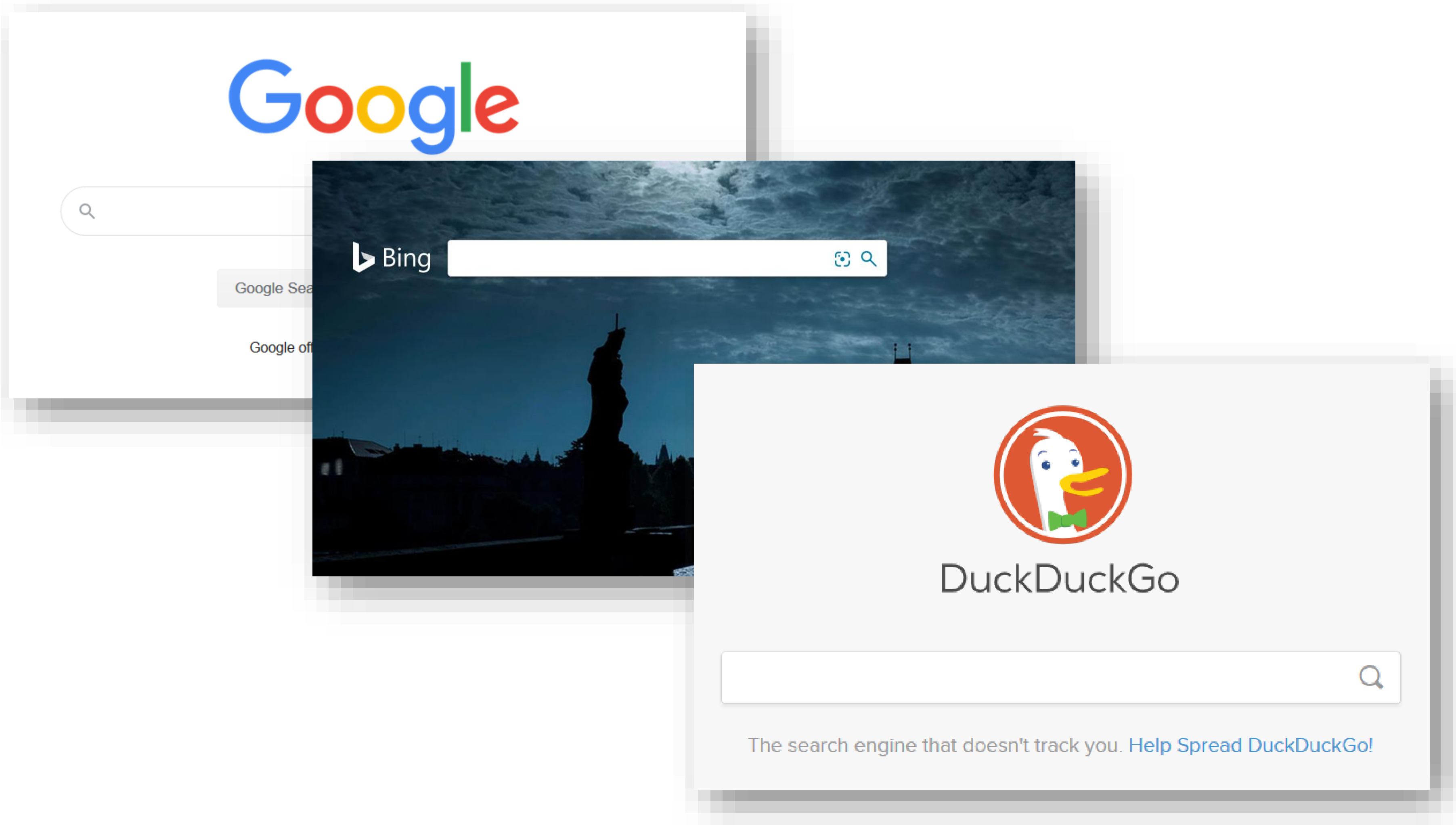
Multimodal Conversational Search

Web Data Mining and Search

Beyond simple search?

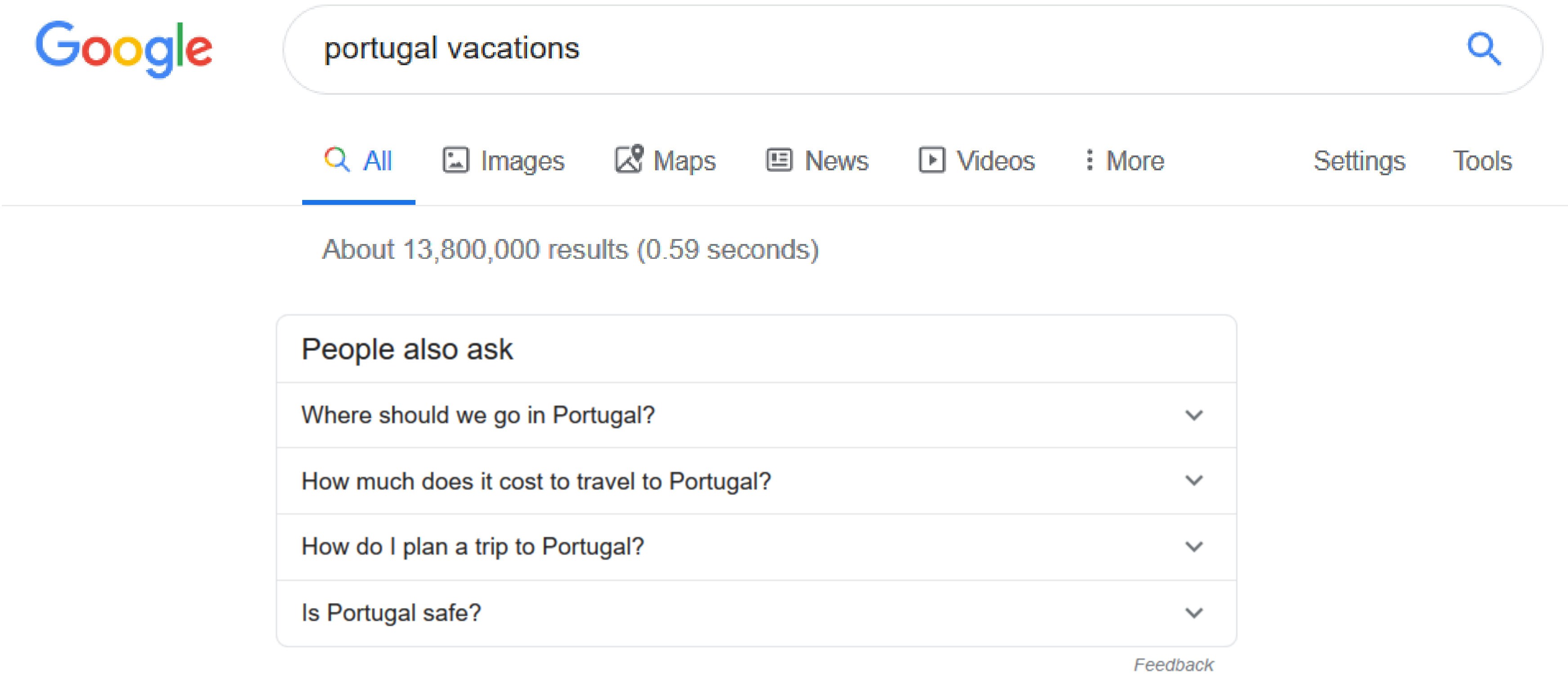


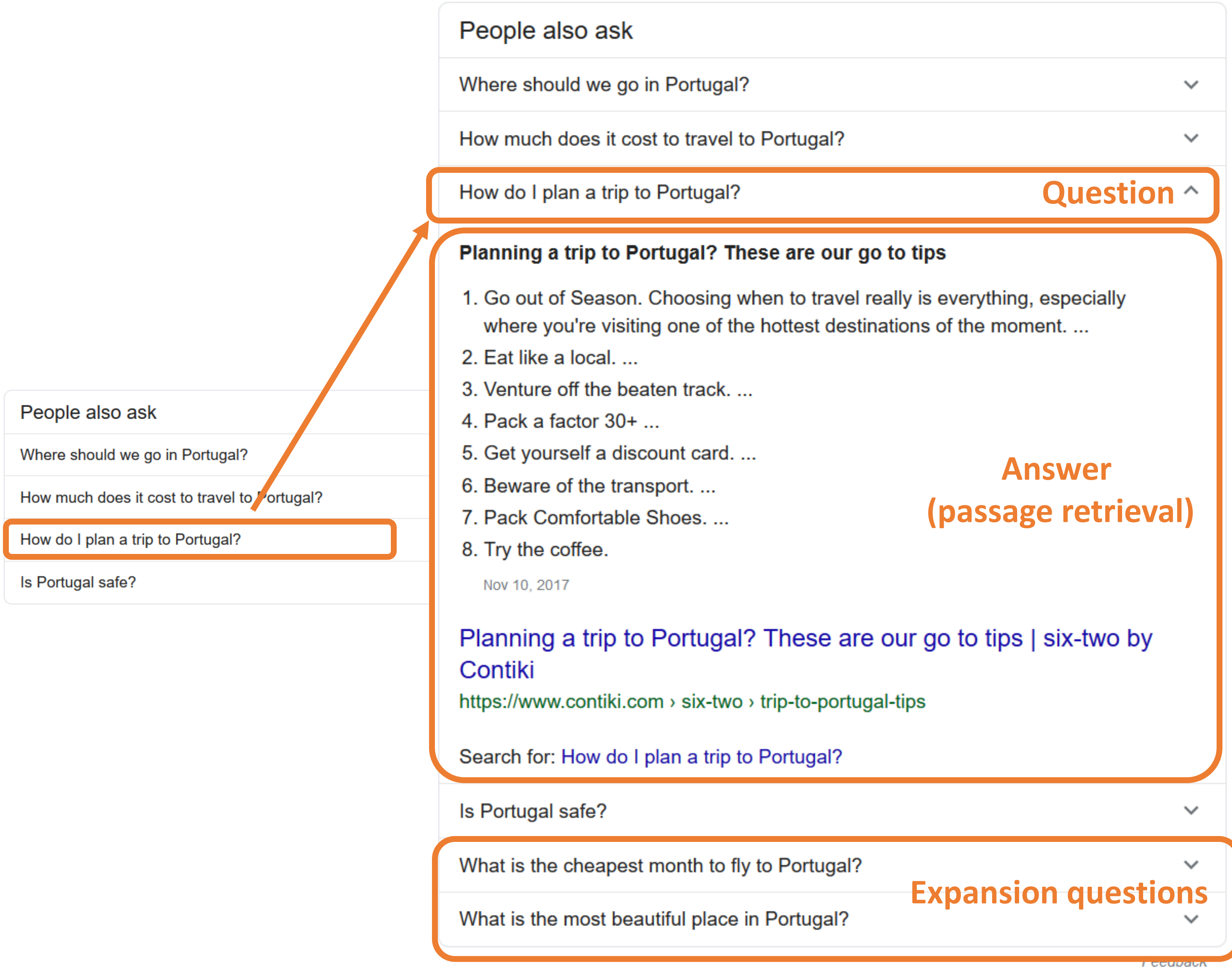
Mobile QA

Move to mobile favors a move to **speech** which favors **natural language information search**

- Will we move to a time when over half of searches are spoken?

Question Answering




People also ask
Where should we go in Portugal?
How much does it cost to travel to Portugal?
How do I plan a trip to Portugal?
Is Portugal safe?
People also ask
Where should we go in Portugal?
How much does it cost to travel to Portugal?
How do I plan a trip to Portugal?
Question
Planning a trip to Portugal? These are our go to tips
1. Go out of Season. Choosing when to travel really is everything, especially where you're visiting one of the hottest destinations of the moment. ...
2. Eat like a local. ...
3. Venture off the beaten track. ...
4. Pack a factor 30+ ...
5. Get yourself a discount card. ...
6. Beware of the transport. ...
7. Pack Comfortable Shoes. ...
8. Try the coffee.
Nov 10, 2017
Answer
(passage retrieval)
Planning a trip to Portugal? These are our go to tips | six-two by Contiki
https://www.contiki.com › six-two › trip-to-portugal-tips
Search for: How do I plan a trip to Portugal?
Is Portugal safe?
What is the cheapest month to fly to Portugal?
What is the most beautiful place in Portugal?
Expansion questions

Conversational Search

U: Tell me about the Neverending Story film.

