: KL-vanishing

$$-\mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|c,z)] + \mathbf{KL}(q_\phi(z|x,c)\|p_\theta(z|c)) \quad (2)$$

RNN decoder $p_\theta(x|c, z)$ is a universal function approximator and tends to represent the distribution without referring to the latent variable.

At the beginning of the training process, when the approximate posterior $q_\phi(z|x, c)$ carries little useful information, it is natural for the model to blindly set $q_\phi(z|x, c)$ closer to the Gaussian prior $p_\theta(z|c)$ so that the extra cost from the KL divergence can be avoided (Chen et al. 2017).

$$-\log \int_z p_\theta(z|c)p_\theta(x|z,c)dz + \mathbf{KL}(q_\phi(z|x,c)||p_\theta(z|x,c)) \quad (3)$$

$$-\log \int_z p_\theta(z|c)p_\theta(x|z,c)dz + \mathbf{KL}(q_\phi(z|x,c)||p_\theta(z|x,c)) \tag{3}$$

Let's first take a look at the first item, $\log \int_z p_\theta(z|c)p_\theta(x|z,c)dz = \log p_\theta(x|c)$. When the family of $p_\theta(x|z,c)$ is complex enough and includes the real distribution of x, the optimal value of this item is $p(x|c)$ and the reliance on z is not necessary. However, reliance on z provides the model with a chance of taking advantage of $z's$ distribution and reduces the complexity requirement for the distribution family $p_\theta(x|z,c)$. For example, suppose $p(x|c) = \mathcal{N}(0,1)$ and $p_\theta(z|c) = \mathcal{N}(3,1)$, modeling $p(x|c)$ accurately without reliance on z requires $p_\theta(x|z,c)$ to include the Gaussian distribution, while by means of the linear mapping between x and z, $p_\theta(x|z,c)$ can describe the real distribution with only linear complexity. When Gaussian distribution is not covered in the family $p_\theta(x|z,c)$, this model has to exploit the relation between x and z to model the real distribution. Likewise, in dialogue generation, although the RNN decoder $p_\theta(x|c)$ can in theory approximate arbitrary function, perfectly fitting the real dialogue distribution is still difficult due to the optimizing challenge, training corpus size and approximating errors. Therefore, to achieve the global optimum, we believe this first item will always prefer utilizing the latent variables, so long as the decoder $p_\theta(x|z,c)$ is not perfect. The weaker the decoder family is, the more it will be biased to utilizing latent variables. A more flexible prior distribution $p_\theta(z)$ will also increase the chance as it provides more possibilities for the utilisation.

$$-\log \int_z p_\theta(z|c)p_\theta(x|z,c)dz + \mathbf{KL}(q_\phi(z|x,c)||p_\theta(z|x,c)) \tag{3}$$

The second item is the KL divergence, whose minimum value is 0 if and only if $q_\phi(z|x,c) = p_\theta(z|x,c)$. According to the Bayes theorem, we can express $p_\theta(z|x,c)$ as:

$$p_\theta(z|x,c) = \frac{p_\theta(x|z,c)p_\theta(z|c)}{p_\theta(x|c)} \tag{4}$$

By ignoring the latent variable z, $p_\theta(x|z,c)$ and $p_\theta(x|c)$ cancels out, setting $q_\phi(z|x,c) = p_\theta(z|c)$ can easily arrive at the global optimum 0. Otherwise, when $p_\theta(z|c)$ is parametrised as a mean-field Gaussian distribution as in VHRED, the real posterior is impossible to fall into the same distribution family. Firstly, the independence relation cannot be satisfied. To make dimensions of $p_\theta(z|x)$ independent with each other, the likelihood $p_\theta(x|z)$ must exactly disentangle the effect of every dimension, which is unrealistic when $p_\theta(x|z)$ is a categorical distribution modelled by the RNN softmax. Secondly, the real posterior distribution can hardly still follow a Gaussian distribution when the likelihood $p_\theta(x|z)$ is based on discrete sequential data. Normally the training process will adjust $p_\theta(x|z)$ to make the real posterior easier to be modelled by $q_\phi(z|x)$ (Hinton et al. 1995). However, when x represents sentences with variable length, the value of $p_\theta(x|z)$ vanishes greatly when the length grows, which makes the adjusting task much more difficult. This implies the second item will always prefer ignoring the latent variables, so long as the approximated posterior is not powerful enough to perfectly match the real posterior. The weaker the approximating posterior distribution family is, the more it will be biased to ignoring latent variables.

Two current approaches: 1. KL annealing 2. word drop regularization

First, the *KL annealing* scales the KL divergence term of Eq. 2 using a KL multiplier λ, which gradually increases from 0 to 1 during training:

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = -\lambda D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] \quad (12)$$

This helps the optimization process to avoid local optima of zero KL divergence in early training.

Second, the *word drop* regularization randomly replaces some conditioned-on word tokens in the RNN decoder with the generic unknown word token (UNK) during training. Normally, the RNN decoder predicts each next word in an autoregressive manner, conditioned on the previous sequence of ground truth (GT) words. By randomly replacing a GT word with an UNK token, the word drop regularization weakens the autoregressive power of the decoder and forces it to rely on the latent variable to predict the next word. The word drop probability is normally set to 0.25, since using a higher probability may degrade the model performance (Bowman et al., 2016).

Still have problems

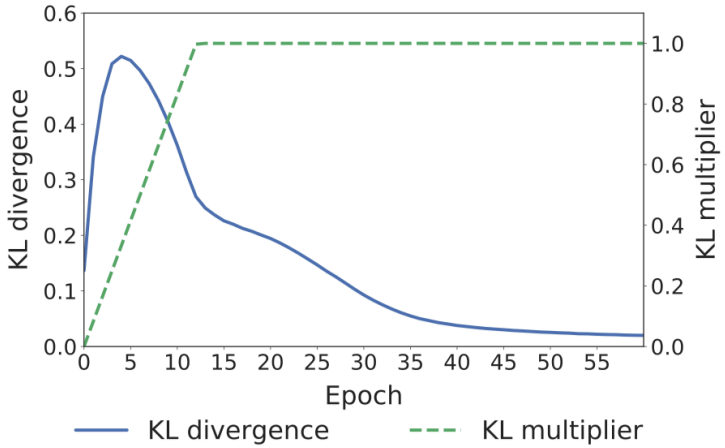


Figure 1: **Degeneration of VHRED**. The KL divergence term continuously decreases as training proceeds, meaning that the decoder ignores the latent variable $\mathbf{z}^{utt}$. We train the VHRED on Cornell Movie Dialog Corpus with word drop and KL annealing.

$\mathbf{h}_t^{\text{cxt}}$ and stochastic $\mathbf{z}^{\text{utt}}$. In order to check whether the presence of deterministic source $\mathbf{h}_t^{\text{cxt}}$ causes the degeneration, we drop the deterministic $\mathbf{h}_t^{\text{cxt}}$ and condition the decoder only on the stochastic utterance latent variable $\mathbf{z}^{\text{utt}}$:

$$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{<t}) = f_{\boldsymbol{\theta}}^{\text{dec}}(\mathbf{x}|\mathbf{z}_t^{\text{utt}}) \tag{13}$$

While this model achieves higher values of KL divergence than original VHRED, as training proceeds it again degenerates with the KL divergence term reaching zero (Fig. 2).