A: ...

U: What is it about?

A: ...

U: Who was the author and when it was published?

A: ...

U: Who are the main characters?

A: ...

User utterance

Agent utterance

Types of conversations

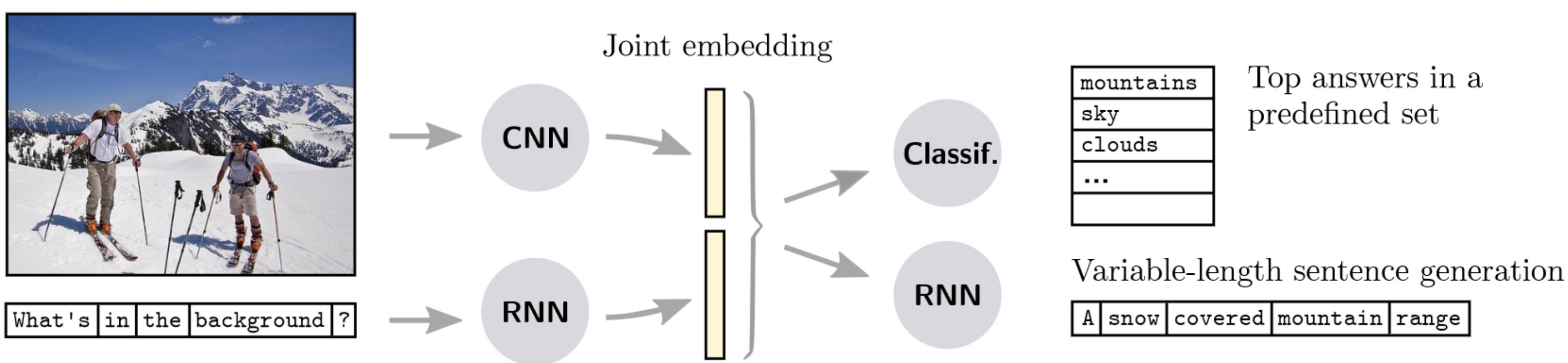
- **Question answering**
  - Information seeking tasks
- **Task oriented conversation**
  - Online shopping, booking, etc.
- **Open ended conversation**
  - Information seeking tasks
  - Chit-chat

Single turn Human-Agent interaction

Multi-turn Human-Agent interaction

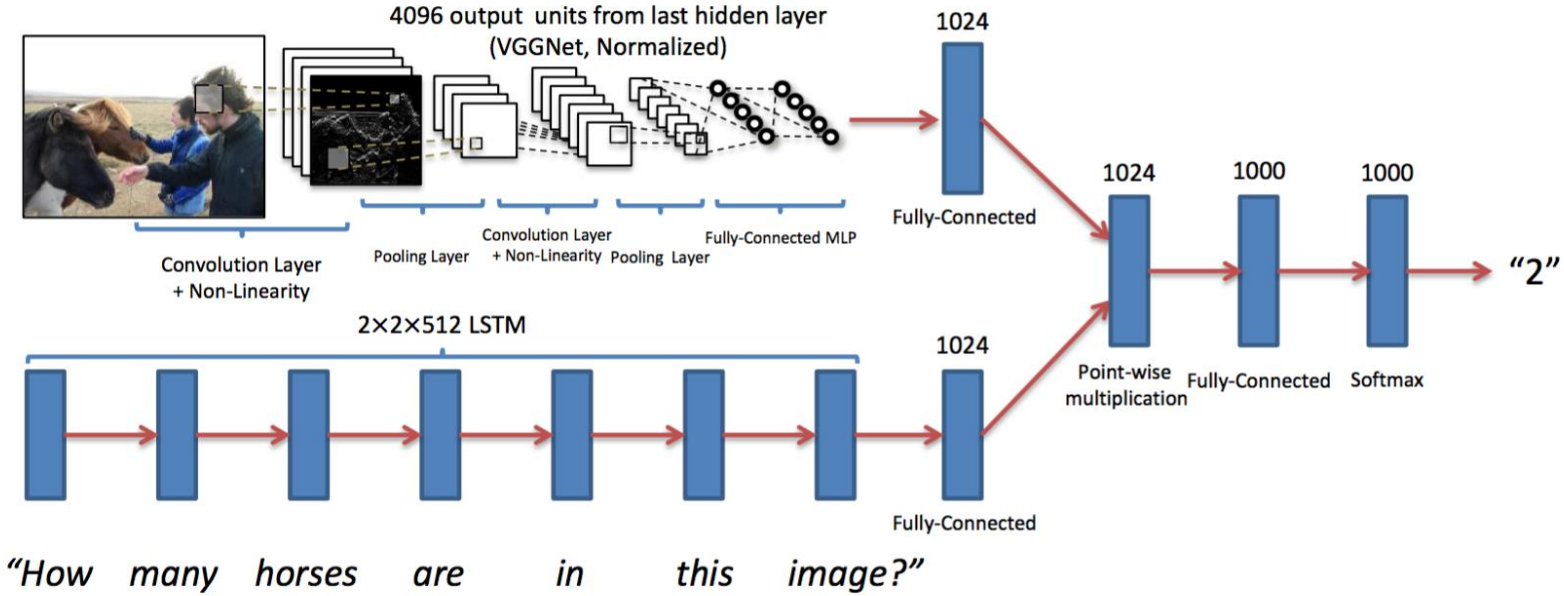
Visual Question Answering

- The traditional VQA architecture maps the input image and the input question into a common embedding space.
- This embedding can be learned jointly or sequentially.
- From this joint embedding projection of the image and question, an answer is generated by an RNN.



Wu, Qi, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. "Visual question answering: A survey of methods and datasets." *Computer Vision and Image Understanding* 163 (2017): 21-40.

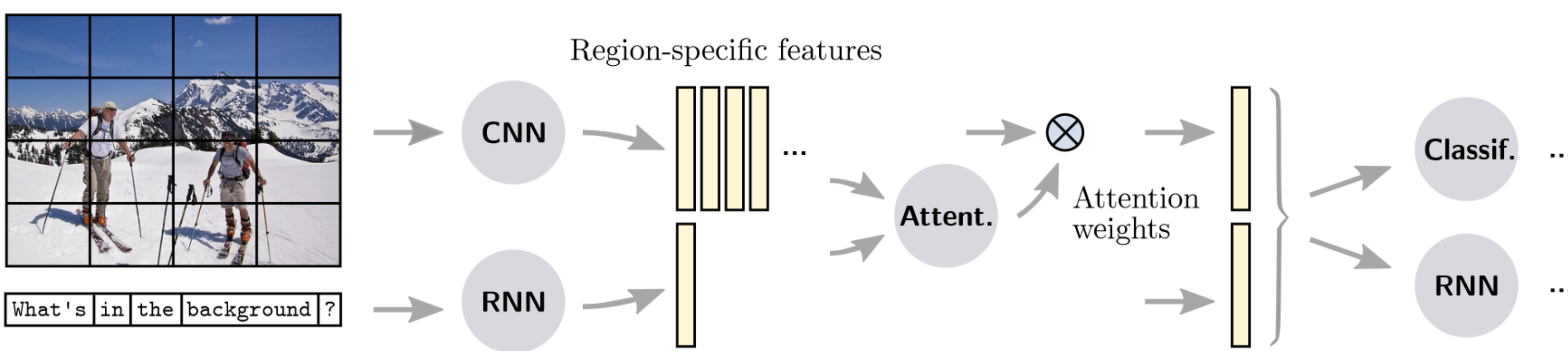
Visual Question Answering



Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. "Vqa: Visual question answering." *IEEE CVPR*. 2015.
https://arxiv.org/pdf/1505.00468.pdf

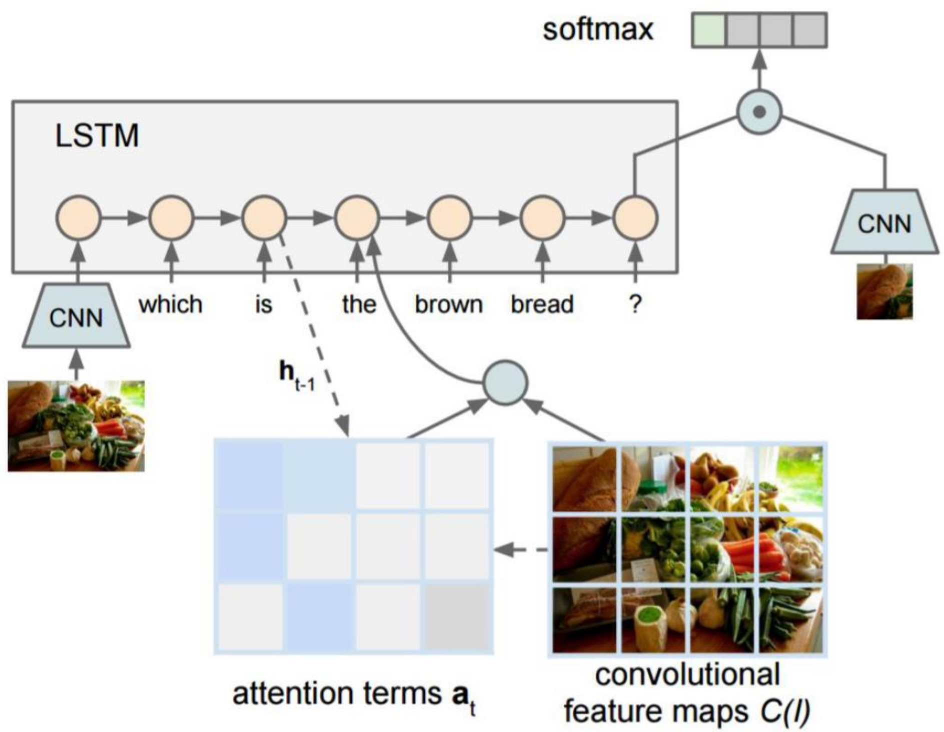
Visual Question Answering with Attention

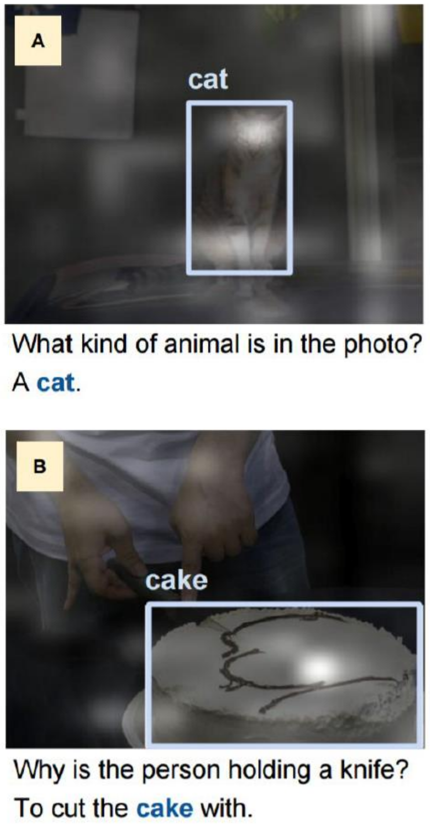
- VQA can be focused on specific image regions.
- Attention weights are computed from the question embedding vector and the region-specific embedding vector.
- The answer is then computed by giving attention to different image regions when generating the different parts of the answer.



Wu, Qi, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. "Visual question answering: A survey of methods and datasets." *Computer Vision and Image Understanding* 163 (2017): 21-40.

Attention based VQA and Grounded VQA





Zhu et al, "Visual 7W: Grounded Question Answering in Images", CVPR 2016

VQA examples

Q: What endangered animal is featured on the truck?

A: **A bald eagle.**
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?

A: **Onto 24 ¾ Rd.**
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?

A: **During a wedding.**
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church

Q: Who is under the umbrella?

A: **Two women.**
A: A child.
A: An old man.
A: A husband and a wife.

Type of conversations

- **Question answering**
  - Information seeking tasks
- **Open ended conversation**
  - Information seeking tasks
  - Chit-chat
- **Task oriented conversation**
  - Online shopping, booking, etc.

Techniques

- **Rule-based**
  - Pattern-action rules (Eliza)
  - + a mental model (Parry)
- **Corpus-based (from large chat corpus)**
  - Information Retrieval
  - Neural network encoder-decoder

Rule-based open-ended conversations

- Sometimes for fun
- But originally designed as a testbed for clinical therapy

  (And still used that way!)

Eliza: Weizenbaum (1966)

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

...

WHO ELSE IN YOUR FAMILY TAKES CARE 0F YOU?

My father

YOUR FATHER

Eliza: Weizenbaum (1966)

You are like my father in some ways.

WHAT RESEMBLANCE DO YOU SEE

You are not very aggressive but I think you don't want me to notice that.

WHAT MAKES YOU THINK I AM NOT AGGRESSIVE

You don't argue with me.

WHY DO YOU THINK I DON'T ARGUE WITH YOU

You are afraid of me.

DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU

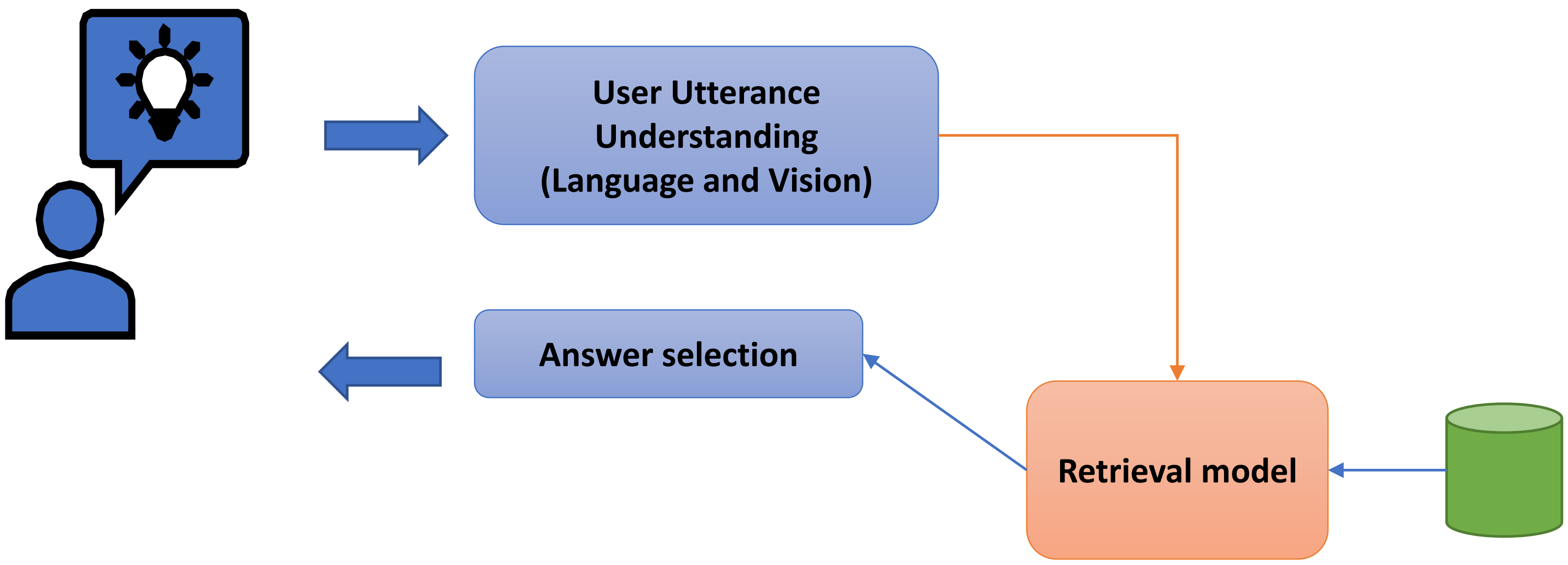
My father is afraid of everybody.

WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

IR based conversations



Corpus-based open-ended conversations

- **Idea**: Mine conversations of human chats or human-machine chats
  - Microblogs: Twitter or Weibo (微博)
  - Movie dialogs
- Cleverbot (Carpenter 2017 http://www.cleverbot.com)
- Microsoft XiaoIce
- Microsoft Tay

A Cleverbot conversation

User:	Do you like Doctor Who?
Cleverbot:	Yes, so funny
User:	What's your favorite show?
Cleverbot:	I don't have one
User:	What's your name?
Cleverbot:	Jason I already told you.
User:	No your real name.
Cleverbot:	Stephanie IS my real name.
User:	I thought it was Jason?
Cleverbot:	How did you come to that conclusion.
User:	Because you said so.
Cleverbot:	I have to do my homework.

Two IR-based chatbot architectures

1. Search for similar questions
   - Take user's turn (*q*) and find a (tf-idf) similar turn *t* in the corpus C

     q = "do you like Doctor Who"

     t' = "do you like Doctor Strangelove"
   - Grab whatever the response was to *t*.

$$r = response\left(\underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}\right)$$

Yes, so funny

2. Search for the most similar turn

$$r = \underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}$$

Do you like Doctor Strangelove

IR-based models of chatbots

- Also fine to use other features like user features, or prior turns

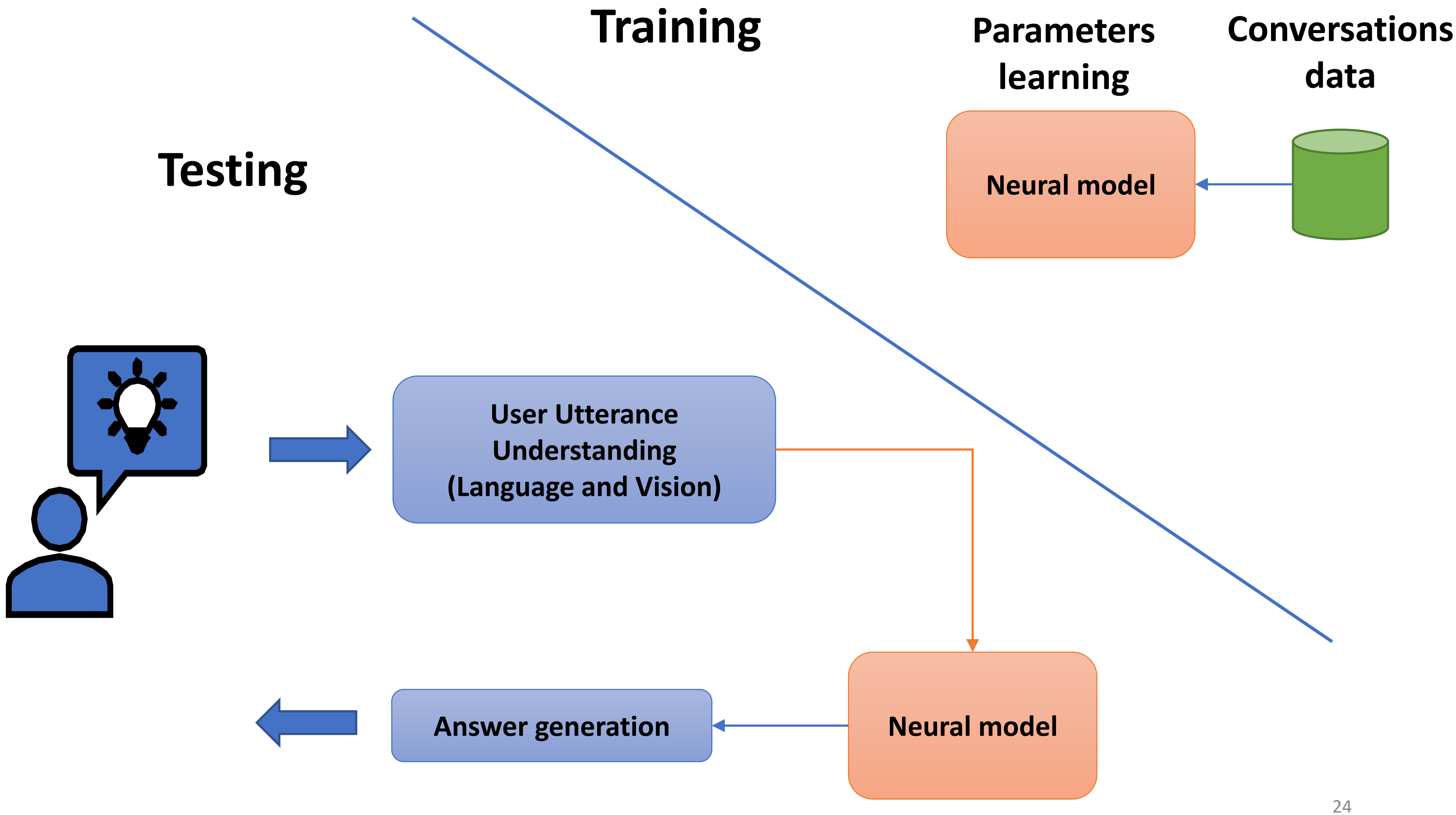
- Or non-dialogue text
  - COBOT chatbot (Isbell et al., 2000)
    - sentences from the Unabomber Manifesto by Theodore Kaczynski, articles on alien abduction, the scripts of "The Big Lebowski" and "Planet of the Apes".
  - Wikipedia text

Neural chatbots

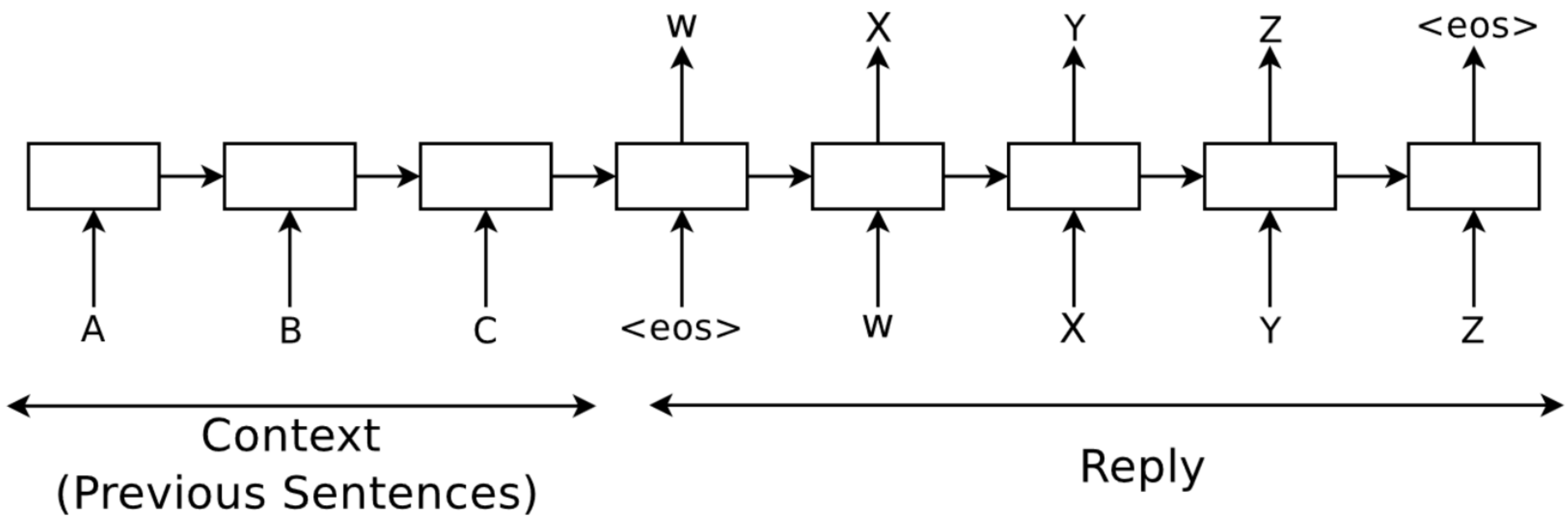
- Think of response generation as a task of *transducing* from the user's prior turn to the system's turn.