To gain an insight of the degeneracy, we examine how the conditional prior $p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t})$ (Eq. 5) of the utterance latent variable changes during training, using the model above (Eq. 13). Fig. 2

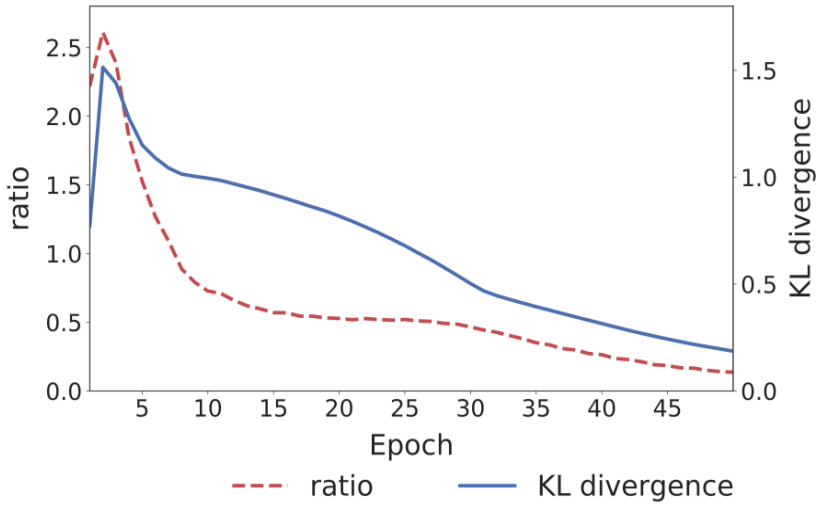


Figure 2: **The average ratio $\mathbb{E}[\boldsymbol{\sigma}_t^2]/\mathbf{Var}(\boldsymbol{\mu}_t)$ when the decoder is only conditioned on $\mathbf{z}_t^{\text{utt}}$.** The ratio drops to zero as training proceeds, indicating that the conditional priors $p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t})$ degenerate to separate point masses.

The ratio gradually falls to zero, implying that the priors degenerate to separate point masses as training proceeds.

Motivation

- Traditional: Hierarchical RNN + VAE
  - First, the expressive power of hierarchical RNN decoders is often high enough to model the data using only its decoding distributions without relying on the latent variables.
  - Second, the conditional VAE structure whose generation process is conditioned on a context, makes the range of training targets very sparse; that is, the RNN decoders can easily overfit to the training data ignoring the latent variables.
- VAE
  - First, latent variables can learn an interpretable holistic representation, such as topics, tones, or high-level syntactic proper- ties.
  - Second, latent variables can model inherently abundant variability of natural language by encoding its global and long-term structure, which is hard to be captured by shallow generative processes (e.g. vanilla RNNs) where the only source of stochasticity comes from the sampling of output words.

Motivation

- VAE with RNN decoder -> degeneration
  - This issue makes VAEs ignore latent variables, and eventually behave as vanilla RNNs.
  - Chen et al. (2017) also note this degeneration issue by showing that a VAE with a RNN decoder prefers to model the data using its decoding distribution rather than using latent variables, from bits-back coding perspective.
- <= KL annealing and word drop regularization
  - fail to prevent the degeneracy in VHRED
- Goal: to alleviate the degeneration problem

Improving Variational Encoder-Decoders in Dialogue Generation

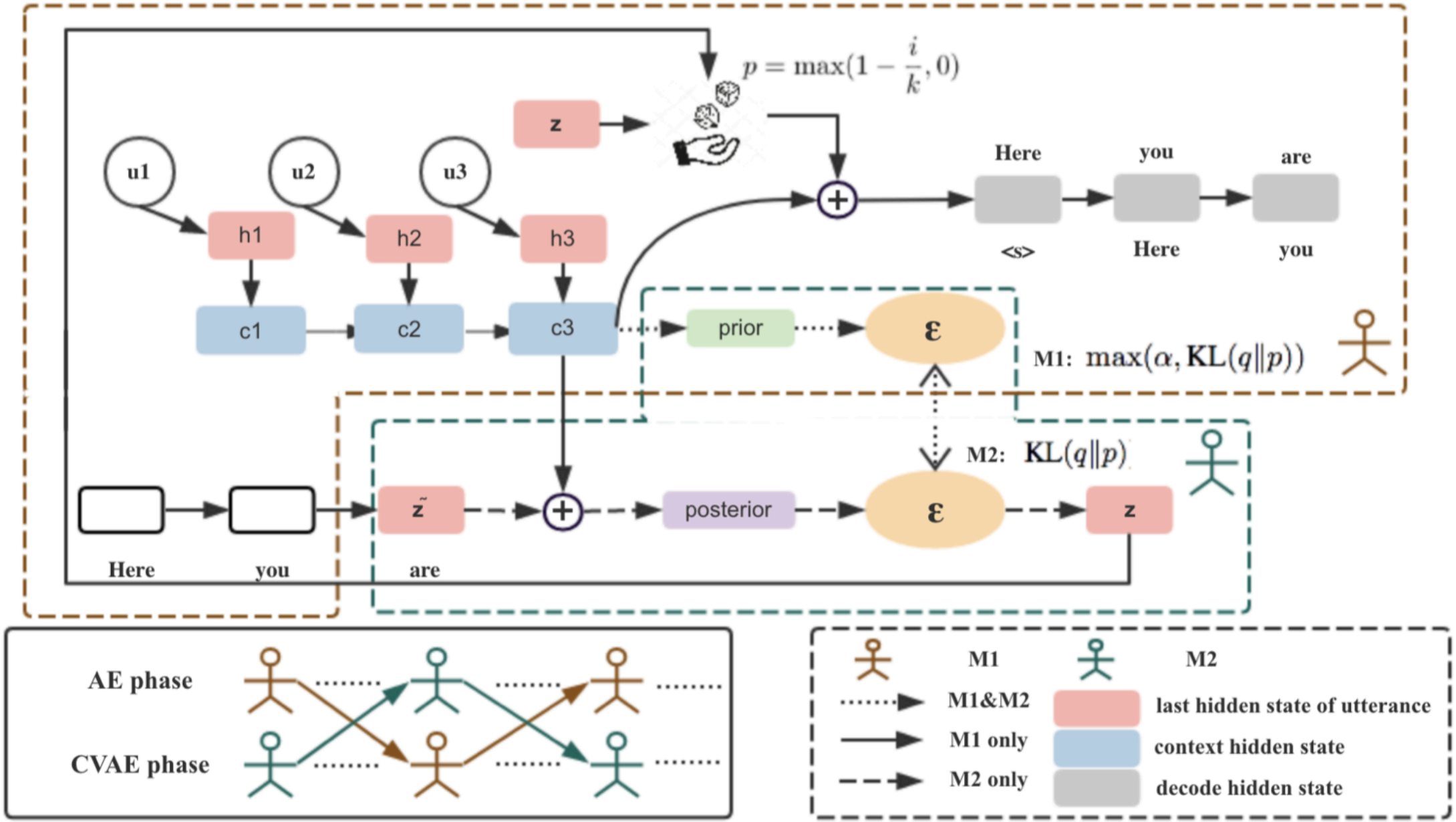


Figure 1: Architecture for collaborative variational encoder-decoder. $\bigoplus$ denotes concatenation of information. $M_1(AE)$ and $M_2(CVAE)$ are represented in brown and green respectively.