- Train on some data to capture all types of conversation iterations:
  - Movie dialogue databases
  - Twitter conversations

Neural based conversations



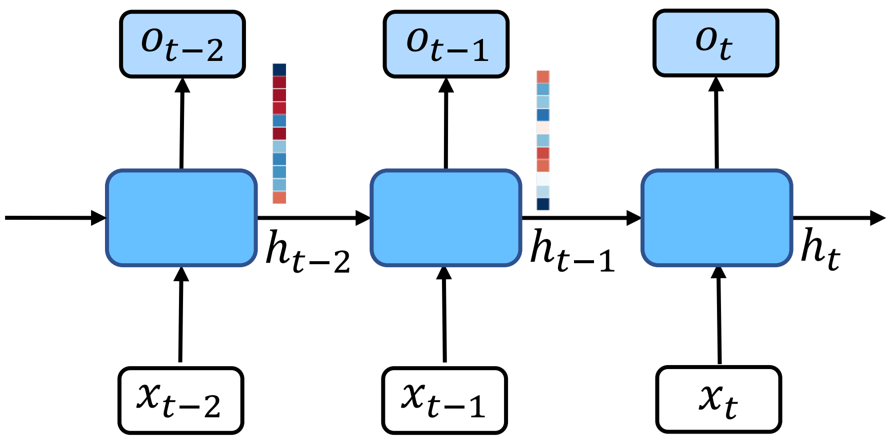
First cut to conversational systems



Vinyals, Oriol, and Quoc Le. "A neural conversational model." ICML Workshop on Deep Learning (2015).
https://arxiv.org/abs/1506.05869
https://sites.google.com/site/deeplearning2015/accepted-papers

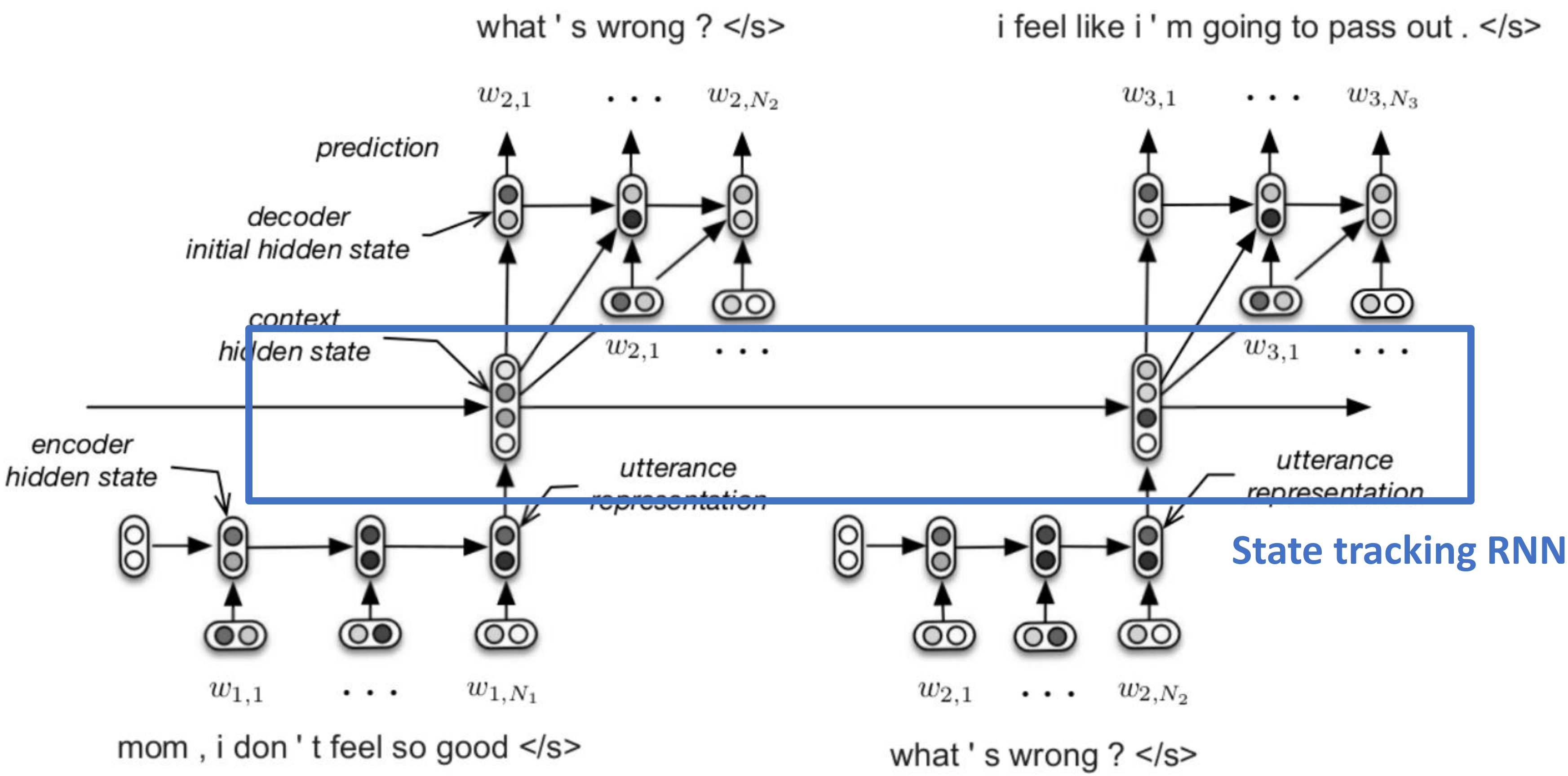
RNN-based dialog state tracking

- A first RNN is used to track the state of the conversation.
- It’s input data are the utterances of the conversation.
- Agent utterances can be generated from the conversation state.

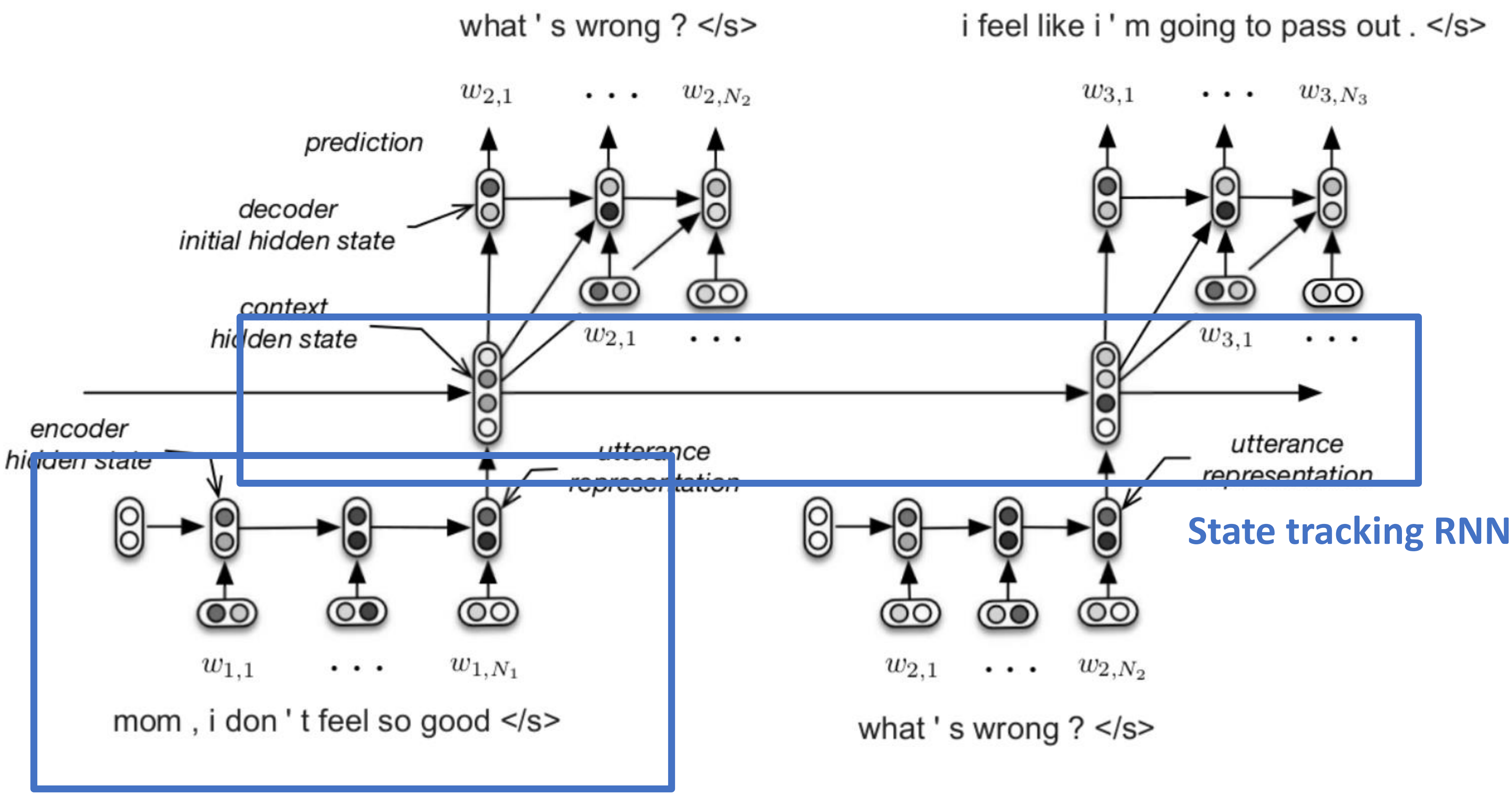


Serban, Iulian V., Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.
http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/11957/12160

Hierarchical Recurrent Encoder-Decoder

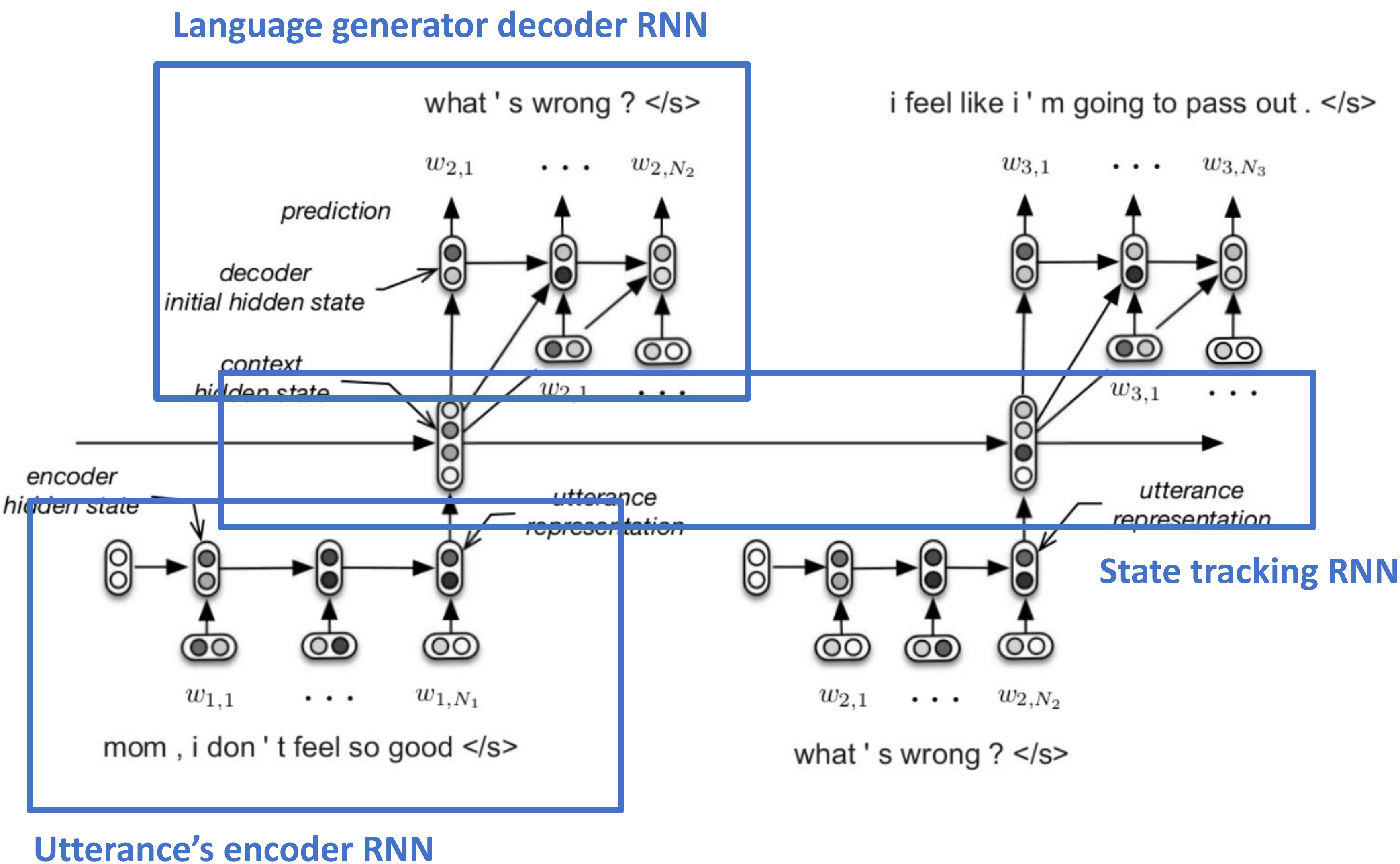


Hierarchical Recurrent Encoder-Decoder



Utterance's encoder RNN

Hierarchical Recurrent Encoder-Decoder



Initialization

- Word embeddings are initialized from word2vec
- Utterance encoders and decoders are initialized with QA from movie subtitles

Sample neural chatbot output

Input	tell me ... how long have you had this falling sickness ?
Vanilla-Seq2Seq	i'm not a doctor.
Adversarial	a few months, i guess .
Input	so i had the doctors test sammy 's response to conditioning .
Vanilla-Seq2Seq	sammy wrote the test sammy wrote the test .
Adversarial	so he took the pills .
Input	they didn 't have much success with this problem commander .
Vanilla-Seq2Seq	they 're not the only ones who have been in the system .
Adversarial	can we find someone else ?

Visual Dialog

- What is a visual dialog?
  - given an image I;
  - a history of a dialog consisting of a sequence of question-answer pairs;
  - a natural language follow-up question;
  - the task for the machine is to answer the question in free-form natural language.

- This task is the visual analogue of the Turing Test.

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

Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "**Visual dialog**." *IEEE CVPR* 2017. https://arxiv.org/pdf/1611.08669.pdf

Tasks comparisons

Captioning

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

VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

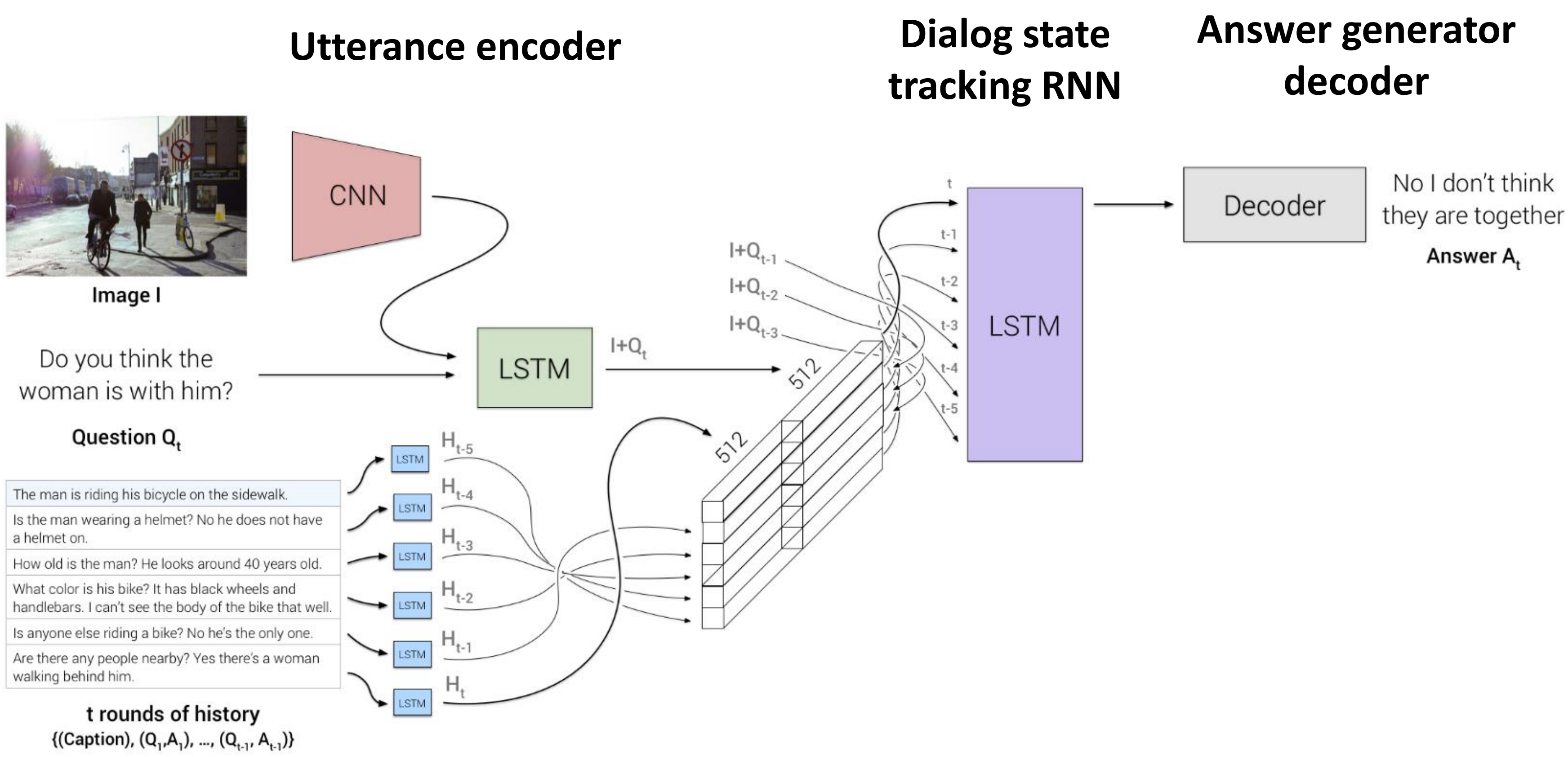
A: One male and one female

Q: Which one is holding a racket ?

A: The woman

Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "**Visual dialog**." *IEEE CVPR* 2017. https://arxiv.org/pdf/1611.08669.pdf

Hierarchical Recurrent Encoder-Decoder



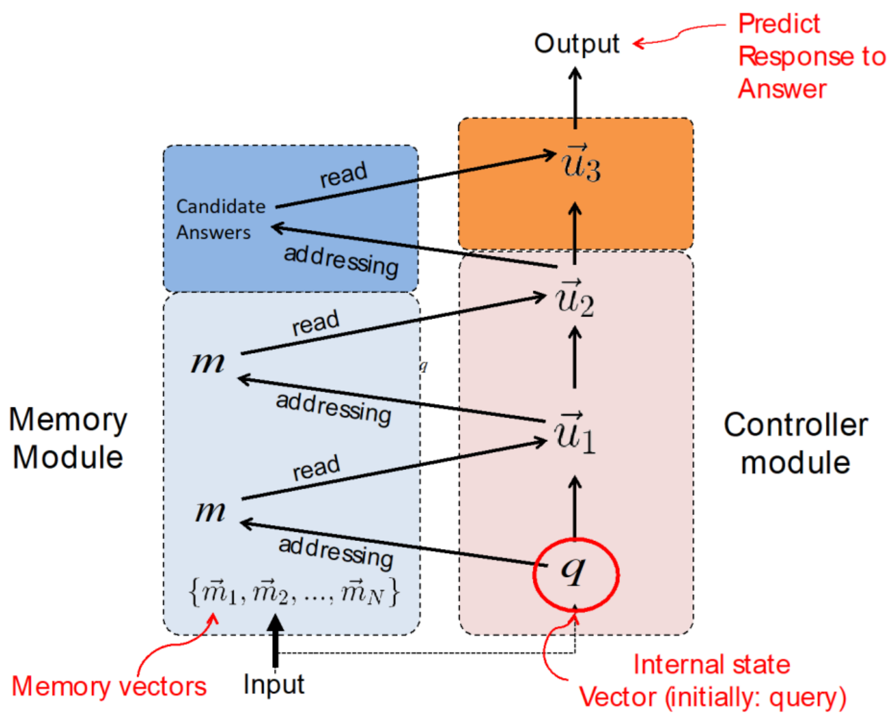
Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "**Visual dialog**." *IEEE CVPR* 2017. https://arxiv.org/pdf/1611.08669.pdf

MemNets-based dialogue state tracking

- Class of models that combine large memory with learning component that can read and write to it.

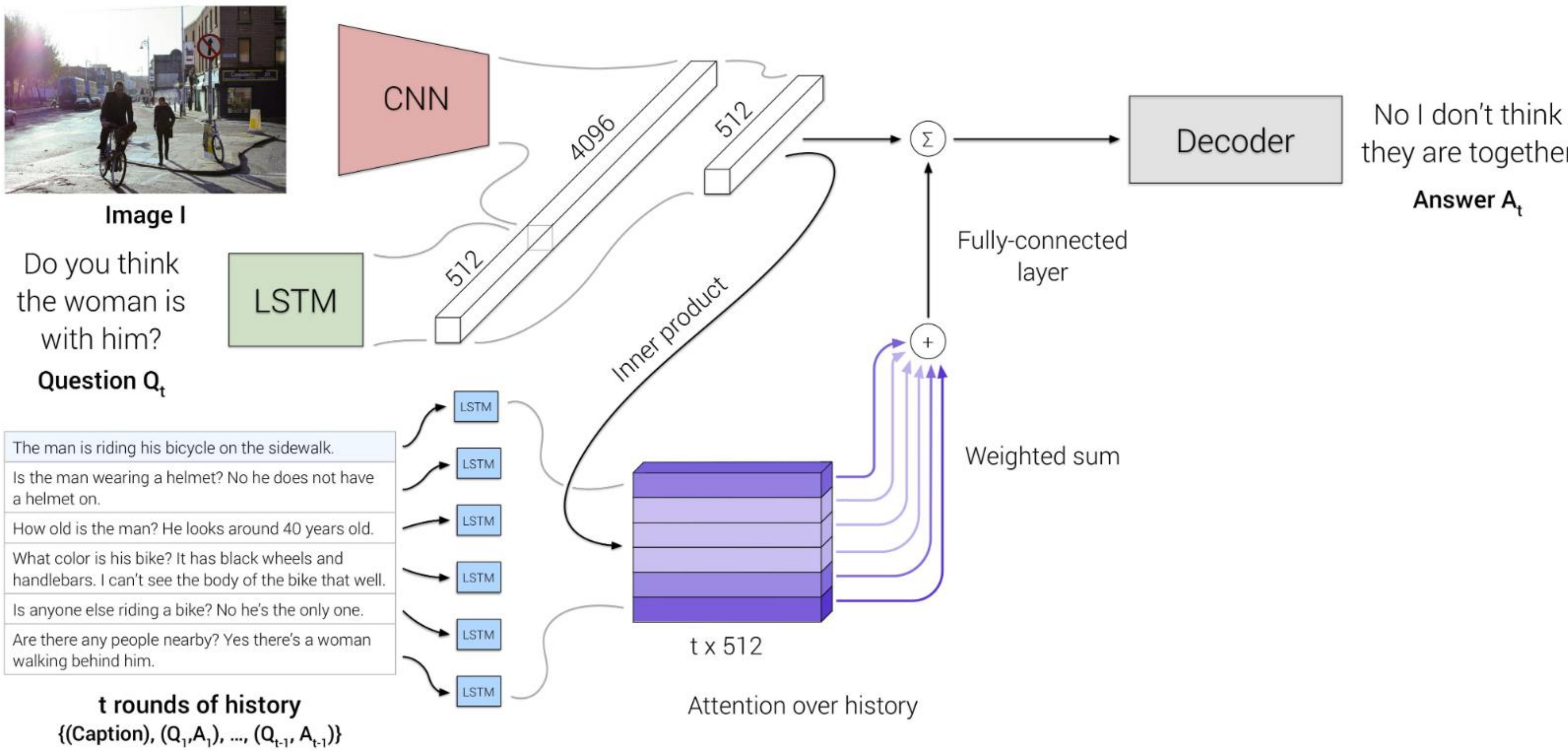
- Incorporates **reasoning** with **attention** over **memory** (RAM).

- Most ML has limited memory which is more-or-less all that's needed for "low level" tasks e.g. object detection.