A Hierarchical Latent Structure for Variational Conversation Modeling

VHCR

$$\mathbf{h}_{t-1}^{\text{enc}} = f_{\boldsymbol{\theta}}^{\text{enc}}(\mathbf{x}_{t-1}) \tag{3}$$
$$\mathbf{h}_{t}^{\text{cxt}} = f_{\boldsymbol{\theta}}^{\text{cxt}}(\mathbf{h}_{t-1}^{\text{cxt}}, \mathbf{h}_{t-1}^{\text{enc}}) \tag{4}$$
$$p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t\mathbf{I}) \tag{5}$$
$$\text{where } \boldsymbol{\mu}_t = \text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}}) \tag{6}$$
$$\boldsymbol{\sigma}_t = \text{Softplus}(\text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}})) \tag{7}$$
$$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{<t}) = f_{\boldsymbol{\theta}}^{\text{dec}}(\mathbf{x}|\mathbf{h}_t^{\text{cxt}}, \mathbf{z}_t^{\text{utt}}) \tag{8}$$

$$\mathbf{h}_t^{\text{enc}} = f_{\boldsymbol{\theta}}^{\text{enc}}(\mathbf{x}_t) \tag{15}$$

$$\mathbf{h}_t^{\text{cxt}} = \begin{cases} \text{MLP}_{\boldsymbol{\theta}}(\mathbf{z}^{\text{conv}}), & \text{if } t = 0 \\ f_{\boldsymbol{\theta}}^{\text{cxt}}(\mathbf{h}_{t-1}^{\text{cxt}}, \mathbf{h}_{t-1}^{\text{enc}}, \mathbf{z}^{\text{conv}}), & \text{otherwise} \end{cases}$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{z}_t^{\text{utt}}, \mathbf{z}^{\text{conv}}) = f_{\boldsymbol{\theta}}^{\text{dec}}(\mathbf{x}|\mathbf{h}_t^{\text{cxt}}, \mathbf{z}_t^{\text{utt}}, \mathbf{z}^{\text{conv}})$$
$$p_{\boldsymbol{\theta}}(\mathbf{z}^{\text{conv}}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \tag{16}$$
$$p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t}, \mathbf{z}^{\text{conv}}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t\mathbf{I}) \tag{17}$$
$$\text{where } \boldsymbol{\mu}_t = \text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}}, \mathbf{z}^{\text{conv}}) \tag{18}$$
$$\boldsymbol{\sigma}_t = \text{Softplus}(\text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}}, \mathbf{z}^{\text{conv}})). \tag{19}$$

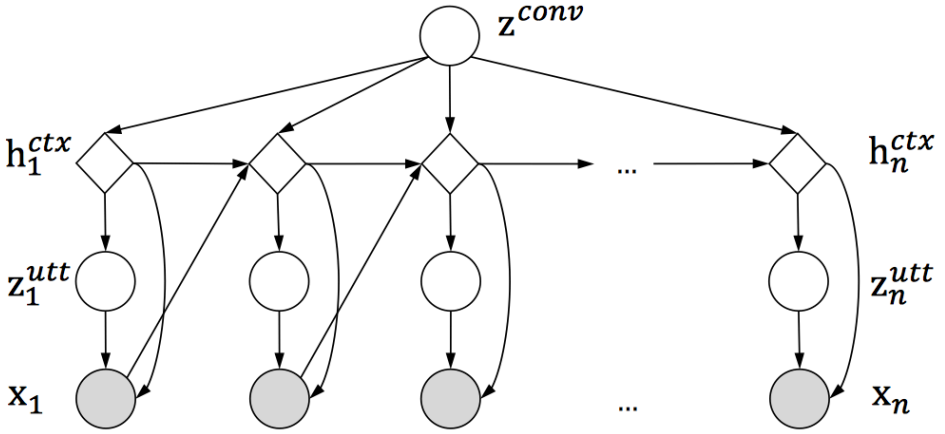
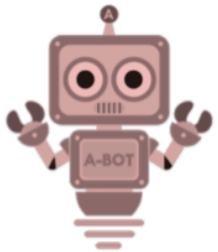


Figure 3: **Graphical representation of the Variational Hierarchical Conversation RNN (VHCR).** The global latent variable $\mathbf{z}^{conv}$ provides a global context in which the conversation takes place.

VQA

Aid ‘situationally-impaired’ analysts



Did anyone enter this room last week?

Yes, 127 instances logged on camera

Were any of them carrying a black bag?

…

Natural language instructions for robots

Is there smoke in any room around you?

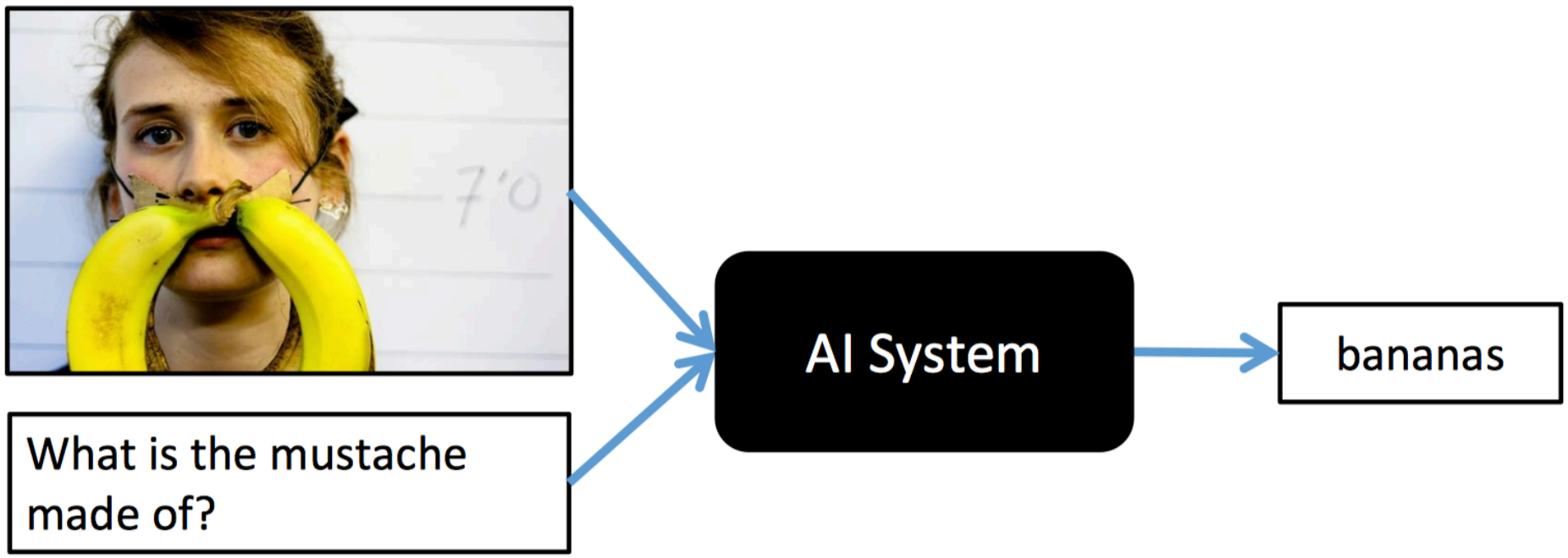
Yes, in one room

Go there and look for people

...

Image Credit: Lockheed Martin; DARPA Robotics Challenge

Visual Question Answering (VQA)



VQA Dataset

>0.25 million images

>0.76 million questions

~10 million answers

[Antol et al., ICCV 2015]

Video QA

TVQA: Localized, Compositional Video Question Answering

00:00.755 --> 00:02.655
(Chandler:) Go to your room!
00:06.961 --> 00:08.622
(Janice:) I gotta go, I gotta go.

00:08.829 --> 00:10.057
(Janice:) Not without a kiss.
00:10.264 --> 00:12.391
(Chandler:) Maybe I won't kiss you so you'll stay.

00:12.600 --> 00:14.761
(Joey:) Kiss her. Kiss her!
00:16.771 --> 00:19.137
(Janice:) I'll see you later, sweetie. Bye, Joey.