Memory Networks for Language Understanding, ICML Tutorial 2016
http://www.thespermwhale.com/jaseweston/icml2016/

Memory Network for Visual Dialogs



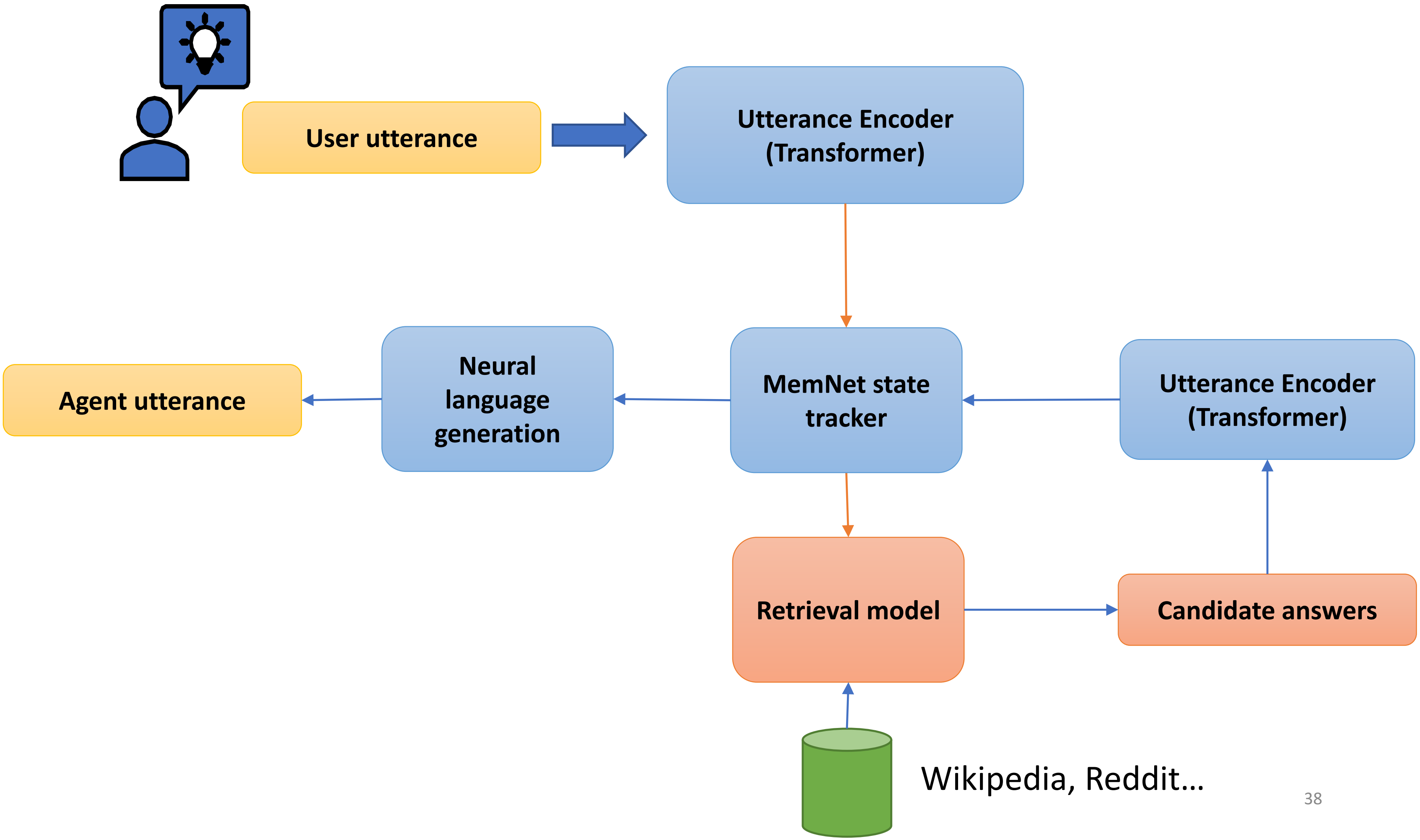
Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "**Visual dialog**." *IEEE CVPR* 2017. https://arxiv.org/pdf/1611.08669.pdf

The Wizard of Wikipedia

- Knowledge Retrieval
  - All utterances are encoded with state-of-the-art methods
  - Retrieval gets a set of candidate answers
- Knowledge Attention
  - The memory network keeps track of past turns
  - Attention mechanism tells which previous turns are more relevant to the current turn
- Response Generation
  - Response is generated from the attention weighted combination of past utterances.

Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. "Wizard of Wikipedia: Knowledge-Powered Conversational Agents." In *ICLR* 2018. https://arxiv.org/abs/1811.01241

The Wizard of Wikipedia



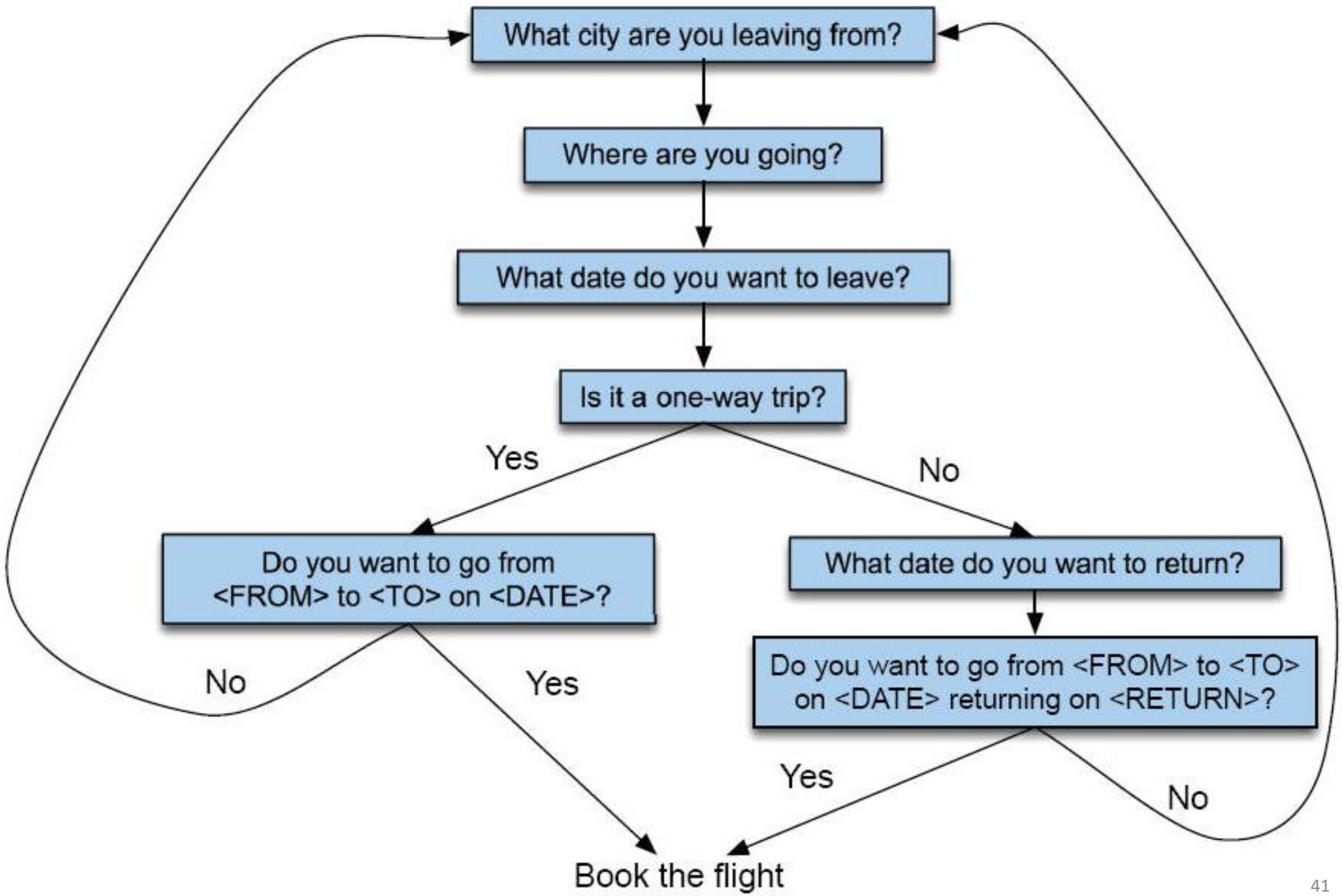
The Wizard of Wikipedia (example)

Model: Retrieval Transformer MemNet	
Topic:	E-book (*seen*)
Human:	i do not know why, but I have never gotten into E-books
Model:	Same here. I have a kindle, but I don't use it for books. It's nicer to hold a real book and turn pages.
Human:	yes, i like the physical feel and smell of a real book
Model:	I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.
Human:	i tried e-books but I do not like reading on a screen, its disorientated for me
Model:	Yes, reading a book on a tablet just isn't the same!
Human:	plus I like owning the physical copy of a real book, I feel like I own it
Model:	I prefer print and newsreels.
Topic:	Toga party (*seen*)
Model:	I love toga parties! I usually make my toga out of a bed sheet and wear my sandals.
Human:	Me too!!! I usually use some flip flops and my blanket!
Model:	Toga parties typically have games involving kegs and excessive drinking!
Human:	Yeah! And getting to wear funny outfits and a crown of leaves!
Model:	I love me some Toga Parties!
Human:	I love me some too!
Model:	Toga parties are so much fun with the Roman or Greek theme.
Human:	Yeah! I especially like Greek themes

Type of conversations

- **Question answering**
  - Information seeking tasks
- **Open ended conversation**
  - Information seeking tasks
  - Chit-chat
- **Task oriented conversation**
  - Online shopping, booking, etc.

Finite State Dialog Manager

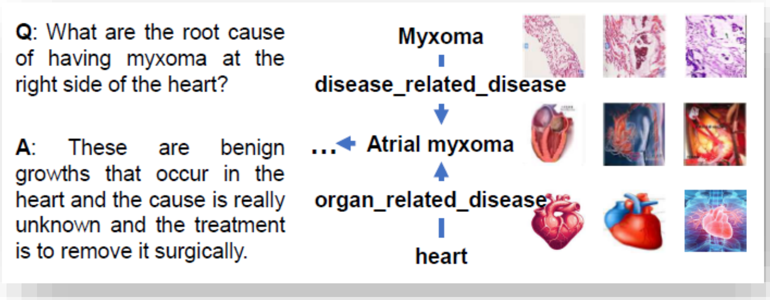


Multimodal conversational systems

- For online shopping, task completion conversational agents will support users with their shopping decisions.
- User preferences and dialogue context are crucial elements.

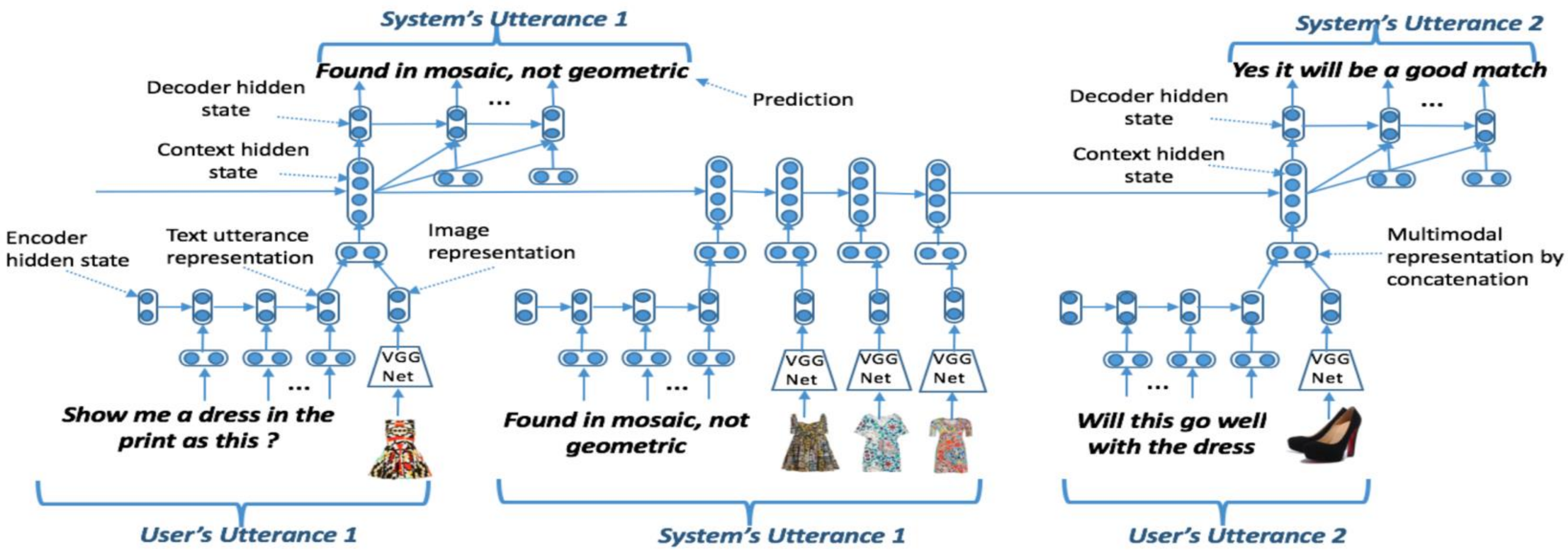
Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. "Towards building large scale multimodal domain-aware conversation systems." *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

Yingying Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2019. Multi-modal Knowledge-aware Hierarchical Attention Network for Explainable Medical Question Answering. *ACM International Conference on Multimedia* (MM '19).