...

00:39.327 --> 00:40.760
(Chandler:) She makes me happy.
00:41.596 --> 00:44.087
(Joey:) Okay. All right.

...

00:00 00:06 00:10 00:17 00:39 00:45 01:04

What is Janice holding on to after Chandler sends Joey to his room?

A Chandler's tie
B Chandler's hands
C Her Breakfast
D Her coat
E Chandler's coffee cup.

Why does Joey want Chandler to kiss Janice when they are in the kitchen?

A Because Joey is glad that Chandler is happy
B Because Joey likes to watch people kiss
C Because then she will leave
D Because Joey thinks Janice is hot
E Because then Chandler will move away from the toast.

What is on the couch behind Joey when he is at the counter?

A A chick
B A soccer ball
C A duck
D A pillow
E Janice's coat

Figure 1: Examples from the TVQA dataset. All questions and answers are attached to 60-90 seconds long clips. For visualization purposes, we only show a few of the most relevant frames here. As illustrated above, some questions can be answered using subtitles or videos alone, while some require information from both modalities.

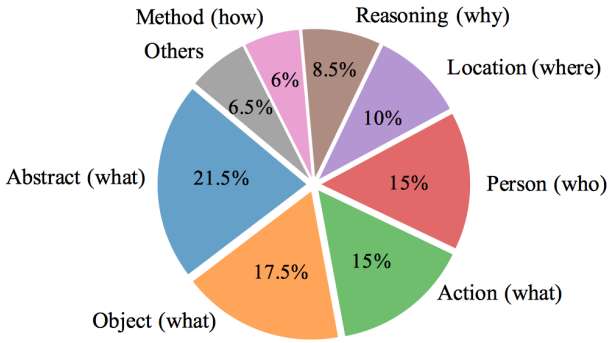
Motivation

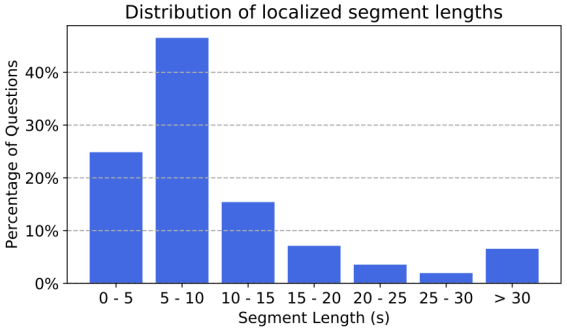
- Less work on video question answering due to the data limitations.
- MovieFIB:
  - video clips are typically short (~4s)
  - collected QAs based on text summarises only, less relevant for visual information
- PororoQA's:
  - catoon-based
- TGIF-QA:
  - used predefined templates for generaion on short GIFs
- TVQA
  - Composational questions
  - more genres (medical dramas, sitcoms, crime shows)
  - larger size

Data Collection

- Composational question
  - [What/How/Where/Why/…] ____[when/before/after]____?
- Answers
  - Workers are asked to answer the question
  - marking the START and END timestamps

QType	#QA	Q. Len.	CA. Len.	WA. Len.
what	84768	13.3	4.9	4.3
who	17654	13.4	3.1	3.0
where	17777	12.5	5.2	4.8
why	15798	14.5	9.0	7.7
how	13644	14.4	5.7	5.1
others	2904	15.2	4.9	4.7
total	152545	13.5	5.2	4.6





152.5K questions

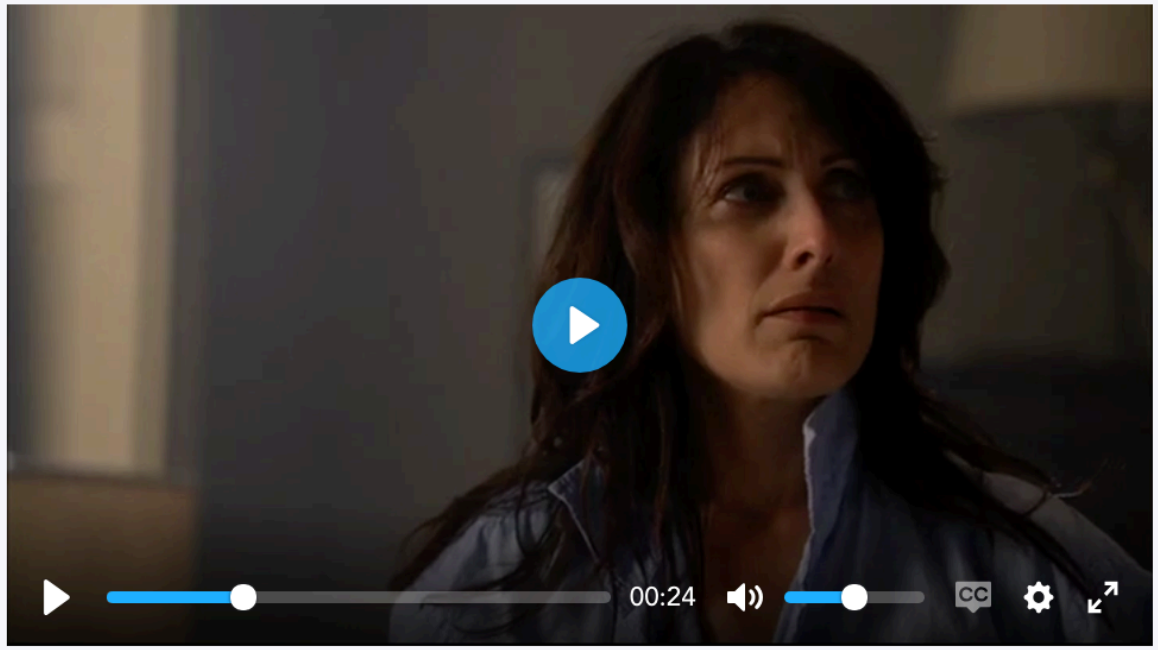
- 84.8K What
- 17.7K Who
- 17.8K Where
- 15.8K Why
- 13.6K How

925 episodes from 6 TV shows

- Sitcoms:
  - *Friends*
  - *The Big Bang Theory*
  - *How I Met Your Mother*
- Medical:
  - *Grey's Anatomy*
  - *House M.D.*
- Crime:
  - *Castle*

Usage

Click the Play Localized button to play the video clip specified by the timestamp annotation. The video come with subtitle, turn it on if you did not see it. Best viewed in Chrome.



00:24

00:47

Play Localized

Question What room was Wilson breaking into when House found him?

Answer 0 The bedroom.
Answer 1 The bathroom.
Answer 2 The living room.
Answer 3 The kitchen.
Answer 4 The dining room.

Show Answer

Last example 1 / 12 Next example

Baseline

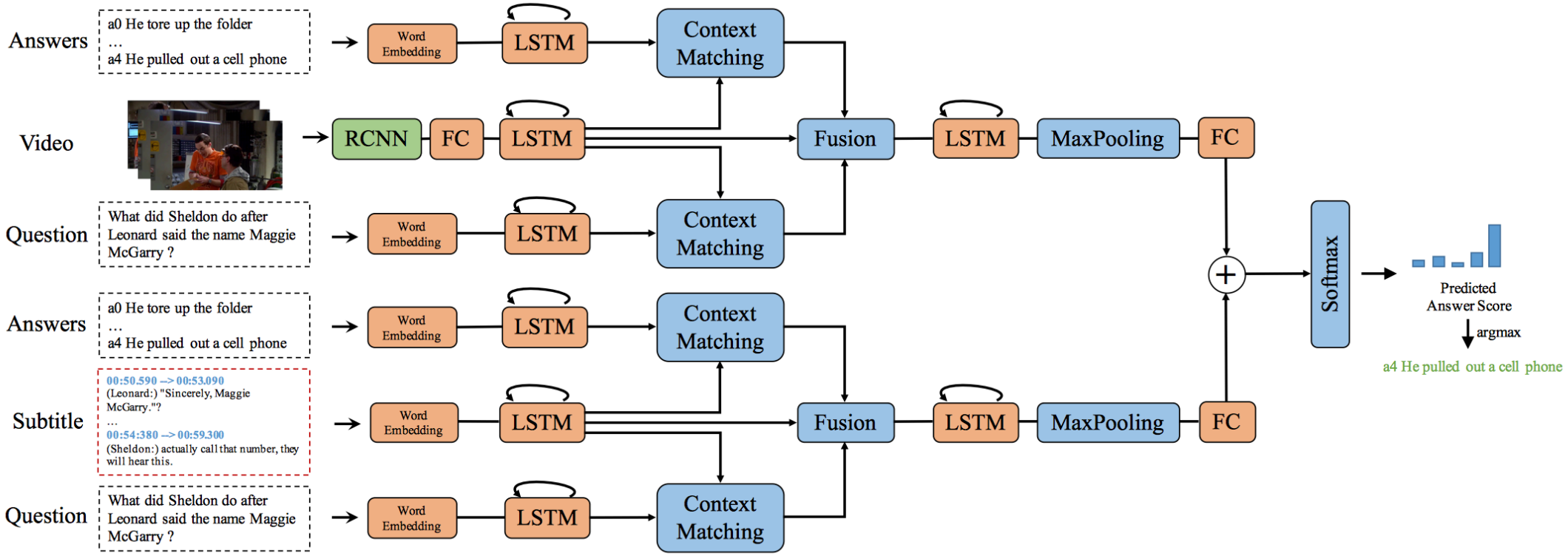


Figure 4: Illustration of our multi-stream model for Multi-Modal Video QA. Our full model takes different contextual sources (regional visual features, visual concept features, and subtitles) along with question-answer pair as inputs to each stream. For brevity, we only show regional visual features (upper) and subtitle (bottom) streams.

Visual Dialog

Visual Dialog: Task

- Given
  - Image I
  - History of human dialog
    $(Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})$
  - Follow-up Question Q_t

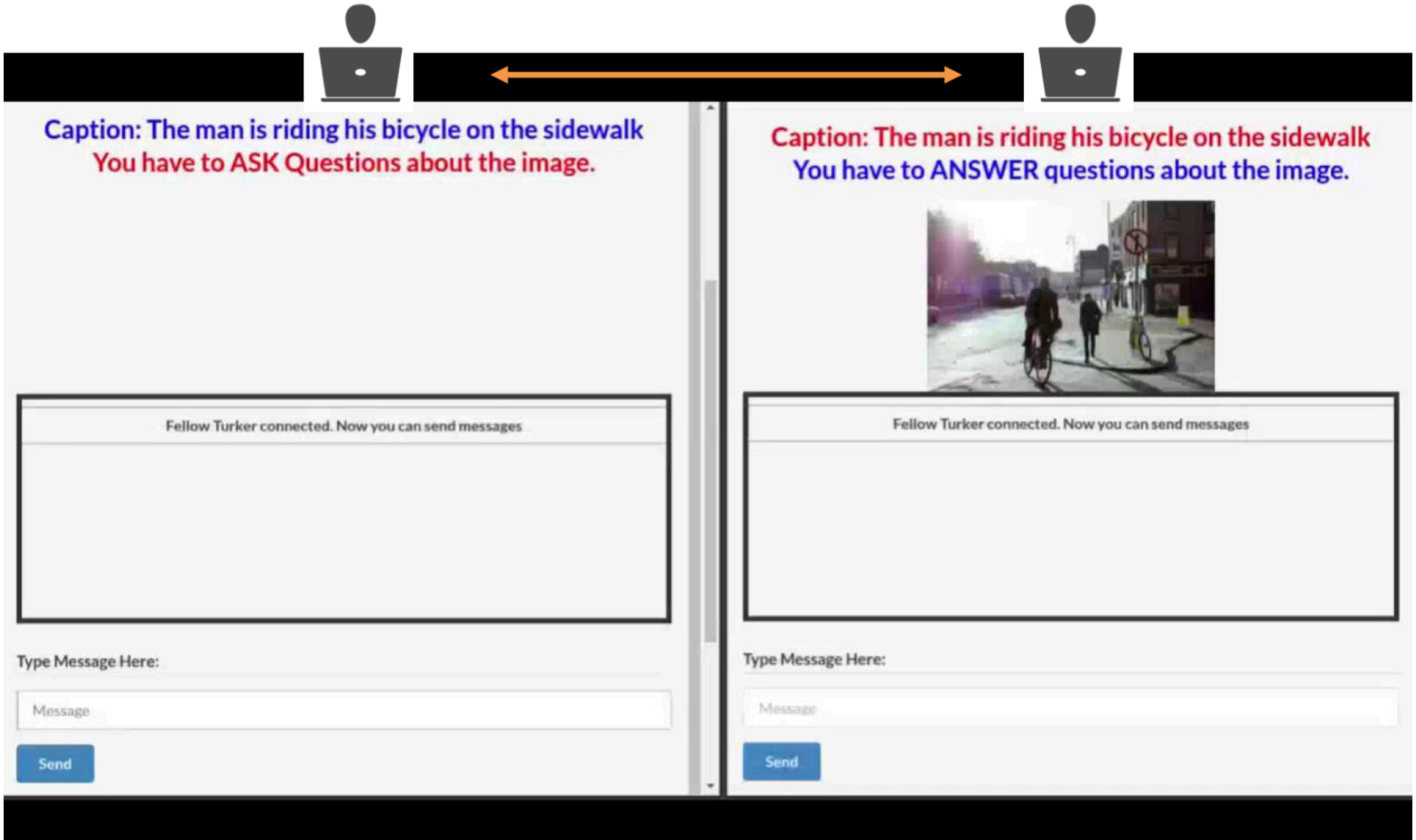
- Task
  - Produce free-form
    natural language answer
    A_t

Visual Dialog

Q: How many people on wheelchairs?
A: Two.
Q : What gender are the people in the wheelchairs?
A : One is female, one is male.
Q : Which one is holding the racket?
A : The female.
Q : Is the other one holding anything?
A : He is not.