Multimodal state-tracking

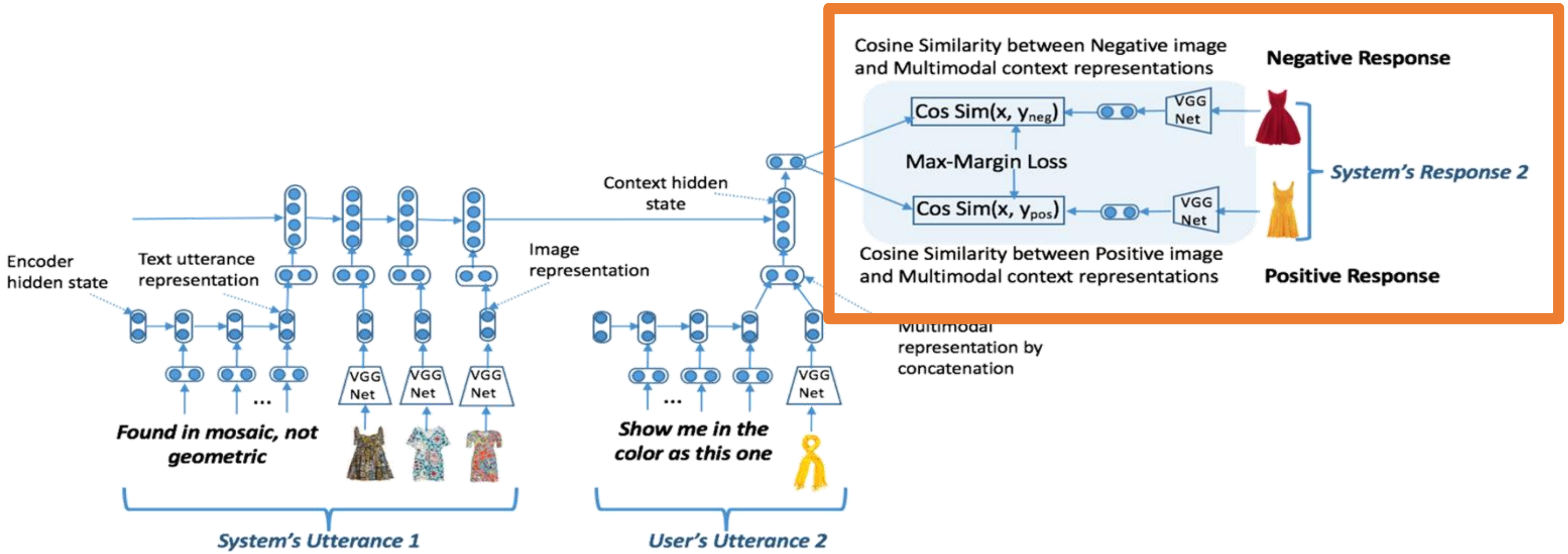
- HRED can be extended to include both language and visual data.
- VGG features are transformed to match the context state representation.
- Both representations can be concatenated.



Saha, Amrita, Mitesh M. Khapra, and Karthik Sankaranarayanan. "**Towards building large scale multimodal domain-aware conversation systems**." In *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018. https://amritasaha1812.github.io/MMD/

Product selection in conversational recommendation/search

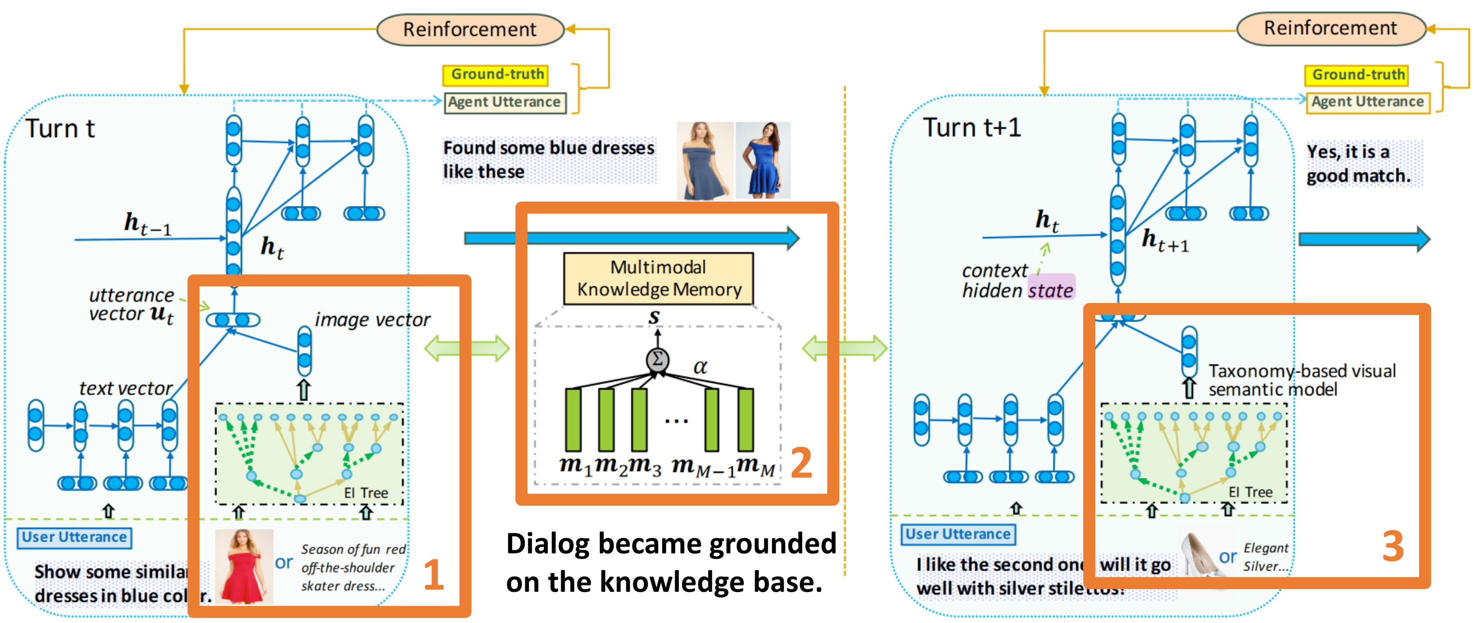
- A straightforward way of selecting products is by a way to compare the context state with existing products
- It requires learning a way of extracting a product representation embedding from the context state.



Saha, Amrita, Mitesh M. Khapra, and Karthik Sankaranarayanan. "**Towards building large scale multimodal domain-aware conversation systems**." In *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018. https://amritasaha1812.github.io/MMD/

Knowledge-based product selection in conversational recommendation/search

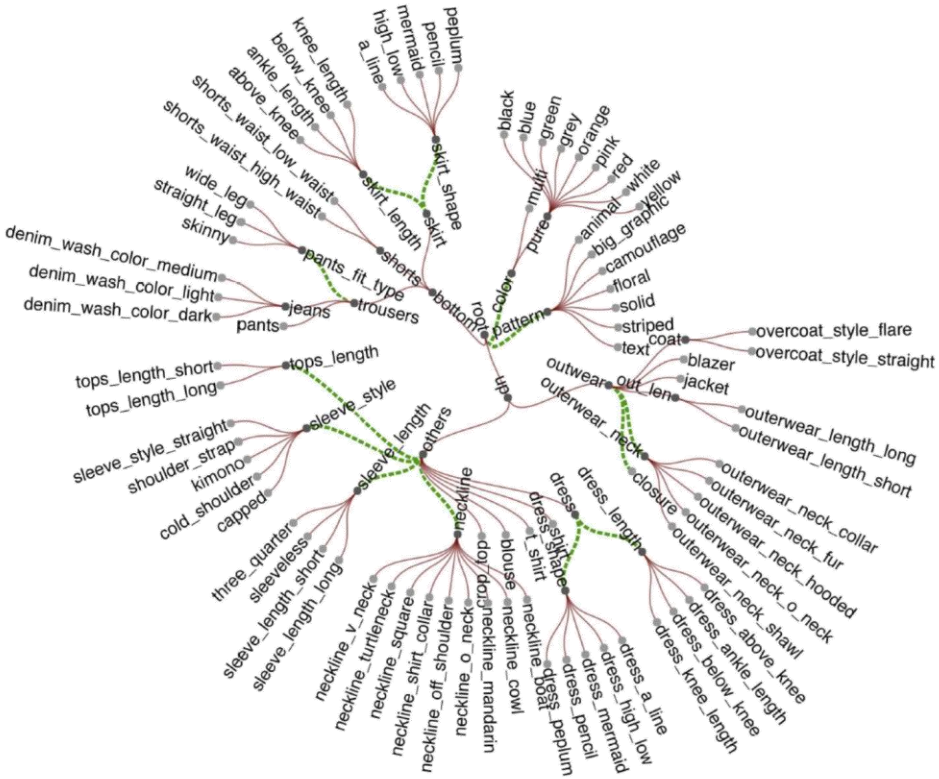
- Products need to be represented in a way that capture their fill set of characteristics
  - Taxonomy-based model have this information.



Liao, Lizi, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. "**Knowledge-aware multimodal dialogue systems**." In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 801-809. 2018. http://staff.ustc.edu.cn/~hexn/papers/mm18-multimodal-dialog.pdf

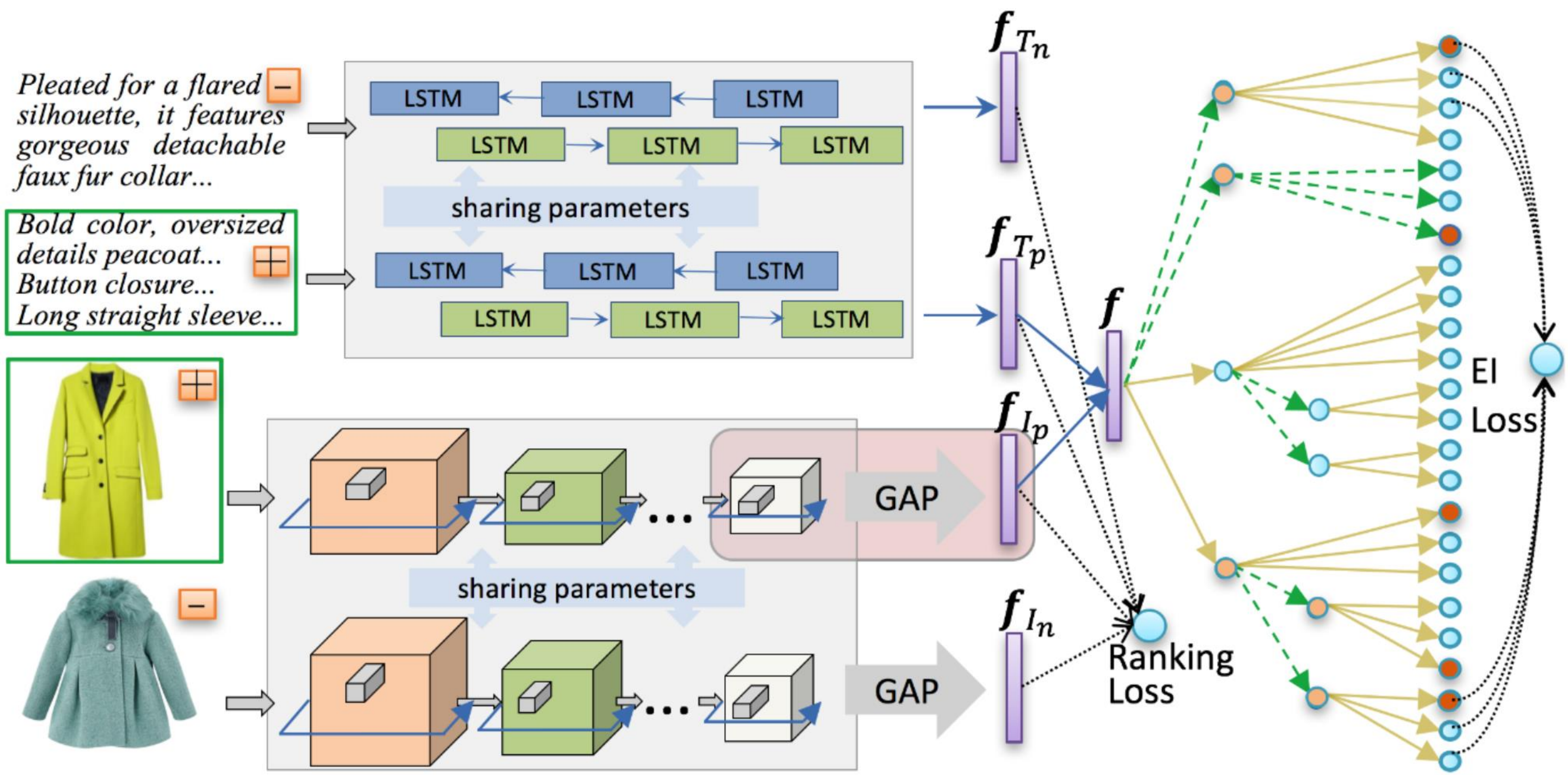
Knowledge base

- Product types
  - Dresses, shoes, shirts, etc.
- Product attributes
  - Color, texture, material, size, style, etc.