VisDial Dataset

Live Two-Person Chat on Amazon Mechanical Turk



VisDial v0.9 Stats

>120k images (from COCO)

1 dialog/image

10 question-answer rounds/dialog

Total of *>1.2 Million* dialog QA pairs

Models for Visual Dialog

Encoder

1. Late Fusion

2. Hierarchical Recurrent Encoder

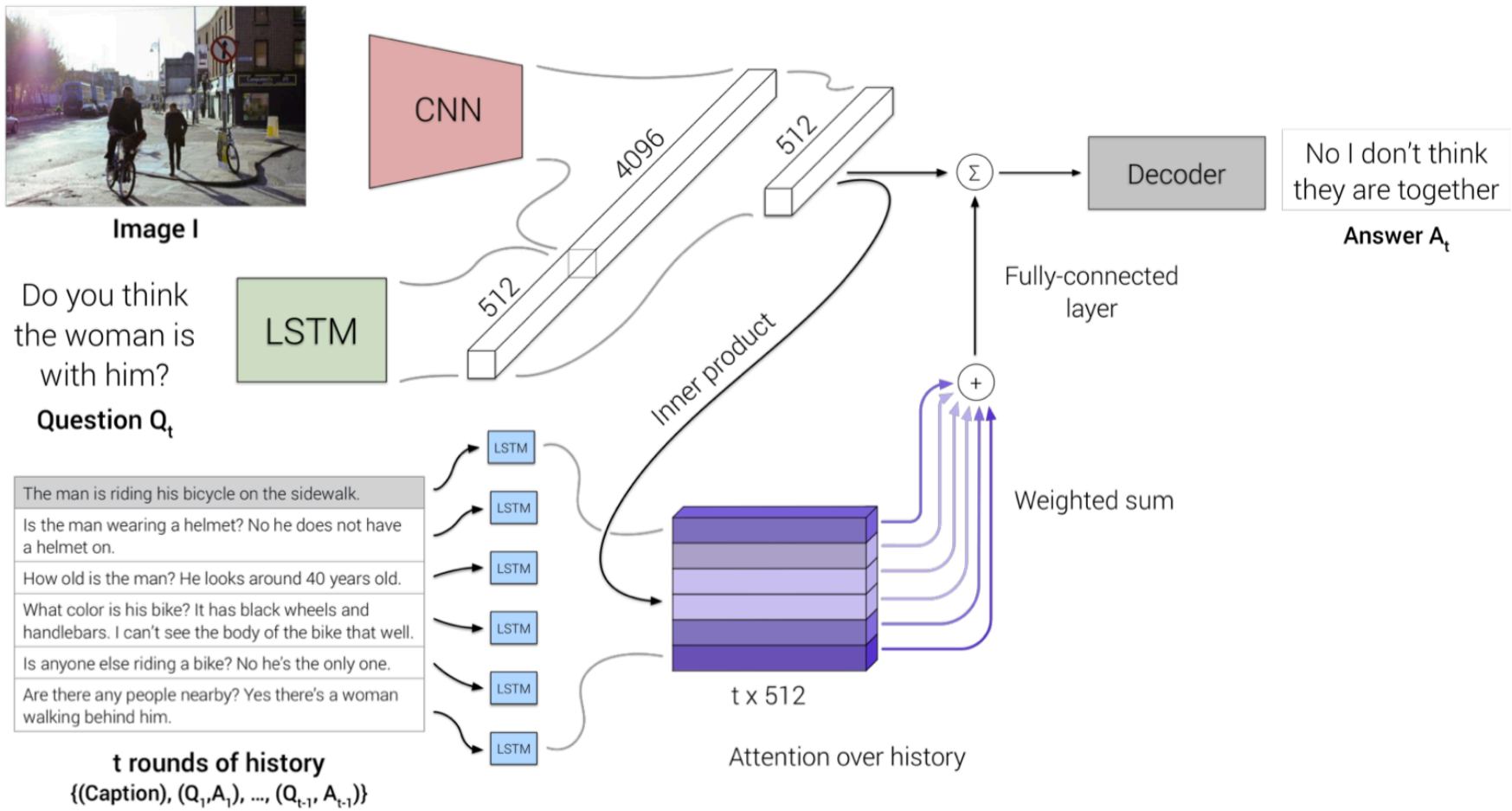
3. Memory Network

Decoder

1. Generative
   - During training, maximizes LL of human response
   - For evaluation, ranks options by LL scores

2. Discriminative
   - Learn to rank 100 options

Visual Dialog Model #3



Memory Network Encoder

Slide credit: Abhishek Das

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

[ICCV '17]

Abhishek Das*
(Georgia Tech)

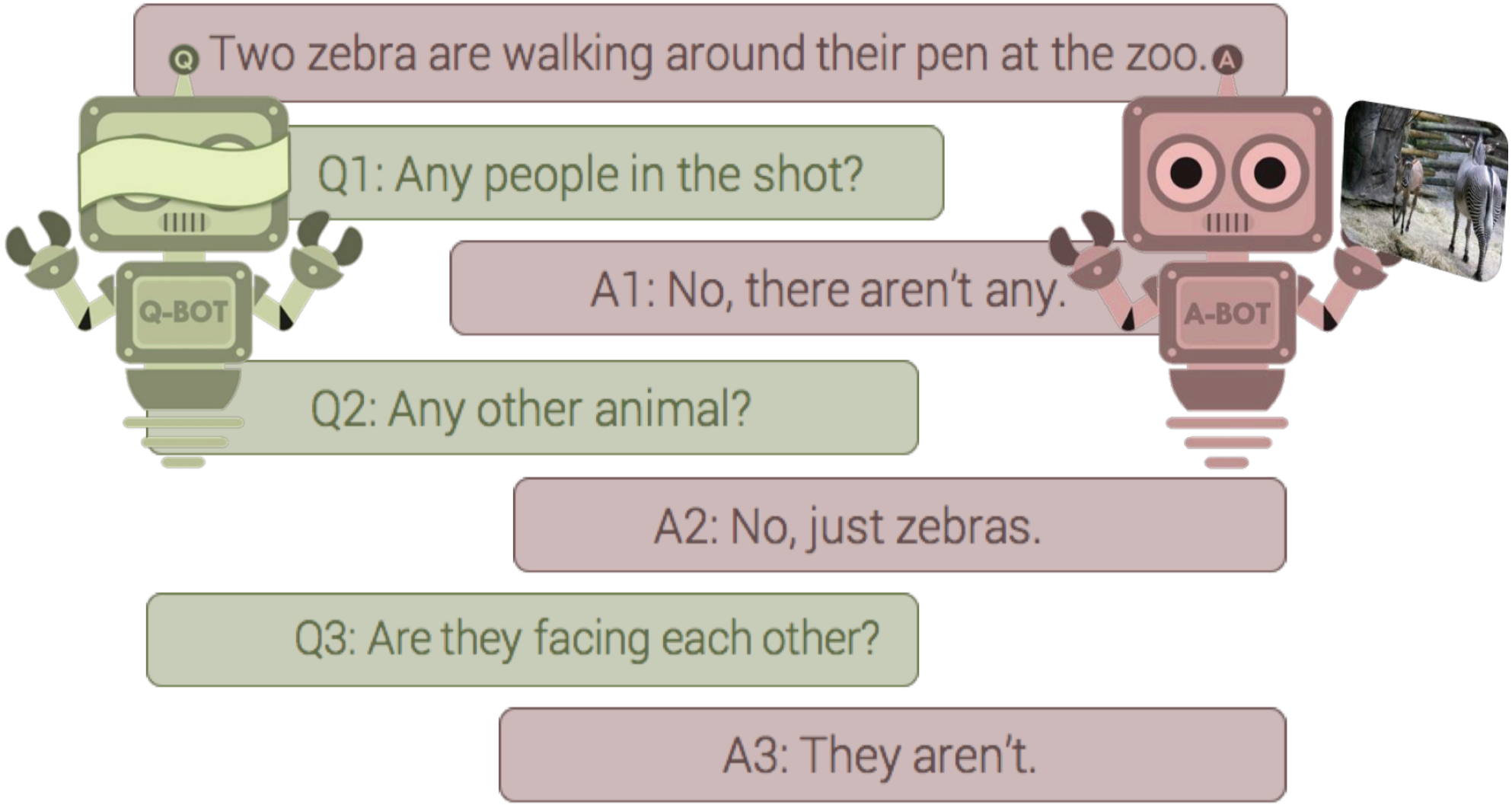
Satwik Kottur*
(CMU)

José Moura
(CMU)

Stefan Lee
(Virginia Tech)

Dhruv Batra
(Georgia Tech)

Image Guessing Game



RL for Cooperative Dialog Agents

- Action:
  - Q-bot: question (symbol sequence)
  - A-bot: answer (symbol sequence)
  - Q-bot: image regression

q_t *Any people in the shot?*

a_t *No, there aren't any.*

$\hat{y}_t \in \mathbb{R}^{4096}$

- State
  - Q-bot: $s_t^Q = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}]$
  - A-bot: $s_t^A = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t]$

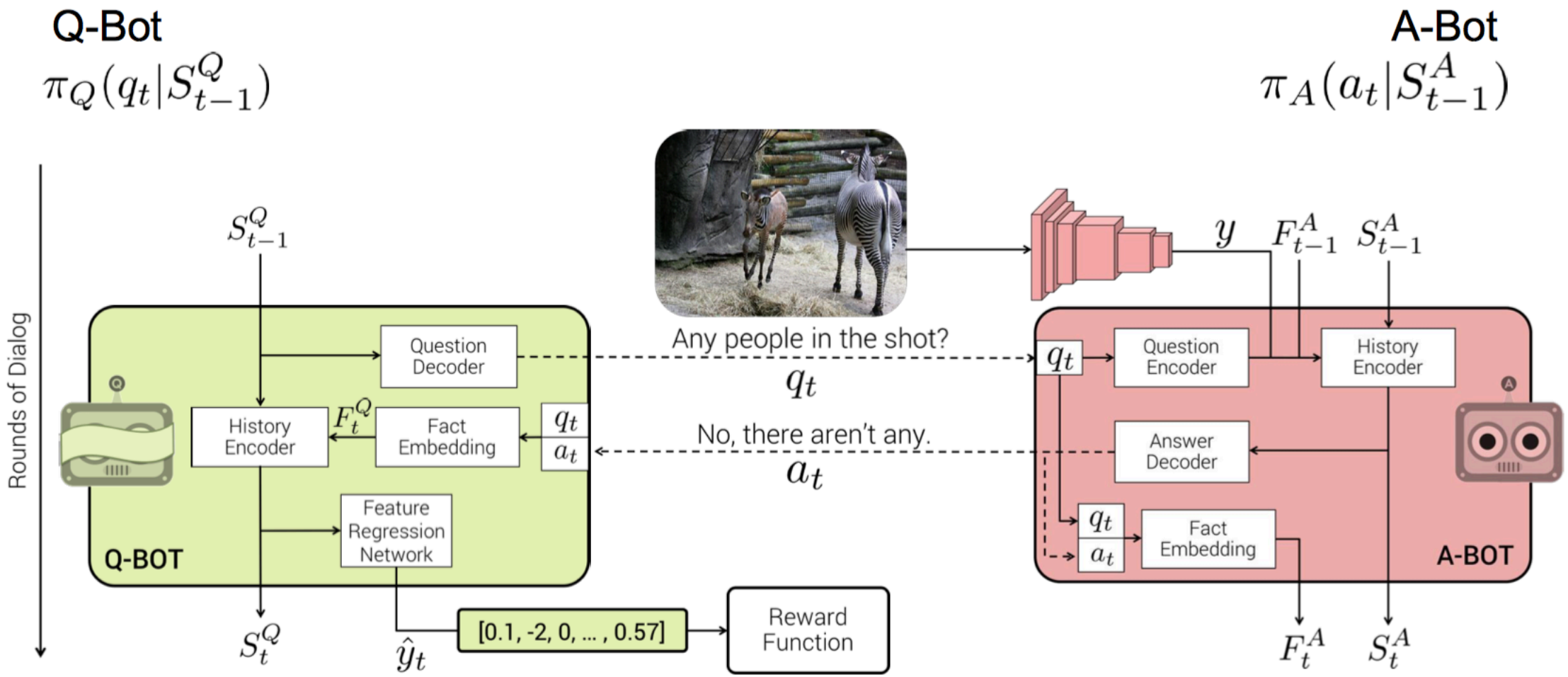
- Policy

Q-bot $\pi_Q(q_t|S_{t-1}^Q)$ A-bot $\pi_A(a_t|S_{t-1}^A)$

- Reward

$$r_t\Big(\underbrace{s_t^Q}_{\text{state}}, \underbrace{(q_t, a_t, y_t)}_{\text{action}}\Big) = \underbrace{\ell\left(\hat{y}_{t-1}, y^{gt}\right)}_{\text{distance at } t\text{-}1} - \underbrace{\ell\left(\hat{y}_t, y^{gt}\right)}_{\text{distance at } t}$$

Policy Networks



Policy Gradients

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

REINFORCE Gradients

$$\nabla_{\theta_Q} J = \nabla_{\theta_Q} \left[\mathop{\mathbb{E}}_{\pi_Q, \pi_A} \left[r_t\left(\cdot\right) \right] \right]$$

$$= \sum_{q_t, a_t} \pi_Q\left(q_t | s_{t-1}^Q\right) \nabla_{\theta_Q} \log \pi_Q\left(q_t | s_{t-1}^Q\right) \pi_A\left(a_t | s_t^A\right) r_t\left(\cdot\right)$$

$$= \mathop{\mathbb{E}}_{\pi_Q, \pi_A} \left[r_t\left(\cdot\right) \nabla_{\theta_Q} \log \pi_Q\left(q_t | s_{t-1}^Q\right) \right]$$

Dialog-based Interactive Image Retrieval

Introduction

Figure 1: In the context of interactive image retrieval, the agent incorporates the user's feedback to iteratively refine retrieval results. Different from existing works which are based on relevance feedback or relative attribute feedback, our approach allows the user to provide feedback in natural language.