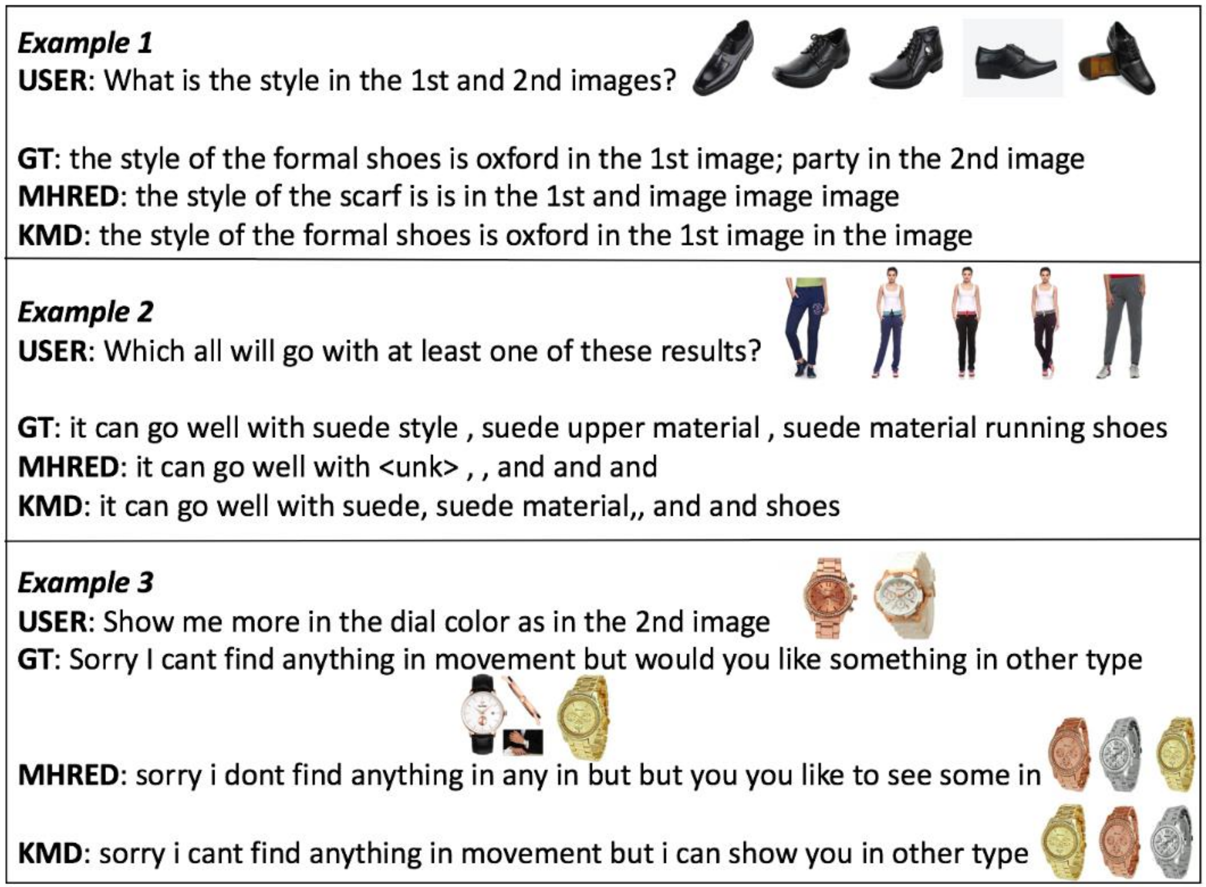
Liao, Lizi, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. "**Knowledge-aware multimodal dialogue systems**." In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 801-809. 2018.
http://staff.ustc.edu.cn/~hexn/papers/mm18-multimodal-dialog.pdf

Knowledge-based product selection



Liao, Lizi, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. "**Knowledge-aware multimodal dialogue systems**." In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 801-809. 2018. http://staff.ustc.edu.cn/~hexn/papers/mm18-multimodal-dialog.pdf

Qualitative results for multimodal conversational AI



Liao, Lizi, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. "**Knowledge-aware multimodal dialogue systems**." In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 801-809. 2018.
http://staff.ustc.edu.cn/~hexn/papers/mm18-multimodal-dialog.pdf

Results for multimodal conversational AI

- Comparison between HRED, Multimodal HRED, MemNets and Knowledge based MHRED.

Method	Text Response		Image Response (k = 5)		
	BLEU	**Diversity (unigram)**	**R@1**	**R@2**	**R@3**
HRED (text-only)	0.3174	0.00369	0.4323	0.6217	0.7486
MemNN	0.5013	0.00435	0.7800	0.8372	0.9091
MHRED	0.5195	0.00426	0.7980	0.8859	0.9345
KMD	**0.6731**	**0.00534**	**0.9198**	**0.9552**	**0.9755**

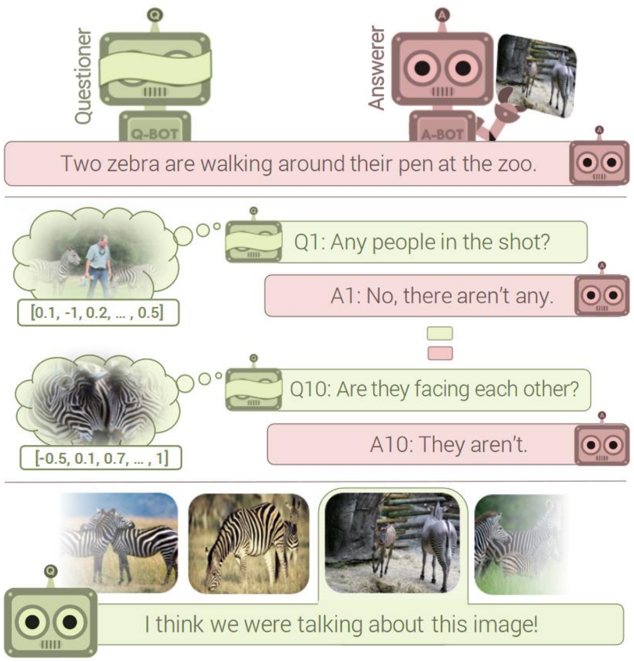
Liao, Lizi, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. "**Knowledge-aware multimodal dialogue systems**." In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 801-809. 2018.
http://staff.ustc.edu.cn/~hexn/papers/mm18-multimodal-dialog.pdf

Dialog policy enforcement

- Image guessing game:
  - Both agents share a database
  - A-bot selects an image at random
  - Q-bot needs to guess that image
  - Agents interact through a dialog in English

- Based on the HRED architecture.
- Uses reinforcement learning to enforce a dialog policy.

- **Interesting**: the agents may develop their own language to solve the task.



Das, Abhishek, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. "**Learning cooperative visual dialog agents with deep reinforcement learning**." IEEE CVPR. https://arxiv.org/abs/1703.06585

Summary

- Visual question answering methods are the base of conversational search.
- Architectural considerations:
  - User utterance processing
  - State tracking
  - Dialog policy enforcement (if goal oriented dialogue)
  - Answer retrieval
  - Answer generation
- State tracking and dialog control:
  - HRED
  - Memory networks
  - Reinforcement learning

Readings: Visual Question Answering

- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. IEEE CVPR.

- Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In Advances in neural information processing systems (pp. 4967-4976).

- Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. IEEE CVPR.

- Liang, Junwei, Lu Jiang, Liangliang Cao, Yannis Kalantidis, Li-Jia Li, and Alexander G. Hauptmann. "Focal visual-text attention for memex question answering." *IEEE PAMI* (2019). https://memexqa.cs.cmu.edu/fvta.html

- Jiang, Lu, Junwei Liang, Liangliang Cao, Yannis Kalantidis, Sachin Farfade, and Alexander Hauptmann. "Memexqa: Visual memex question answering." *arXiv preprint arXiv:1708.01336* (2017). https://arxiv.org/pdf/1708.01336.pdf

Readings: Conversational Search

- Surveys
  - Jannach, Dietmar, Ahtsham Manzoor, Wanling Cai, and Li Chen. "**A Survey on Conversational Recommender Systems.**" arXiv preprint arXiv:2004.00646 (2020). https://arxiv.org/pdf/2004.00646.pdf
  - Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. **Challenges in Building Intelligent Open-domain Dialog Systems**. ACM Trans. Inf. Syst. 38, 3, Article 21 (May 2020)

- Conversation state tracking:
  - HRED: Serban, Iulian V., Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. "**Building end-to-end dialogue systems using generative hierarchical neural network models**." In Thirtieth AAAI Conference on Artificial Intelligence. 2016. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPaper/11957

  - Memory Networks: Sukhbaatar, Sainbayar, Jason Weston, and Rob Fergus. "**End-to-end memory networks**." In Advances in neural information processing systems, pp. 2440-2448. 2015. http://papers.nips.cc/paper/5846-end-to-end-memorynetworks

Readings: Multimodal Conversational Search

- Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "**Visual dialog**." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326-335. 2017. https://arxiv.org/pdf/1611.08669.pdf

- Saha, Amrita, Mitesh M. Khapra, and Karthik Sankaranarayanan. "**Towards building large scale multimodal domain-aware conversation systems**." In *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018. https://amritasaha1812.github.io/MMD/

- Liao, Lizi, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. "**Knowledge-aware multimodal dialogue systems**." In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 801-809. 2018. http://staff.ustc.edu.cn/~hexn/papers/mm18-multimodal-dialog.pdf

- Das, Abhishek, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. "**Learning cooperative visual dialog agents with deep reinforcement learning**." IEEE CVPR. https://arxiv.org/abs/1703.06585

- Shah, Sanket, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. "**Kvqa: Knowledge-aware visual question answering**." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8876-8884. 2019. http://malllabiisc.github.io/resources/kvqa/

- Nie, Liqiang, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. "**Multimodal Dialog System: Generating Responses via Adaptive Decoders**." In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1098-1106. 2019. https://acmmultimedia.wixsite.com/magic

Readings: Open-ended Conversational Agents

- Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. "Wizard of Wikipedia: Knowledge-Powered Conversational Agents." In International Conference on Learning Representations. 2018. https://arxiv.org/abs/1811.01241

- Dalton, Jeffrey, Chenyan Xiong, and Jamie Callan. "Cast 2019: The conversational assistance track overview." In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC*, pp. 13-15. 2019.

Datasets

- QA:
  - TREC QA, SquAD, QUAC
  - VQA
  - 7w
  - CLEVR
  - KVAQ

- Conversational
  - CoQA
  - TREC-CAST
  - MMDS