Framework

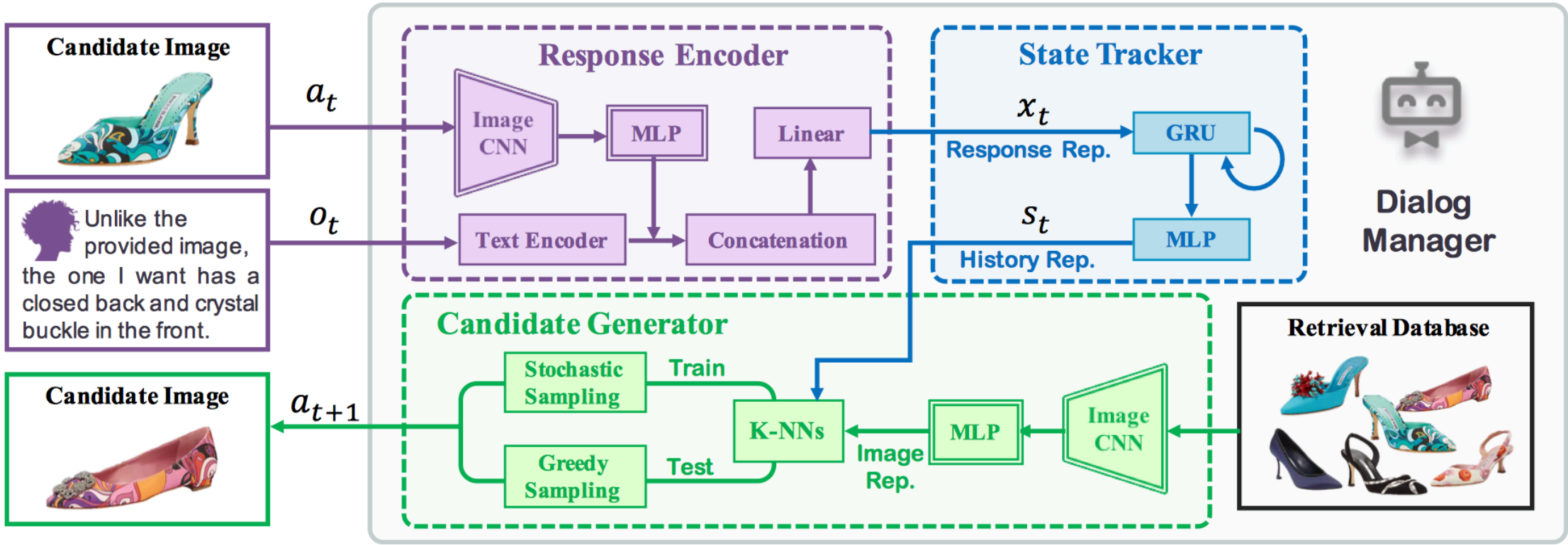
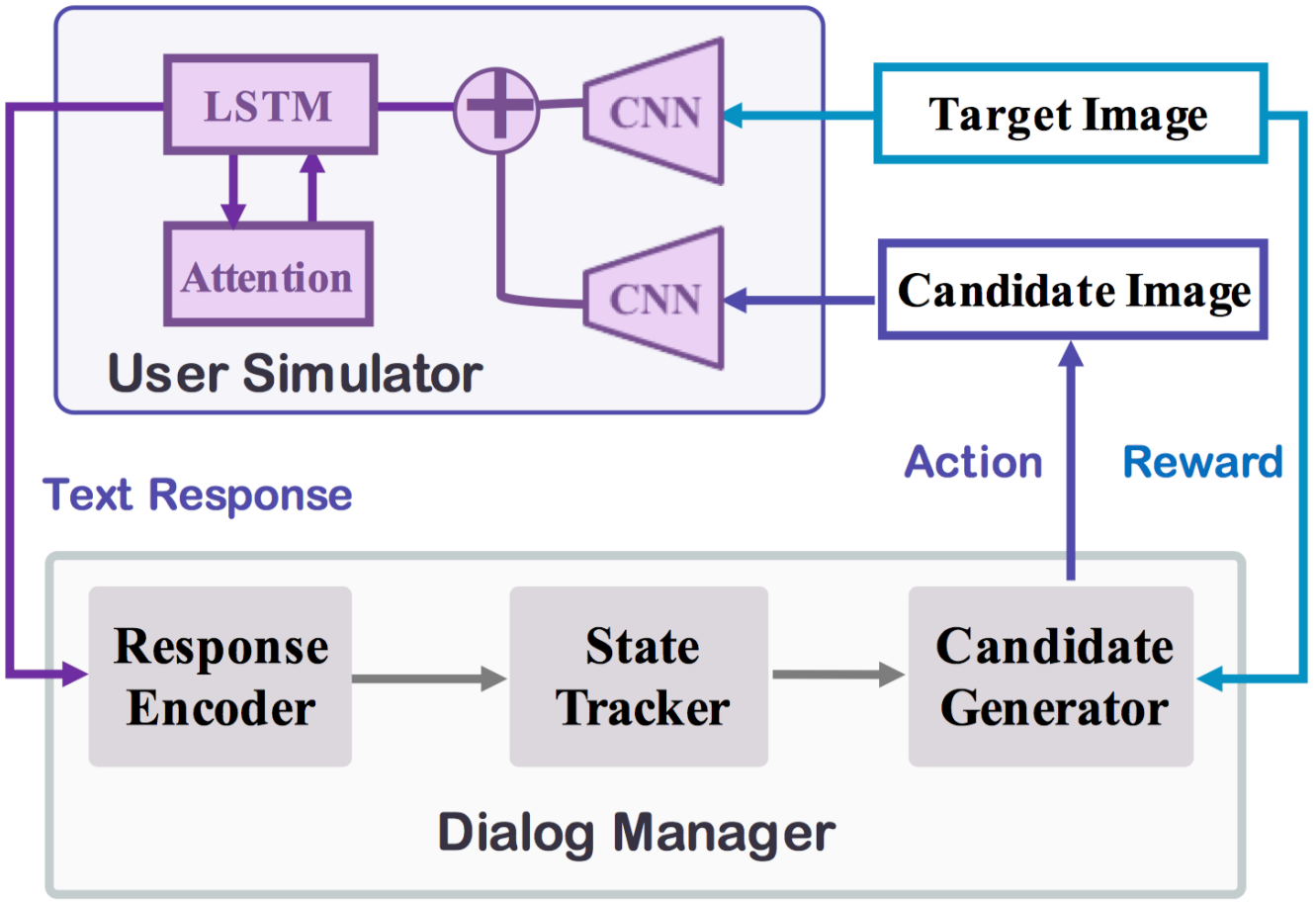


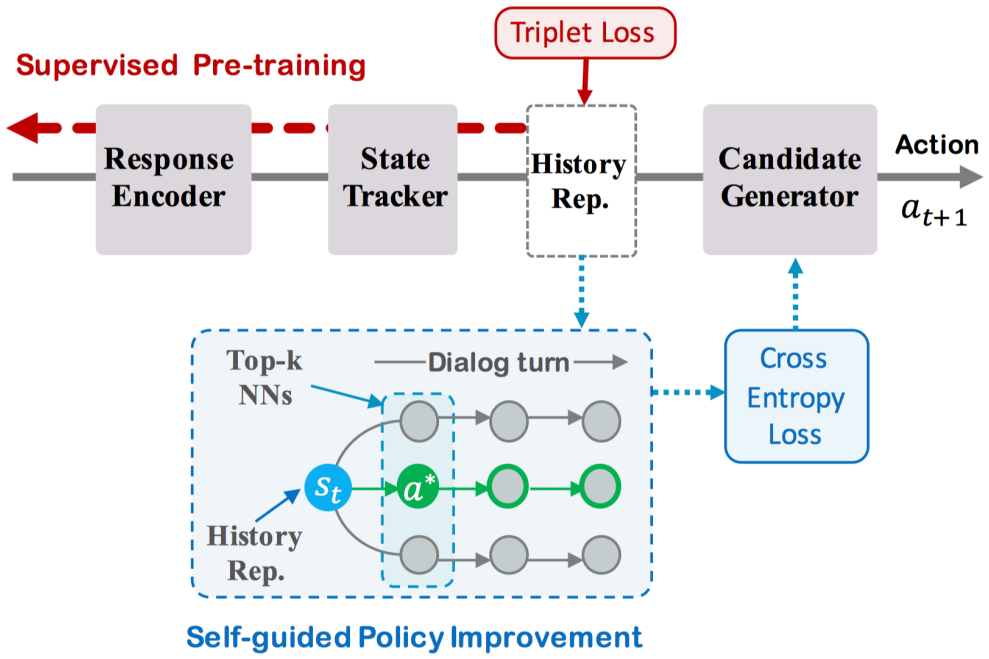
Figure 2: The proposed end-to-end framework for dialog-based interactive image retrieval.

User simulator (relative captioning)



(a)

Learning



(b)

- Pretraining

$$\mathcal{L}^{\text{sup}} = \mathbb{E}\Big[\sum_{t=0}^{T} \max(0, \|s_t - x^+\|_2 - \|s_t - x^-\|_2 + \text{m})\Big]$$

- Model-Based Policy Improvement

$$Q^{\pi}(h_t, a_t) = \mathbb{E}\big[\textstyle\sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} | \pi\big]$$

$$\pi'(h_t) \equiv a_t^* = \arg\max_a Q^{\pi}(h_t, a)$$

$$\mathcal{L}^{\text{imp}} = \mathbb{E}\Big[-\sum_{t=0}^{T} \log\Big(\pi(a_t^*|h_t)\Big)\Big]$$